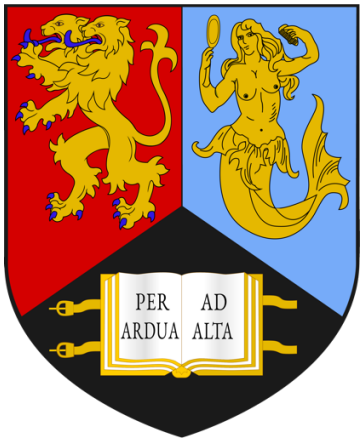


Investigating Novel Routes to High Intratumoural Immunity

Exploring the Immune Landscape in Low Mutation Burden Cancer

Authored By

Jack Luke McMurray

"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

November 2021

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Immune checkpoint blockade (ICB) has revolutionised treatment for high-mutation burden cancers that have deficient mismatch repair systems, and typically exhibit strong intratumoural immunity. However, extending this success to low-mutation burden settings that dominate many cancers remains a major challenge. For unknown reasons, low mutational cancer cohorts exhibit substantial heterogeneity in both the strength of intratumoural immunity, and in clinical prognosis. Colorectal cancer (CRC) represents a robust model to investigate such phenomena, since it comprises distinct subgroups that are ICB-responsive (microsatellite instability-high, MSI-H) and non-responsive (microsatellite stable, MSS) which exhibit high and low mutational burden, respectively. Utilising a T_H1-centric signature, I demonstrate that MSS patients have a highly variable intratumoural immunity and that MHC Class II expression can stratify a high-immunity MSS subgroup (MSS-hi-CIRC). Subsequently, using multispectral immunofluorescence, I demonstrate that MSS-hi-CIRC patients have an immune infiltrate comparable to ICB-responsive CRC patients (MSI-H), despite substantially lower mutational loads and no specific genetic determinants. Finally, I demonstrate differences in the microenvironmental immunobiology of MSS-hi-CIRC and MSI-H patients that indicate distinct routes to establishing high intratumoural immunity and suggest MSS-hi-CIRC immunobiology results from microbiome-driven adjuntivisation of intratumoural immunity. Furthermore, by mining data from other low-mutation burden cancers, I demonstrate that MSS-hi-CIRC immunobiology defined here might occur across several cancer types. My results therefore indicate that MHC-II level stratifies patients with immunobiology like that of MSI-H patients, albeit arising from distinct mechanisms. The analyses and mechanistic model proposed have implications for both the understanding of differential clinical prognosis in MSS cancer patients, and for the development of new immunotherapies and microbiome-based therapies focussed on this grouping.

Dedication

I would like to dedicate this thesis to my late Grandparents Douglas, and Irene McMurray. Your enthusiasm kept me strong; I will be forever grateful for your love for I know I would not be writing this if it wasn't for you both. Oh, and Grandad… I finally did it.

Acknowledgements

I am indebted to my primary supervisor Professor Benjamin Willcox for providing me with advice and guidance throughout, and even prior to, my PhD. I feel honoured to have done a PhD in Ben's group, and I thank him for the trust he placed in me to lead my own project. I would also like to extend this thanks to my co-supervisors, Professor Gary Middleton, and Dr Heather Long. Gary, I thank you for your intuitive discussions and ideas, and Heather, I thank you for your support.

In addition to my supervisors, I would like to acknowledge a few other key individuals. My Masters supervisor, Dr Duncan Murray. Duncan, you passed on your passion for programming to me which has proved vital to this project. I thank you for your patience in teaching me. I'd like to thank current, and previous, employees of the Birmingham CRUK Centre, the BTA, and the HBRC. In particular, I would like to thank Kelly Hunter, Ana Tedeosio, Joe Flint, Amy Hadley-Garry, and Charlotte Gray, who worked tirelessly to support me. I would like to thank the CRUK bioinformatician - Boris Noyvert, and Roland Arnold for their guidance with various bioinformatic analyses and Toritseju Sillo for granting me access to her data once she had finished her PhD. I thank the entire Willcox group for their resolute support throughout my PhD. I would like to pay a special regard to two ex-Willcox members, Neeraj Lal and Ghaleb Goussous. Neeraj, I thank you for your guidance and for completing the work that provided the basis of my project. Ghaleb, I thank you for your patience with me when I first started this project. I also thank Thomas Elliott who will take this project forward. I thank you for the insightful discussions we had towards the end of my PhD. I want to pay a special thanks to all members, creators, and moderators of StackOverflow. This resource has proved invaluable to me during my PhD.

I would like to thank my funders, Cancer Research UK for providing support to this project, and the patients who so selflessly donated to make this work possible. I thank Genomics England, who granted me access to their database, and the patients who populated it.

Finally, I would like to extend thanks to my family and friends, all of whom have been encouraging throughout my PhD. My father, David McMurray, and stepmother, Jackie McMurray, I thank you for your kindness throughout all my studies. We're finally here, and I wouldn't be here without your backing. I would like to pay a special regard to Kirsty Bugajski and her family. Kirsty, you have been a pillar of support during my PhD, and I thank you for your encouragement as well as for maintaining my sense of humour and preserving my sanity when I thought I'd lost them both. I was struggling when I met you, and your love steered me back on course. I will be forever grateful for your help.

Last but certainly not least, a thanks to my lifelong friend - James Tear. You've been a mainstay of support for me for as long as I can remember, thank you for your unwavering friendship.

Table of Contents

Chapter 1: Introduction 1

1.1 Introduction 2

1.2 Barriers to immune action 2

1.2.1 Targeting resistance to immune action 4

1.3 Immunogenicity and genetics of cancer 7

1.3.1 Using CRC as a model for investigating the immunobiology of low-mutation burden cancer 10

1.4 Colonic anatomy and function 11

1.4.1 Microscopic anatomy 13

1.5 The colonic microbiome 16

1.5.1 Functions of the colonic bacteria 17

1.5.2 Keeping bacteria at bay 18

1.6 Colorectal Cancer 30

1.6.1 CRC Carcinogenesis 30

1.6.2 CRC Diagnosis & Classification 33

1.6.3 Stratification approaches towards CRC 34

1.6.4 Management of CRC 38

1.6.5 CRC Immunology 39

1.7 Project aims 43

Chapter 2: Materials and Methods 45

2.1 Introduction 46

2.2 Public Data Collection and Processing 46

2.2.1 TCGA-COADREAD 50

2.2.2 TCGA-STAD and TCGA-UCEC 57

2.2.3 Affymetrix datasets 57

2.2.4 Agilent dataset 60

2.2.5 Geneset enrichment analysis 60

2.3 Local Cohort Collection and Processing 62

2.3.1 Birmingham cohort 62

2.3.1 100K Genomes Colorectal Cancer Dataset 68

2.3.2 Wales Cancer Bank Cohort 70

Chapter 3: A Redefinition of High Immunity CRC using the CIRC Signature 80

3.1 Introduction 81

3.2 Results 84
3.2.1 Patient characteristics 84
3.2.2 Validation of the CIRC as an immunity signature 87
3.2.3 Variability of the CIRC signature amongst MSS patients 90
3.2.4 Utilising the MSI Template from CMScaller to define MSS-hi-CIRC 93
3.2.5 Use of an MSI-H based cut-off to define MSS-hi-CIRC 96
3.2.6 Determining an 'MSS-hi-CIRC template' 100
3.2.7 Application of Random Forest to classify MSS-hi-CIRC 106
3.2.8 Cross-platform testing of the classifier 108
3.2.9 The Definition of MSS-hi-CIRC 114
3.2.10 Comparison of clinical characteristics between MSS-hi-CIRC, MSI-H and MSS-lo-CIRC 117
3.2.11 Condensing the CIRC to an immunohistochemistry stain 122
3.2.12 MSS-hi-CIRC patients have a comparable immune infiltrate to MSI-H patients 130
3.3 Discussion 133
Chapter 4: Investigating the Genetic Basis of MSS-hi-CIRC CRC 136
4.1 Introduction 137
4.2 Results 140
4.2.1 MSS-hi-CIRC and MSS-lo-CIRC cancers display similar levels of mutational burden 140
4.2.2 MSS-hi-CIRC patients display no specific enrichments of mutation signatures 143
4.2.3 MSS-hi-CIRC and MSS-lo-CIRC share recurrently mutated genes 146
4.2.4 MSS-hi-CIRC and MSS-lo-CIRC are indiscernible in terms of copy number alterations 149
4.2.5 Three mitochondrial SNPs were differentially enriched between the MSS-hi-CIRC and MSS-lo-CIRC cohorts 152
4.3 Discussion 158
Chapter 5: Investigating Microbial Adjuvantisation of Intratumoural Immunity in MSS-hi-CIRC CRC 163
5.1 Introduction 164
5.2 Results 168
5.2.1 T_H17 cell abundance differs between MSS-hi-CIRC, MSS-lo-CIRC, and MSI-H 168
5.2.1 MSI-H cancers have a slight increase in overall bacterial abundance 172
5.2.2 Pathway analysis indicates increased responses to Gram-negative bacteria in the MSS-hi-CIRC cohort 176
5.2.3 Response to reactive oxygen species is different between MSS-hi-CIRC and MSS-lo-CIRC 179

5.2.4 MSS-hi-CIRC does not exhibit increased overall abundance of Gram-negative bacteria ... 182
5.2.5 MSS-hi-CIRC patients exhibit a distinct microbial profile enriched in *Gammaproteobacteria* ... 184
5.3 Discussion ... 190
Chapter 6: Bioinformatic Investigation of MSS-hi-CIRC Biology in other Low-Mutation Burden Cancers ... 195
6.1 Introduction ... 196
6.2 Results ... 198
6.2.1 TCGA-STAD and TCGA-UCEC demonstrate similar percentages of MSS-hi-CIRC patients to CRC patients ... 198
6.2.2 No differences in mutational burden exist between MSS-hi-CIRC and MSS-lo-CIRC in both TCGA-STAD and TCGA-UCEC cohorts ... 201
6.2.3 Mutation signatures did not alter between MSS-hi-CIRC and MSS-lo-CIRC cohorts ... 203
6.2.4 MSS-hi-CIRC patients in the TCGA-STAD cohort demonstrate significant enrichment of microbes ... 206
6.2.1 MSS-hi-CIRC patients in the TCGA-UCEC cohort demonstrate significant enrichment of microbes ... 210
6.2.2 TCGA-UCEC MSS-hi-CIRC patients demonstrate increased responses to reactive oxygen species ... 215
6.3 Discussion ... 220
Chapter 7: Overall Discussion ... 223
7.1.1 The lack of MSS-hi-CIRC in the recurrence setting ... 224
7.1.2 Mechanistic drivers of the MSS-hi-CIRC grouping ... 228
7.1.3 The link with obesity ... 231
7.1.1 The MSS-hi-CIRC hypothesis ... 234
7.1.2 Future directions ... 238
Chapter 8: Collaborative Statement ... 240
Chapter 9: Reference List ... 241
Chapter 10: Appendices ... 262

List of Figures

Figure 1.1 – The spectrum of mutational burden load in various cancers 8

Figure 1.2 – The crypt organisation of normal colonic mucosa 15

Figure 1.3 – Genetic and epigenetic changes that occur in CIN and MSI routes to CRC carcinogenesis 32

Figure 2.1 – Combat batch correction for sequencing platform in the TCGA-COADREAD cohort..... 51

Figure 2.2 – Delineation of the number of mutation signatures present in the TCGA-COADREAD cohort by Negative Matrix Factorisation decomposition 54

Figure 2.3 – Combat batch correction for study in the Affymetrix cohorts 59

Figure 2.4 – The main steps taken in InForm analysis of multispectral immunofluorescence images 67

Figure 2.5 – The main steps in Visiopharm analysis of singleplex IHC data 78

Figure 3.1 – Relationships between the CIRC gene signature and other immune meta-gene signatures..... 88

Figure 3.2 – CIRC enrichment scores across MMR status 91

Figure 3.3 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and the MSI-H template 95

Figure 3.4 – CIRC enrichment scores across phenograph clusters of TCGA-COADREAD MSI-H patients 98

Figure 3.5 – UMAP projections of TCGA-COADREAD subtypes and CIRC enrichment scores..... 99

Figure 3.6 – Differential gene expression between temporarily assigned MSS subtypes in the TCGA-COADREAD cohort 101

Figure 3.7 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and the 'MSS-hi-CIRC template' 105

Figure 3.8 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and a RandomForest classifier 107

Figure 3.9 – CIRC enrichment scores of TCGA-STAD subtypes as defined by MMR status and a RandomForest classifier 109

Figure 3.10 – CIRC enrichment scores of Affymetrix CRC subtypes as defined by MMR status and a RandomForest classifier 111

Figure 3.11 – CIRC enrichment scores of Affymetrix CRC subtypes as defined by MMR status and a RandomForest classifier trained on differentially expressed genes in the GSE39582 cohort 113

Figure 3.12 – CIRC enrichment scores of Affymetrix subtypes as defined by MMR status and an MSI-H derived CIRC enrichment score cut-off 115

Figure 3.13 – CIRC enrichment scores of Agilent subtypes as defined by MMR status and the average number of MSS-hi-CIRC patients found in TCGA-COADREAD and Affymetrix datasets 116

Figure 3.14 – Kaplan-Meier curves of relapse-free survival stratified by CRC subgroups 121

Figure 3.15 – Frequency of overlapping differentially expressed genes between MSS-hi-CIRC and MSS-lo-CIRC in Affymetrix and TCGA-COADREAD datasets..... 123

Figure 3.16 – ROC curve demonstrating the accuracy of HLA Class II genes to predict MSS-hi-CIRC patients from MSS-lo-CIRC patients in the TCGA-COADREAD cohort 126

Figure 3.17 – Relationships between MHC Class II protein and MHC Class II transcript or CIRC score 128

Figure 3.18 – The range of MHC Class II staining in MSS colorectal cancer in the Wales Cancer Bank cohort..... 129

Figure 3.19 – Representative multispectral immunofluorescence images of CRC subtypes..... 131

Figure 3.20 – Principal component analysis of immune cell densities characterised by multispectral immunofluorescence..... 132

Figure 4.1 – Non-synonymous mutation burdens of TCGA-COADREAD patient subgroups..... 141

Figure 4.2 – Frameshift mutation burdens of TCGA-COADREAD patient subgroups 142

Figure 4.3 – Hierarchical clustering of the five mutation signatures found in the TCGA-COADREAD subtypes 144

Figure 4.4 – Schematic of the *APC* gene demonstrating the frequency of mutations in TCGA-COADREAD MSS subgroups 147

Figure 4.5 – Schematic of the *TP53* gene demonstrating the frequency of mutations in TCGA-COADREAD MSS subgroups 148

Figure 4.6 – Degree of chromosomal instability in the genomes of TCGA-COADREAD subgroups . 150

Figure 4.7 – The frequency of chromosomal copy number alterations across TCGA-COADREAD subtypes 151

Figure 4.8 – Autosomal SNP enrichment analysis between TCGA-COADREAD MSS-hi-CIRC and MSS-lo-CIRC patients 153

Figure 4.9 – Mitochondrial SNP enrichment analysis between TCGA-COADREAD MSS-hi-CIRC and MSS-lo-CIRC patients 154

Figure 4.10 – CIRC enrichment scores across mitochondrial SNP homozygosity status in TCGA-COADREAD MSS patients 157

Figure 5.1 – Two-dimensional hierarchical clustering of the cellular density of phenotypes using multispectral immunofluorescence 169

Figure 5.2 – Cellular densities of T-cell subsets across subgroups using multispectral immunofluorescence 170

Figure 5.3 – A comparison of T_H17 geneset enrichment scores across TCGA-COADREAD subgroups 171

Figure 5.4 – A comparison of the percentage abundance of microbes across Genomics England subgroups 173

Figure 5.5 – A comparison of bacterial presence between CRC subgroups 174

Figure 5.6 – A comparison of bacterial and viral diversity indexes across the subgroups in the Genomics England Dataset 175

Figure 5.7 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC CRC patients using GO database search terms related to microbes 177

Figure 5.8 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-COADREAD cohort using pathways relating to the production of ROS (GO:0072593, GO:2000379, GO:1903428) or RNS (GO:2001057) 180

Figure 5.9 – Relationships between the intensity of HLA Class II and 4HNE IHC staining 181

Figure 5.10 – A comparison of the percentage abundance of bacteria according to Gram status between the subgroups in the Genomics England Dataset 183

Figure 5.11 – Frequency of overlapping differentially enriched microbes between MSS-hi-CIRC vs MSI-H and MSS-hi-CIRC vs MSS-lo-CIRC comparisons in the Genomics England dataset 185

Figure 5.12 – Sankey plot demonstrating the phyla of differentially enriched microbes in MSS-hi-CIRC compared to MSS-lo-CIRC 188

Figure 5.13 – Sankey plot demonstrating the phyla of differentially enriched microbes in MSS-hi-CIRC compared to MSI-H 189

Figure 6.1 – CIRC enrichment scores across phenograph clusters of MSI-H patients from TCGA-STAD and TCGA-UCEC cohorts 199

Figure 6.2 – CIRC enrichment scores of TCGA-STAD and TCGA-UCEC subtypes as defined by MMR status and CIRC enrichment score thresholds 200

Figure 6.3 – Non-synonymous mutation burdens of patient subgroups in TCGA-STAD and TCGA-UCEC cohorts 202

Figure 6.4 – Hierarchical clustering of the three mutation signatures found in the TCGA-STAD subtypes 204

Figure 6.5 – Hierarchical clustering of the three mutation signatures found in the TCGA-UCEC subtypes 205

Figure 7.1 – Schematic denoting the theory of MSS-hi-CIRC 235

List of Tables

Table 1.1 – Non-exhaustive list of human Pattern Recognition Receptors (PRRs) 24

Table 1.2 – Criteria for separating patients based on microsatellite instability loci 36

Table 1.3 – Genes within the initial gene list, including those which formed the Coordinate Immune Response Cluster (CIRC) 42

Table 2.1 – Meta data for the datasets used in this thesis 48

Table 2.2 – Frequency of patients for each MMR status 49

Table 2.3 – Number of patients in each subgroup with SNP arrays completed on paired blood 56

Table 2.4 – Primary antibodies used for multispectral immunofluorescence 64

Table 2.5 – Phenotypes delineated from the combination of multispectral immunofluorescence stains 66

Table 2.6 – Forward and reverse primers for the detection of microsatellite instability 72

Table 2.7 – PCR conditions for the detection of microsatellite status 73

Table 2.8 – Primary antibodies for singleplex staining 75

Table 2.9 – Leica BondMax Autostainer Reagents 76

Table 3.1 – Clinicopathological characteristics of the COADREAD datasets 85

Table 3.2 – Relationships between the CIRC signature and signatures of cell types from the ConsensusTME 89

Table 3.3 – Measures of variability between MSS and MSI-H patients 92

Table 3.4 – The 'MSS-hi-CIRC' template 103

Table 3.5 – Clinicopathological characteristics of COADREAD datasets stratified by threshold-derived subtypes 118

Table 3.6 – Multinomial logistic regression of clinicopathological characteristics between CRC subtypes 119

Table 3.7 – List of the top 10 most differentially expressed genes across TCGA-COADREAD and Affymetrix cohorts 124

Table 4.1 – Cosine similarity values between TCGA-COADREAD mutation signatures and Alexandrov signatures 145

Table 4.2 – Percentage of patients demonstrating homo- or heterozygosity at significant SNPs .. 155

Table 5.1 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H CRC patients using GO terms relating to microbes 178

Table 5.2 – List of the top 20 differentially enriched microbes common to MSS-hi-CIRC vs MSS-lo-CIRC and MSS-hi-CIRC vs MSI-H comparisons in the Genomics England dataset 186

Table 6.1 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC gastric cancer patients using GO terms relating to microbes 207

Table 6.2 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H gastric cancer patients using GO terms relating to microbes 209

Table 6.3 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC endometrial cancer patients using GO terms relating to microbes 211

Table 6.4 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H endometrial cancer patients using GO terms relating to microbes 213

Table 6.5 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC gastric cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS) 216

Table 6.6 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H gastric cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS) 217

Table 6.7 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC endometrial cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS) 218

Table 6.8 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H endometrial cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS) 219

List of Abbreviations

AIG	Autoimmune Chronic Gastritis
ANICCA	Assessing Nivolumab In class II expressing MSS Colorectal Cancer
APC	Adenomatous Polyposis Coli
ASC	Apoptosis-associated Speck-like protein containing a CARD
ATP	Adenosine Triphosphate
AUC	Area Under Curve
BCG	Bacillus Calmette-Guerin
BMI	Body Mass Index
CAR	Chimeric Antigen Receptor
CD	Cluster of Differentiation
CIN:	Chromosomal Instability
CIRC:	Co-ordinate Immune Response Cluster
CMS:	Consensus Molecular Subtypes
COADREAD	Colorectal Adenocarcinoma Rectal Adenocarcinoma
COAD	Colorectal Adenocarcinoma
CRC:	Colorectal Cancer
CQN	Conditional Quantile Normalisation
CRCSC	Colorectal Cancer Subtyping Consortium
CRIS:	CRC Intrinsic Subtypes
CT:	Tumour Core
CTLA4:	Cytotoxic T-Lymphocyte-Associated Protein 4
DAB	Diaminobenzene
DAMP	Danger-Associated Molecular Pattern
DNA:	Deoxyribonucleic Acid
DUOX	Dual Oxidase
EGF:	Epidermal Growth Factor
EGFR:	Epidermal Growth Factor Receptor
ELISA	Enzyme-linked Immunosorbent Assay
EMSA	Electrophoretic Mobility Shift Assay
eQTL	Expression Quantitative Trait Loci
ETBF	Enterotoxigenic Bacteroides Fragilis
FC	Fold Change
FDA	Food and Drug Administration
FFPE	Formalin-Fixed Paraffin-Embedded
FMT	Faecal Microbiota Transplantation
FOLFIRI	Folinic acid, fluorouracil, and irinotecan
FOXP3	Forkhead Box P3
FPKM:	Fragments per Kilobase per Million
fRMA	Frozen Robust Multiarray Analysis
GDC:	Genomic Data Commons
GeCIP	Genomics England Commons Information Portal
GEO:	Gene Expression Omnibus
GF	Germ-Free
GI	Gastrointestinal
GISTIC	Genome Identification of Significant Targets In Cancer
GO	Gene Ontology
GSEA:	Gene Set Enrichment Analysis

GSH	Glutathione
GvHD	Graft versus Host Disease
gSNP	Germline Single Nucleotide Polymorphism
HGNC:	HUGO Gene Nomenclature Committee
HLA:	Human Leukocyte Antigen
HRP:	Horseradish Peroxidase
IBD	Inflammatory Bowel Disease
iCAM	Immune Checkpoint Activating Mutation
ICB	Immune Checkpoint Blockade
ICR:	Immune Checkpoint Receptors
ID	Identifier
IFN:	Interferon
IEC	Intestinal Epithelial Cell
IEL	Intestinal Epithelial Lymphocyte
Ig	Immunoglobulin
IHC:	Immunohistochemistry
IL:	Interleukin
IL23R	Interleukin 23 Receptor
IM:	Invasive Margin
IRF	Interferon Regulatory Factor
ISC	Intestinal Stem Cell
ISH	In Situ Hybridisation
KEGG	Kyoto Encyclopaedia of Genes and Genomes
KRAS:	Kirsten Rat Sarcoma Virus
LPL	Lamina Propria Lymphocyte
LPS	Lipopolysaccharide
mAb:	Monoclonal Antibody
MAD:	Median Absolute Deviation
MAPK	Mitogen-Activated Protein Kinase
mCRC:	Metastatic Colorectal Cancer
MDP	Muramyl Dipeptide
MHC:	Major Histocompatibility Complex
MMR:	Mismatch Repair
mRNA:	Messenger Ribonucleic Acid
MSI-H:	Microsatellite Instability (High)
MSI-L:	Microsatellite Instability (Low)
MSS:	Microsatellite Stable
MSS-hi-CIRC:	Microsatellite Stable high CIRC Score
MSS-lo-CIRC	Microsatellite Stable low CIRC Score
NCBI	National Centre for Biotechnology Information
NIH	National Institutes of Health
NFκB	Nuclear Factor κ B
NK	Natural Killer
NLR	NOD-like Receptor
NMF:	Negative Matrix Factorisation
NOD	Nucleotide Binding Oligomerisation Domain Containing 2
NOX	NADPH Oxidase
NSCLC:	Non-Small Cell Lung Carcinoma
OOB	Out of Box

OS	Overall Survival
PAMP	Pathogen-Associated Molecular Pattern
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
PD-1:	Programmed Cell Death Protein 1
PD-L1:	Programmed Death Ligand 1
PD-L2:	Programmed Death Ligand 2
PRR	Pattern Recognition Receptor
qPCR	Quantitative Polymerase Chain Reaction
RCC	Renal Cell Carcinoma
READ	Rectal Adenocarcinoma
RFS	Relapse Free Survival
RNA:	Ribonucleic Acid
RNS	Reactive Nitrogen Species
ROC	Repeater Operator Curve
ROS	Reactive Oxygen Species
RSEM	RNA Sequencing by Expectation Maximisation
RSS	Residual Sum of Squares
RT:	Room Temperature
SCFA	Short Chain Fatty Acid
S-IgA	Secretory Immunoglobulin A
sITH	Spatial Intratumoural Heterogeneity
SNP	Single Nucleotide Polymorphism
SOD	Superoxide Dismutase
STAD	Stomach Adenocarcinoma
TAC	Transit Amplifying Cell
TCGA:	The Cancer Genome Atlas
TCR:	T Cell Receptor
TGFβ	Transforming Growth Factor β
T_H:	T Helper
TIL	Tumour Infiltrating Lymphocyte
TLR	Toll-like Receptor
T_{REGS}:	Regulatory T Cells
TNF:	Tumour Necrosis Factor
TNM:	Tumour-Node-Metastasis
TMNI	Tumour-Node-Metastasis-Immunoscore
TSS	Transcriptional Start Site
UCEC	Uterine Corpus Endometrial Carcinoma
UMAP	Uniform Manifold Approximation Projection
UPC	Universal Expression Code
UV	Ultraviolet
WCB	Wales Cancer Bank
WGS	Whole Genome Sequencing
ZAP-70	Zeta Chain Associated Protein-70

Chapter 1: Introduction

1.1 Introduction

Cancer immunotherapy is now a mainstay in the cancer therapy arsenal. Multiple accounts of prolonged responses are consistently being reported (Hodi et al., 2010, Borghaei et al., 2015), with reduced/more manageable side effects than those described by more conventional therapies. The initial success demonstrated by immunotherapies, such as immune checkpoint blockade (ICB) and adoptive transfer of CD19-specific chimeric antigen receptor (CAR) T-cells, in basic and translational science has accelerated their introduction into the clinic and has now resulted in a plethora of varied and efficacious treatments in development (Zhang and Zhang, 2020). Immunotherapies can be primarily split into two groups, passive or active. Whereas passive immunotherapies, such as antibody immunotherapy, have been used in the clinic for years in the treatment of haematological malignancies or the targeting of specific receptors on solid malignancies (Cetuximab, Herceptin, etc.), active immunotherapies were initially challenging to implement into clinical practice, due to the multiple barriers that an immune response must overcome to be efficacious against cancer.

1.2 Barriers to immune action

Despite previous attempts, arguably the first successful demonstration of active immunotherapy occurred in 1891 by the surgeon William Coley. Coley, having studied the literature, noticed an association between sarcoma patients developing erysipelas, an acute skin infection, and their spontaneous remission. To validate this association, Coley developed a heat-inactivated concoction of gram-positive *Streptococcus* and gram-negative *Serratia marcescens* to 'treat' inoperable soft tissue sarcomas. After injecting the toxin directly into the sarcomas, Coley reported a success rate of higher than 10% (Wiemann and Starnes, 1994). The application of this toxin was later shown to have produced an immune response at the site of injection – killing the bacteria and bystander tumour cells alike (Mellman et al., 2011), and demonstrated the ability of the immune system in eliminating

cancer**.** However, despite Coley demonstrating that immune cells could kill cancer, it was long argued that cancer alone could not induce an immune response. Several lines of evidence suggested the potential for anti-tumour immunity, such as epidemiological studies of patients and immunodeficient mice strains demonstrating a relationship between immunodeficiency and increased tumour incidence [reviewed in (Pardoll, 2003)]. Furthermore, the co-existence of cancer and immune effectors was demonstrated by Rosenberg and colleagues (1988). This experiment demonstrated that cell suspensions from metastases could exert cytolytic activity on primary tumours *in vitro* without the aid of exogenous agents. Rosenberg's study highlighted that not only was cancer immunogenic, but it could utilise mechanisms to inactivate responding immune cells. Despite researchers reporting difficulty in recapitulating the results of this study, other accounts of natural immunogenicity against cancer have been published (Van der Bruggen et al., 1991). Moreover, murine studies have demonstrated a vital role for the immune system in shaping the immunogenicity of tumours. Tumours grown in immunocompromised mice were shown to be highly immunogenic when adoptively transferred to immunocompetent mice, leading to their rapid clearance (Kaplan et al., 1998). Together, these studies, amongst many others, have provided clear evidence for the theory of cancer immunosurveillance and immunoediting. These concepts, broken down into three phases of *Elimination*, *Equilibrium*, and *Escape* state that cancer and the immune system are in constant communication. Initially, as cancer develops it presents its mutated peptidome to the immune system. These antigens are recognised as non-self, eliciting an anti-tumour immune response and elimination of these cells before they develop further. If *Elimination* does not complete, it is succeeded by *Equilibrium*, the second phase and, perhaps, the longest phase of this process. *Equilibrium* involves a Darwinian-esque cycle in which continual growth of cancer cells and subsequent elimination by the immune system shapes the cancer cell pool to demonstrate non-immunogenic phenotypes. Once established, cancer has reached the last phase where it has outwitted the immune system - *Escape*. Here, cancer proliferates and gathers more mutations which

increase its ability to survive (Dunn et al., 2004). Underpinning this, it is now abundantly clear that to exist in an immunocompetent host, cancer must employ mechanisms to evade the host's immune system (Hanahan and Weinberg, 2011).

1.2.1 Targeting resistance to immune action

For an immune response to be deemed successful, it must remove the pathogen but also limit harm to healthy tissue. Regarding the T-cell compartment, the immune system has developed an immunosuppressive arm (T_{REGS}) and a hard-wired cell-surface system in the form of immune checkpoint receptors (ICRs) to ensure appropriate restraint of immune action. ICRs are upregulated on T-cell activation to dampen responses by inhibitory signalling and, after chronic TCR stimulation, to ensure control through two primary mechanisms: T-cell exhaustion and T-cell apoptosis. T-cell exhaustion is characterised by both an upregulation of ICRs, such as Cytotoxic T-Lymphocyte Associated protein-4 (CTLA-4) or Programmed Death-1 (PD-1), and a reduction in cytokine release (Li et al., 2015). In contrast, T-cell apoptosis occurs through Fas-Ligand acting on its cognate receptor, among other mechanisms (Patil et al., 2015). In addition to providing a homeostatic mechanism of control, ICRs are often exploited by cancers to inactivate cancer-specific T-cells.

1.2.1.1 CTLA4

CTLA-4 was, arguably, the first described ICR. Expressed exclusively on T-cells, CTLA-4 binds to the same ligands as the co-stimulatory receptor CD28, albeit to a much higher affinity. Although the downstream pathways of CTLA-4 action are still yet to be fully elucidated, the most robust evidence suggests that CTLA-4 binds ZAP-70 and SYK to prevent tyrosine phosphorylation of the T-cell receptor (TCR) (Walker and Sansom, 2015). Therapeutic targeting of CTLA-4 is limited to the receptor, rather than including its ligands, due to the promiscuity of the ligands in producing both stimulatory and inhibitory responses. Despite this, therapeutically targeting CTLA-4 has shown dramatic success in

advanced cancers with Ipilimumab representing the first compound targeting immune checkpoints to enter the clinic (Thallinger et al., 2018). Initially, Ipilimumab was assumed to act by blocking the ability of CTLA-4 to sequester CD28 ligands from providing the second signal for effector T-cell activation. However, subsequent experiments have shown that CTLA-4 molecules are vastly increased on T_{REGS} in comparison to effector populations. Furthermore, Ipilimumab, as an IgG1 antibody, has a very high affinity for Fc receptors used to remove cells by antibody dependent cell-mediated cytotoxicity (Arce Vargas et al., 2018). Finally, murine studies demonstrated a reduced T_{REG} infiltration post-Ipilimumab and combined, these results demonstrate that Ipilimumab may act by removing inhibitory T_{REGS} from the microenvironment rather than inhibiting CTLA-4-mediated ligand sequestration. However, these results are yet to be recapitulated in human studies, with no difference in T_{REG} density pre- and post-Ipilimumab treatment in melanoma, bladder, and prostate cancer (Sharma et al., 2019).

1.2.1.2 PD-1

PD-1 is a type I transmembrane protein consisting of an Ig-like extracellular region and a cytoplasmic tyrosine-based inhibitory motif acting to inhibit ZAP-70, a CD3ζ chain-binding protein required to potentiate the TCR signal (Sheppard et al., 2004). PD-1 has two cognate ligands, PD-L1 and PD-L2, both of which belong to the B7 superfamily. Whereas PD-L1 is expressed ubiquitously on haematopoietic and non-haematopoietic tissue alike, PD-L2 is more restricted in its expression to haematopoietic tissue (Latchman et al., 2001). Expression of PD-1 ligands, however, is not limited to healthy tissue, with expression by tumours reported in a wide variety of malignancies. Tumours utilising this mechanism of control target and inactivate T-cells in their immediate vicinity, providing a means of immune evasion. Therefore, perhaps unsurprisingly, therapeutically targeting PD-1 or its ligands are of great clinical interest. Currently, the FDA has approved five antibodies that

therapeutically target PD-1 or its ligands in cancer, but there are many more in clinical trials (Syed, 2017, Alsaab et al., 2017, Kim, 2017).

1.2.1.3 The current state of checkpoint blockade in cancer

Collectively, the targeting of ICRs, such as CTLA-4 and PD-1, or their ligands is known as ICB. While these antibodies have transformed the outcome of a subset of patients, there remains a group of patients who quickly relapse due to acquired resistance, or upon whom ICB has little to no impact. Therefore, pivotal to the field is the synergism of ICB with other therapies, including other immunotherapies to maximise efficacy. In the Checkmate 012 trial, researchers investigated the synergism between Nivolumab (anti-PD-1) and Ipilimumab (anti-CTLA-4) as first-line therapy in non-small cell lung carcinoma (NSCLC) and demonstrated significant efficacy over single-agent therapy. The combination was first approved for first-line treatment of metastatic melanoma in 2015, it has since been approved in renal cell carcinoma and is currently being evaluated in several Phase III trials for NSCLC (Motzer et al., 2018, Hellmann et al., 2017). Although the authors caveated the Checkmate 012 trial by suggesting it was completed on patients with high PD-L1 expression and therefore performed on those expected to respond, more recent evidence suggests that PD-L1 is less effective at predicting response to ICB than first thought. Inconsistencies in the PD-L1 positivity cut-off value across trials and the realisation that the cytokine microenvironment profoundly influences PD-L1 expression, have all dampened PD-L1's status as a biomarker of ICB success. Few predictive biomarkers have been validated to predict response to ICB; however, one that has consistently separated patients into responders and non-responders is a high mutational load or a deficiency in the mismatch repair (MMR) system (Thallinger et al., 2018).

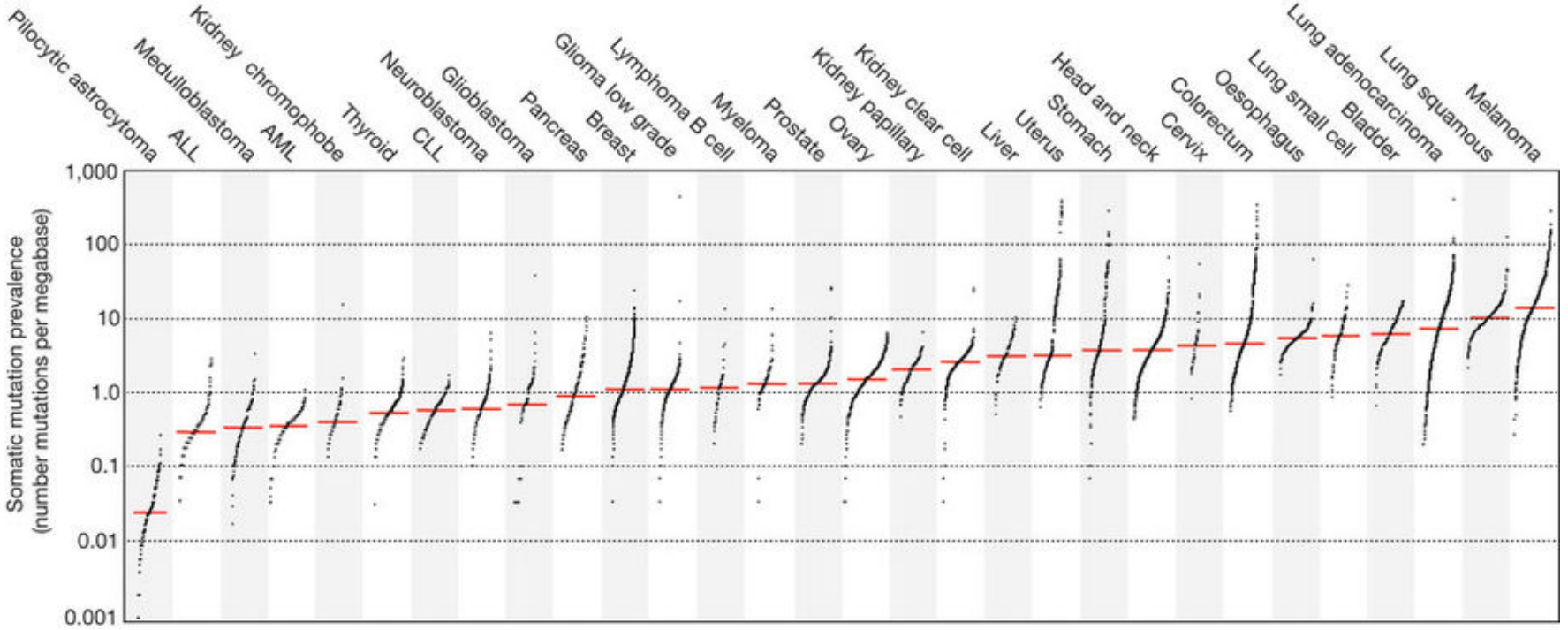
1.3 Immunogenicity and genetics of cancer

Natural adaptive immunity relies on the presentation of peptides derived from intracellular proteins by nucleated cells via major histocompatibility complex (MHC)-I molecules or the capability of specialised antigen presenting cells to process and present extracellular antigens via MHC-II to T-cells (Blum et al., 2013). With regards to MHC-I-restricted peptides, there is a highly diverse repertoire of self-peptides that a cell can present, to which self-tolerance is established, such that these antigens won't induce an immune reaction (Bianconi et al., 2013). However, non-self-antigens which have an altered sequence, such as those of bacterial origin, will produce a response.

Critical to the field of tumour immunology is the genetic instability that characterises cancer. Genetic instability, arising from inaccurate duplication of genetic data in cells under severe replication stress, results in a plethora of mutations which may be subsequently processed and presented as neoantigens. Therefore, one of the most significant predictors of an immune response in cancer is a high mutational load. When compiling mutation loads for various cancers, Alexandrov *et al*. (2013) found there was a high variability between and within tumours. While tumours which are chronically exposed to mutagenic agents, such as lung (tobacco smoke) and melanoma (UV radiation), were highest regarding mutation burden, they still displayed a high variability among patients (Figure 1.1).

Figure 1.1 – The spectrum of mutational burden load in various cancers

Mutation burdens are plotted for 30 different cancers. Each dot represents a patient with the red line representing the median number of mutations in each cancer type. The vertical axis represents the number of mutations per megabase (log transformed). The horizontal axis shows cancer types ordered by median mutation rates with the highest burden being seen in Melanoma. Figure and legend were adapted from Alexandrov, et al. (2013).



Perhaps unsurprisingly, the cancers which showed the highest mutation burden (melanoma and lung cancer) also display the highest sensitivity to ICB, with these therapies becoming front-line in these cancers (Alsaab et al., 2017). Attempts have been made to establish a mutation burden cut-off that predicts patient response to checkpoint blockade, to expand the cancer types that may usefully receive ICB. Rizvi *et al.* (2015) separated lung cancer patients into high-mutation burden and low-mutation burden using a cut-off of 200 nonsynonymous mutations and demonstrated prognostic differences, however, this cut-off has not been validated clinically and only represents a median number of mutations within this cohort. Secondly, Panda *et al*., (2017) attempted to find a mutation burden threshold by investigating whole-exome sequencing and RNA sequencing data from pan-cancer TCGA patients. Out of the 33 solid cancers investigated, 8 had mutation burden thresholds associated with signatures of immune activation. Patients with scores above these thresholds, termed immune checkpoint-activating mutation (iCAM) thresholds by the authors, were shown to have histological evidence of immune infiltration and a significantly better response to ICB therapy than those with lower scores. Importantly, iCAM thresholds could be predicted using pre-existing clinical sequencing assays, and therefore could directly impact clinical practice. However, despite this, iCAM thresholds were largely found to be surrogates for measures already used to determine whether one is eligible for ICB or not, such as microsatellite status for CRC. Although iCAM thresholds suggested more patients may be eligible for ICB therapy, there is still no established mutation burden cut-off used in clinical practice (Panda et al., 2017).

In addition to a comprehensive investigation of mutational load across multiple cancers, Alexandrov *et al*. (2013) devised an algorithm to extract mutational signatures from whole genome data. Hypothesising that cancers occurring by a similar process will give rise to distinct mutation patterns and utilising both the base change and the flanking nucleotides either side of the alteration, they investigated the frequency of 96 possible single-base changes. Initially, 21 distinct signatures could

be delineated. However, this was subsequently extended and validated to 29 signatures each with distinct aetiologies and frequencies in different cancers (Alexandrov et al., 2015). Despite not being evaluated concerning a response to ICB, one would hypothesise that patients with comparable mutational signatures may respond similarly, however, other factors such as MHC haplotype, the microbiome, etc., are likely to be significant. Although this is unlikely to influence clinical practice directly, due to the expense and time-consumption of sequencing complete genomes, the stratification of patients by mutational signatures may highlight some unknown processes that give rise to an ICB-sensitive tumour.

Although a high mutational burden is a strong predictor of the efficacy of immunotherapy, it is not the sole factor that impacts immunogenicity and, therefore, response to immunotherapies. Defective priming and antigen presentation, local immunosuppression by extracellular metabolites, and gut flora alterations have all been shown to impact responses to immunotherapies (Routy et al., 2018). Therefore, perhaps one of the most pressing issues in cancer immunology is to understand these other factors better, with a view to extending the success of immunotherapies to the low-mutation burden setting.

1.3.1 Using CRC as a model for investigating the immunobiology of low-mutation burden cancer

One cancer in which immunotherapy has produced mixed results is CRC. Considering its high status concerning overall mutation burden (Figure 1.1), CRC also shows one of the highest variabilities. While there are patients with very high mutation burdens, there are also patients with very low levels of mutations, likely arising from the distinct pathways in CRC carcinogenesis. Importantly, ICB has been shown to be successful in a distinct group of patients with microsatellite instability-high (MSI-H) cancers in comparison to the low mutation burden, microsatellite stable (MSS) and

chromosomal instable (CIN) patients (Le et al., 2015). Both success and failure of ICB at the same anatomical site means CRC is an attractive human system in which to study the immunobiology of both settings. Other cancers also display deficiencies in MMR, most notably endometrial and gastric cancer, where they account for ~10% of tumours and share the objective response to ICB seen in MSI-H CRC. In fact, a subsequent investigation by Le *et al*. (2017) assessed the ability of MMR status to predict response to ICB across 12 separate tumour types, including endometrial and gastric cancers, and demonstrated that MSI-H cancers objectively respond better than MSS cancers to ICB. However, whilst collectively CRC MSS patients are deemed non-responders to ICB, 11% of CRC MSS patients have been shown to respond. Previous studies completed by this group (Lal, 2016), and others (Galon et al., 2006, Mlecnik et al., 2016), have demonstrated that intratumoural immunity is highly variable within the MSS grouping, meaning CRC is an appealing model to investigate novel routes to increased immunity within low-mutation burden cancer. There is clearly a strong link between high mutational burden, strong intratumoural immunity, and response to ICB. However, it is also clear that there is great heterogeneity in intratumoural immunity, even in the low-mutation burden settings. The drivers of immunity in such settings are unclear, however, understanding the basis for this will be important for extending the success of ICB/immunotherapy beyond high mutation burden settings – an issue that is highly pertinent to multiple tumour types.

1.4 Colonic anatomy and function

The human colon is part of the gastrointestinal (GI) tract, which begins at the mouth, finishes at the rectum and functions in the final stages of the digestive process. It serves four main roles; reabsorption of water, production and reabsorption of vitamins, movement of faeces towards the rectum, and to provide a home for an estimated 100 trillion microbial cells constituting the microbiome, which are vital for metabolism of polysaccharides, lipids, and peptides that pass

undigested through the small intestine. Anatomically, the colon is split into six main parts which make up its characteristic 'broken-rectangle' shape. The colon begins at the cecum, a pouch-shaped structure connected to the appendix. Although the appendix derives from the same embryological origin as the cecum, it isn't considered a part of the large intestine as it does not aid in digestion. The next section of the colon, the ascending colon, travels up the right side of the abdomen, connecting to the transverse colon at the hepatic flexure. The cecum, ascending colon, hepatic flexure, and the proximal two thirds of the transverse colon make up the right-side of the colon and cancers which manifest here are collectively known as right-sided cancers. The transverse colon curves at the splenic flexure emptying into the descending colon. This structure travels down the left side of the abdomen reaching the S-shaped sigmoid colon, where it meets the rectum and anal canal. These structures are also grouped together and cancers manifesting here are collectively known as left-sided cancers. A common embryological ancestor is one of the main reasons that splits the structures into being right and left, with those structures manifesting on the right-side of the colon originating from the midgut and those on the left arising from the hindgut (Mik et al., 2017).

The various segments of the colon are tasked with specific jobs. At the point of entry into the colon, most of the nutrients and up to 90% of the water has already been reabsorbed by the small intestine. Despite this, the main function of the ascending colon is to reabsorb water. This concentrates any undigested material into stool for storage in the distal colon. In mammalian species, water is largely reabsorbed through osmosis of which the gradient for is set up by the selective absorption of sodium ions via channels. These channels show a heterogeneous distribution, being in abundance in the proximal colon and decreasing along its length thus restricting water reabsorption to the proximal colon. In a reverse pattern, the mucus layer which coats the lumenal-most layer (mucosa) is much denser as the colon progresses. This is for two main reasons. Firstly, to protect the colonic epithelium

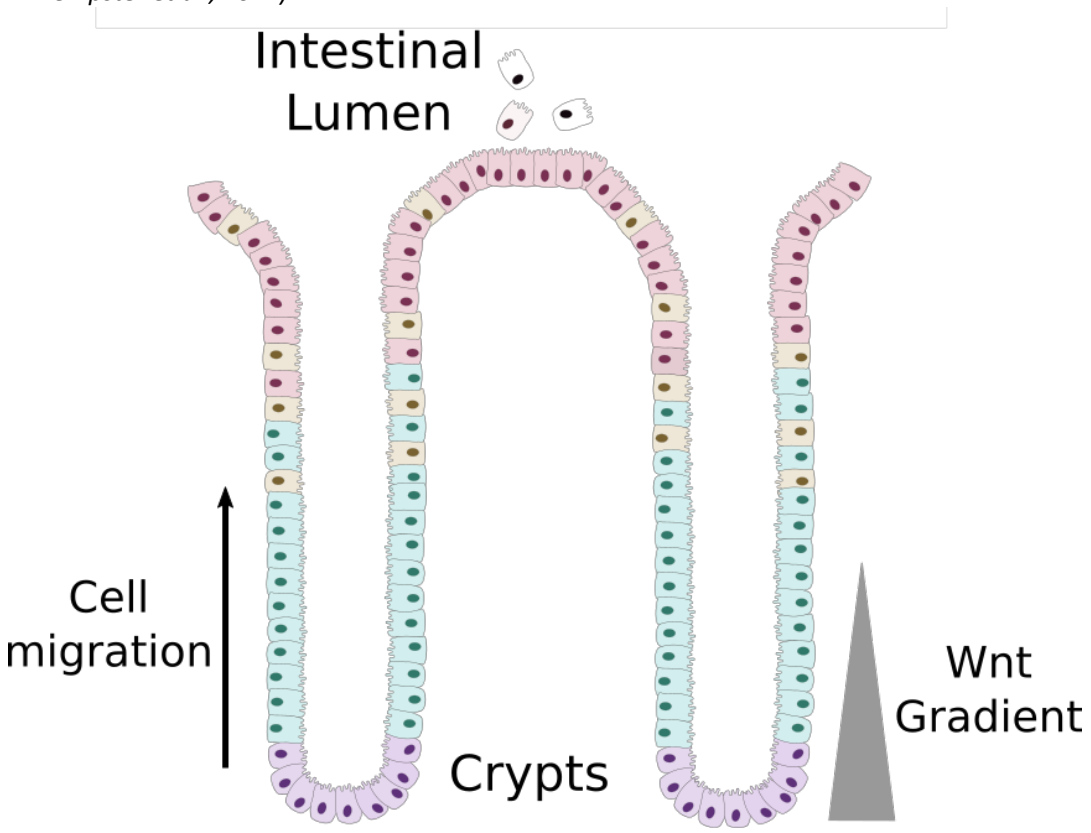
from the rough food particles as they harden along the length, and secondly, to provide a protection from the increased number of bacterial species living in the distal colon.

1.4.1 Microscopic anatomy

The colon is organised into four main layers: the mucosa, submucosa, a muscle layer, and the serosa. Most lumenal of these is the colonic mucosa. This layer is further divided into two sublayers; a single layer of columnar epithelial cells which face the colonic lumen, and an underlying lamina propria. The lamina propria is loose connective tissue, housing many different cell types including those from the immune system such as B-cells, T-cells, and antigen-presenting cells. Whilst the colonic epithelial layer has a much lower immune presence than the lamina propria, it also contains lymphocytes which home to this region without antigen-specific activation. Unlike the small intestine, where the epithelium is organised into villi and crypts, the colon only has deep glands called the crypts of Lieberkühn and is interspersed by flat epithelium (Figure 1.2). At the base of such crypts reside the intestinal stem cells (ISCs), characterised by LGR5 expression (de Sousa e Melo et al., 2017), which are responsible to support the rapid turnover of cells in the colon. ISCs maintain high cell turnover by giving rise to transit amplifying daughter cells (TACs), which undergo rapid proliferation and migrate up the crypt wall as they differentiate into all the cell types of the colon. ISCs are few, estimated at around 10 per crypt, however, TACs can make up around 25% of cells in the crypt. In terms of differentiated cells, there are three major types in the colon. The fewest in number, comprising ~1% of differentiated cells, are the enteroendocrine cells. These cells differentiate directly from the stem cell compartment rather than the TACs, and function to secrete hormones. In contrast, colonocytes, the water-absorbing cells, are the most abundant cell type and account for ~75% of differentiated cells. Goblet cells are the last major cell type present in the colon and function to release mucus to ease the passage of faeces through the colon, and account for ~24% of differentiated cells (Ingham-Dempster et al., 2017). Mucus-producing goblet cells are not only vital to easing the passage of

faeces, but also to maintaining a physical distance between bacteria and the epithelium. Microbes increase in density along the length of the colon, which is reflected by a thickness/viscosity of the mucus layer, and by extension an increasing abundance and activity of Goblet cells along the length of the colon. These microbes, collectively termed the colonic microbiome, function to digest otherwise indigestible material, providing a source of energy in exchange for hospitality in the colon.

Figure 1.2 – The crypt organisation of normal colonic mucosa
Schematic of the normal organisation of 'healthy' colonic mucosa. Intestinal stem cells (purple) reside at the base of the crypt. Wnt, and other factors, are secreted by cells surrounding the crypt (not shown), generating a high density of Wnt within the crypt and gradient along the crypt axis. The ISCs are slow dividing and maintain the transit amplifying cell (TAC) population (cyan) which are responsible for cell replacement. As TACs migrate up the crypt towards the villi they differentiate into either colonocytes (red) or goblet cells (yellow). Enteroendocrine cells (not shown) are repopulated directly from the stem cell compartment. Dying or dead cells (pale red, or no colour) are shed outside of the crypt through a process referred to as anoikis. The arrow depicts the direction of cell migration as cells differentiate. Wnt gradient is depicted by grey triangle, where the highest Wnt concentration is at the base and reduces across the axis (Ingham-Dempster et al., 2017).



1.5 The colonic microbiome

The human microbiome refers to all species of microorganisms, bacteria, viruses, protozoa, and fungi colonising the skin, oral cavity and GI tract, placenta, and vagina. In 2008, the NIH Human Microbiome Project was launched to identify all members of the human microbiome and to investigate whether a common 'healthy' microbiome existed. Using high throughput multi-omic approaches, microbial communities were delineated at 18 different body sites in 300 healthy individuals at a single time point and generated over 32 terabytes worth of data (Human Microbiome Project Data Portal, 2021, The Human Microbiome Consortium, 2012a, The Human Microbiome Consortium, 2012b). These data demonstrated that each microbial site had distinct microbial communities and that the microbiomes of two healthy people differed vastly from one another. Interestingly, over the sampling period, the oral cavity microbiome changed the most drastically whereas at other sites, such as the vagina and colon, the microbiome seemed to be more constant (Ding and Schloss, 2014). The microbiome has become a 'hot topic' in biology, likely due to the increasing associations reported with various disease states. In fact, at the time of writing, a PubMed search for "microbiome" yields 86,320 results with ~25% of those arising between 2020 and 2021 (PubMed, 2021). It seems now more than ever scientists are interested in our invisible friends.

In addition to the body sites housing varying compositions of microorganisms, they also differ in abundance (The Human Microbiome Consortium, 2012b). The colon is home to between 10^{10} and 10^{12} microbial cells, many orders of magnitude more than anywhere else in the body, and these express over 165 times the number of genes than expressed by the human genome. Interestingly, 90% of the gut microbiota could be classified by just two bacterial Phyla – *Firmicutes*, and *Bacteroidetes* (Arumugam et al., 2011, Rinninella et al., 2019). Given the bacterial kingdom's sheer abundance in the colon, it is no surprise that it is also the best studied.

1.5.1 Functions of the colonic bacteria

Colonic bacteria primarily reside in the outer mucus layer and the lumen. Here they maintain a symbiotic relationship with the host, providing metabolic support in exchange for nutrients in the form of undigested/indigestible dietary material or epithelial cells shed by anoikis. Their major substrate - indigestible carbohydrates, is primarily processed by saccharolytic fermentation to synthesise short-chain fatty acids (SCFAs). Although the proportions of SCFAs vary greatly depending on the specific composition of the gut microbiota; acetate, butyrate, and propionate almost always comprise the majority (Davis-Richardson and Triplett, 2015). Despite formate and lactate also being produced during fermentation, little is known of the action of formate in the colon, and lactate is often further metabolised to acetate, butyrate, or propionate by cross-feeding bacteria (Morrison and Preston, 2016). Whilst acetate is widely produced by multiple species of bacteria, the production of the other main SCFAs (propionate and butyrate) seem to be more exclusive to specific bacterial genera. Unfortunately, knowledge of the specific species producing propionate is currently lacking, however, butyrate is known mainly to be produced by *clostridia*, *eubacteria*, and *roseburia* microbes (Reichardt et al., 2014, Nicholson et al., 2012). All SCFAs are metabolised rapidly in the colon, where they act as a major fuel for colonocytes and constitute ~70% of their energy requirements. Despite representing the minority of the most common SCFAs, butyrate is the 'number one choice' of fuel in the colon with it being preferentially oxidised even when competing fuels such as glucose or glutamine are present (Liu et al., 2018, Roediger, 1980). Butyrate, and other SCFAs, are well reported to have multiple systemic effects (Van Der Hee and Wells, 2021), however, the importance of local SCFA action, in particular butyrate, was demonstrated by Donohoe and colleagues in germ-free (GF) mice. Colonocytes in GF mice were shown to be in an energy-deprived state in comparison to the colonocytes of conventionally raised mice, exhibiting marked decreases in oxidative phosphorylation and ATP production, with concomitant increases in colonocyte autophagy and colonic barrier

breakdown (Donohoe et al., 2011). Interestingly, the addition of butyrate alone was sufficient to rescue mitochondrial dysfunction and prevent autophagy, restoring barrier integrity. Despite the colonic microbiome serving the host by providing energy among other vital functions, colonic microbes still have the capacity to cause infection/intestinal inflammation in case of a barrier breach. However, multiple mechanisms exist to prevent such an interaction between the microbiome and the immune system.

1.5.2 Keeping bacteria at bay

Mechanisms designed to protect from microbial encroachment are largely separated into segregation and mediation. Broadly, "mediation" refers to the cocktail of signals exchanged between gut microbes and host immune cells. These signals, in the form of cytokines or chemokines, can activate T-cell responses which in turn release various cytokines, such as IL-17, which increase production of antimicrobial molecules from colonic epithelial cells to prevent bacterial encroachment. On the other hand, "segregation" refers to the physical or chemical barriers that spatially separate microbes and the immune cells in the underlying lamina propria.

1.5.2.1 Physical and chemical barriers

Colonic epithelium is covered in a dense mucus produced by goblet cells. Secretion of mucus is tightly regulated by the host sensing the presence of microbes or their metabolites such as SCFAs. As such, this layer is much thicker in the colon than in the small intestine to compensate for the increased bacterial presence. Colonic mucus forms a bilayered structure in which the outer layer is more viscous and serves as a habitat to microbes, and the inner is highly structured and largely impermeable. The major constituent of both layers is the highly O-glycosylated Muc2 mucin protein, with the differences in microbe hospitality and viscosity between the layers owing to the net-shape of the Muc2 protein. On release from the goblet cells, Muc2 self-assembles into sheet-like structures.

Moving upwards, Muc2 undergoes proteolytic cleavages by the host or by bacterial species that inhabit the outer mucin layer. These cleavages do not destabilise the mucin protein, but allow it to expand up to 4 times in volume – giving rise to a structure with increased viscosity (Johansson et al., 2008). Although the inner mucin layer is largely impermeable to microbes, in the absence of dietary fibre the selection for Muc2-degrading bacteria occurs. These microbes use the polysaccharide-dense Muc2 as fuel, degrading the epithelial cells' protective barrier and gaining access to the epithelial layer (Desai et al., 2016). In addition to physically restraining microbe movement, the inner mucus layer is filled with various antimicrobial molecules which chemically impede movement. Although the colon lacks the dedicated antimicrobial peptide-making Paneth Cell, the colonic epithelium can produce antimicrobial agents such as defensins and secretory immunoglobulin A (S-IgA). Although expression of defensins is much lower in the colon than the small intestine, they work to induce microbe cell death. Defensins are small, cationic peptides which electro-attract to microbes, which by and large, are always negatively charged. Microbial membranes have high expression of the phospholipids phosphatidylglycerol and phosphatidylserine, two phospholipids carrying a high negative charge due to their acidic nature. Furthermore, these negative charges are further exacerbated by the inclusion of lipopolysaccharide (LPS) in gram-negative bacteria, or abundance of teichoic or teichuronic acids in gram-positive bacteria. In contrast, mammalian bilayers contain phospholipids which have a much more neutral charge, causing antimicrobial agents to repel away from the host and attract to sites of infection. After attraction, defensins integrate into microbial membranes creating pores which 'empty' microbes of their necessities to life, such as metabolites and ions, ensuring decoupling of membrane-coupled respiration and death of the microbe (Cobo and Chadee, 2013).

In contrast to defensins, the colonic mucus is more densely packed with S-IgA antibodies, although still not to the degree of the small intestine. These antibodies, produced by the underlying plasma

cells residing in the lamina propria, are transported to the lumen by transcytosis after binding the polymeric immunoglobulin receptor on the basolateral surface of epithelial cells. Once in the lumen, S-IgA binds exclusively in the less-dense outer mucus layer (Rogier et al., 2014). Here, it is perfectly poised to neutralise any microbe attempting to make epithelial contact. In contrast to other immunoglobulin molecules, S-IgA is perfectly adapted for mucosal defence. Firstly, as mucosal barriers are coated in helpful microbes the translation of microbial presence to the onsite immune system would cause unwarranted immune activation. S-IgA are weak activators of the complement system - the messenger between S-IgA and an immune response. This means that S-IgA can neutralise microbes, but the translation into a potentially chaotic, inflammatory response is prevented. Secondly, the majority of S-IgA, especially in the gut, is cross-reactive meaning that S-IgA can neutralise a wide variety of pathogens (Mak et al., 2014). S-IgA predominantly neutralises bacteria by simply preventing their attachment to the epithelium. The coating of bacteria with S-IgA immobilises bacteria, impeding their ability to invade. This method of immobilisation, rather than direct killing, allows for the S-IgA-coated-microbiome to still function, however, limits the ability of pathogens to infect colonic epithelial cells (Bunker and Bendelac, 2018).

Should the mucus layer be penetrated, and the anti-microbial agents be bypassed, epithelial cells are covered in a dense glycocalyx and express multiple tight junction proteins which physically prevent bacteria from getting too close or invading paracellularly, respectively. Furthermore, increased goblet cell mucus production, defensin and S-IgA production, and increased expression of the tight junction proteins; claudin-1 and Zonula Occludens-1, have been documented on SCFA application (Wang et al., 2012, Willemsen, 2003, Zhao et al., 2018). These mechanisms act as a positive feedback loop which increases barrier function the closer bacteria get.

1.5.2.2 Intestinal Epithelial Lymphocytes (IELs)

Should pathogens breach the barriers, the colon has a dense immune system. In fact, well over half of the cells of the immune system are concentrated in the gut (de Kivit et al., 2014) . Most of these cells reside underlying the epithelium in the lamina propria, however, some are interspersed within the colonic epithelium. These latter cells, referred to as intestinal epithelial lymphocytes (IELs), are almost exclusively antigen-experienced T-cells and act as a first-line defence to any microbes that have made it past the many barriers previously outlined. Colonic IELs consist predominantly of the 'conventional' (αβ TCR and CD8 αβ heterodimer or CD4) subtype and, in adulthood, are mostly induced from naïve T-cells (Mowat and Agace, 2014). These cells, having been primed in secondary lymphoid structures such as the colonic patches or the mesenteric lymph node by $CD103^{POS}$ dendritic cells, gain their colon-homing capacity by predominantly upregulating the integrin αEβ7 (Dzutsev et al., 2017). Once in place, these cells are perfectly poised to provide immediate clearance of any pathogen. However, damage must also be limited. These potent inducers of inflammation are in contact with a fragile barrier, and therefore must demonstrate dual inflammatory and regulatory function. Whilst $CD8^{POS}$ IELs are almost exclusively inflammatory, $CD4^{POS}$ IELs are also present. In the steady state, these cells are almost solely forkhead box P3 (FOXP3) positive T_{REG} cells and release the anti-inflammatory cytokines, TGFβ and IL-10. In fact, even under inflammatory conditions where these CD4 IELs are polarised to a more effector T_H fate and are participating in host immunity, these cell types are capable of producing IL-10 so as to dampen their T_H1 or T_H17 inflammatory behaviour and limit damage to the vulnerable barrier (Cheroutre et al., 2011).

1.5.2.3 Relaying the message

In addition to deploying a physical barrier in the mucus layer to keep microbes away, the colonic epithelium features prevalent expression of pattern recognition receptors (PRRs) responsible for relaying pro-inflammatory or anti-inflammatory signals to the underlying lamina propria immune

cells. These receptors recognise conserved pathogen-associated molecular patterns (PAMPs), or intracellular patterns associated with perturbations in homeostasis termed danger associated molecular patterns (DAMPs). The main families of PRRs with examples of their ligands are shown in Table 1.1. Signals conveyed by PRRs in response to their ligands are dependent on their subcellular localisation within the polarised IEC layer. By growing human IEC cell lines on microporous inserts, one can create a polarised layer of epithelial cells analogous to the human colon. Studies completed by Lee et al., (2006) using these polarised layers highlighted that the application of certain toll-like receptor (TLR) ligands to the basolateral membrane caused an upregulation of the inflammatory cytokine, IL-8, and NF-κB activation - the downstream targets of which are responsible for epithelial homeostasis and repair. Interestingly, when the same ligands were applied on the apical side of the membrane, no release of IL-8 or activation of NF-κB could be measured by ELISA and EMSA, respectively (Lee et al., 2006). Furthermore, when cells were not polarised, IECs reacted similarly to application of ligand on the basolateral membrane potentially owing to the loss of polarisation of IECs on a barrier breach, demonstrating the importance of ligand localisation in determining the appropriate response to commensals versus pathogens.

Perhaps the second most well-described family of PRRs is the NOD-like receptor (NLR) superfamily. Unlike TLRs, which have polarised downstream signalling, NLRs are intracellular sensors and therefore don't require polarisation for activation (Al Nabhani et al., 2017). NLRs can be split into two main groups. The first of these groups function analogously to basolaterally-stimulated TLRs and includes NOD2. NOD2, an intracellular sensor of the Gram-positive and Gram-negative bacterial peptide MDP, was the first described susceptibility locus for Crohn's disease (Hugot et al., 2001, Ogura et al., 2001). It is expressed widely in the human colon in various cell types, including those of the immune system as well as enterocytes and goblet cells. Whilst NOD2 activation by MDP is poorly understood, once activated NOD2 signals via NF-κB, mitogen-activated protein kinases (MAPK), or

interferon regulatory factors (IRFs) to increase barrier function (Ramanan et al., 2014) and induce inflammatory cytokine production (Kobayashi et al., 2005).

In contrast, the second group of NLRs act as molecular sensors, and cooperate with other adaptor proteins to form an 'inflammasome'. These proteins recruit ASC and pro-caspase-1 to form the inflammasome complex. This complex, capable of cleaving pro-caspase 1 into its active form, permits the cleavage of pro-IL-1β and pro-IL-18 into the active pro-inflammatory cytokines, IL-1β and IL-18 (Li and Wu, 2021). Whereas IL-1β has been shown to be capable of inducing T_H17 differentiation, promoting inflammation, IL-18 has been linked to epithelial repair, and the induction of MHC class II on IECs (Van Der Kraak et al., 2021).

Table 1.1 – Non-exhaustive list of human Pattern Recognition Receptors (PRRs)

Condensed from Table 1 and the main text in Li and Wu (2021).

Abbreviations - dsRNA; double-stranded RNA, ssRNA; single-stranded RNA, iE-DAP; γ-D-glu-meso-diaminopimelic acid, MDP; muramyl dipeptide, MyD88; myeloid differentiation factor 88, TRIF; TIR domain-containing adaptor protein-inducing interferon β, RIP2; receptor-interacting serine threonine protein 2, TAK1; transforming growth factor-β-activated kinase 1, NF-κB; nuclear factor κB, MAVS; mitochondrial antiviral signalling protein, TRAF6; tumour necrosis factor receptor-associated factor, TBK1; TANK-binding kinase 1.

Family	***PRR***	***PAMP***	***Sources***	***Signalling Pathway***
Toll-like Receptors (TLRs)	TLR1 (TLR1-TLR2)	Triacyl lipopeptide	Bacteria	MyD88-dependent
	TLR2 (TLR1-TLR2, TLR2-TR6)	Lipoteichoic acid	Bacteria	MyD88-dependent
		Arabinomannan	Mycobacterium	
		Peptidoglycan	Bacteria	
		Zymosan	Fungi	
		Lipoprotein	Mycoplasma	
		Pore Protein	Neisseria	
	TLR3	dsRNA	Virus	TRIF-dependent
	TLR4 (MD-2/CD14)	Lipopolysaccharides (LPS)	Bacteria	MyD88-dependent and TRIF-dependent
		Heat-shock proteins	Host	
	TLR5	Flagellin	Bacteria	MyD88-dependent
	TLR6 (TLR2-TLR6)	Lipoteichoic acid	Bacteria	MyD88-dependent
		Peptidoglycan	Bacteria	
	TLR7	ssRNA	Virus	MyD88-dependent
		Imidazoquinoline	Artificially synthesised	
	TLR8	ssRNA	Virus	MyD88-dependent
	TLR9	Non-methylated CpG DNA	Bacteria and Virus	MyD88-dependent
	TLR10	dsRNA	Virus	MyD88-dependent
Nucleotide-binding oligomerization domain-like receptors (NLRs)	NOD1	IE-DAP	Gram-negative bacteria	RIP2-TAK1-NF-κB pathways
	NOD2	MDP	Gram-negative bacteria and Gram-positive bacteria	
	NLRP1	Anthrax toxin	Bacteria	Inflammasome Complex
	NLRC4	Flagellin	Bacteria	
	NLRP3	Pore-forming toxins, M2 protein, Hemozoin	Bacteria, viruses, parasites	

		DAMPs - (ATP, Hyaluronan, Alum)		
RIG-I-like receptors (RLRs)	RIG-I	5’ triphosphorylated RNA, short-chain dsRNA	Virus	MAVS-TRAF6- NF-κB/TBK1 pathways
	MDA5	Poly IC, long chain dsRNA	Virus	
	LGP2	dsRNA	Virus	
C-type lectin receptors (CLRs)	Dectin-1	β-Glucan	Fungus	Tyrosine kinase-dependent and non-tyrosine kinase-dependent pathways
	Dectin-2	α-Mannan	Fungus	
Absent in melanoma-2-like receptors (ALRs)	ALRs	dsDNA	Bacteria	Inflammasome-pyroptosis

Finally, the production of reactive oxygen species (ROS) can have varied effects depending on the cellular source, the enzyme that has produced it, and the primary oxygen species produced. For example, ROS produced by the phagocyte-limited NADPH oxidase (NOX) 2 enzyme will produce large quantities of superoxide after engulfing microbes. Superoxide will be subsequently converted to hydrogen peroxide and then to the highly reactive hypochlorous acid which destroys bacterial proteins and inhibits bacterial cell division (McKenna and Davies, 1988, Aviello and Knaus, 2018). In contrast, ROS production by IECs is vital to ensuring barrier integrity. Production of ROS by colonic IECs is chiefly performed by NOX1 and dual oxidase (DUOX) 2. These enzymes, activated by upstream PRR signalling, produce ROS, which inactivate redox-sensitive tyrosine phosphatases, leading to the formation of focal adhesions - a vital process in wound healing, and restoration of barrier function (Swanson et al., 2011). In addition to their role as intracellular messengers, ROS release from IECs can have roles as intercellular messengers, most notably by communicating with the dense immune presence in the underlying lamina propria. ROS action on T-cells is complicated. Whilst low concentrations of intracellular ROS are vital for T-cell activation, proliferation, and differentiation, high levels of extracellular ROS such as those released by immunosuppressive cells, are toxic to T-cells (Yarosz and Chang, 2018). Likely reflecting internal expression of antioxidants, such as superoxide dismutase (SOD1) or glutathione (GSH), various T-cell subsets have differing sensitivities to a high ROS microenvironment (oxidative stress). By subjecting freshly isolated T_{REG} cells to increasing hydrogen peroxide concentrations, the Kiessling laboratory demonstrated increased viability of T_{REG} cells in comparison to conventional ($CD4^{pos}$ $CD25^{neg/low}$) T-cells. In fact, T_{REG} cells survived and maintained their suppressive activity at H_2O_2 concentrations which killed half of the conventional cell subset (Mougiakakos et al., 2009). This selective ability of T_{REGS} to survive, and function, under severe oxidative stress, selectively dampens the inflammatory response and acts as an important negative feedback mechanism to limit tissue damage.

1.5.2.4 Lamina propria lymphocytes (LPLs)

The colonic lamina propria is filled with an abundance of immune cells. Previously, studies attempting to characterise this compartment have been limited to animal models or tissue that has been surgically removed due to disease. However, in 2013, a collaboration between the New York Organ Donor Network and the Organ Procurement Organisation offered the chance to investigate the immune composition of 8 different tissues in 24 individuals with healthy immune systems. Samples were taken prior to organ transplantation from organ donors who died suddenly from traumatic causes. These samples were subsequently digested and subjected to flow cytometry for analysis. Analysis of colonic tissue demonstrated that whilst the colon had an enrichment of B-cells in comparison to other tissue areas, colonic lymphocytes were around 8 times more likely to be of the T-cell subset rather than B-cells (Sathaliyawala et al., 2013). These T-cells, predominantly of an effector memory phenotype, can be both $CD4^{POS}$ and $CD8^{POS}$ but, in contrast to the IEL compartment, are twice as likely to be $CD4^{POS}$. This $CD4^{POS}$ LPL compartment is highly diverse, with T_H1, T_H2, $FOXP3^{POS}$ T_{REGS}, and T_H17 cells all being reported. Interestingly, whilst the distribution of T_H1 and T_H2 cells didn't seem to vary along the length of the colon, there seems to be an inverse relationship between T_H17 and T_{REGS}, with later stages of the colon demonstrating larger T_{REG} populations and a lower T_H17 cell abundance (Mowat and Agace, 2014) so as to preserve the anti-inflammatory state of colonic homeostasis. Despite differing inflammatory actions, T_H17 and T_{REGS} take a similar journey in their differentiation from naïve CD4 T-cells. Both have a requirement for TGFβ in their differentiation, although it may be replaced by IL-1β for IL-6-mediated T_H17 differentiation (Ghoreschi et al., 2010). Therefore, in an environment rich in TGFβ and the absence of IL-6, naïve CD4 T-cells will take on a T_{REG} fate, whereas if both TGFβ and IL-6 are present, T-cells will take on a T_H17 fate. The dual promoting behaviour of TGFβ likely lies in a concentration-dependent functionality. At low concentrations, it can synergise with IL-6 to promote T_H17 differentiation through inducing the IL-23 receptor (IL23R) whereas at higher concentrations TGFβ represses IL-23R and promotes FOXP3

expression. The differentiation into these two subtypes, therefore, is highly dependent on the surrounding cytokine milieu, in which a microbe-rich (high in IL-6) microenvironment favours T_H17 differentiation over T_{REG} fates. T_H17 $CD4^{POS}$ cells, as well as directing $CD8^{POS}$ T-cells via IL-21, have direct antimicrobial function (Cai et al., 2016). IL-26, thought to be characteristic of highly differentiated T_H17 cells, has a cationic amphipathic structure working analogously to the defensin molecules (Pène et al., 2008, Meller et al., 2015). Whilst it is generally accepted that T_H17 cells are useful for the protection against extracellular pathogens, they also kill intracellular parasites. In mouse models of *Trypansoma cruzi* infection, Cai and colleagues showed that allogenic adoptive transfer of parasite specific T_H17 cells conferred a greater protection from *T. cruzi* mortality than those from a T_H1 differentiation status. Furthermore, this action was linked predominantly to their IL-17A-dependent induction of NOX enzymes on macrophages (Cai et al., 2016). The production of ROS, and T_H17 cells are therefore intertwined. Whilst the cytolytic ability of T_H17 seems to predominantly reside in the utilisation of ROS, the presence of ROS negatively feedbacks to inhibits the action of T_H17 and other effector-T_H subsets. These feedback mechanisms are vital for resolving inflammation, and ensuring colonic homeostasis is maintained.

1.5.2.5 The importance of being immunologically idle

The colonic epithelial layer represents one of the most incredible immunobiological and microenvironmental balancing acts that exist. It is in constant contact with potential pathogens, but it must remain one of the most fragile barriers to permit nutrient uptake and function effectively. This combination means that harmonising inflammation and tolerance is vital. Uncontrolled inflammation is not only potentially devastating to barrier function such as that apparent in inflammatory bowel disease (IBD), but also in terms of the potential transformation of ISCs, driving CRC. Unlike the small intestine in which the ISCs are buried deep in crypts, the crypts of the colon are much shallower and are therefore more at risk of inflammation-induced transformation. These

differences are translated to increasing abundances of tolerance-inducing cell types along the length of the intestine (Mowat and Agace, 2014), but also to increasing abundances of mucus producing goblet cells as you enter the crypts, enabling increased protection of the ISCs from microbial encroachment. Protection, and maintenance of the ISC population is vital for healthy colonic function.

1.5.2.6 Maintenance of the ISC niche

ISCs give rise to all differentiated cells of the colonic epithelium and therefore ISC maintenance is tightly regulated. Maintenance of the ISC niche is performed by a multitude of cell types as well as the physical action of the extracellular matrix. To investigate which signals were important for ISC maintenance, Kosinski *et al*. (2007) characterised gene expression profiles in various microanatomical locations of the human colonic crypt. By separating the crypt into cells located at the top and at the base, they could perform differential gene expression analysis to determine which pathways are being regulated in the stem cell compartment. Unsurprisingly, "Cell Cycle" and "DNA replication" gene ontology (GO) terms were detected as significant for cells in the base of the crypt with respect to the top, consistent with these cells being proliferative ISCs. Interestingly, *Wnt*-related genes were significantly increased in the base of the crypt, and those that are induced by interruption of Wnt signalling were induced at the colon tops. This provided key evidence that Wnt pathways are a major determinant of cell behaviour, by differentially regulating pathways such as cell proliferation, migration, and differentiation across the colon crypt axis (Kosinski et al., 2007, Neumann et al., 2014). The action of Wnt on ISCs is of special interest in CRC as inactivating mutations in adenomatous polyposis coli (APC) are highly prevalent. Inactivating *APC* mutations are often a first port of call in many tumours because of APC's normal role in regulating destruction of the intracellular mediator of Wnt signalling, β-catenin. β-catenin is a constitutively produced transcription factor that, when not destroyed, acts to increase transcription of Wnt-target genes -

most notably *c-Myc* which controls cell proliferation. In the absence of Wnt, β-catenin is destroyed by the Axin complex of which APC is an integral part (MacDonald et al., 2009). In the presence of Wnt, however, this complex is disrupted by recruitment of scaffolding proteins to the membrane and β-catenin accumulates. Inactivating APC mutations, therefore, mimic the presence of Wnt by causing disruption of the Axin complex leading to increased proliferation of the epithelium. Although heterozygous mutations in APC can be hereditary in a condition called Familial Adenomatous Polyposis, most APC mutations are spontaneous. In both pathways, however, the second allele is often lost through loss of heterozygosity. The progression of CRC from adenoma to invasive carcinoma is, perhaps, one of the best characterised pathways in multistep carcinogenesis.

1.6 Colorectal Cancer

CRC, also known as bowel cancer, refers to tumours arising in both the large intestine and rectum. Despite these cancers having clear progression routes and prognoses, they are primarily classified together due to their shared basic biology (Perez-Ruiz and Berraondo, 2016). Despite incidence and mortality declining over the past few decades, likely due to better screening measures and improved awareness of preventable risk factors, CRC remains the second most common cause of cancer-related death in the United States when both genders are combined and, therefore, continues to be a significant clinical burden (Siegel et al., 2020).

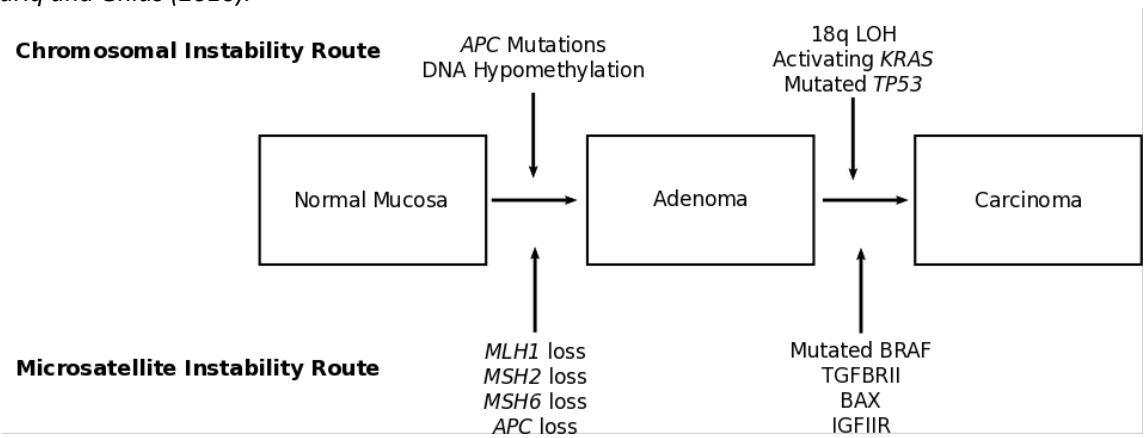
1.6.1 CRC Carcinogenesis

CRC development often first involves benign growths of the mucosa (polyps). These growths can be classified into two main types, the most common of which are adenomatous polyps and commonly referred to as adenomas. Whilst polyps are usually harmless and can be routinely removed during a colonoscopy, if left over time, they can grow into the wall of the colon or rectum and become cancer.

The genetic model for progression from adenoma to invasive carcinoma in CRC was first postulated by Fearon & Vogelstein in 1990 and follows two main routes from normal epithelium to carcinoma: chromosomal instability (CIN) and microsatellite instability (MSI) (Figure 1.3).

Figure 1.3 – Genetic and epigenetic changes that occur in CIN and MSI routes to CRC carcinogenesis

Schematic depicting the stepwise gain of genetic/epigenetic changes that must occur in colorectal cancer progression. Alterations are split by the two main routes: chromosomal instability (CIN) and microsatellite instability (MSI) and outline the changes most associated with each route. Adapted from Tariq and Ghias (2016).



Whilst both pathways (MSI and CIN) demonstrate a tendency for inactivating APC mutations, mutated APC has been suggested to be particularly key in the development of CIN tumours. Although an imperfect relationship exists between APC mutations and the likelihood of CIN-driven CRC, experiments have demonstrated the role for APC in ensuring dividing cells receive equal numbers of chromosomes during division. After engineering mouse embryonic stem cells to express a defective, truncated version of APC, Kaplan *et al*. (2001) found that daughter cells often acquired unequal numbers of chromosomes. Furthermore, they later showed that APC localised to the kinetochore with wildtype, but not truncated, APC also binding the positive end of microtubules. These findings place APC centrally in regulating the segregation of chromosomes during mitosis, and thus, it is not too speculative to suggest that defective APC plays a large role in subsequent development of CIN. Comprising the remaining 15-20% of CRCs are those developing via MSI. MSI tumours can arise through a genetic predisposition – Lynch Syndrome, although the majority of MSI CRC cases are sporadic. Lynch syndrome patients possess a germline mutation in the MMR system used to correct DNA mutations that may be generated during DNA replication. These tumours, therefore, exhibit large quantities of mutations scattered across the genome. These two biological subtypes of CRC are important, as they often have differing clinical features, including prognoses and potential treatments (Tariq and Ghias, 2016).

1.6.2 CRC Diagnosis & Classification

Early diagnosis of CRC has vastly improved over the last few years, most notably due to the introduction of population-based CRC screening programs. In the UK, these screening programs invite men and women between the ages of 60 and 74 years to take part. They have drastically improved mortality rates due to the ability to diagnose CRC in its asymptomatic form by completing home-test kits or routine colonoscopies in high-risk individuals. Conversely, early diagnosis of CRC in symptomatic patients poses an issue in that many of the early signs of CRC are shared with non-

neoplastic conditions. Whilst severe abdominal symptoms, such as rectal bleedings, pain, or a change in bowel habits may be an indication of CRC, these symptoms are nonspecific and insufficient for a diagnosis of CRC. Despite this, clinicians will often perform a colonoscopy for patients reporting abdominal symptoms on the suspicion of CRC. A colonoscopy is an examination in which a thin tube with a small camera and light is inserted into the rectum and moved along the length of the colon to check for any abnormalities that may indicate CRC. If abnormal areas are encountered, then a biopsy will be taken and analysed by a pathologist (CRUK, 2015). Once a diagnosis of CRC is confirmed, staging must be undertaken. Staging in the UK largely follows the tumour-node-metastasis (TNM) staging criteria first set out in 1946 (Denoix, 1946) and last re-evaluated for CRC in 2017 (Weisenberg, 2017).

1.6.3 Stratification approaches towards CRC

While current clinical assessment mainly relies on tumour-node-metastasis (TNM) staging (Weisenberg, 2017), there have been several attempts to introduce stratification approaches to improve the prediction of clinical outcome or the response to particular therapies. Perhaps the most widely adopted stratification criterion to complement TNM staging is based on MMR status, now a standard test in most hospitals. MMR is a highly conserved system which repairs erroneous insertions and deletions that can occur during DNA replication. Testing of MMR status can be completed using immunohistochemistry (IHC) on paraffin-embedded tumour tissue, by staining for the MMR proteins (MLH, MSH2, MSH6, and PMS2), however, this is often validated by PCR (fragment analysis) if the results are inconclusive (Lindor et al., 2002). Generally considered more accurate, fragment analysis can be completed by utilising a panel of five validated nucleotide repeat sequences (two mononucleotide repeats in BAT26 and BAT26, and three dinucleotide repeats in D2S123, D5S346, and D17S259), termed the Bethesda panel. Whilst the utilisation of this panel did standardise the procedure, the inclusion of dinucleotide repeats has been questioned due to these

being less sensitive and specific for the identification of MMR-deficient cancers. It is now advised that MSI-H be called using a pentaplex panel of mononucleotide repeats (BAT25, BAT26, NR21, NR24, and NR27) (Murphy et al., 2006). Irrespective of whether the Bethesda panel or the pentaplex PCR assay are used, the number of loci which demonstrate instability stratifies the results of fragment analysis into distinct groups (Table 1.2). Although an MSI-L group is also defined in these criteria as containing one marker of instability, subsequent studies have found no qualitative reason for the distinction between MSI-L and MSS (Pawlik et al., 2004). In contrast, MSI-H tumours have a specific clinical profile. They have presentation features such as being localised to the proximal (right) colon, poor differentiation, an abundant lymphocytic infiltrate, and are eligible for ICB therapies (Boland and Goel, 2010, Buckowitz et al., 2005). MSI-H tumours also largely have a favourable prognosis, especially in those with localised disease with these tumours demonstrating a higher disease free survival, regardless of stage. The opposite is true for recurrent MSI-H tumours, however. Post-recurrence survival was much worse for MSI-H patients than MSS patients and therefore a significance in overall survival between MSI-H and MSS is rarely seen. The reasons for a reduced recurrence survival are likely many-fold. However, a major reason is due to the lack of a curative option for recurrent MSI-H tumours. These tumours have an inherent resistance to chemotherapies, due to the lack of an intact MMR system, and therefore aren't eligible for this type of therapy (Kim et al., 2016b).

Table 1.2 – Criteria for separating patients based on microsatellite instability loci

Microsatellite status is determined based on the number of microsatellites detected by PCR. This can be based on either the Bethesda panel (BAT25, BAT26, D2S123, D5S346, D17S250), or a pentaplex panel of mononucleotide repeats (BAT25, BAT26, NR21, NR24, and NR27).

Microsatellite Status	***Description***
Microsatellite Stable (MSS)	No markers show instability
Microsatellite Instability – Low (MSI-L)	One of the five markers show instability
Microsatellite Instability – High (MSI-H)	Two or more markers show instability

Many researchers have attempted to develop classifiers that alter clinical practice, however, alongside the Immunoscore classification (see below), perhaps the closest to doing so are the Consensus Molecular Subgroups (CMS groups) suggested by Guinney et al (2015). Guinney and colleagues identified that discrepant results between studies was the main reason biological stratifications had struggled to access the clinic. They aimed to create a 'methodological framework' in which a set of subgroups could be delineated across various data processing and sample preparation methodologies, within diverse patient cohorts, and across gene expression platforms. By establishing the "CRC Subtyping Consortium (CRCSC)", six independent classifiers and 18 separate data sets spanning multiple RNA sequencing platforms were used to build a classifier. This classifier converged on four, distinct consensus molecular subgroups (CMS). By incorporating other data sources such as mutation, copy number, methylation, microRNA, and proteomics, it was shown that each of these groups had distinct characteristics. 14% of CRC had a CMS1 subtype. These cancers represented the conventional MSI-H subtype with hypermutation, immune infiltration, and poor post-relapse survival characteristics being their defining features. The majority of CRC, however, fell in the CMS2 subtype, with 37% of cancers having high levels of copy number alterations and high signalling through the *wnt* and *myc* pathways. In contrast, CMS3 cancers represent the fewest in abundance, representing only 13% of CRC. These cancers are made up from both MSI-H and MSS cancers and have *KRAS* mutations, low levels of copy number alterations, and metabolic dysregulation as defining features. Lastly, CMS4 cancers or the "mesenchymal subgroup", made up 23% of CRC. These cancers have high levels of stromal immune infiltration, TGFβ activation, and increased angiogenesis. Multiple studies have highlighted the CMS4 subgroup as having the worst relapse-free survival and overall survival out of all the CMS subgroups and are therefore in dire need of targeted therapeutics (Guinney et al., 2015). Subsequent trials, such as the randomised MAX trial, have attempted to target the hallmark of angiogenesis in CMS4 cancers by combining chemotherapy with the anti-VEGF drug – bevacizumab, although to no avail. The authors found that CMS2 and

CMS3 benefitted much more from the addition of bevacizumab than CMS1 and CMS4 (Mooi et al., 2018). These therapies, however, were only tested in metastatic disease. CMS subtyping is completed on primary tumours; therefore, it is conceivable that CMS subtype could change between primary and advanced disease (Pfeiffer and Qvortrup, 2020). Despite attempts to utilise CMS subgroups clinically, at the time of writing, they are yet to influence any practice.

1.6.4 Management of CRC

Current management of CRC is mainly dependent on the stage of disease. Generally, regarding tissue-limited cancer, patients undergo a colectomy where the affected tissue is surgically removed. After CRC develops to the point of presentation with lymphatic metastases, patients receive additional adjuvant chemotherapy administered over six months. Progression to metastatic CRC (mCRC) means palliative treatment is generally used unless the metastases can be surgically resected. Palliative treatment for stage IV CRC patients consists of the combination of surgery, chemotherapy, and radiotherapy, although ICB therapies have recently been added as an option for a select group of patients (Cancer.Net Editorial Board, 2012). Patients who have progressed after treatment with chemotherapy, or have mCRC and test positive for MMR deficiency are currently eligible for Pembrolizumab (PD-1 blockade) as second or third-line therapy (Fellner, 2017). Recent trials have demonstrated Pembrolizumab has high success in MSI-H settings, with a progression-free survival at 20 weeks of 78% in the MSI-H setting, compared to 11% in MSS (Le et al., 2015). After successful Phase II trials in MSI-H CRC patients (Le et al., 2020), a recently closed Phase III clinical trial (KEYNOTE-177) investigated Pembrolizumab as a first-line therapy in cancers with deficient MMR systems (ClinicalTrials.gov, 2020). Whilst analyses regarding overall survival are still ongoing, initial data suggest that pembrolizumab was superior to chemotherapy with regards to a longer relapse free survival and a reduced likelihood of treatment-related adverse events. These results could well justify pembrolizumab to be escalated to front-line therapy soon (Shiu et al., 2021).

1.6.5 CRC Immunology

Despite the success of immune checkpoint blockade being primarily limited to a subset of CRC patients to date (MSI-H), much emphasis has been placed on delineating the role of the immune infiltrate in CRC. Over the past 15 years, studies have consistently reported the improved prognosis of patients with a denser immune infiltrate (Galon et al., 2006). However, the type, location, and abundance of the various effectors are likely to impact prognosis. While reports for some effectors, such as eosinophils and mast cells may reflect the stage of disease rather than an active immune response to a tumour, reports of CD8 and CD4 αβ T-cells are more reliably considered as indicative of a systemic immune response (Dalerba et al., 2003).

1.6.5.1 Immune landscape in CRC

Functionally, αβ T-cells are divided into two main groups – effector and regulatory. The effector population, primarily comprising $CD4^{POS}FOXP3^{NEG}$ and $CD8^{POS}$ T-cells, are further subdivided based on function, whereas the regulatory compartment (T_{REGs}) is typically defined as $CD4^{POS}FOXP3^{POS}$. Expression of CD4 or CD8 discriminates function in the effector compartment. Whereas $CD8^{POS}$ T-cells are conventionally regarded as classical "effectors" in that they are cytolytic and act to remove cells expressing aberrant/non-self-peptides on MHC-I molecules, $CD4^{POS}$ T-cells are generally classified as helper T-cells (T_H). T_H cells canonically act to orchestrate an immune response by adopting functionally distinct phenotypes. Historically, only two T_H subsets were defined – T_H1 and T_H2. T_H1, which can be phenotypically delineated by their expression of *T-bet*, act to clear intracellular pathogens by secretion of IFNγ and activation of $CD8^{POS}$ T-cells, whereas T_H2 cells, characterised by *GATA3* expression and release of IL-4, are biased towards an extracellular response and act to increase humoral immunity. From a cancer perspective, a strong emphasis has been placed on the importance of T_H1 cells in the anti-tumour response, since the action of these cells in increasing

intracellular presentation of processed antigen would be expected to result in tumour clearance due to the increased number of non-tolerised antigens arising from cancer mutation. More recently, however, other T_H phenotypes have been delineated, of which possibly the most famous is T_H17. T_H17 cells, identified by their expression of RORγt and named for their characteristic production of IL-17 cytokines (IL-17A-F), normally function to maintain mucosal barrier function and contribute to pathogen clearance at these sites (Curtis and Way, 2009). While the impact of T_H1 and T_H2 cells in cancer is relatively well established, the effects of T_H17 cells in cancer are still controversial, with many arguing that their impact is mostly dependent on the cancer site. The existence of these distinct phenotypes of T_H cells, however, does highlight the importance of delineating the phenotype of immune cells when linking immune infiltrate with outcome.

Arguably the most comprehensive study completed on immune presence in CRC was carried out by Galon and colleagues (2006). This study was one of the first to demonstrate the impact that an immune-based stratification technique could have in CRC. By first identifying a cluster of T_H1-centric genes whose expression was inversely correlated with tumour recurrence, Galon *et al*. (2006) demonstrated that functional polarisation towards a T_H1 fate was vital for patients to gain a clinical benefit. However, despite highlighting the importance of this functional phenotype for patient outcome, Galon *et al*. went on to determine the influence of geographical location of immune subsets in determining patient outcome. Utilising IHC stains for CD8 and CD3, they separated $CD8^{POS}$ and $CD3^{POS}$ T-cells in patient biopsies into those in the tumour core (CT) and those at the invasive margin (IM). Based on these metrics, they found they could stratify patients into various groups, each with distinct disease-free survival and overall survival statistics. Furthermore, the combination of CT and IM metrics into a single score increased the accuracy of survival prediction and led to the development of the "Immunoscore". Recently, Mlecnik *et al*. (2016) demonstrated that the Immunoscore was more prognostic than MMR status although its utility as a predictive biomarker for

ICB is still being investigated. The ease of staining for two markers will likely accelerate the adoption of the Immunoscore into the clinic; however, it is likely to be incorporated into TNM staging as a revised TNMI staging paradigm, rather than a standalone test (Donnem et al., 2016).

1.6.5.2 Coordinated immune response cluster (CIRC)

Adopting a similar approach to that of Galon and colleagues, Lal *et al*. (2015) utilised a self-curated set of immune-related genes as a surrogate for immunity for comparison across publicly available CRC patients' microarray data (The Cancer Genome Atlas, 2012). After defining clusters using two-dimensional hierarchical clustering, they found a central group of genes which were co-ordinately expressed across patients (Table 1.3), and subsequently termed this the Coordinate Immune Response Cluster (CIRC). This geneset was extended to include 4 more genes which, by visual analysis, followed the trend of the central group determined by hierarchical clustering. Whilst the addition of these genes is unlikely to affect a strong expression of the CIRC signature, it does introduce noise and deflects from the unbiased analysis in which the geneset was produced. CIRC expression was almost uniformly increased in MSI-H patients, whereas MSS patients had a more variable expression of the CIRC signature with MSS patients being split across all four clusters (Lal et al., 2015). Subsequently, by comparing CIRC scores between MSS patients and MSI-H patients, Lal (2016) suggested an arbitrary cut-off based on MSI-H patients to highlight a subset of MSS patients with a similar CIRC score. As MSI-H tumours are well documented to have a dense immune infiltrate, it was hypothesised that these patients might also share a comparable immune infiltrate and thus may be eligible for ICB therapies, however, the nature of this infiltrate is currently unclear.

Table 1.3 – Genes within the initial gene list, including those which formed the Coordinate Immune Response Cluster (CIRC)

All genes in the table represent those that Lal (2016) curated to investigate an immune response. Those highlighted in red resemble the 24 genes which were co-ordinately expressed and formed the CIRC signature. Adapted from Lal (2016).

Initial gene list used to delineate the CIRC signature			
ACTB	*CXCL9*	*HLA-DRB5*	*RAET1E*
CCL11	*GNLY*	*ICAM1*	*RAET1G*
CCL2	*GZMB*	*ICOS*	*STAT1*
CCL5	*HAVCR2*	*IFNγ*	*STAT3*
CD247	*HLA-A*	*IL12RB2*	*TBX21*
CD274	*HLA-B*	*IL17A*	*TNFRSF14*
CD3D	*HLA-C*	*IL18RAP*	*TNFSF4*
CD3E	*HLA-DMA*	*IL7R*	*ULBP1*
CD3G	*HLA-DMB*	*IRF1*	*ULBP2*
CD4	*HLA-DOA*	*KLRK1*	*ULBP3*
CD80	*HLA-DOB*	*LAG3*	*VCAM1*
CD86	*HLA-DPA1*	*MADCAM1*	*VTCN1*
CD8B	*HLA-DPB1*	*MICB*	
CTLA4	*HLA-DQA1*	*PDCD1*	
CX3CL1	*HLA-DQA2*	*PDCD1LG2*	
CXCL10	*HLA-DRA*	*PROCR*	

1.7 Project aims

In this thesis, I investigate the potential for immune presence in low-mutation burden cancers using CRC as a model human cancer. Possibly the most significant challenge in the field of cancer immunology is to extend the success of ICB to the low mutation burden setting. To date, ICB has had success in cancers with high mutation burdens, such as melanoma and lung cancer, or subsets of cancers which have deficiencies in the MMR system. However, most cancers, on average, display much lower mutational burdens and, therefore, are not eligible for ICB based on current stratification approaches. While it is clear a high mutation burden is pivotal in driving strong immunity in the tumour microenvironment, there is clearly considerable heterogeneity in intratumoural immunity in low mutational settings. Despite this, underlying mechanisms driving an immune response in such settings are currently unclear. Although multiple factors other than mutation have been suggested, none have been incorporated into current stratification approaches (Routy et al., 2018).

CRC presents a compelling case to act as a model in which to explore the optimisation of immunotherapy in the low-mutation burden cancer setting. The application of ICB in CRC is currently limited to ~15% of patients who have deficiencies in the MMR system. Most CRC patients, however, exhibit very poor responses to ICB (Le et al., 2015). CRC, therefore, demonstrates both success and failure of ICB at the same anatomical site, and is thus poised to explain the immunobiology of both settings. Furthermore, while a high mutational load is highly prognostic for CRC, high intratumoural immunity is, perhaps, more so (Mlecnik et al., 2016). Although collectively MSS patients are deemed non-responders to ICB, 11% of patients do respond and previous studies completed by this group (Lal, 2016), and others (Galon et al., 2006, Mlecnik et al., 2016), have demonstrated that intratumoural immunity is highly variable within the MSS grouping.

In this thesis, I aim to systematically define a subgroup of CRC patients which are genetically MSS but have a high intratumoural immunity and explore underlying mechanistic drivers of the immune response in this subgroup. I will utilise previously published signatures, which are surrogates for an increased immune response (CIRC signature), and MMR status to evaluate various clustering mechanisms to stratify patients. Next, I will exploit the wealth of publicly available data to interrogate differences in clinical parameters, cancer and host genetics, and microbiotal compositions between high-immune MSS, low-immune MSS, and MSI-H cohorts. Here, I hypothesise that the MSS cohort will display differences in one or more of these factors, offering explanation for the differing intratumoural immunity seen within this patient subgroup. Lastly, I will investigate whether patterns outlined in CRC extend to other barrier-based cancers by completing similar analyses using public data for gastric and endometrial cancer.

Finally, this thesis benefits from being aligned with a trial driven by the University of Birmingham and in collaboration with Bristol Myers-Squibb which investigates the potential for HLA-Class II in stratifying patients who could benefit from Nivolumab (ANICCA). This trial is due to finish February 2023 and is currently recruiting (ClinicalTrials.gov, 2019).

Chapter 2: Materials and Methods

2.1 Introduction

The bespoke scripts used in this thesis can be accessed at: https://github.com/JackLMc/ThesisAnalysis. The R packages; `tidyverse` and `ggstatsplot` were used throughout (Wickham et al., 2019, Kassambara, 2020, Patil, 2021). To ensure the versions of all R packages used in this thesis were consistent throughout analyses, all R packages were updated to the latest releases and analyses were rerun prior to submission. A collection of bespoke functions was also used throughout these analyses and were bundled in a local package, `GeneralFunctions`, which is available on demand. To ensure repeatability of subsequent runs, the seed for all analyses was set to 2021, unless otherwise stated.

2.2 Public Data Collection and Processing

The datasets used in Guinney et al., (2015) were used as inspiration for this thesis. However, the public availability of clinical data dictated the datasets which were used (Table 2.1). Eleven datasets from the Affymetrix microarray platform, one from the Agilent microarray platform, and the TCGA-COADREAD RNA sequencing dataset were used. Clinical and phenotype data for those with a GSE identifier were accessed from the Gene Expression Omnibus (GEO) database using the R package `GEOquery` (Davis and Meltzer, 2007). Clinical data for GSE2109 unsuccessfully parsed in R and, therefore, columns were manually adjusted in Excel Ltd. TCGA-COADREAD clinical data was downloaded using the R package `TCGAbiolinks` (Colaprico et al., 2016, Mounir et al., 2019, Silva et al., 2016).

Samples were filtered to only include those annotated as primary colorectal cancer. Where MMR data were not available, it was inferred separately for RNA sequencing and microarray data using Entrez IDs and the `CMSCaller` R package using processed data (Eide et al., 2017, R Core Team,

2018). Any patients for whom an MMR status call could not be delineated were discarded. The number of patients belonging to each MMR group is shown in Table 2.2. CMS and CRIS subgroups were also called using this method.

Table 2.1 – Meta data for the datasets used in this thesis

List of datasets used in this thesis, including a breakdown of the number of patients from each study after applying filtering steps. Filtering steps included the availability of clinical data, samples belonging to primary CRC, the availability of sequencing data, samples passing outlier criteria, and the availability of data for MMR status data. TCGA data were downloaded from two sources. RSEM data was accessed from Synapse (synapse repository IDs: syn1910201 [TCGA-COAD], syn1910239 [TCGA-READ]), and all other TCGA data were accessed using TCGAbiolinks. Studies prefixed by 'GSE' were downloaded from their respective gene expression omnibus accession databases.

Study Code	***Platform***	***No. Patients with Clinical Data***	***Primary CRC with Clinical Data***	***No. Patients with Clinical and Sequencing Data***	***No. Patients after Outlier Removal***	***No. Patients with MMR Data***
TCGA-COADREAD	RNA Sequencing	633	633	591	543	542
GSE13067	Affymetrix	74	74	74	69	49
GSE13294	Affymetrix	155	155	155	149	149
GSE14333	Affymetrix	290	290	290	276	216
GSE17536	Affymetrix	177	177	177	163	128
GSE20916	Affymetrix	145	101	101	99	38
GSE2109	Affymetrix	366	365	279	247	193
GSE23878	Affymetrix	59	35	35	31	23
GSE33113	Affymetrix	96	90	90	77	64
GSE35896	Affymetrix	62	62	62	57	57
GSE37892	Affymetrix	130	130	130	117	86
GSE39582	Affymetrix	585	566	566	512	504
GSE42284	Agilent	188	188	188	164	145
Total		2,960	2,866	2,738	2,504	**2,194**

Table 2.2 – Frequency of patients for each MMR status

Number of patients belonging to each MMR status. Data shown is for all available datasets.

Subtype	***Frequency***
MSI-H	569
MSS	1,625
Total	2,194

2.2.1 TCGA-COADREAD

Various data for TCGA-COAD and TCGA-READ projects were accessed from multiple sources. Guinney *et al.* (2015) utilised RNA sequencing data normalised by RNA sequencing by expectation maximisation (RSEM) and reported no batch effects between TCGA-COAD and TCGA-READ samples, therefore, RSEM-normalised RNA sequencing data were used only for outlier removal and batch effects were not checked for between TCGA-COAD and TCGA-READ samples here. The patients deemed outliers in RSEM were removed from the raw count data below.

2.2.1.1 RNA sequencing by Expectation Maximisation-normalised (RSEM) data

TCGA-COADREAD RNA sequencing data were downloaded from Synapse (synID; syn1910201 [TCGA-COAD], syn1910239 [TCGA-READ]) using the SynapseClient in a recursive manner. RSEM data for both the Illumina HiSeq and Illumina Genome Analyzer (GA) were read into R. If a patient was run on both platforms, the Genome Analyzer sequenced sample was discarded. Expression data were subjected to $\log_2(x + 1)$ transformation, where x represents RSEM values. PCA revealed strong batch effects between GA and HiSeq samples, which were removed by the ComBat method from the `sva` R package (Leek et al., 2020) (Figure 2.1). Outlier detection was performed using two separate approaches; PCA and the `arrayQualityMetrics` R package (Kauffmann et al., 2009). A sample was deemed an outlier if it failed criterion set out in Guinney *et al.* (2015). All outliers were removed from the analyses. These data were used for outlier removal only.

Figure 2.1 – Combat batch correction for sequencing platform in the TCGA-COADREAD cohort

RNA sequencing of the TCGA-COADREAD cohort was performed via two separate platforms (HiSeq – blue, and GenomeAnalyzer – red). Single value decomposition used all available genes to evaluate batch effects between sequencing platforms. Prior to batch correction, the sequencing platforms showed strong separation (left). After Combat batch correction, the separation of samples based on sequencing platform was removed (right).



2.2.1.2 Count data processing

HT-seq count data for TCGA-COADREAD patients were downloaded and processed using the R packages *TCGAbiolinks* (Colaprico et al., 2016, Mounir et al., 2019, Silva et al., 2016) and *SummarizedExperiment*, respectively (Morgan et al., 2020). Where patients had multiple tumours sequenced, the tumour with the largest library size was taken. Ensembl gene identifiers (Ensembl IDs) were converted to Entrez gene identifiers (Entrez ID) using the *org.Hs.eg.db* R package (Carlson, 2020). In the 233 occasions where one Ensembl IDs mapped to multiple Entrez IDs, the first instance was taken. Whereas in the 67 occasions that multiple Ensembl IDs mapped to one Entrez ID, the Ensembl ID with the largest median absolute deviation (MAD) across the cohort was taken.

For geneset enrichment analysis (see below) counts were normalised using conditional quantile normalisation (CQN) and the *cqn* R package (Hansen et al., 2012). To inform the start and end sites of genes, and their percentage G+C content, a biomaRt object was created using the *biomaRt* R package (Durinck et al., 2005, Durinck et al., 2009). 9,413 Entrez IDs didn't have information in the biomaRt object and were discarded. 1,273 genes had duplicate information for start and end sites and/or percentage G+C content. To ameliorate this issue, the data was averaged for individual genes. Gene length information was rounded to the nearest integer. In total, 24,699 genes were available for interrogation. Raw counts were used for differential gene expression analysis (see Section *2.2.5*).

2.2.1.3 UPC Transformation

The processed count file matrix of the TCGA-COADREAD cohort was subjected to UPC transformation using the *SCAN.UPC* R package (Piccolo et al., 2012, Piccolo et al., 2013). Data were normalised using a negative-binomial model, and correction for GC content and length was performed using the data frame generated in Section *2.2.1.2*. A maximum of 100 iterations were allowed for the data to reach the default convergence threshold of 0.01.

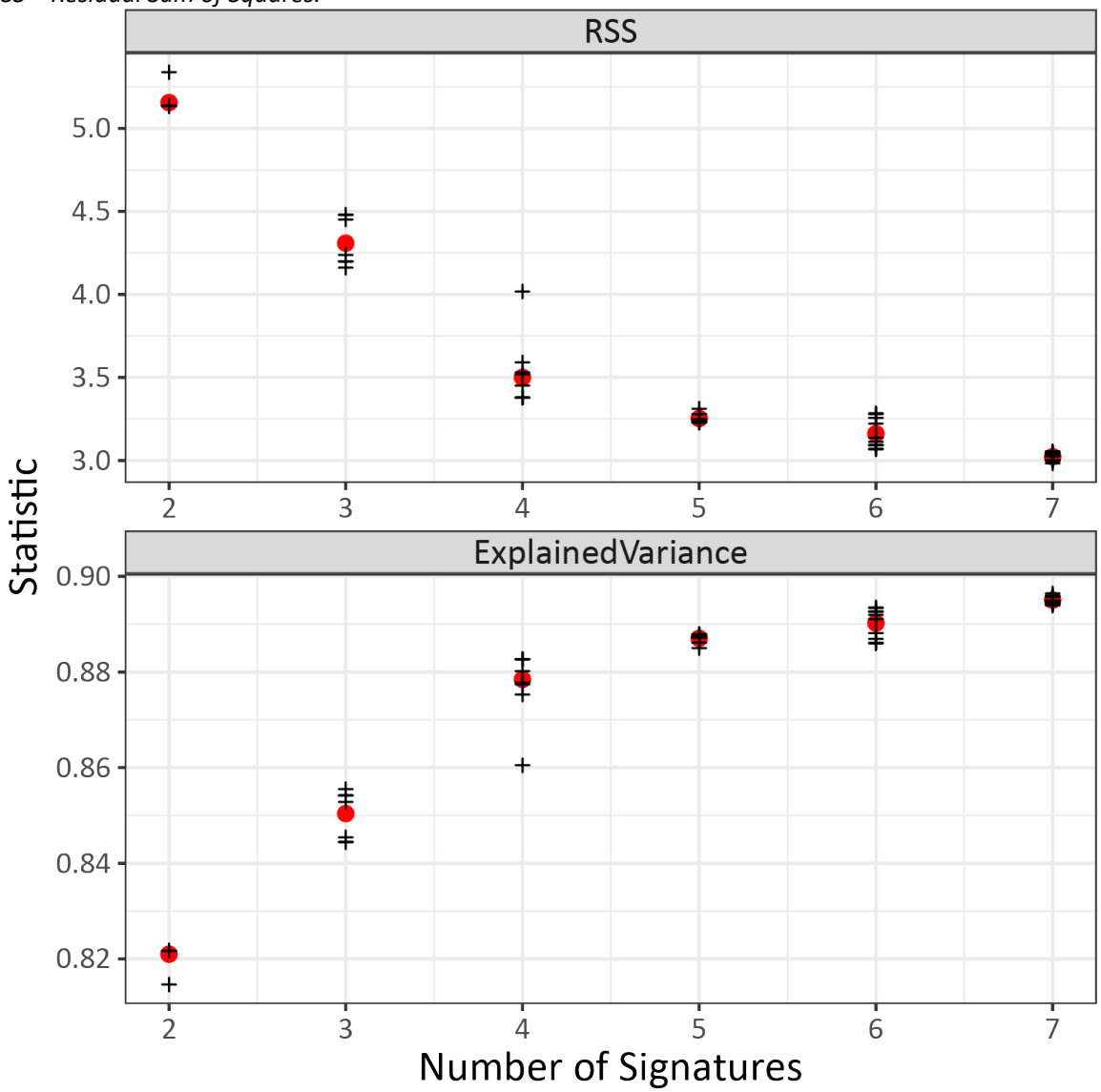
2.2.1.4 Mutation data collection and analysis

Mutect-called mutation data was downloaded using the `TCGAbiolinks` R package (Colaprico et al., 2016, Mounir et al., 2019, Silva et al., 2016). Silent mutations were removed and mutation numbers for each variant type and patient were tabulated. Mutation spectrum analysis utilised all SNP mutations, regardless of whether they were silent and used the `SomaticSignatures` and `BSGenome.Hsapiens.UCSC.hg38` R packages (Gehring et al., 2015, Team, 2020). Data were processed to produce a matrix of the frequency of base changes for each patient. Negative matrix factorisation (NMF) decomposition was utilised to determine the number of signatures present in the cohort, replicating the analysis ten times. Data converged on five signatures (Figure 2.2), and, therefore, five signatures were chosen to segregate the patients. The contribution of individual base changes to the delineated signatures are shown in Appendix 1. The delineated signatures were compared to those described in Alexandrov *et al*. (2015) using cosine similarity. Alexandrov signatures are present as a matrix in the package `SomaticSignatures`. Recurrent mutations were calculated utilising the `maftools` R package (Mayakonda et al., 2018).

Figure 2.2 – Delineation of the number of mutation signatures present in the TCGA-COADREAD cohort by Negative Matrix Factorisation decomposition

A matrix of the frequency of single nucleotide base changes was used as input to Negative Matrix Factorisation decomposition to test the number of mutation signatures present in the TCGA-COADREAD cohort. Data points converged on five mutation signatures. An average of 10 repeats demonstrated a low average error statistic of 3.25 (Residual Sum of Squares [RSS]) and a high fraction of explained variance (average of 88.7%). Top panel – Y axis represents the RSS statistic. Bottom panel – Y axis represents Fraction of Explained Variance.

RSS = Residual Sum of Squares.



2.2.1.5 Copy number data collection and analysis

GISTIC and Segment Mean copy number files were downloaded for the TCGA-COADREAD cohort using the `TCGAbiolinks` R package (Colaprico et al., 2016, Mounir et al., 2019, Silva et al., 2016). Data visualisation utilised both the `svpluscnv` and `CNVRanger` R packages (Garcia, 2021, Da Silva et al., 2019).

2.2.1.6 SNP data collection and analysis

Access to controlled SNP arrays from TCGA-COADREAD patients was gained on 28/10/2019. SNP arrays were downloaded using the GDC download client (Grossman et al., 2016). SNP arrays contain ~1,000,000 common SNPs. SNPs not located on Chromosomes 1-22, or the mitochondrial genome were removed, leaving 868,026 autosomal SNPs and 107 mitochondrial SNPs. A meta data file was downloaded from the GDC repository and read into R using the `rjson` package (Couture-Beil, 2018). Data were filtered to include only data that were from normal paired blood data. Not all patients had SNP arrays completed. The number of patients for each subgroup can be seen in Table 2.3. SNP arrays were concatenated together using the terminal, and association tests were used to determine the enrichment of SNPs with the subgroups. For mitochondrial germline SNPs, a genotype test was used whereas autosomal germline SNPs utilised allele tests due to the diploid nature of autosomes. Annotation of SNP locations were downloaded on 21/11/2019 from: http://www.affymetrix.com/Auth/analysis/downloads/na35/genotyping/GenomeWideSNP_6.na35.annot.csv.zip. SNP meta information regarding associated genes and population frequency was gained using `rsnp` R package (Gustavsen et al., 2020). Autosomal SNPs passing a phred score of 60 and mitochondrial SNPs passing a phred score of 23 were displayed using the `karyotypeR` package (Gel and Serra, 2017). For mitochondrial SNPs, a custom genome was built using `karyotypeR` and information gained from `biomaRt` (Durinck et al., 2005, Durinck et al., 2009). FDR correction was completed on BlueBEAR using 100 random shuffles of the clusters and repeating the analyses.

Table 2.3 – Number of patients in each subgroup with SNP arrays completed on paired blood

Subtype	**Frequency**
MSI-H	56
MSS	304
MSS-hi-CIRC	80
Total	**440**

2.2.2 TCGA-STAD and TCGA-UCEC

TCGA Stomach Adenocarcinoma (STAD) and the TCGA Uterine Corpus Endometrial Carcinoma (UCEC) datasets were downloaded and processed in a similar manner as the TCGA-COADREAD dataset (Section 2.2.1). Unfortunately, TCGA-UCEC could not be exposed to RSEM-based outlier removal due to a lack of data availability.

Mutation analysis of both TCGA-UCEC and TCGA-STAD were chosen to have three signatures. With an RSS statistic of 4.5 (TCGA-STAD) and 8 (TCGA-UCEC) and an explained variance 0.78 (TCGA-STAD) and 0.73 (TCGA-UCEC).

2.2.3 Affymetrix datasets

2.2.3.1 fRMA normalisation

Data analysed on Affymetrix platforms were redownloaded from each of their GEO accession databases on 12/05/2021. These data were normalised in R using the frozen robust multiarray analysis (fRMA) method. Unpacked .tar files were read into R using the `affy` package (Gautier et al., 2004), and normalised using the `frma` package for R (McCall et al., 2010), taking the robust weighted average. Despite Guinney *et al.* (2015) having stated that there existed overlapping samples between GSE14333 and GSE17536, I did not find any overlapping samples between these datasets. As datasets were from different laboratories and at different time points, they displayed strong batch effects (Figure 2.3). The ComBat method from the `sva` R package (Leek et al., 2020) was used to remove the batch effects on a study basis, followed by outlier removal following the same parameters previously outlined for TCGA data (Figure 2.3). Affymetrix microarrays utilise probesets, which are made up of multiple different probes. For this reason, one probeset may cover more than one locus. In this dataset, however, there were no single probes which covered different Entrez IDs as determined by

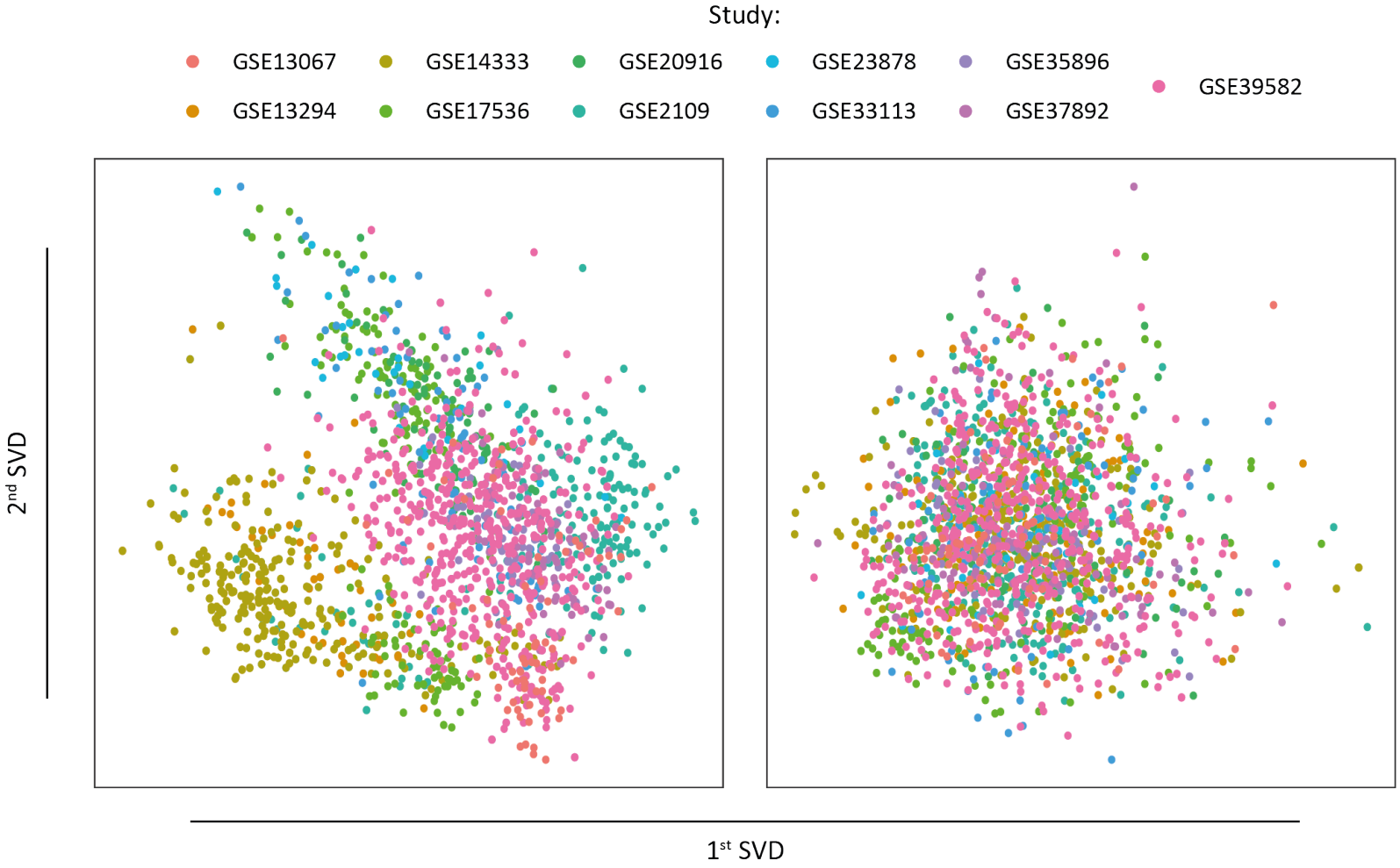
the *hgu133plus2.db* R package (Carlson, 2016). On the other hand, 20,161 Entrez genes had multiple probes mapping to them. Therefore, the probe with the largest MAD across patients were used to represent individual Entrez IDs. This ensured a 1:1 mapping between probes and Entrez IDs.

2.2.3.2 UPC Transformation

The ".CEL.gz" files for non-outlier samples were copied to the BlueBEAR cluster. Files were normalised on an individual study basis using the *SCAN.UPC* R package (Piccolo et al., 2012, Piccolo et al., 2013) and BlueBEAR. Probes were converted to Entrez IDs using the 1:1 mapping delineation created in Section *2.2.3.1*.

Figure 2.3 – Combat batch correction for study in the Affymetrix cohorts

Affymetrix datasets were gathered from studies performed at various timepoints and displayed significant batch effects based on study (left). Single value decomposition used all available genes to evaluate batch effects between Affymetrix cohorts. After Combat batch correction the separation of samples based on study was removed (right).



2.2.4 Agilent dataset

The GSE42284 dataset was downloaded on the 21/07/2020. Data were read into R and normalised using the *limma* package (Ritchie et al., 2015). Data were normalised using global loess normalisation, and background correction was performed resulting in average $\log_2$(values) taken forward. Missing data were imputed using the *impute* package (Hastie et al., 2020). Eight rows had over 50% of data missing, and the mean imputation was used for these values. Outlier detection was completed as outlined previously for the other cohorts. The supplied mapping file in the .tar archive was first used to map the probes to HUGO symbols. These symbols were then used to find Entrez identifiers using the *org.Hs.eg.db* R package (Carlson, 2020). Any HUGO symbols, and therefore probes, which didn't map to an Entrez ID were removed. In accordance with Affymetrix microarray data, Agilent microarray also utilise probesets, and probesets can cover more than one gene. A total of nine probesets were found to cover more than one genos in this dataset, and the first instance was taken to avoid duplication. The reverse can also be true in that a single Entrez identifier can be covered by more than one probe. 6,369 Entrez identifiers were covered by more than one probe. To avoid multiple probes mapping to the same entrez ID, the probe with the maximum MAD was taken to represent the Entrez ID. The resultant data frame had 15,071 genes.

2.2.5 Geneset enrichment analysis

Genesets were collated from various sources. Where only HUGO gene symbols were available, these were converted to Entrez IDs using the *CMScaller* package. If genes did not exist in the data frame, they were discarded from the signature. Gene signatures were converted to lists and the *GSVA* R package was used to determine enrichment scores (Hänzelmann et al., 2013).

The CIRC signature and initial gene list were taken from Lal *et al.,* (2015)(Table 1.3). Genesets reflecting different cell types were largely taken from, and processed, using the R package, `consensusTME` (Jiménez-Sánchez et al., 2019). Patients were split into COAD and READ samples for processing. Those which didn't have this information and were simply referred to as "colorectal" for primary site were removed from these analyses. Enrichment scores were again calculated using GSVA. HGNC Symbols are required for `consensusTME`, and therefore, Entrez IDs were converted using the `CMScaller` R package (Eide et al., 2017). Signatures of T_{Helper} subtypes were taken from multiple sources. T_H1 and T_H2 signatures were taken from the study performed by Bindea and Colleagues (2013). This study also contained a T_H17 signature, however, this signature was likely not collated in the same manner due to the considerable overlap between T_H1 and T_H17 cells. and was, therefore, deemed too simple to be reminiscent of an entire cell type. A T_H17 signature was, therefore, taken from Castro *et al*., (2017) who had skewed naïve $CD4^{POS}$ T-cells to either a T_H17 or a T_H1 fate and completed a subtractive assay to ensure that the genes were specific to T_H17, rather than being shared with those upregulated in T_H1. Only one gene overlapped between Castro T_H17 and Bindea T_H1 signatures – *IL22*.

Significance in pathways was tested on the TCGA-COADREAD cohort using the R package, `edgeR`. Following guidance in the user manual, transcripts with a lower count than 6.5 counts in at least 69 samples were removed. 69 samples were chosen as a cut-off as this represented the smallest of the three groups investigated. This filtering step removed 9,513 transcripts from analysis. The `gage` R package was used to search the GO database for buzzwords for genesets of interest (Luo et al., 2009). Only genesets of length between and including 5 and 200 genes were analysed. The R packages; `RCurl` and `RCy3` were used to communicate with, and plot in, Cytoscape using the enrichmentMap module (Temple Lang, 2020, Gustavsen et al., 2019, Shannon, 2003, Merico et al.).

2.3 Local Cohort Collection and Processing

2.3.1 Birmingham cohort

2.3.1.1 Patient characteristics

Multiplex immunohistochemistry data from this cohort was kindly donated by Ghaleb Goussous and Neeraj Lal for analysis via computational methods. In short, 30 patient slides were collated from Heartlands or the Queen Elizabeth Hospital (Birmingham, United Kingdom). Where available, MMR status was taken from clinical data which had been investigated by IHC staining for MSH2, MSH6, MLH1, and PMS2. Those who were negative for all four proteins were deemed MSI-H. Where MMR status had not been clinically investigated, or data were not available, MMR status was investigated in-house by fragment analysis (Lal, 2016).

MSS-hi-CIRC patients from the Queen Elizabeth hospital were previously delineated from a cohort of 95 MSS patients using singleplex IHC for MHC Class II (ABCAM, clone: CR3/43J). After ranking patients for epithelial MHC Class II expression, as quantified by Definiens Tissue Studio Software's H score, the top 10 patients in terms of highest epithelial MHC Class II expression were assigned as MSS-hi-CIRC. MSS patients from Heartlands, however, hadn't been subjected to the same singleplex MHC Class II IHC protocol and, therefore, were re-evaluated using a mean of the multiplex immunofluorescence MHC Class II metric on entire cells phenotyped as either "Class II" or "Other" from InForm (see Section *2.3.1.3ii,* below). Two MSS patients from Heartlands had a mean MHC Class II expression higher than the 75% quartile of the MSI-H and MSS-hi-CIRC patients combined (246.31) and were relabelled as MSS-hi-CIRC.

2.3.1.2 Multispectral Immunofluorescence

FFPE sections were stained with the antibodies listed in Table 2.4, in addition to DAPI. Using the Leica Bond RX platform, staining was automated to run overnight. Staining utilised a heat-based step between each antibody to remove any antibodies before the subsequent antibody was added. Spectral deconvolution was achieved by completing a series of single stains (Table 2.4) on serial sections of an individual patient's biopsy. Autofluorescence was compensated for by scanning an unstained slide, which was subsequently used for DAPI staining. All antibodies were either validated in-house or had sufficient IVD-CE/CST based validation (Goussous, 2020). Post-staining, whole slides were scanned by the Vectra 3.0 Automated Quantitative Pathology Imaging system.

Table 2.4 – Primary antibodies used for multispectral immunofluorescence

Experimental details for primary antibodies used in multispectral immunofluorescence experiments. Clone and company details are provided.

Stain	***Channel***	***Dilution***	***Company***	***Secondary Species***	***Clone***
CD4	Opal 520	1:50	Leica	Mouse	4B12
CD8	Opal 540	1:400	Leica	Mouse	4B11
FOXP3	Opal 570	1:400	Abcam	Mouse	236A/E7
T-Bet	Opal 620	1:400	CST	Rabbit	D6N8B
MHC Class II	Opal 650	1:200	Abcam	Mouse	CR3/43J
RORγt	Opal 690	1:4000	Merck	Rabbit	6F3.1

2.3.1.3 Image Analysis

i PhenoChart

Whole slide images were uploaded to PhenoChart (version 1.0.5) in a ".QPTIF" file format to choose areas for high power magnification scanning. Six areas of interest were randomly chosen for each patient. These regions of interest were subsequently scanned by the Vectra 3.0 at high power magnification (20X).

ii InForm

A spectral library was constructed for the stains using the InForm 2.3 software and single slide images. An InForm project was set up which utilised trainable tissue and cell segmentation, and a cell phenotyping algorithm. The unstained, autofluorescence slide image was imported into the project, allowing correction for autofluorescence in the subsequent images. Images were processed in batches, to include one patient from each subtype; however, only the first batch was used to produce the algorithm. Tissue categories were separated into "Epithelium", "Stroma", and "Background", with a minimum segment size set at 1000 pixels. All training images were checked for any obvious mis-segmentation and were corrected up to, but no more, than three times. Cell segmentation was performed, segmenting both nuclei and membrane. Only the nuclei were used to aid detection of cells and cells were discarded if they touched the edge of an image. For nuclei aid, DAPI was chosen as a counterstain, whereas membrane detection was aided by the membrane stains MHC Class II and CD8. Cell phenotyping included choosing at least five cells for each of the phenotypes shown in Table 2.5. The algorithm produced was then used on all images. The main steps and representative images for InForm analysis are shown in Figure 2.4.

Table 2.5 – Phenotypes delineated from the combination of multispectral immunofluorescence stains

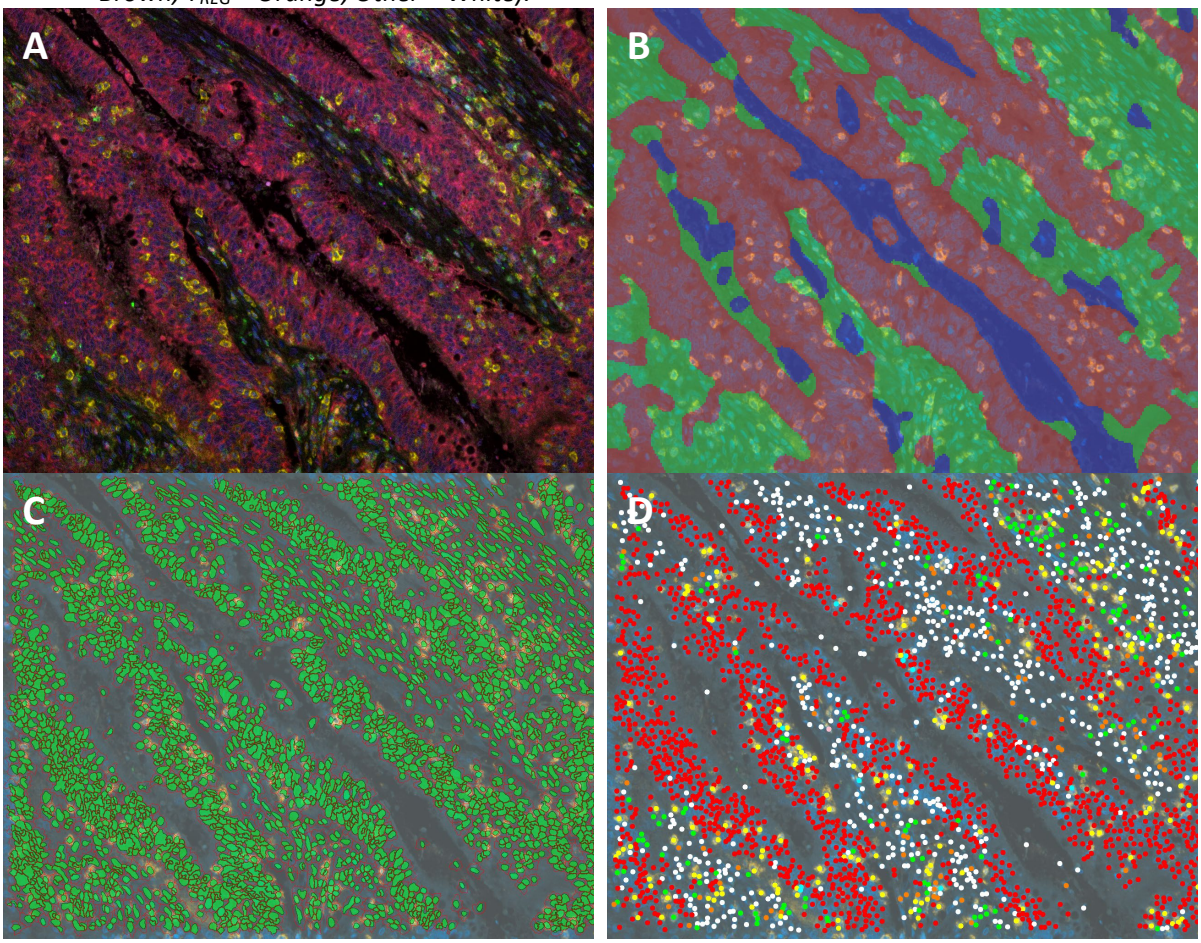
Eight phenotypes were delineated based on the combination of available stains. All phenotypes included the DAPI stain.

Stain Combination	***Name***
CD4 Only	CD4 Only
CD8 Only	CD8 Only
CD4 & FOXP3	T_{REG}
CD4 & T-Bet	T_H1
CD4 & RORγt	T_H17
CD8 & T-Bet	T_C1
MHC Class II	Class II
Any other combination or DAPI alone	Other

Figure 2.4 – The main steps taken in InForm analysis of multispectral immunofluorescence images

Composite images are subjected to trainable tissue and cell segmentation algorithms. Cells were then phenotyped based on the combination of stains (Table 2.5).

A. *Composite Vectra image. Image shows all stains after autofluorescence correction. (Red – MHC Class II, Yellow – CD8, Green – CD4, Cyan – T-bet, Orange – FOXP3, RORγt – Magenta, DAPI – Blue).*
B. *Tissue segmentation map outlining epithelium (Red), stroma (Green) and background (Blue).*
C. *Cell segmentation map demonstrating outline of cells (Green – cells, Red - membrane).*
D. *Cell phenotyping map demonstrating the various phenotypes trained by the phenotyping algorithm (Red – Class II, Yellow, CD8 Only, Green – CD4 Only, T_H1 – Pink, T_C1 – Cyan, T_H17 – Brown, T_{REG} – Orange, Other – White).*



iii R analysis

Data were exported as cell summary data. All cells present in the Background category were removed. By taking the average density per megapixel as unit, the densities of the various phenotypes were compared.

2.3.1 100K Genomes Colorectal Cancer Dataset

Access to the 100,000 genomes colorectal cancer dataset was granted on: 10/02/2021. Those patients which had been withdrawn in the GeCIP were removed from the analysis. These data had associated clinical data, which was accessed by Toritseju Sillo using LabKey in 2019. MMR status was combined from two columns – PathMMR, which was the MMR status completed by Pathologists at the QE hospital and GELMMR, an MMR call by the GeCIP. The GELMMR column was used as the reference, using the PathMMR column to fill any NAs that existed. Three patients had duplicate information in the provided clinical datasheet – P016250, P018987, and P019993. In all duplicated patients, most of the clinical information matched, therefore the first instance of each patient was taken.

2.3.1.1 RNA Sequencing

3' single-end RNA sequencing data of 203 FFPE CRC patients was kindly donated by Toritseju Sillo in the form of raw FASTQ files. Samples were sequenced on the NextSeq and merged. FASTQ files were aligned to Ensembl GRCh38, downloaded from: ftp://ftp.ensembl.org/pub/release-84/fasta/homo_sapiens/dna, using HISAT2 (version 2.2.1) (Kim et al., 2019). Bam files were read into R using the `Rsubread` package, and counts were exported (Liao et al., 2019). 23 sequenced samples were withdrawn from the GeCIP, and thus were removed. Of the remaining samples, one patient had the same sample sequenced twice, and 7 patients had multiple samples sequenced. To avoid a one-to-many mapping issue regarding samples, the sample with the largest library size was taken

forward. Six patients had no data for MMR status and were removed. During CQN normalisation, data for gene length was taken from the Rsubread object for annotations, and percentage G-C content was taken from a biomaRt object. 982 genes had duplicate data for percentage G+C content and were dealt with as outlined previously (Section 2.2.1.2). Outliers were determined using CQN normalised counts. Five patients were deemed outliers by PCA only. These pre-processing steps left 146 patients for analysis, consisting of 36 MSI-H patients, and 110 MSS patients. For the final cohort, average successful alignment was low (18.9%) however, this was attributed to samples being from FFPE origin. Samples had an average library size of 281,197.

2.3.1.2 DNA Sequencing

Access to whole genome sequencing (WGS) data from the 100,000 genomes project was granted on 10/02/2021. Samples were sequenced by and sorted by Illumina, and subsequently aligned to GRch38. Sorted bam files were subjected to bacterial calling by Centrifuge (version 1.0.4)(Kim et al., 2016c). Default settings were used other than the minimum hit length (`--min-hitlen`) being increased to 35 (default = 22) and the host tax ID (`--host-taxids`) was set to 9606 (Homo sapiens). The reference database included: the human genome, prokaryotic genomes, and viral genomes including the 106 SARS-CoV-2 complete genome (h+p+v+c, found here: https://ccb.jhu.edu/software/centrifuge/). These initial analyses were performed by Boris Noyvert, and access was granted to output files on 06/05/2021 for downstream analyses.

i Downstream microbial analyses

Output files were annotated by NCBI taxonomic ID. These values were used as input to the *taxonomizr* R package (Sherrill-Mix, 2021) to gain taxonomy data for each of the microbes. Gram staining data was annotated for Bacteria using the gram staining rules (Chamberlain, 2019). 49 microbial species lacked taxonomy data and were discarded, leaving 5,753 species for investigation.

Diversity calculations utilised Shannon diversity indices. These functions were taken from the *vegan* package for R (Oksanen et al., 2020). Differential microbe analysis utilised the *metagenomeSeq* R package for the zero-inflated model they employ (Paulson et al., 2013b, Paulson et al., 2013a). 2,500 features were deemed rare and were removed from the matrix prior to normalisation. Sankey diagrams were created utilising the *networkD3* R package (Allaire et al., 2017).

2.3.1.3 RNAscope for eubacteria 16s rRNA probe

30 samples were requested from the HBRC tissue bank. These consisted of 10 patients from each subgroup (MSS-hi-CIRC, MSS-lo-CIRC, MSI-H). To distinguish between MSS-hi-CIRC and MSS-lo-CIRC, patients were scored and ordered by CIRC enrichment score. The top 10 MSS patients were labelled as MSS-hi-CIRC and the bottom 10 patients, MSS-lo-CIRC. The top 10 CIRC score expressors of MSI-H patients were also requested. Unfortunately, only 9 MSI-H patients, 7 MSS-lo-CIRC, and 7 MSS-hi-CIRC tumour samples could be sourced. Matched normal sections were also requested from this original cohort. Healthy tissue for 5 MSI-H, 8 MSS-lo-CIRC, and 7 MSS-hi-CIRC were available for analyses.

FFPE slides were freshly cut for staining by the eubacteria 16s rRNA probe [EB-16S-rRNA] (bio-techne, catalogue #464468) with the RNAscope 2.5 LS Reagent Kit-BROWN (bio-techne, Catalogue #322100). Positive and negative controls utilised matched normal tissue. Slides were loaded onto the Leica Bond Max. A 150µL dispense volume was set, following the pre-set “ISH Detection” protocol.

2.3.2 Wales Cancer Bank Cohort

2.3.2.1 Patient Characteristics

FFPE slides and scrolls were obtained from Wales Cancer Bank (WCB), project number – 19/003 and ethics REC number – 13/WM/0339. WCB did not have MMR status information readily available. To

inform the natural distribution of MSS and MSI-H tumours, the preferential localisation of MSI-H tumours to the right-side of the colon was used to get an 80:20 split of MSS:MSI-H. To avoid any bias in batch preparation, patient identifiers were randomised into batches as to maintain the MSS:MSI-H ratio using R (R Core Team, 2018). For MHC Class II staining, the seed for randomisation was set at 1, whereas for all other analyses it was set at 2020.

2.3.2.2 Mismatch Repair Status Delineation

DNA was extracted from FFPE scrolls as per manufacturer instructions using the Maxwell RSC DNA FFPE Kit (Promega – AS1450). The extraction protocol was run on the Maxwell RSC instrument. Once finished, DNA concentration was determined using the Qubit dsDNA BR assay kit (Fisher Scientific - 10625443) as per manufacturer instructions, and DNA was frozen. DNA was thawed prior to polymerase chain reaction (PCR) amplification of the quasimonomorphic markers - BAT-25, BAT-26, NR-24, NR-21, and NR-27 (primer sequences in Table 2.6). Extracted DNA underwent PCR with reagents listed in Table 2.7. Primers were first delyophilized by resuspending in Tris-EDTA (TE) buffer to make a 100µM stock. Primer mix included 8µL of all forward primers, 8µL of NR-27 and BAT-26 reverse primers, 6µL of NR-21, NR-24, and BAT-25 reverse primers, and 26µL of water to make up to 100µL of primer mix. The PCR reaction started with 2 minutes at 95^{o}C to denature. This was followed by 95^{o}C for 15 seconds, 55^{o}C for 15 seconds, and 72^{o}C for 50 seconds for 33 cycles. Prior to fragment analysis, PCR product was cleaned using a MinElute PCR purification kit as per manufacturer instructions (Qiagen - 28004) and frozen. PCR products were thawed and 5µL was submitted in a 96-well plate to the University of Birmingham Gene Facility for processing on the Tapestation High Sensitivity D1000. Microsatellite status in each sample were assessed using the Tapestation Analysis Software (A.02.02) under the guidance of Professor Andrew Beggs.

Table 2.6 – Forward and reverse primers for the detection of microsatellite instability
Primer sequences are given in a 5' to 3' orientation. Primer sequences were gained from Buhard et al., (2004) and orientation checked via IGVviewer BLAT and FindMotif searches. Microsatellites are ordered by fragment size. Bp = Base pairs. All primers were purchased from Merck.

Microsatellite	***Gene***	***Primer Sequences***	***Estimated Fragment Size***
BAT26	*MSH2*	**F:** CTGCGGTAATCAAGTTTTTAG	183 Bp
		R: AACCATTCAACATTTTTAACCC	
BAT25	*KIT*	**F:** TACCAGGTGGCAAAGGGCA	153 Bp
		R: TCTGCATTTTAACTATGGCTC	
NR24	*ZNF2*	**F:** GCTGAATTTTACCTCCTGAC	131 Bp
		R: ATTGTGCCATTGCATTCCAA	
NR21	*SLC7A8*	**F:** CTGGTCACTCGCGTTTACAA	109 Bp
		R: GAGTCGCTGGCACAGTTCTA	
NR27	*BIRC3*	**F:** AACCATGCTTGCAAACCACT	87 Bp
		R: CGATAATACTAGCAATGACC	

Table 2.7 – PCR conditions for the detection of microsatellite status
MyTaq reagents gained from MyTaq DNA Polymerase kit (Meridian Bioscience – BIO21105).

Reagent	***Volume per sample (µL)***
Water	19
Primer Mix	0.6
5x MyTaq Reaction Buffer	5
MyTaq DNA Polymerase	0.2
Sample (DNA)	1.0

2.3.2.3 Immunohistochemistry

Paraffin-embedded CRC sections from WCB were single-stained for various antigens. Information regarding staining conditions for all antibodies can be seen in Table 2.8. Antigens were stained in two batches using the Leica BondMax Autostainer. Dispense volume was set at 150µL and the staining mode was set at "Single" and "Routine". IHC Protocol F (DAB) was used for staining. Leica Bond reagents can be seen in Table 2.9. Images were scanned in using Leica Aperio AT2 digital pathology scanner. Although no dedicated positive controls were utilised, in the two batches of staining for 4-HNE I saw no evidence of batch effects across the cohort (P-value = 0.087).

Table 2.8 – Primary antibodies for singleplex staining

Experimental details for primary antibodies used in IHC experiments. Clone and company details are provided.

Antibody Target	***Company***	***Clone***	***Secondary***	***Epitope Retrieval Method***	***Dilution***
MHC Class II	Abcam	CR3/43	Mouse	ER1, 20 minutes	1:100
4-HNE	Abcam	HNEJ-2	Mouse	ER1, 20 minutes	1:100

Table 2.9 – Leica BondMax Autostainer Reagents

Reagent	***Supplier***	***Catalogue Number***	***Comments***
Bond Dewax Solution	Leica	AR9222	NA
Bond Primary Antibody Diluent	Leica	AR9352	NA
Bond Epitope Retrieval Solution 1	Leica	AR9961	pH6
Bond Epitope Retrieval Solution 2	Leica	AR9640	pH9
Bond Wash Solution 10X Concentrate	Leica	AR9590	10% dilution in Type 2 deionised water before use

i Singleplex Visiopharm analysis

Scanned images were loaded into Visiopharm (2021.02). The WCB slides stained for HLA-Class II were used to build applications for tissue detection, and the delineation of tumour versus stroma.

a) Tissue detection

To reduce as much background as possible, a classifier was produced to detect tissue. This classifier utilised the RGB-G channel, taking a minimum of filter size 1x1. Output was then manually checked to ensure no off-tissue staining or air bubbles were being taken as tissue (Figure 2.5 – left).

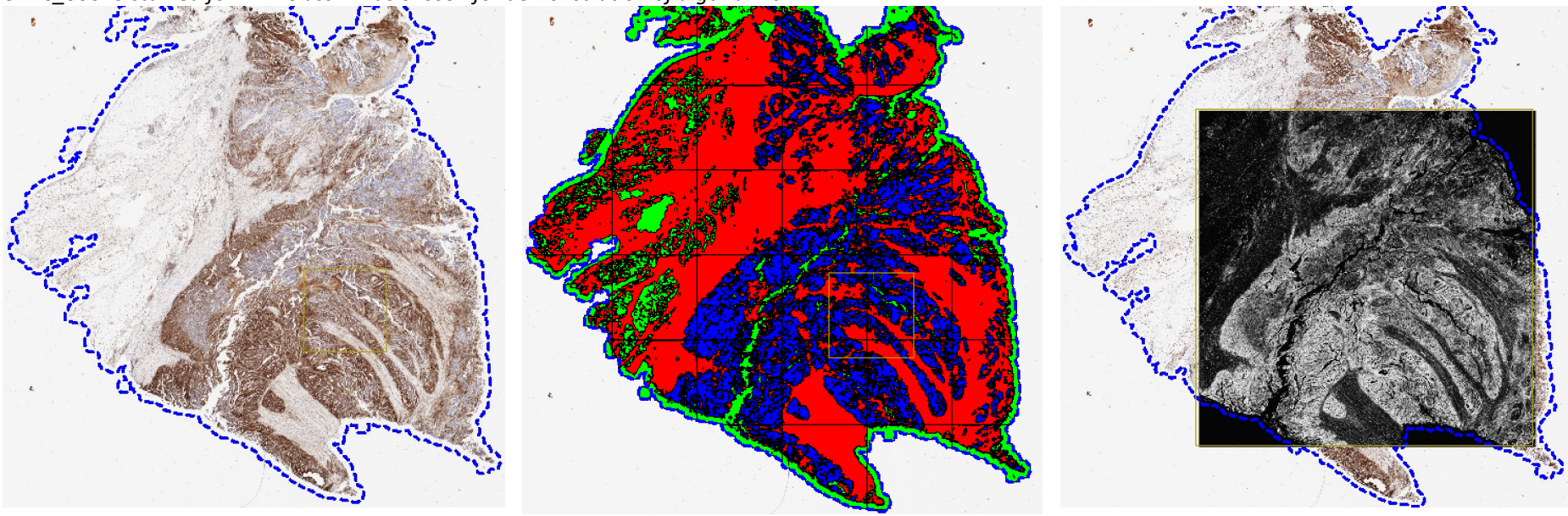
b) Stroma versus epithelium

A classifier to delineate stroma from epithelium was created and trained using the U-NET deep learning module. To inform the algorithm of areas of epithelium, stroma, and background, the patients which failed MMR testing were used as input – CRAC_0035, CRAC_0066, CRAC_0100, and CRAC_0113. CRAC_0100 was subsequently removed from the study due to no epithelium being present in the slide. CRAC_0009 was added to the training pool to replace CRAC_0100 due to it having an epithelium with a high-Class II expression, a variable the other slides lacked. Five fully annotated ROIs were used for initial training, with an average diameter of 457µm. ROIs were chosen as to prioritise stromal and epithelial areas. 2,500 iterations were allowed for the training process, reaching an error rate of 4.9%. This algorithm was then tested on four separate ROIs of average diameter 4,681µm in this cohort. It was then tested on two whole slides – CRAC_0015 and CRAC_0684 for refinement of post-processing steps.

DAB stain was detected by the HDAB- feature (Figure 2.5 - right). Mean intensity of this feature was calculated in both epithelium and stroma and for the total slide. This value was used for computing differences.

Figure 2.5 – The main steps in Visiopharm analysis of singleplex IHC data

Analysis was performed in a stepwise manner from left to right. Tissue was first detected by constructing a tissue-detection algorithm (Left panel, dotted blue line resembles tissue outline). Next, stroma versus epithelium was detected by an AI-trained algorithm (Centre panel, Green – Background, Red – Stroma, Blue – Epithelium). Lastly, the DAB stain was detected using the HDAB feature (Right panel, White = high DAB intensity, Black = low intensity of DAB). Patient CRAC_00015 stained for HLA-Class II was chosen for demonstration of algorithms.



These applications were also applied to the Genomics England cohort for MHC Class II and 16s rRNA RNAscope quantification. An additional 3,000 iterations of training on four annotated ROIs were given to this cohort. An initial 2,000 iterations utilised the slides - S362480, and S363141 which were chosen at random. A further 1,000 iterations of training were given using S363951 to incorporate epithelium with high staining.

Chapter 3: A Redefinition of High Immunity CRC using the CIRC Signature

3.1 Introduction

At the start of my involvement in the project, previous analyses of CRC on the TCGA database from our research group had revealed a cohort of MSS patients who had an increased immune infiltrate. These patients were determined by utilising hierarchical clustering on microarray Z-scores using a self-curated geneset to find a group of genes which were co-ordinately expressed. This set of genes, termed the Coordinate Immune Response Cluster (CIRC), was then used to 'cut' MSS patients into high and low expression (Lal, 2016). Whilst this analysis was useful in informing the project design going forward, many of the decisions to delineate these patient groups were subjective. These included the application of 'visual analysis' and using arbitrary thresholds for delineating MSS-hi-CIRC patients. The aim of this chapter, therefore, was to ameliorate this, and to robustly define the MSS-hi-CIRC cohort, in an objective and evidence-based manner.

The potential importance of stratifying cancer patients by immune infiltration has been extensively demonstrated. Multiple papers by multiple groups, and across different cancers, have repeatedly demonstrated that patients who have an increased intratumoural immune infiltration have an increased survival than their lower infiltrated counterparts (Mlecnik et al., 2016, Clemente et al., 1996, Schumacher et al., 2001). In comparison to molecular stratification, which has become increasingly important in clinical management, immunological stratification is still in its infancy. Patients are now routinely administered single-agent EGFR mAbs following chemotherapy failure provided they have a wildtype *RAS* gene, however, the administration of ICB has only recently been approved for use after chemotherapy failure and if a patient has microsatellite instability (National Institute for Clinical Excellence, 2017, National Institute for Clinical Excellence, 2021). Tumours with microsatellite instability have a dense immune infiltration (De Smedt et al., 2015) . The basis for the increased immune infiltration observed in these patients is thought to arise from their increased

mutational burden, especially as a similar level of immune infiltration is observed in *POLE* and *POLD1* mutants, which also give rise to an increased mutation burden (Van Gool et al., 2015, Chae et al., 2018). The ideology behind an increased mutation burden driving an increased immune infiltration is relatively easy to comprehend. The greater number of mutations, the increased chances that a non-tolerised mutated peptide will be displayed on MHC Class I and induce an immune response. However, this does not exclude other factors causing an immune response in mutation-low cancer. Multiple studies have now demonstrated high variability of intratumoural immunity within MSS cancers, however, those patients with the highest infiltrates will be refused immunotherapy on the basis that their MMR system is intact, despite arguably having the direct target of the therapy – an increased immune infiltrate (Mlecnik et al., 2016, Millen et al., 2020). Trial data initially demonstrated a reduced efficacy of ICB in MSS cancers, however, more recent data demonstrates a small percentage of MSS patients benefit from ICB (Le et al., 2015). One would assume these patients to have an increase intratumoural immunity, therefore, it seems bizarre that, despite the means to directly test for an increased immune infiltration, the surrogate of microsatellite instability is still used. A worldwide taskforce is attempting to ameliorate this issue by the introduction of the 'Immunoscore' first proposed by Galon and colleagues (Galon et al., 2006). This score, as discussed briefly in the thesis introduction, centres on IHC staining of CD3 and CD8. Their density is based on two areas of the tumour – the tumour core, and the invasive margin. Subsequently, these densities are combined into a central score which distinguishes the 'immunological-heat' of a tumour, and which has been shown to be more prognostically accurate than MMR status (Mlecnik et al., 2016). The simplicity of this scoring procedure will likely increase chances of clinical implementation, especially considering that IHC is still routine for MMR status delineation, and this utilises just two IHC stains. A likely obstacle for applying Immunoscore into clinical practice, however, is its analysis. Analysis of Immunoscore requires density counts of CD8 and CD3 stains at both the invasive margin and in the tumour core, which if carried out conventionally and counted by pathologists across an

entire slide may be laborious and slow. However, the incorporation of new, digital pathology analyses into this space would greatly ameliorate these speed issues, although these techniques would still require pathologist oversight.

Despite Immunoscore likely impacting clinical practice, it offers little to furthering biological insight (Galon et al., 2014, Galon et al., 2012). For example, in non-MSI-H cancers, where the reason for the increased immune infiltrate is unclear, Immunoscore will provide little understanding as to why. Incorporation of other key mediators of immunity, such as other cell types, or cytokines may give more insight into the origin of increased immunity, however, will complicate clinical implementation. In contrast to Immunoscore, the CMS subgroups proposed by Guinney et al. (2015) offer much to be investigated, however, currently little clinical translatability. The delineation of four subgroups of patients within CRC, each with distinct mechanisms (discussed in detail in thesis introduction) is of high interest to fundamental science. In particular, the delineation of CMS4 is of high interest to this project, as this subgroup displays a high immune signature. The delineation of CMS groups, however, requires RNA sequencing – a technique requiring expertise in analysis and appropriate sample material.

To define an MSS cohort of CRC patients which have an increased intratumoural immune infiltrate, I utilised public databases of CRC sequencing data. I combined this with local cohorts to ensure that these patients could be delineated across multiple platforms, by various approaches. These analyses resulted in generation of a central repository of cross-platform CRC sequencing data of high utility, including for hypothesis generation in this project and in the future. Furthermore, these analyses provide a bioinformatic framework, ultimately of strong relevance across different cancer types, to investigate reasons as to why some cancer patients have an increased immune infiltrate despite a canonically low-mutation burden background.

3.2 Results

3.2.1 Patient characteristics

The CRC datasets from Guinney et al. (2015) were used as input for this thesis. Datasets which had available clinical data were first filtered for only primary samples. Clinical characteristics across all cohorts were split by MMR status and available clinical data can be visualised in Table 3.1. Unfortunately, data were largely incomplete for clinical characteristics, however, MSS and MSI-H were largely comparable in terms of age groups, gender, tendency for being colon or rectal cancer, and cancer stage. Notable differences were seen where MSI-H patients had a greater tendency to be right-sided, as documented extensively elsewhere (Tariq and Ghias, 2016). Furthermore, CMS and CRIS distribution varies greatly between the two groups. MSI-H patients were enriched for CMS1 and CRIS-A subtypes, again in accordance with published literature (Guinney et al., 2015, Sawayama et al., 2020).

Table 3.1 – Clinicopathological characteristics of the COADREAD datasets

Results are stratified by MMR status. The frequency of patients belonging to either MSS or MSI-H in each clinical characteristic subcategory are displayed, and the percentage abundance is given in brackets. Data were taken from GEOquery and TCGAbiolinks. MMR status, CMS subgroups, and CRIS subgroups were taken either from raw clinical data or inferred using CMScaller. Whether a tumour was deemed left or right was based on descriptions from Mik et al., (2017). Stages were condensed to stages I-IV. Non-integers are rounded to 3 significant figures. 'Not Classified' exists in 'Colon or rectal cancer' category due to cancer being classified as "colorectal" in clinical data.

Characteristics		***MMR Status***	
		MSS (%) ***n = 1,625***	**MSI-H (%)** ***n = 569***
Age Group (%)	*20-30*	8 (0.492)	4 (0.703)
	30-40	27 (1.66)	18 (3.16)
	40-50	126 (7.75)	33 (5.80)
	50-60	256 (15.8)	63 (11.1)
	60-70	431 (26.5)	106 (18.6)
	70-80	411 (25.3)	153 (26.9)
	80-90	165 (10.2)	79 (13.9)
	90-100	2 (0.123)	3 (0.527)
	NA	199 (12.2)	110 (19.3)
Gender (%)	*Female*	670 (41.2)	271 (47.6)
	Male	843 (51.9)	208 (36.6)
	NA	112 (6.89)	90 (15.8)
Colon or rectal cancer (%)	*Colon*	1226 (75.4)	394 (69.2)
	Rectum	148 (9.11)	10 (1.76)
	Not Classified	251 (15.4)	165 (29.0)
Histological Subtype (%)	*Adenocarcinoma*	773 (47.6)	229 (40.2)
	Mucinous Adenocarcinoma	52 (3.20)	15 (2.64)
	Adenoma	2 (0.123)	17 (2.99)
	Carcinoma	2 (0.123)	4 (0.703)
	Mucinous Carcinoma	2 (0.123)	15 (2.64)
	NA	794 (48.9)	289 (50.8)
Location (%)	*Left*	578 (35.6)	76 (13.4)
	Rectum	180 (11.1)	12 (2.11)
	Right	370 (22.8)	218 (38.3)
	NA	497 (30.6)	263 (46.2)
Stage (%)	*I*	165 (10.2)	58 (10.2)
	II	436 (26.8)	139 (24.4)
	III	374 (23.0)	79 (13.9)
	IV	137 (8.43)	12 (2.11)
	NA	513 (31.6)	281 (49.4)
Consensus Molecular Subtypes (%)	*CMS1*	55 (3.38)	285 (50.1)
	CMS2	576 (35.4)	4 (0.703)
	CMS3	203 (12.5)	115 (20.2)
	CMS4	568 (35.0)	110 (19.3)
	NA	223 (13.7)	55 (9.67)
CRIS Subtypes (%)	*CRIS A*	271 (16.7)	278 (48.9)
	CRIS B	129 (7.94)	95 (16.7)

	CRIS C	481 (29.6)	4 (0.703)
	CRIS D	218 (13.4)	9 (1.58)
	CRIS E	226 (13.9)	6 (1.05)
	NA	300 (18.5)	177 (31.1)

3.2.2 Validation of the CIRC as an immunity signature

Although the CIRC signature has been validated as a measure of immunity elsewhere (Lal et al., 2015), these analyses were based on the original 28-gene meta-gene signature, including the 4 genes which did not objectively separate using hierarchical clustering analyses. Furthermore, these analyses were correlations with signatures of cell types rather than signatures which are indicative of 'overall immunity'. Here, I used the HTG geneset – a well-documented meta-gene signature which contains 745 genes split over 15 functional groups associated with a general immune response. The CIRC signature showed a highly significant correlation with the HTG dataset on all platforms (Figure 3.1), supporting its role as a surrogate measure of immunity. Furthermore, I utilised the "Immune Function" section of the *ConsensusTME* R package and show much stronger correlations. These data suggest that the CIRC signature acts as a good surrogate for a general immune response, at least to the degree of the HTG and/or the Immune Function section of the *ConsensusTME* R package. Lastly, *ConsensusTME* also contains multiple gene signatures of cell types. The. CIRC signature was highly correlated with all signatures (>0.6 Pearson correlation estimate), except for the non-immune cell types – endothelial cells and fibroblasts (Table 3.2). Interestingly, γδ T-cells showed the greatest correlation out of all cell types present in the analysis. However, from other work, it is extremely difficult to distinguish γδ T-cells from other effector cell types from transcriptional data, and therefore, unlikely to be evidence of a γδ T-cell infiltrate (McMurray et al., 2022).

Figure 3.1 – Relationships between the CIRC gene signature and other immune meta-gene signatures
Enrichment scores for the CIRC gene signature were plotted against enrichment scores of the HTG gene signature (Top row) and the 'Immune Score' geneset from ConsensusTME R package (Bottom row). All three platforms (TCGA-COADREAD [RNA sequencing – Left column], Affymetrix Microarray [Centre column], and Agilent Microarray [Right column]) demonstrated highly significant positive Spearman correlations between the CIRC signature and immune meta-gene signatures (>0.67 Spearman statistic). Bin widths were set according to the Freedman-Diaconis rule.

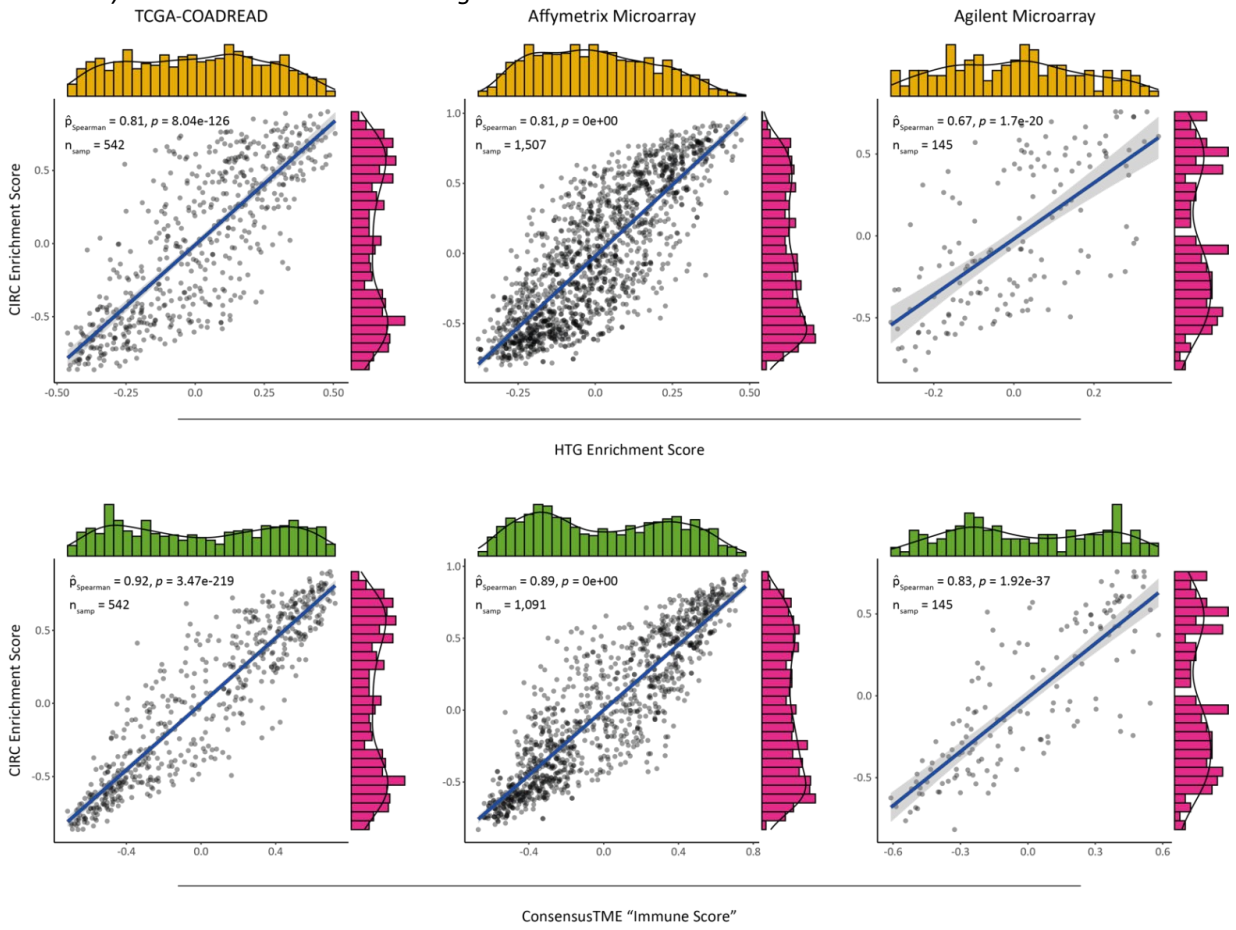


Table 3.2 – Relationships between the CIRC signature and signatures of cell types from the ConsensusTME

CIRC enrichment scores were correlated with enrichment scores of genesets of various cell types from the ConsensusTME R package. The CIRC signature demonstrated highly positive Pearson correlations with all signatures of immune cell types (>0.6) in the ConsensusTME R package on all three platforms. Cell types are ordered by the highest average Pearson correlation estimates across the cohorts. Correlations are rounded to 3 significant figures. Number of samples per group were as follows. TCGA-COADREAD: COAD; 393, READ; 149 Affymetrix: COAD; 1,058, READ; 6. Agilent: COAD; 145, READ; 0.

Cell Type	***Correlation with CIRC Signature (TCGA-COADREAD)***	***Correlation with CIRC Signature (Affymetrix)***	***Correlation with CIRC Signature (Agilent)***	***Average Pearson Correlation Statistic***
γδ T-cells	0.937	0.894	0.878	0.903
NK cells	0.909	0.880	0.895	0.895
CD8 T-cells	0.925	0.896	0.860	0.894
CD4 T-cells	0.908	0.880	0.841	0.876
M1 Macrophages	0.899	0.891	0.783	0.858
Dendritic cells	0.899	0.881	0.784	0.855
T_{REGS}	0.895	0.856	0.776	0.842
Cytotoxic cells	0.857	0.845	0.795	0.832
Macrophages	0.816	0.802	0.704	0.774
M2 Macrophages	0.814	0.783	0.641	0.746
Monocytes	0.806	0.768	0.650	0.741
Mast cells	0.780	0.797	0.644	0.740
Eosinophils	0.799	0.746	0.555	0.700
B cells	0.785	0.666	0.649	0.700
Neutrophils	0.756	0.709	0.608	0.691
Plasma cells	0.739	0.678	0.605	0.674
Endothelial cells	0.456	0.420	0.314	0.397
Fibroblasts	0.368	0.298	0.0598	0.242

3.2.3 Variability of the CIRC signature amongst MSS patients

To investigate the differences in CIRC expression across MMR status that Lal *et al*., (2015) report, I plotted CIRC enrichment scores for each of the MMR subgroups (Figure 3.2). Significant differences were found between groups on all platforms, demonstrating a higher immune-phenotype in MSI-H patients, as reported elsewhere (Zhang *et al.*, 2017). However, while significance did exist between these groups, it was apparent from the violin plots that the MSS setting displayed a higher variation in contrast to MSI-H (Figure 3.2). Statistical tests investigating the homogeneity of variances and their distribution were conducted between the subgroups and demonstrated significance (Table 3.3). The high degree of variability between patients in the MSS grouping suggested that there were patients within this subtype which also display a high CIRC signature expression as Lal and colleagues (2016) previously reported. This serves as evidence that an MSI-like MSS grouping does exist and warrants investigation into the best method of stratifying these patients.

Figure 3.2 – CIRC enrichment scores across MMR status

The enrichment score for the CIRC signature was calculated using RNA sequencing (TCGA-COADREAD - left), Affymetrix (centre) and Agilent (right) microarrays and compared across MMR status. Results are stratified by dataset. The number of patients per MMR status for each of the platforms are shown on the X axis.

Statistical analysis used Mann Whitney Tests for comparisons, and were Holm corrected. TCGA-COADREAD; $CI_{95\%}$ [0.44, 0.66], n_{samp} = 542. Affymetrix; $CI_{95\%}$ [0.34, 0.45], n_{samp} = 1,507. Agilent; $CI_{95\%}$ [0.30, 0.69], n_{samp} = 145.

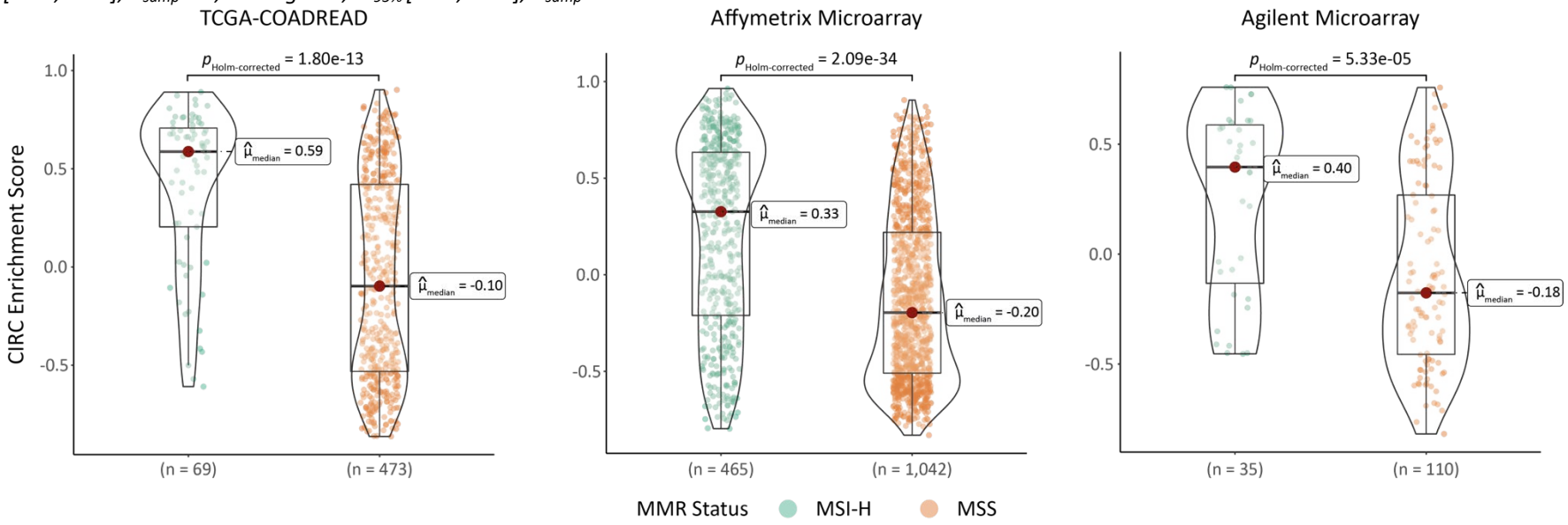


Table 3.3 – Measures of variability between MSS and MSI-H patients

Tests investigating the homogeneity of variances (Fligner test) and differences in distribution (Kolmogorov-Smirnov) were completed between MSI-H and MSS patients using CIRC enrichment scores as input. P-values are stratified by cohort and are rounded to 3 significant figures.

Cohort	***Fligner Test P-value***	***Two-sample Kolmogorov-Smirnov Test P-value***
TCGA-COADREAD	0.00472	6.08e-10
Affymetrix	1.11e-05	< 2.2e-16
Agilent	0.783	0.000924

3.2.4 Utilising the MSI Template from CMScaller to define MSS-hi-CIRC

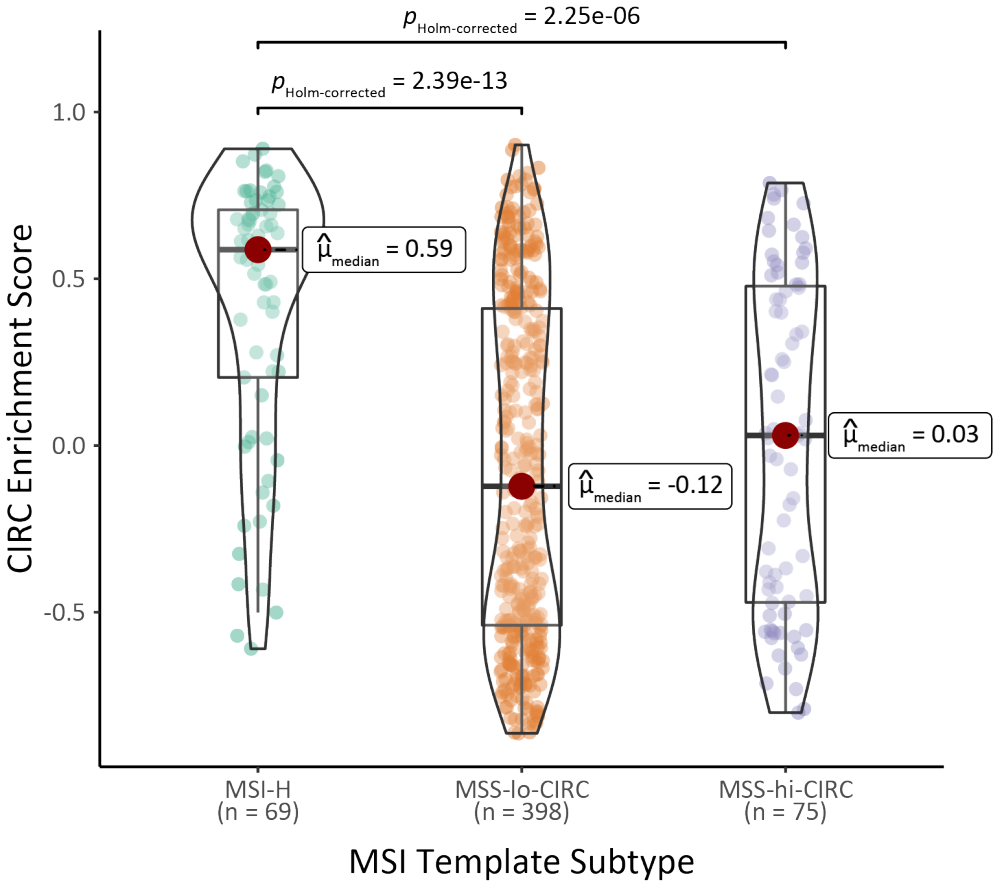
During the imputation of MMR status on each of the platforms to reduce the degree of missing values, I noticed that some patients distinguished as MSS on IHC or by fragment analysis were being determined as MSI-H by the CMScaller algorithm. Therefore, I hypothesised that the algorithm was capturing this MSI-like cohort of MSS patients. By utilising this template of genes and the classifier I aimed to find the patients who had similar expression of the CIRC geneset to the MSI-H cohort – and call these patients MSS-hi-CIRC (Figure 3.3). This method found 79 MSS-hi-CIRC patients, representing 16.7% of MSS patients. This delineation, however, included four *POLE* or *POLD1* mutants, known for their high immune infiltration and therefore these were removed from the subtype of MSS-hi-CIRC. Whilst this method did indeed stratify a noteworthy figure of MSS-hi-CIRC patients (15.9% of MSS patients), it failed to capture the highest expressors of the CIRC signature and find a significant difference between MSS-hi-CIRC and MSS-lo-CIRC. Interestingly, the distribution of MSS-hi-CIRC and MSS-lo-CIRC were indiscernible , with a two-sample Kolmogorov-Smirnov test revealing no significant difference (pVal = 0.506).

Importantly, this template was created by comparing MSI-H patients and MSS patients, inclusive of the MSS-hi-CIRC patients. This would, therefore, likely reduce the number of genes associated with immunity due to the high proportion of MSS patients which demonstrate increased immunity. Also of note, is the distribution of the MSI-H patients. Whilst there are patients with a very high immune score, there also exists a section of patients who have a very low score, again blurring the lines in terms of immune genes being enriched in this geneset. In fact, on closer inspection of this geneset only 5 out of 50 genes could be considered immune-related (*CD55* – complement cascade, *CTSE* – microbial activity, *BNIP3L* – pro-apoptosis cascade, *PBK* – lymphoid activation, and *PSME1* – proteasome activation), with all other genes being associated with DNA damage repair, or ER-stress responses. This geneset, therefore, is unlikely to capture those patients who have been defined

based on their increased expression of an immune signature. Using clustering mechanisms on the MSI-H patients, may ameliorate issues with MSI-H distribution in CIRC score and allow a more direct 'cutting' of MSS patients into MSS-hi-CIRC and MSS-lo-CIRC.

Figure 3.3 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and the MSI-H template

MSS patients in the TCGA-COADREAD dataset were split into MSS-hi-CIRC and MSS-lo-CIRC using the MSI-H template from CMScaller and the CMScaller algorithm. MSS patients misclassified as MSI-H were labelled MSS-hi-CIRC. CIRC enrichment scores were calculated using RNA sequencing data and compared across subtypes. The number of patients per subgroup are shown on the X axis. Statistical analysis used a Kruskal-Wallis test, p-value = 7.52e-13. $CI_{95\%}$[0.06, 0.16]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



3.2.5 Use of an MSI-H based cut-off to define MSS-hi-CIRC

MSI-H patients were used as input to clustering analyses to find a CIRC score cut-off for MSS patients to be split into hi-CIRC and lo-CIRC. The initial geneset described in Lal *et al.,* (2015) was used as the variables and phenograph clustering was employed to determine the number of clusters that appeared within the MSI-H patients, using the `RPhenograph` package in R (R Core Team, 2018, Chen, 2015). According to the modularity rule, three clusters were the optimum number of clusters (modularity value – 0.367 [3 significant figures]). Using these clusters, the CIRC enrichment score highlighted Cluster 1 and Cluster 3 had a much higher CIRC enrichment score than Cluster 2 (Figure 3.4). The number of MSI-H patients that fell into either of these clusters amounted to 42, representing 60.9% of MSI-H patients. Therefore, the CIRC enrichment score representing the 39th (rounded to 2 significant figures) quantile was taken as the cut-off for MSS patients to be deemed MSS-hi-CIRC (CIRC Enrichment score greater than or equal to 0.483 [rounded to 3 significant figures]). This delineation found 93 MSS-hi-CIRC patients (19.7% of MSS patients). The equation to calculate MSS-hi-CIRC patients is shown below.

$$Quantile\ for\ cutoff\ (n)\ =\ 1 - {}^{No.(MSI-H\ patients\ in\ Cluster\ 1\ or\ 3)}/_{Total\ No.MSI-H\ Patients}$$

$$MSS\ patients\ (POLE\ and\ POLD1\ WT)\ CIRC\ score > nth\ quantile\ of\ MSI-H\ patients\ CIRC\ score$$

As previously alluded to, this dataset also included data on *POLE* and *POLED1* mutation status. Both *POLE* and *POLD1* mutants are known for their high mutation burden and by extension, a high intratumoural immunity (Van Gool et al., 2015, Chae et al., 2018) and therefore all patients were checked for non-silent *POLE* or *POLD1* mutations. Three MSS-hi-CIRC patients had non-synonymous mutations in *POLE* or *POLD1* and were removed from this subgroup. By utilising a T_H1-centric gene signature and cut-offs defined based on the MSI-H patients with the highest immune signatures, I report approximately 19.0% of MSS patients demonstrate a high CIRC enrichment score and

therefore, is an indication that these patients may have a denser immune infiltrate than their native MSS grouping.

Figure 3.4 – CIRC enrichment scores across phenograph clusters of TCGA-COADREAD MSI-H patients

Phenograph clustering of TCGA-COADREAD MSI-H patients using the initial geneset revealed three clusters. CIRC enrichment scores were calculated using RNA sequencing data and compared across the clusters. The number of patients per cluster are shown on the X axis.
Statistical analysis used a Kruskal-Wallis test, p-value = 1.11e-11. $CI_{95\%}$[0.68, 0.81]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

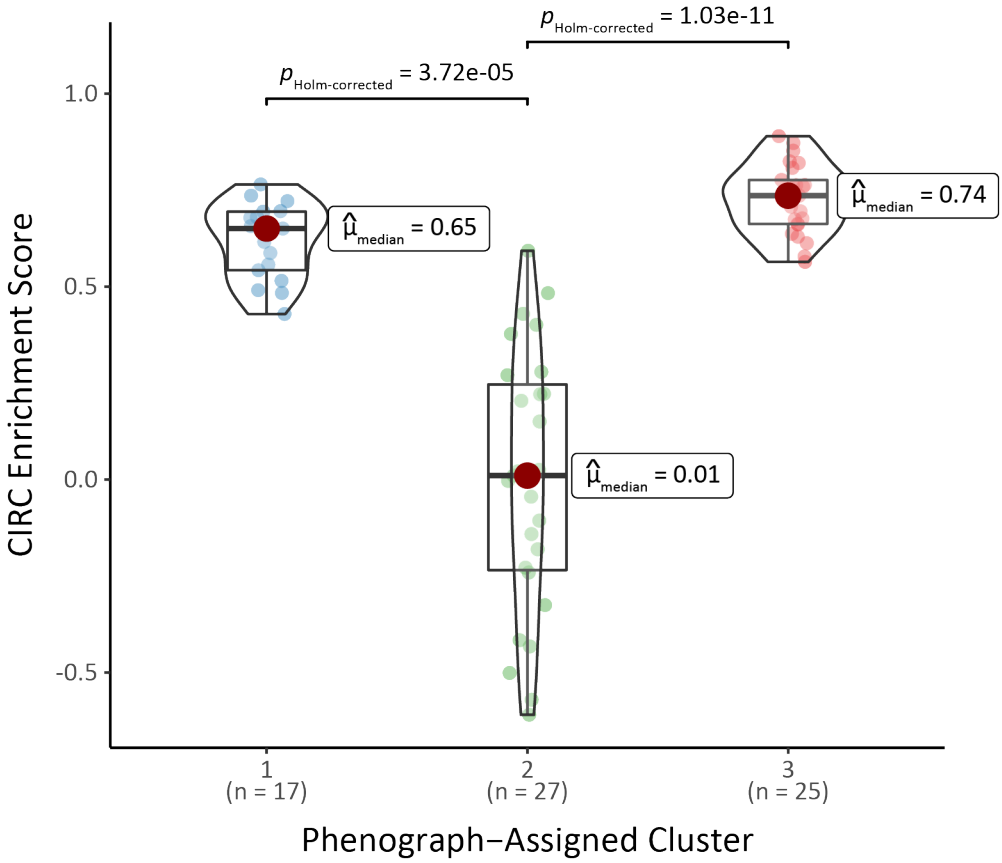
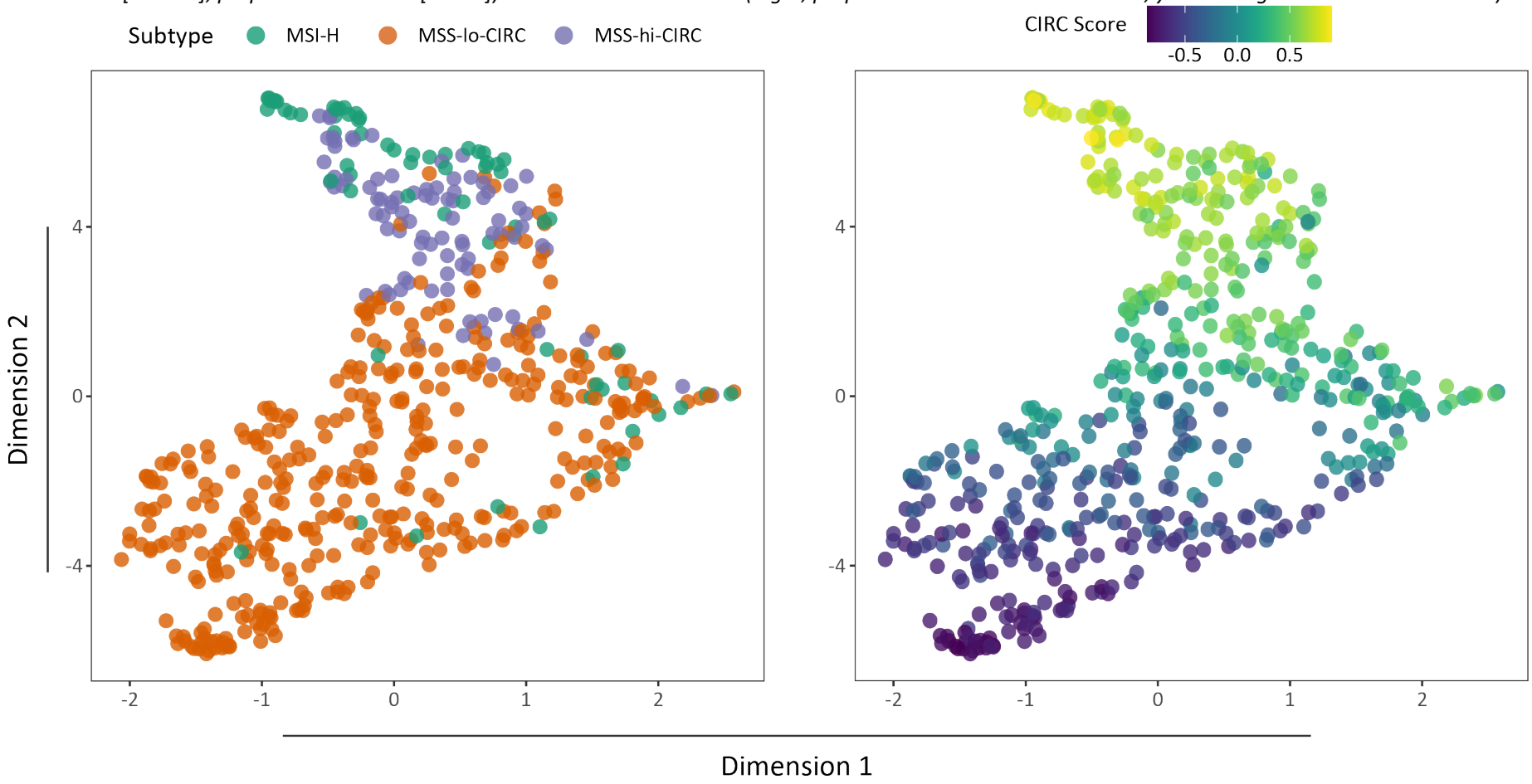


Figure 3.5 – UMAP projections of TCGA-COADREAD subtypes and CIRC enrichment scores

MSS patients in the TCGA-COADREAD cohort were split into MSS-hi-CIRC and MSS-lo-CIRC based on an MSS status delineation by fragment analysis and a CIRC score above 0.483 (MSS-hi-CIRC). UMAP projections based on the genes in the CIRC signature were coloured by subgroup (Left, green = MSI-H [n = 69]; orange = MSS-lo-CIRC [n = 383]; purple = MSS-hi-CIRC [n = 90]) or CIRC enrichment score (Right, purple = low CIRC enrichment score; yellow = high CIRC enrichment score).



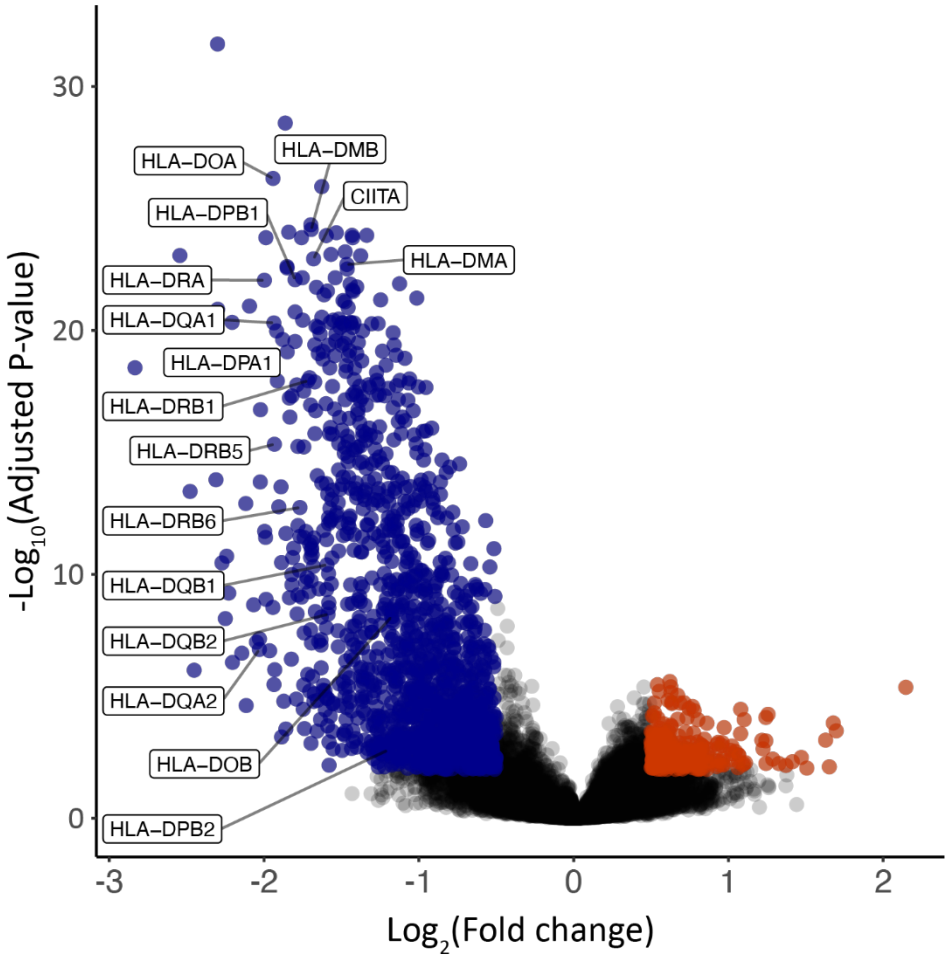
Whilst this represents a good method of distinguishing those patients with the highest CIRC scores, it depends on having enough MSI-H patients sequenced to delineate a cut-off specific to each cohort. Whilst a similar cut-off point was also delineated in the Affymetrix cohort (CIRC enrichment score >= 0.429), a similar cut-off point could not be delineated in the Agilent dataset due to a lack of MSI-H patients. Furthermore, this method also relies on having mutation data present to remove the high immune *POLE*/*POLD1* mutants. These issues raised the possibility of using the TCGA-COADREAD cohort as a 'discovery' group to define a bespoke template which could delineate an 'MSS-hi-CIRC profile' that could be used cross-platform.

3.2.6 Determining an 'MSS-hi-CIRC template'

To allow calling of MSS-hi-CIRC patients, I subjected TCGA-COADREAD patients to differential gene expression analysis to produce a 'MSS-hi-CIRC template' that could be used to train a classifier. Using a randomly chosen half of the MSS patients for training (43 MSS-hi-CIRC patients, 192 MSS-lo-CIRC patients) and the previously defined subtypes delineated from Section 3.2.5, I performed differential gene expression between MSS-lo-CIRC and MSS-hi-CIRC patients and found 2,412 differentially expressed genes (adjusted P-value <= 0.01) (Figure 3.6). Most of the differentially expressed genes were upregulated in the MSS-hi-CIRC setting (1,753) and contained all Class II genes and their master regulator, *CIITA* (Figure 3.6).

Figure 3.6 – Differential gene expression between temporarily assigned MSS subtypes in the TCGA-COADREAD cohort

Volcano plot of differentially expressed genes between MSS-lo-CIRC and MSS-hi-CIRC as defined by an MSI-H derived CIRC enrichment score threshold. Analysis revealed 2,412 differentially expressed genes, the majority of which were upregulated in MSS-hi-CIRC direction. Of these genes, all Class II genes were differentially expressed in addition to their master regulator - CIITA (labelled). Blue points indicate a significant gene in the MSS-hi-CIRC direction, whereas red points indicate a significant enrichment of the gene in the MSS-lo-CIRC setting. Genes are only coloured if the gene has an FDR < 0.01 and a Log_2(Fold change) above 0.5 (red) or below -0.5 (blue). Any genes that didn't pass the threshold are coloured black.



Using the *CMScaller* MSI template as a guide, I next took the top 50 upregulated and top 50 downregulated genes, despite MSS-hi-CIRC and MSS-lo-CIRC patients not being predicted to exist in equal proportions in the patient cohort based on the temporary subtypes. This led to the creation of an 'MSS-hi-CIRC template' that could delineate between MSS-lo-CIRC and MSS-hi-CIRC patients (Table 3.4). I next applied our 'MSS-hi-CIRC template' across the TCGA-COADREAD cohort to call CIRC status. 77.0% of patients were determined successfully as MSS-lo-CIRC (185 patients) or MSS-hi-CIRC (179 patients) with confidence below a 0.005 FDR. 109 patients returned an "NA" score, and therefore, were deemed MSS-lo-CIRC. Five *POLE* or *POLD1* mutants were included in this MSS-hi-CIRC delineation and were removed. With regards to the original delineation of MSS-hi-CIRC via phenograph clustering, only 61.9% of patients kept the same notation. Using CMScaller and a template based on temporary subtypes from MSI-based phenograph clustering, 36.8% of MSS patients could be deemed MSS-hi-CIRC. However, only 84 patients had a CIRC score above the threshold previously set based on the MSI-H patients. Furthermore, whilst this improved on a difference to MSS-lo-CIRC, the "tail" of low expressing MSI-H patients is still recapitulated in this MSS-hi-CIRC grouping (Figure 3.7). To reduce the 'tail' of MSS-hi-CIRC patients, I decided to test other clustering mechanisms.

Table 3.4 – The 'MSS-hi-CIRC' template

The HUGO Symbols, Entrez ID, and FDR values of the top 50 upregulated (right) and downregulated (left) genes between temporarily assigned MSS-hi-CIRC and MSS-lo-CIRC (upregulation/downregulation indicates enrichment in MSS-hi-CIRC). These genes were taken as a template to delineate MSS-hi-CIRC patients. NA where data was not available in the `org.Hs.eg.db` *R package gene annotation.*

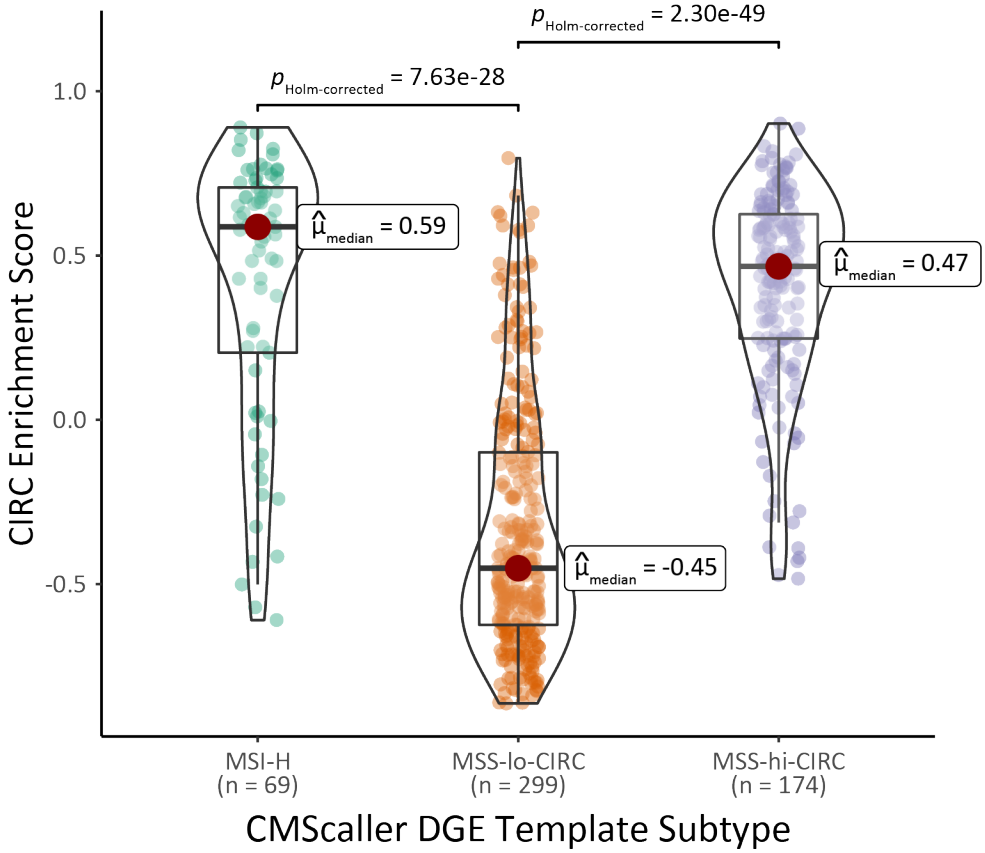
Downregulated in MSS-hi-CIRC			***Upregulated in MSS-hi-CIRC***		
HUGO Symbol	**EntrezID**	**FDR**	**HUGO Symbol**	**EntrezID**	**FDR**
NA	283335	2.50E-06	*GBP5*	115362	1.81E-32
LINC00662	148189	3.35E-06	*GBP4*	115361	3.15E-29
C5orf34	375444	4.11E-06	*HLA-DOA*	3111	5.85E-27
ZNF26	7574	4.18E-06	*IL2RB*	3560	1.27E-26
TRIM71	131405	4.29E-06	*SLAMF8*	56833	4.68E-25
CCT6P3	643180	6.29E-06	*HLA-DMB*	3109	7.25E-25
MNX1	3110	6.97E-06	*CCL5*	6352	9.34E-25
CBR3-AS1	100506428	9.17E-06	*CCR5*	1234	9.94E-25
MARVELD2	153562	1.15E-05	*ARHGAP30*	257106	1.26E-24
HYKK	123688	1.45E-05	*WAS*	7454	1.26E-24
SPC24	147841	1.81E-05	*ITGAL*	3683	1.29E-24
LEAP2	116842	1.84E-05	*GBP1*	2633	1.57E-24
ZNF620	253639	1.88E-05	*LILRB4*	11006	1.57E-24
C21orf58	54058	1.92E-05	*TIGIT*	201633	1.57E-24
GPR89B	51463	1.93E-05	*LCP2*	3937	6.00E-24
ARHGEF38	54848	1.98E-05	*IL10RA*	3587	7.69E-24
DUS4L	11062	2.04E-05	*CXCL9*	4283	8.46E-24
METAP1D	254042	2.10E-05	*HCLS1*	3059	8.61E-24
MNX1-AS1	645249	2.61E-05	*CIITA*	4261	1.16E-23
ANAPC7	51434	2.67E-05	*HLA-DMA*	3108	1.97E-23
NA	100506098	3.13E-05	*PLEK*	5341	2.43E-23
C2orf70	339778	3.43E-05	*CD8A*	925	2.77E-23
ZNF77	58492	3.80E-05	*CYTH4*	27128	3.04E-23
NUP62CL	54830	3.91E-05	*CD84*	8832	6.93E-23
PLCH1	23007	4.77E-05	*CMKLR1*	1240	6.95E-23
APTR	100505854	5.27E-05	*HLA-DPB1*	3115	8.34E-23
WFDC10A	140832	5.48E-05	*HLA-DRA*	3122	8.75E-23
MCM3AP-AS1	114044	6.33E-05	*ARHGAP25*	9938	1.22E-22
RBM14-RBM4	100526737	6.55E-05	*IGSF6*	10261	1.25E-22
MTERF2	80298	7.12E-05	*SLA*	6503	1.63E-22
DOK7	285489	7.41E-05	*IL2RA*	3559	1.71E-22
RCCD1	91433	7.41E-05	*DOK2*	9046	2.20E-22
BNIPL	149428	7.60E-05	*DOCK2*	1794	2.42E-22
ARL6IP4	51329	7.94E-05	*PIK3R5*	23533	2.48E-22
NVL	4931	7.96E-05	*SIGLEC10*	89790	3.49E-22
IQCH-AS1	100506686	8.11E-05	*IL18BP*	10068	4.72E-22
C14orf93	60686	8.64E-05	*AIF1*	199	5.61E-22
DGUOK-AS1	100874048	8.67E-05	*LAIR1*	3903	5.92E-22
LRRC36	55282	9.04E-05	*NA*	388813	6.68E-22
PKP4-AS1	100129029	9.14E-05	*SLAMF7*	57823	1.00E-21
PCBD2	84105	0.000103	*HCK*	3055	1.16E-21
ZNF138	7697	0.000104	*CXCL10*	3627	1.36E-21
F2	2147	0.000125	*FPR3*	2359	1.74E-21
NUTM2G	441457	0.000127	*CD53*	963	2.89E-21
LSG1	55341	0.000131	*CD86*	942	3.43E-21

TRIM45	80263	0.000133	*LILRB1*	10859	3.79E-21
NA	100288748	0.000153	*PTPRC*	5788	3.79E-21
OSGEPL1	64172	0.000156	*CD74*	972	4.68E-21
THUMPD2	80745	0.000158	*IDO1*	3620	4.68E-21
GOLGA6L9	440295	0.000171	*SASH3*	54440	4.68E-21

Figure 3.7 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and the 'MSS-hi-CIRC template'

MSS patients in the TCGA-COADREAD dataset were split into MSS-hi-CIRC and MSS-lo-CIRC using the 'MSS-hi-CIRC template' (Table 3.4) and the CMScaller algorithm. CIRC enrichment scores were calculated using RNA sequencing data and compared across the subtypes. The number of patients per subgroup are shown on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 2.47e-60. $CI_{95\%}$[0.45, 0.57]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



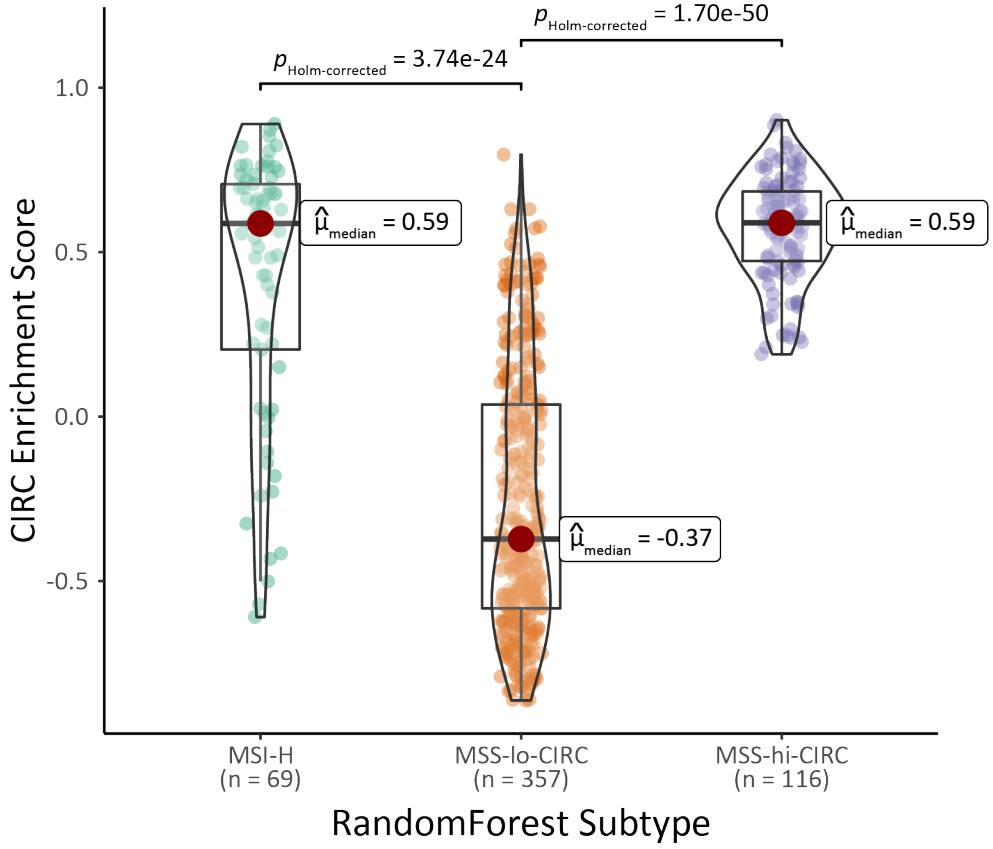
3.2.7 Application of Random Forest to classify MSS-hi-CIRC

Using the subtypes from the previous sections, and the differentially expressed genes between the subtypes. The top 100 most differentially expressed genes between the temporary assigned subtypes in the TCGA-COADREAD cohort were used to train a random forest classifier using 75% of the data to train and the rest to test. This classifier also utilised UPC-transformed data, rather than the CQN transformed data as used previously. To produce a cross-platform classifier, a method of normalisation which would put all data on the same scale is necessary. It is also necessary to not have this normalisation be dependent on the samples within the data. The out of box (OOB) error for the random forest classifier was 9.86% (9.41% for MSS-lo-CIRC, and 11.8% for MSS-hi-CIRC). With regards to the temporary defined subtypes, the classifier was 86.4% accurate. The classifier found 119 MSS patients that could be deemed MSS-hi-CIRC. Three patients were determined as *POLE* or *POLD1* mutants, therefore, 116 patients were determined as MSS-hi-CIRC or 24.5% of the cohort (Figure 3.8). This classifier significantly reduced the 'tail' of MSS-hi-CIRC patients seen with other classifiers, and therefore, was tested on other cohorts.

Figure 3.8 – CIRC enrichment scores of TCGA-COADREAD subtypes as defined by MMR status and a RandomForest classifier

MSS patients in the TCGA-COADREAD dataset were split into MSS-hi-CIRC and MSS-lo-CIRC using a trained RandomForest algorithm based on the 'MSS-hi-CIRC template' (Table 3.4). CIRC enrichment scores were calculated using RNA sequencing data and compared across subtypes. The number of patients per subgroup are shown on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 1.84e-61. $CI_{95\%}$[0.47, 0.56]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



3.2.8 Cross-platform testing of the classifier

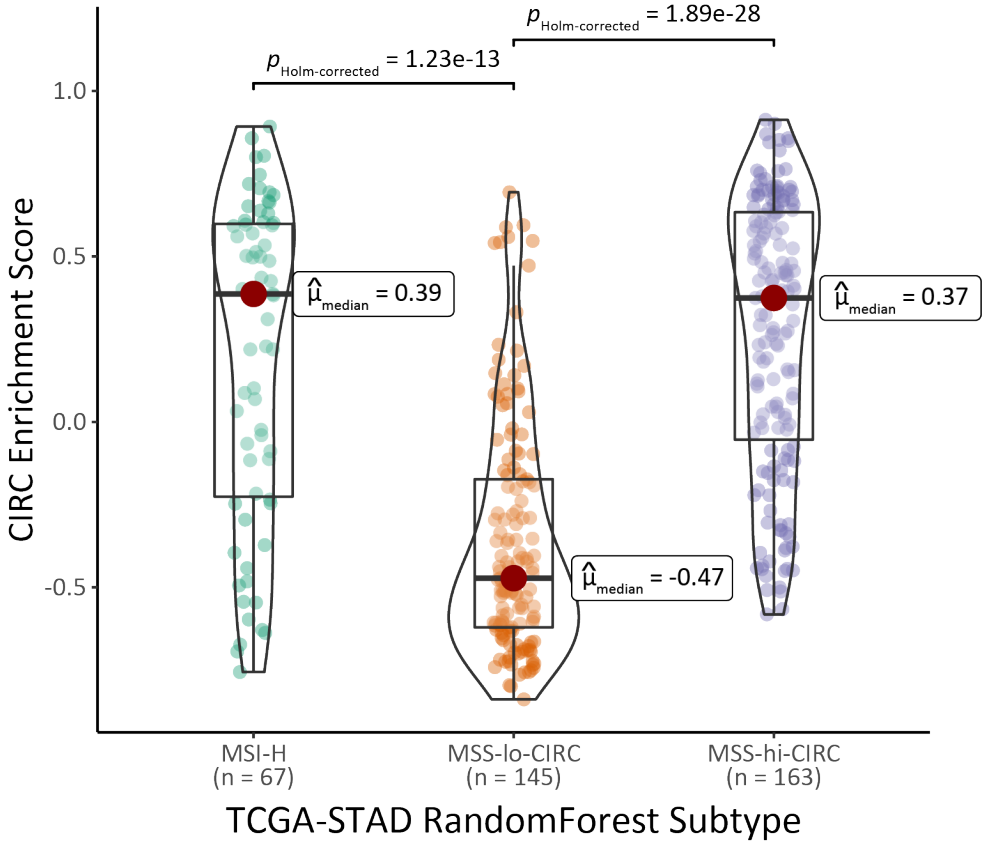
3.2.8.1 TCGA-STAD

First, I tested this data on the TCGA-STAD cohort – a cohort of patients which has been processed in the exact same manner as the TCGA-COADREAD cohort, however, in a separate disease setting. This classifier successfully delineated patients as MSS-hi-CIRC, however, not to the success that it worked with TCGA-COADREAD. This classifier drastically overestimates the number of MSS-hi-CIRC patients in the TCGA-STAD cohort (Figure 3.9), as seen by the large variability in CIRC score across this cohort. These differences likely arise due to the difference in primary site between the TCGA-COADREAD and TCGA-STAD cohorts and therefore, if using a classifier pan-cancer a core consensus number of genes would be required.

Figure 3.9 – CIRC enrichment scores of TCGA-STAD subtypes as defined by MMR status and a RandomForest classifier

MSS patients in the TCGA-STAD dataset were split into MSS-hi-CIRC and MSS-lo-CIRC by TCGA-COADREAD trained RandomForest algorithm and the 'MSS-hi-CIRC template' (Table 3.4). CIRC enrichment scores were calculated using RNA sequencing data and compared across subtypes. The number of patients per subgroup are shown on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 4.95e-30. $CI_{95\%}$[0.29, 0.45]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



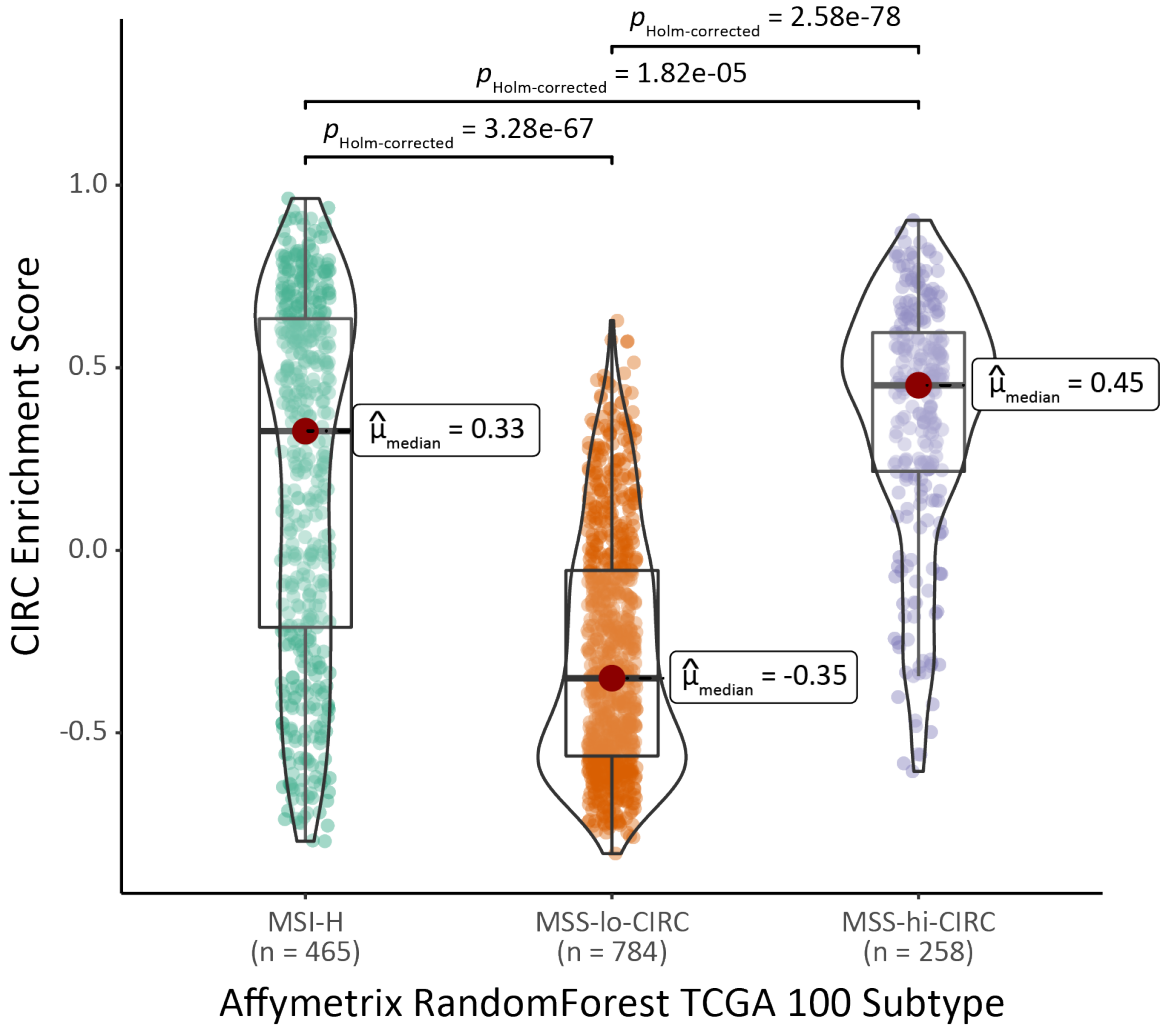
3.2.8.2 Affymetrix Microarray

Unfortunately, 3 genes out of the top 100 differentially expressed genes were not present in the Affymetrix cohort. Therefore, the Random Forest classifier that was trained on the TCGA-COADREAD cohort could not be used due to these missing values. Instead, I proceeded to retrain the classifier using the top 97 genes and the largest Affymetrix cohort - GSE39582. This required the temporary subtypes to again be defined using the cut-offs - in similarity to TCGA-COADREAD training. Affymetrix MSI-H patients split into three clusters using Phenograph, with those MSS patients having a higher CIRC score than 0.429 being considered MSS-hi-CIRC. These temporary subtypes were used to determine how accurate this new classifier was. The OOB error for this classifier was 9.97% (MSS-lo-CIRC = 9.27%, MSS-hi-CIRC = 13.5%). However, this classifier failed to capture the same success that was seen with the TCGA-COADREAD dataset (Figure 3.10), again producing a large 'tail' of low CIRC signature expressing MSS-hi-CIRC patients.

Figure 3.10 – CIRC enrichment scores of Affymetrix CRC subtypes as defined by MMR status and a RandomForest classifier

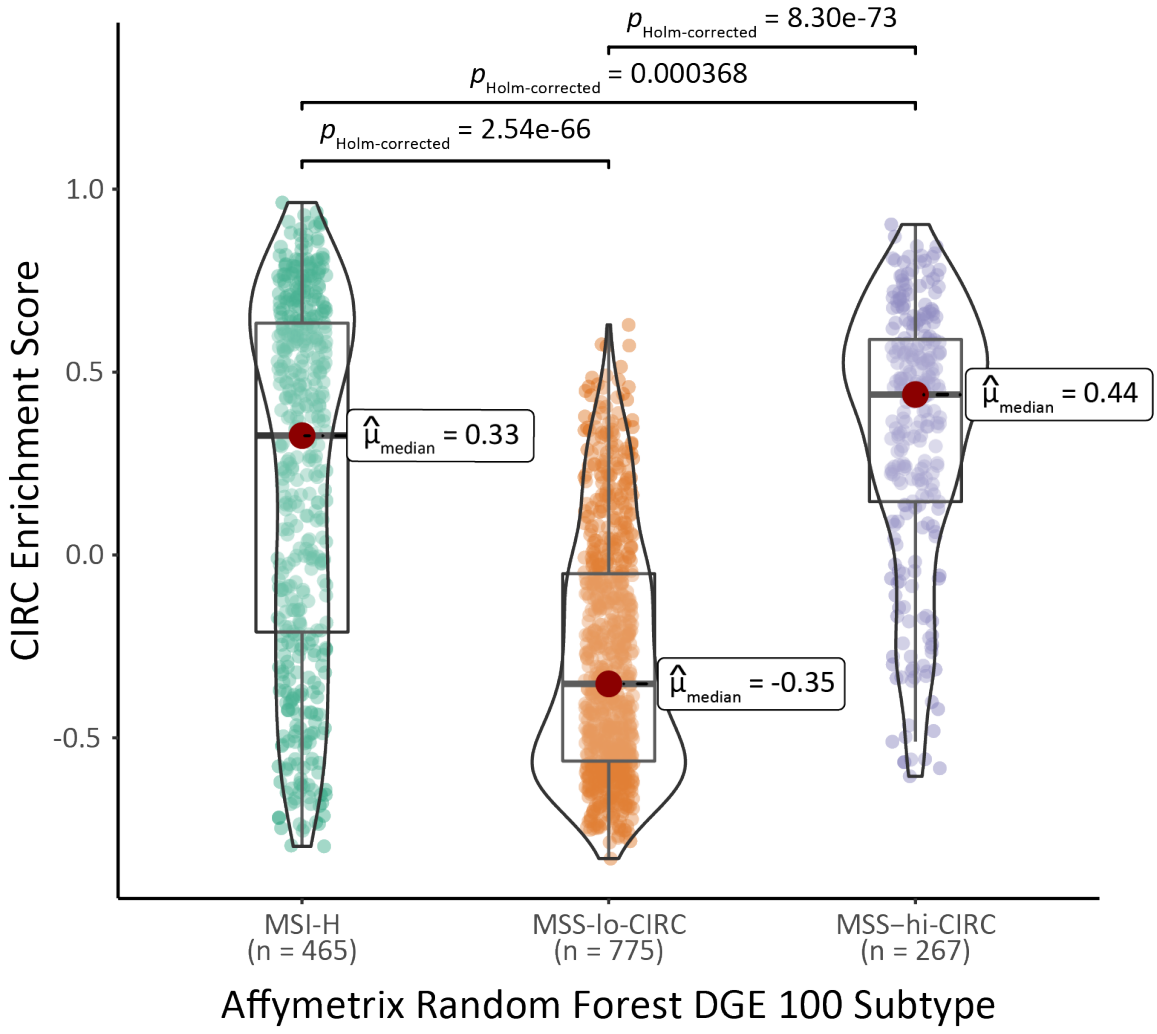
MSS patients in Affymetrix datasets were split into MSS-hi-CIRC and MSS-lo-CIRC using a TCGA-COADREAD trained RandomForest algorithm and the top 97 genes from the 'MSS-hi-CIRC template' (Table 3.4). CIRC enrichment scores were calculated using Affymetrix microarrays and compared across subtypes. The number of patients per subgroup are shown on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 6.56e-110. $CI_{95\%}$[0.31, 0.38]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



I hypothesised that different genes may work better on the separate platforms, and therefore, I recapitulated the process of creating the random forest classifier that I did on the TCGA-COADREAD cohort. Again, the GSE39582 cohort was used for training, however, the top 100 differentially expressed genes between the Affymetrix subtypes were used for input. This classifier performed worse than the previous using the TCGA-COADREAD 97 genes with an OOB of 10.61% (8.49% for MSS-lo-CIRC, and 21.1% for MSS-hi-CIRC) (Figure 3.11). The overestimation of MSS-hi-CIRC by both classifiers is not acceptable when trying to determine a cohort of patients which will share underlying biology.

Figure 3.11 – CIRC enrichment scores of Affymetrix CRC subtypes as defined by MMR status and a RandomForest classifier trained on differentially expressed genes in the GSE39582 cohort
MSS patients in Affymetrix datasets were split into MSS-hi-CIRC and MSS-lo-CIRC using a trained RandomForest classifier based on the top 100 genes differentially expressed between temporarily assigned subtypes in the GSE39582 cohort. CIRC enrichment scores were calculated using Affymetrix microarrays and compared across subtypes. The number of patients per subgroup are shown on the X axis.
Statistical analysis used a Kruskal-Wallis test, p-value = 2.03e-104 $CI_{95\%}$[0.29, 0.36]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



3.2.9 The Definition of MSS-hi-CIRC

Our definition of MSS-hi-CIRC, therefore, used the phenograph-derived cut-offs using MSI-H patients from each of the platforms (Section 3.2.5). Although this method relies on having MSI-H patients present, it will reliably stratify the MSS patients who are the highest expressors of the CIRC signature. Whilst this means that in the Agilent dataset there are not enough MSI-H patients to reliably call clusters, I used the average number of MSS-hi-CIRC patients I found in the other two cohorts to take the top X expressors of the CIRC signatures as MSS-hi-CIRC. In the TCGA-COADREAD cohort, 19.0% of MSS patients were determined MSS-hi-CIRC, whereas in the Affymetrix dataset this was much lower in that 14.8% of MSS patients were considered MSS-hi-CIRC (Figure 3.12). Therefore, MSS-hi-CIRC could be represented as the top 16.9% of MSS patients in terms of CIRC score. Using this as a delineation, I determined a cut-off for the Agilent dataset, and future datasets (Figure 3.13).

Figure 3.12 – CIRC enrichment scores of Affymetrix subtypes as defined by MMR status and an MSI-H derived CIRC enrichment score cut-off

MSS patients in Affymetrix datasets were split into MSS-hi-CIRC and MSS-lo-CIRC based on a CIRC enrichment score cut-off as determined by Phenograph clustering on MSI-H patients from the GSE39582 cohort. CIRC enrichment scores were calculated using Affymetrix microarrays and compared across subtypes. The number of patients per subgroup are shown on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 3.05e-116 $CI_{95\%}$[0.33, 0.39]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

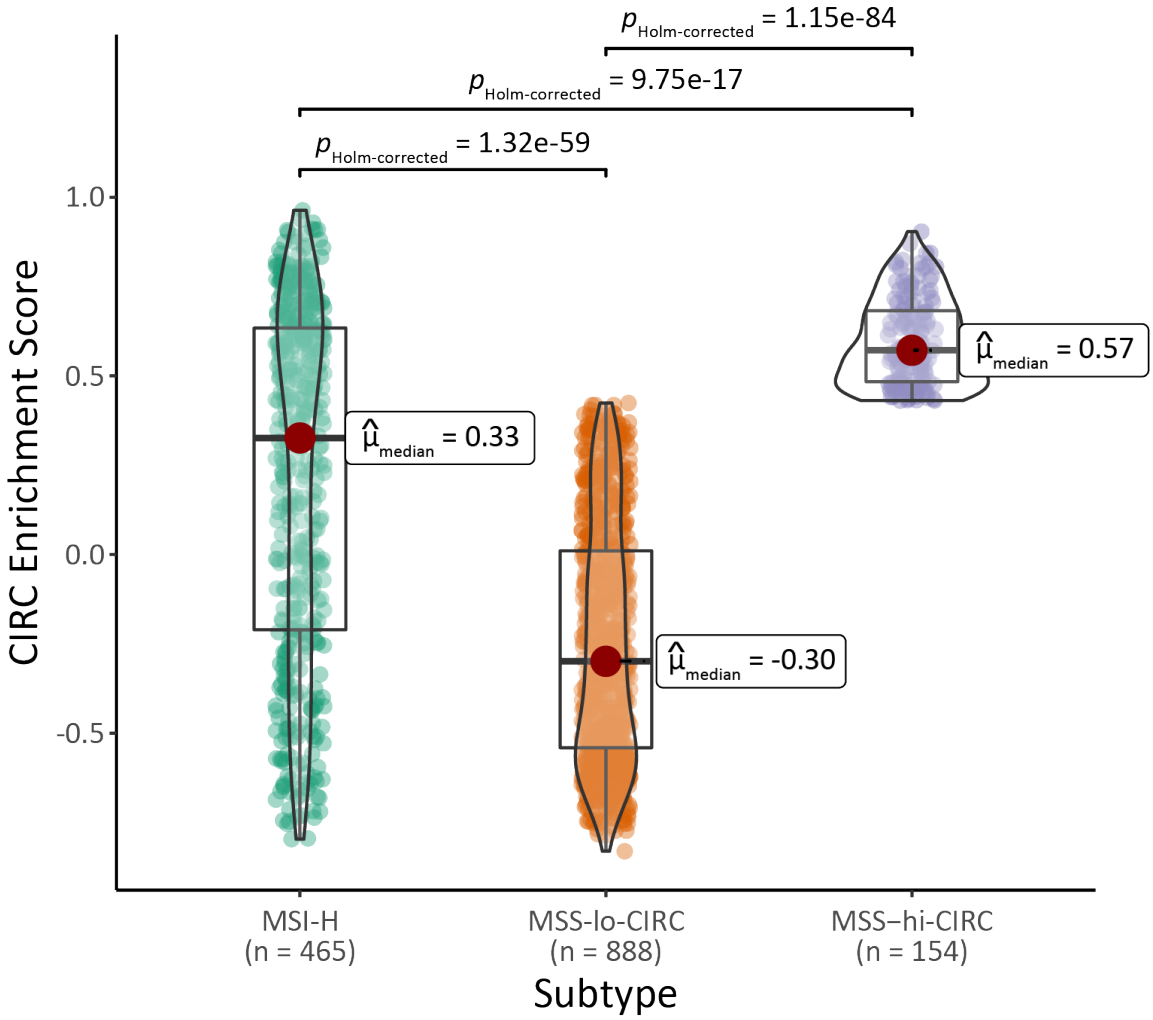
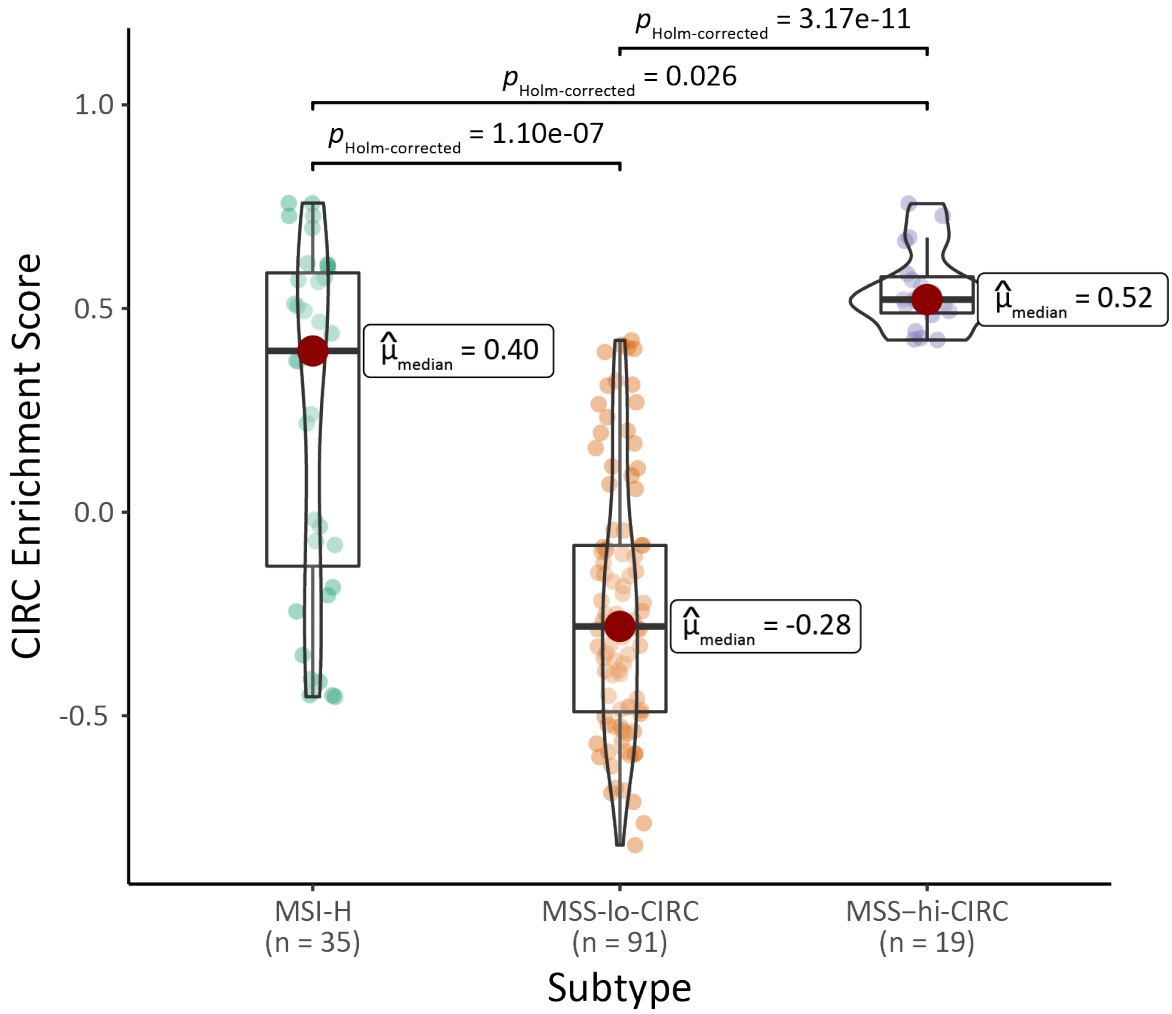


Figure 3.13 – CIRC enrichment scores of Agilent subtypes as defined by MMR status and the average number of MSS-hi-CIRC patients found in TCGA-COADREAD and Affymetrix datasets
MSS patients in the Agilent dataset were split into MSS-hi-CIRC and MSS-lo-CIRC by taking the top 16.9% of MSS CIRC enrichment scores as MSS-hi-CIRC patients. CIRC enrichment scores were calculated using Agilent microarrays and compared across subtypes. The number of patients per subgroup are shown on the X axis.
Statistical analysis used a Kruskal-Wallis test, p-value = 2.59e-14 $CI_{95\%}$[0.33, 0.54]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



3.2.10 Comparison of clinical characteristics between MSS-hi-CIRC, MSI-H and MSS-lo-CIRC

To determine whether the subgroups demonstrated any significant differences with regards to clinical characteristics I carried out multinomial linear regression. To ensure no clinical characteristics could separate MSS-hi-CIRC and MSS-lo-CIRC patients, MSS-hi-CIRC patients were set as the 'baseline' and the reference groups for each of the variables were set to the group which harboured most MSS-hi-CIRC patients (for example, CMS4 represented the most MSS-hi-CIRC patients out of all the CMS subgroups, and therefore, was set as the reference for these comparisons) (Table 3.5 and Table 3.6).

Table 3.5 – Clinicopathological characteristics of COADREAD datasets stratified by threshold-derived subtypes

Results are stratified by subgroup. The frequency of patients belonging to MSI-H, MSS-lo-CIRC, or MSS-hi-CIRC in each clinical characteristic subcategory are displayed, and the percentage abundance is given in brackets. Data were taken from GEOquery and TCGAbiolinks. Data were only included if it were used as input for multinomial logistic regression. MMR status, CMS subgroups, and CRIS subgroups were taken either from raw clinical data or inferred using CMScaller and their respective templates. Whether a tumour was deemed left or right was based on descriptions from (Mik et al., 2017). Stages were condensed to stage I-IV. Non-integers are rounded to 3 significant figures. Data for age group were condensed here for subsequent analyses.

Characteristics		***Subtype***		
		MSI-H (%) ***n = 569***	**MSS (%)** ***n = 1,362***	**MSS-hi-CIRC (%)** ***n = 263***
Age Group (%)	*Under 40*	22 (3.87)	28 (2.06)	7 (2.66)
	40-50	33 (5.8)	107 (7.86)	19 (7.22)
	50-60	63 (11.1)	212 (15.6)	44 (16.7)
	60-70	106 (18.6)	355 (26.1)	76 (28.9)
	70-80	153 (26.9)	348 (25.6)	63 (24.0)
	Over 80	82 (14.4)	136 (9.99)	31 (11.8)
	NA	110 (19.3)	176 (12.9)	23 (8.75)
Gender (%)	*Female*	271 (47.6)	547 (40.2)	123 (46.8)
	Male	208 (36.6)	707 (51.9)	136 (51.7)
	NA	90 (15.8)	108 (7.93)	4 (1.52)
Location (%)	*Left*	76 (13.4)	490 (36.0)	88 (33.5)
	Rectum	12 (2.11)	149 (10.9)	31 (11.8)
	Right	218 (46.2)	301 (22.1)	69 (26.2)
	NA	289 (50.8)	422 (31.0)	75 (28.5)
Stage (%)	*I*	58 (10.2)	138 (10.1)	27 (10.3)
	II	139 (24.4)	366 (26.9)	70 (26.6)
	III	79 (13.9)	310 (22.8)	64 (24.3)
	IV	12 (2.11)	117 (8.59)	20 (7.60)
	NA	281 (49.4)	431 (31.6)	82 (31.2)
Consensus Molecular Subtype (%)	*CMS1*	285 (50.1)	32 (2.35)	23 (8.75)
	CMS2	4 (0.703)	528 (38.8)	48 (18.3)
	CMS3	115 (20.2)	185 (13.6)	18 (6.84)
	CMS4	110 (19.3)	414 (30.4)	154 (58.6)
	NA	55 (9.67)	203 (14.9)	20 (7.60)
CRIS Subtype (%)	*CRIS A*	278 (48.9)	219 (16.1)	53 (20.2)
	CRIS B	95 (16.7)	87 (6.39)	42 (16.0)
	CRIS C	4 (0.703)	421 (30.9)	59 (22.4)
	CRIS D	9 (1.58)	192 (14.1)	26 (9.89)
	CRIS E	6 (1.05)	194 (14.2)	32 (12.2)
	NA	177 (31.1)	249 (18.3)	51 (19.4)

Table 3.6 – Multinomial logistic regression of clinicopathological characteristics between CRC subtypes

*MSS-hi-CIRC patients were taken as the reference variable and compared to MSI-H and MSS-lo-CIRC using the variables displayed. Categories demonstrating the highest proportion of MSS-hi-CIRC patients were chosen as reference categories. Values are rounded to 3 significant figures. Lower confidence represents 2.5% and upper confidence represents 97.5%. Significance codes: * P < 0.05, ** P-value < 0.01. P-values are in both bold and red font if P < 0.001. Ref. – Reference category; OR – odds ratio.*

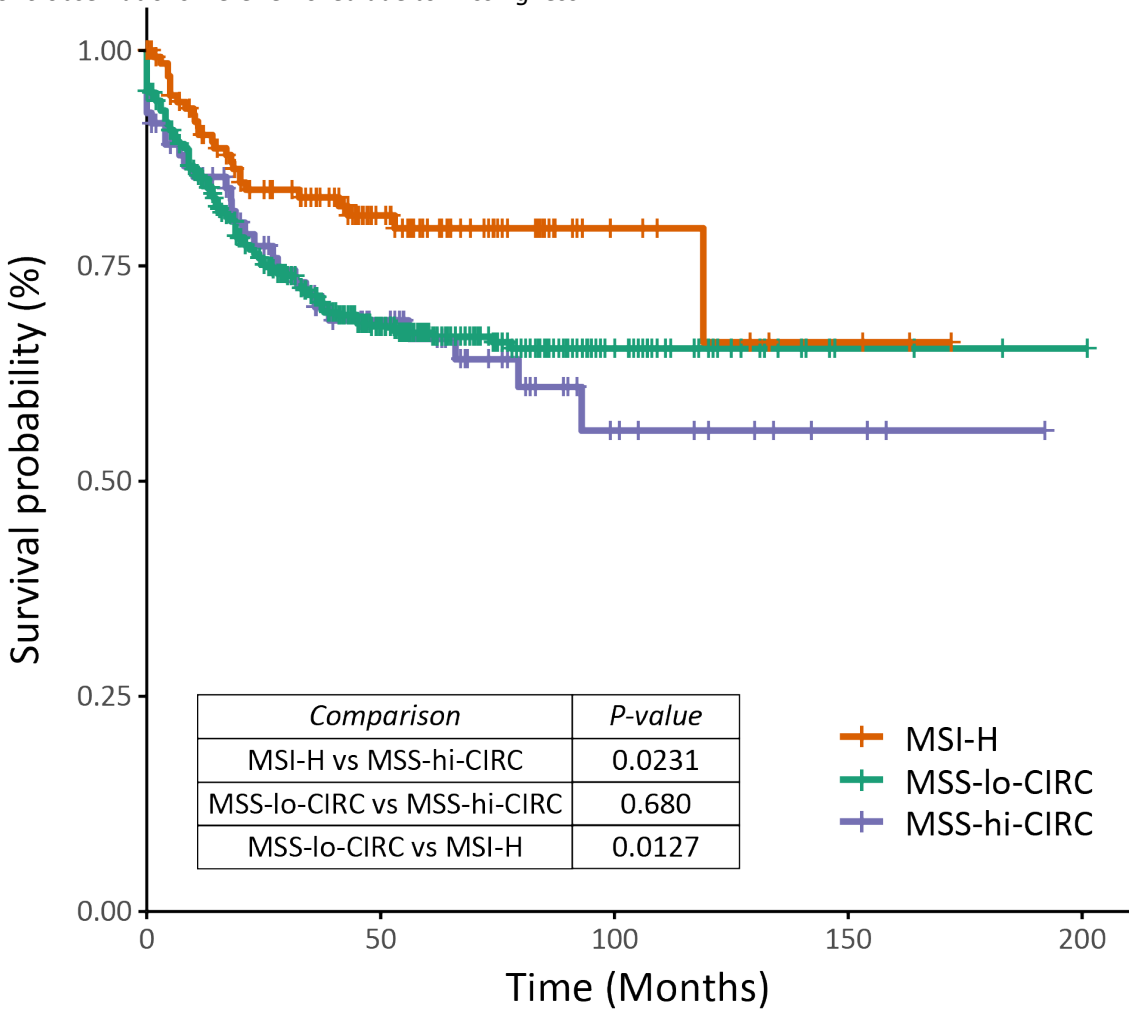
Variables	***Categories***	***MSI-H OR (p-value)***	***95% Confidence Interval***		***MSS-lo-CIRC OR (p-value)***	***95% Confidence Interval***	
			Lower	***Upper***		***Lower***	***Upper***
Intercept	-	0.0589 (**7.73e-04**)	0.00113	0.307	2.77 (0.00131) **	1.49	5.16
Age Group	*Under 40*	2.94 (0.304)	0.375	23.0	1.90 (0.445)	0.366	9.86
	40-50	1.28 (0.666)	0.415	3.97	1.34 (0.446)	0.631	2.85
	50-60	1.30 (0.586)	0.508	3.32	1.08 (0.814)	0.578	2.01
	60-70$^{Ref.}$	-	-	-	-	-	-
	70-80	1.33 (0.479)	0.607	2.91	1.44 (0.193)	0.832	2.48
	Over 80	2.08 (0.163)	0.745	5.79	1.35 (0.463)	0.604	3.03
Gender	*Female*	1.02 (0.959)	0.560	1.84	0.938 (0.769)	0.611	1.44
	Male$^{Ref.}$	-	-	-	-	-	-
Stage	*1*	0.868 (0.779)	0.324	2.33	0.639 (0.224)	0.311	1.32
	2$^{Ref.}$	-	-	-	-	-	-
	3	0.663 (0.227)	0.343	1.29	1.10 (0.708)	0.672	1.79
	4	0.314 (0.0513)	0.0980	1.01	1.07 (0.855)	0.540	2.10
Left/Right/Rectum	*Left*$^{Ref.}$	-	-	-	-	-	-
	Right	1.60 (0.194)	0.788	3.24	0.734 (0.241)	0.438	1.23
	Rectum	0.435 (0.173)	0.131	1.44	0.628 (0.118)	0.350	1.13
CMS Group	*CMS1*	8.23 (**2.15e-06**)	3.44	19.7	1.09 (0.845)	0.463	2.57
	CMS2	0.550 (5.92e-01)	0.0620	4.88	3.82 (**4.76e-06**)	2.15	6.78
	CMS3	5.83 (**2.57e-04**)	2.27	1.50	7.54 (**4.22e-07**)	3.45	16.5
	CMS4$^{Ref.}$	-	-	-	-	-	-
CRIS Group	*CRIS A*	8.84 (0.00894) **	1.73	45.3	0.555 (0.0887)	0.282	1.09
	CRIS B	5.85 (0.0384) *	1.10	31.1	0.384 (0.00766) **	0.190	0.776
	CRIS C$^{Ref.}$	-	-	-	-	-	-
	CRIS D	1.61 (0.713)	0.128	20.1	1.92 (0.0931)	0.897	4.09
	CRIS E	2.69e-06 (**0**)	2.69e-06	2.69e-06	0.889 (0.752)	0.429	1.84

Interestingly, the combination of clinical variables showed differences between the subtypes with both MSI-H and MSS-lo-CIRC being significantly different from MSS-hi-CIRC. Interestingly, though, the significance only existed in the consensus molecular subtypes and the CRIS subtypes, which are subtypes created based on shared underlying biology. There was no propensity for age, gender, stage, or, surprisingly, left/right/rectum sidedness being different between the subtypes. In terms of overall survival (OS), the subtypes showed no differences, in keeping with when MSS-hi-CIRC and MSS-lo-CIRC patients were shown as a collective group. However, there did exist a significant difference in relapse free survival (RFS) between the subtypes (Figure 3.14). With MSS-hi-CIRC demonstrating a significantly shorted RFS compared to MSI-H, consistent with MSS-hi-CIRC demonstrating an enrichment of CMS4-type cancers. It is, however, important to note that majority of patients lacked data for OS and RFS.

Figure 3.14 – Kaplan-Meier curves of relapse-free survival stratified by CRC subgroups

Relapse free survival data for all public cohorts were collated. Comparisons were completed using a Cox proportional hazards regression model.

Likelihood ratio test = 7.82 on 2 df, p = 0.02001. Number of patients = 668, number of events = 190. 1,526 observations were removed due to missingness.



Comparison	*P-value*
MSI-H vs MSS-hi-CIRC	0.0231
MSS-lo-CIRC vs MSS-hi-CIRC	0.680
MSS-lo-CIRC vs MSI-H	0.0127

3.2.11 Condensing the CIRC to an immunohistochemistry stain

To investigate the potential to call hi-CIRC status between MSS patients on IHC, I performed differential gene expression between MSS-hi-CIRC patients in both the TCGA-COADREAD and Affymetrix cohorts. This revealed many genes which were significantly different to an adjusted P-value of less than 0.01 on the two platforms (4,195 on TCGA-COADREAD, and 6,423 on Affymetrix). A large proportion of these genes were shared between the platforms (Figure 3.15). All Class II genes, along with their master regulator, *CIITA*, were present in this shared list of differentially expressed genes, however, represented 12 out of 2,606. We, therefore, took an average of the adjusted P-values across the two cohorts and ordered these by the lowest P-value to the highest. Out of the top 10, Class II genes and *CIITA* represented 50%. Other notable members included the IFNγ-induced guanylate-binding proteins, *GBP1*, *GBP4*, and *GBP5*, as well as the as well as the tryptophan-metabolising enzyme, *IDO1,* and the chemokine ligand *CCL5* (Table 3.7).

Figure 3.15 – Frequency of overlapping differentially expressed genes between MSS-hi-CIRC and MSS-lo-CIRC in Affymetrix and TCGA-COADREAD datasets

Differential gene expression analysis performed between MSS-hi-CIRC and MSS-lo-CIRC in the Affymetrix cohort revealed 6,423 transcripts (green). Of these, 2,324 genes were also differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC patients in the TCGA-COADREAD dataset (orange).

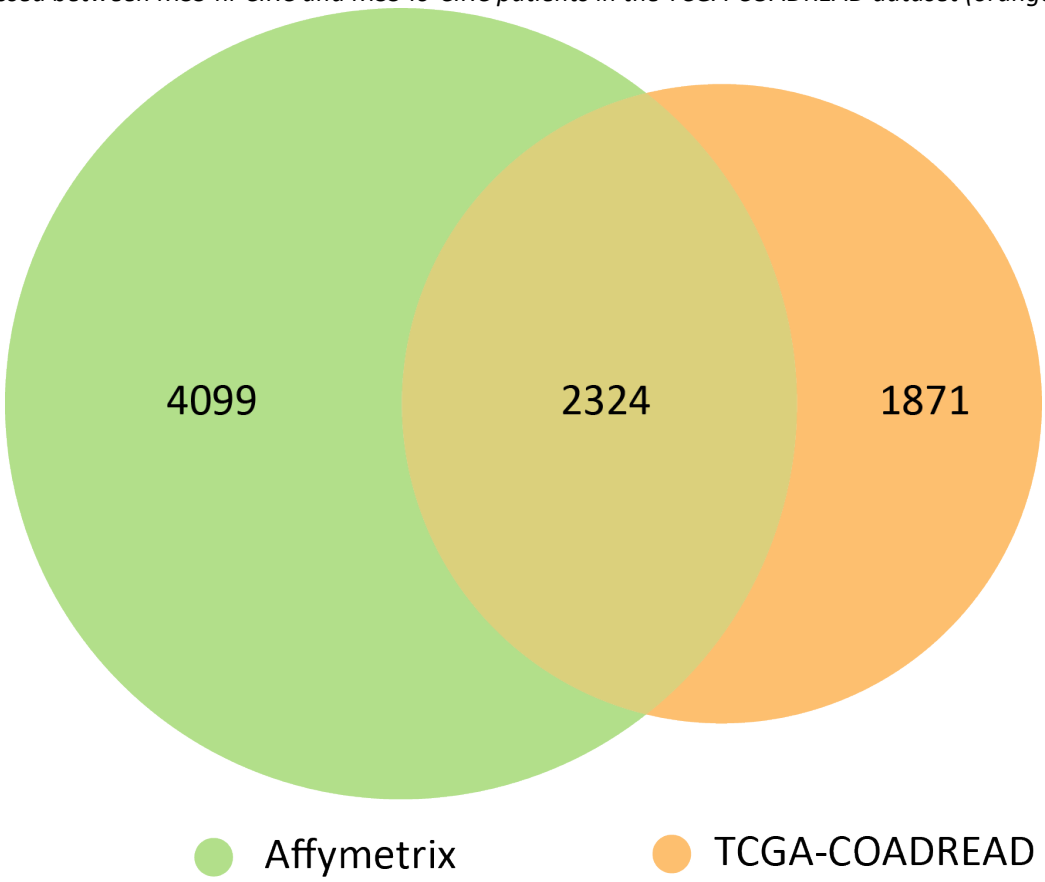


Table 3.7 – List of the top 10 most differentially expressed genes across TCGA-COADREAD and Affymetrix cohorts

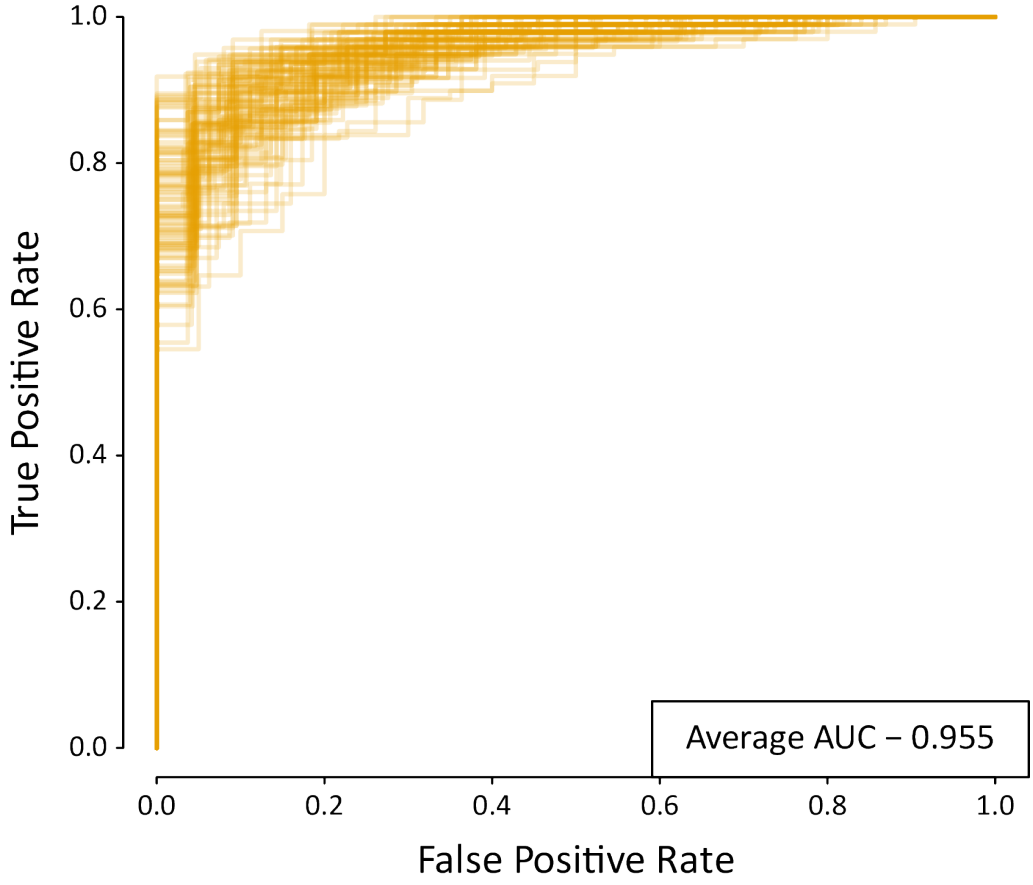
P-values for differentially expressed genes between MSS-hi-CIRC and MSS-lo-CIRC in Affymetrix and TCGA-COADREAD cohorts were averaged and ordered lowest to highest. The top 10 most significant differentially expressed genes are shown and stratified by cohort. An enrichment in the MSS-hi-CIRC patients with regards to MSS-lo-CIRC is indicated by a positive Log2(fold change). Values were rounded to 3 significant figures.

Symbol	Entrez ID	TCGA-COADREAD		Affymetrix Datasets		Average adjusted P-value
		Adjusted P-Value	*Log_2(Fold Change)*	*Adjusted P-Value*	*Log_2(Fold Change)*	
GBP5	115362	3.07e-61	2.20	4.05e-107	1.23	1.54e-61
HLA-DOA	3111	1.72e-56	1.92	2.18e-96	0.944	8.61e-57
HLA-DMB	3109	7.00e-56	1.75	2.03e-90	1.30	3.50e-56
GBP4	115361	1.53e-52	1.79	9.11e-103	1.23	7.67e-53
HLA-DMA	3108	2.17e-52	1.47	4.13e-81	1.48	1.09e-52
IDO1	2620	4.13e-50	2.51	1.17e-98	1.79	2.07e-50
CIITA	4261	5.02e-50	1.73	1.12e-53	0.571	2.51e-50
GBP1	2633	1.13e-49	1.40	5.68e-72	1.54	5.65e-50
CCL5	6352	1.47e-49	1.75	4.90e-72	1.73	7.34e-50
HLA-DRA	3122	3.75e-49	2.00	7.25e-66	1.66	1.87e-49

Due to the high number of Class II genes in this result, I decided to test a machine learning algorithm which assessed whether MHC Class II expression alone could correctly stratify those patients who were MSS-hi-CIRC from those who were MSS-lo-CIRC. TCGA-COADREAD data were filtered to contain only MSS patients and randomly split into two sets 100 times, training on 75% of the patients and testing the naïve bayes model on the remaining 25% (Meyer et al., 2017). A ROC curve was utilised to assess the ability of MHC Class II expression to predict whether the patient was MSS-hi-CIRC or not (MSS-lo-CIRC) (Sing et al., 2005). Figure 3.16 clearly shows that MHC Class II expression was very accurate at predicting MSS-hi-CIRC status with regards to MSS (AUC – 0.955, 95.5% of patients correctly predicted). These results suggest that MHC Class II can successfully delineate MSS-hi-CIRC patients from MSS patients and supports their central location in the heatmap of genes in Lal *et al*., (2015).

Figure 3.16 – ROC curve demonstrating the accuracy of HLA Class II genes to predict MSS-hi-CIRC patients from MSS-lo-CIRC patients in the TCGA-COADREAD cohort

A naïve bayes classifier was trained using Class II genes on TCGA-COADREAD MSS patients and tested 100 times to detect MSS-hi-CIRC patients from MSS-lo-CIRC patients. For each permutation, the Area Under the Curve (AUC) was calculated. The average AUC was taken and displayed on the plot (95.5%).



Using paired RNA sequencing and an HLA-Class II antibody, which stains HLA-DP, HLA-DQ, and HLA-DR, I demonstrated that the enrichment of *HLA-DPB1*, *HLA-DRB1*, and *HLA-DQB1* and the CIRC signature correlate with HLA Class II protein, albeit modestly (Figure 3.17). However, I noted that much of the cohort was negative for MHC Class II, suggesting issues with staining. In contrast, the IHC expression of Class II was highly variable across a separate MSS cohort from Wales Cancer Bank (WCB) (Figure 3.18). This, combined with references of other cohorts demonstrating similar variabilities in MHC Class II expression indicated that MHC Class II expression may capture the variability in the CIRC score shown in RNA-based data (Lal, 2016, Goussous, 2020).

Figure 3.17 – Relationships between MHC Class II protein and MHC Class II transcript or CIRC score

FFPE slides from MSS patients were stained for HLA class II. Enrichment scores for Class II transcripts and the CIRC signature were calculated using RNA sequencing data. Positive trends are shown between the intensity of Class II IHC staining with Class II transcript enrichment scores (Left) and with the CIRC enrichment score (Right) in the 100K genomes cohort. Statistical analysis used Spearman correlations. Bin widths were set according to the Freedman-Diaconis rule.

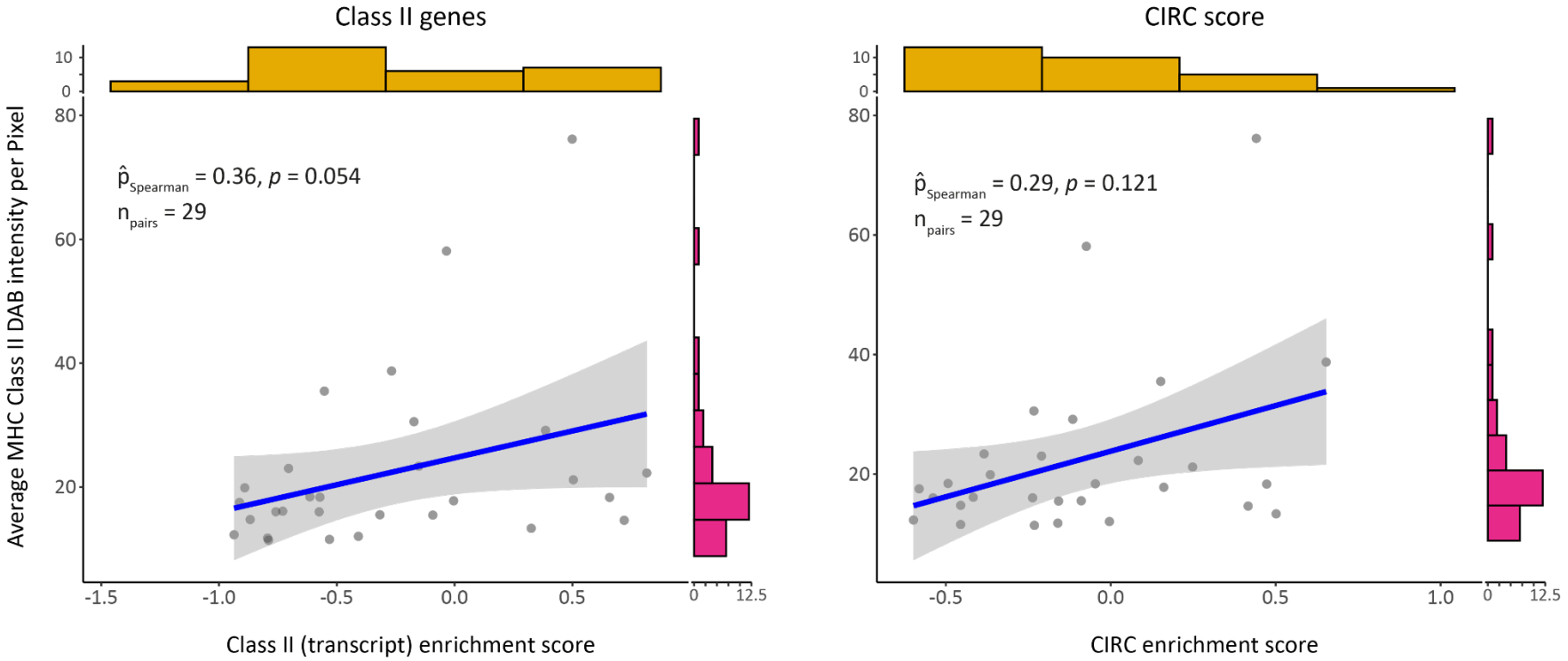
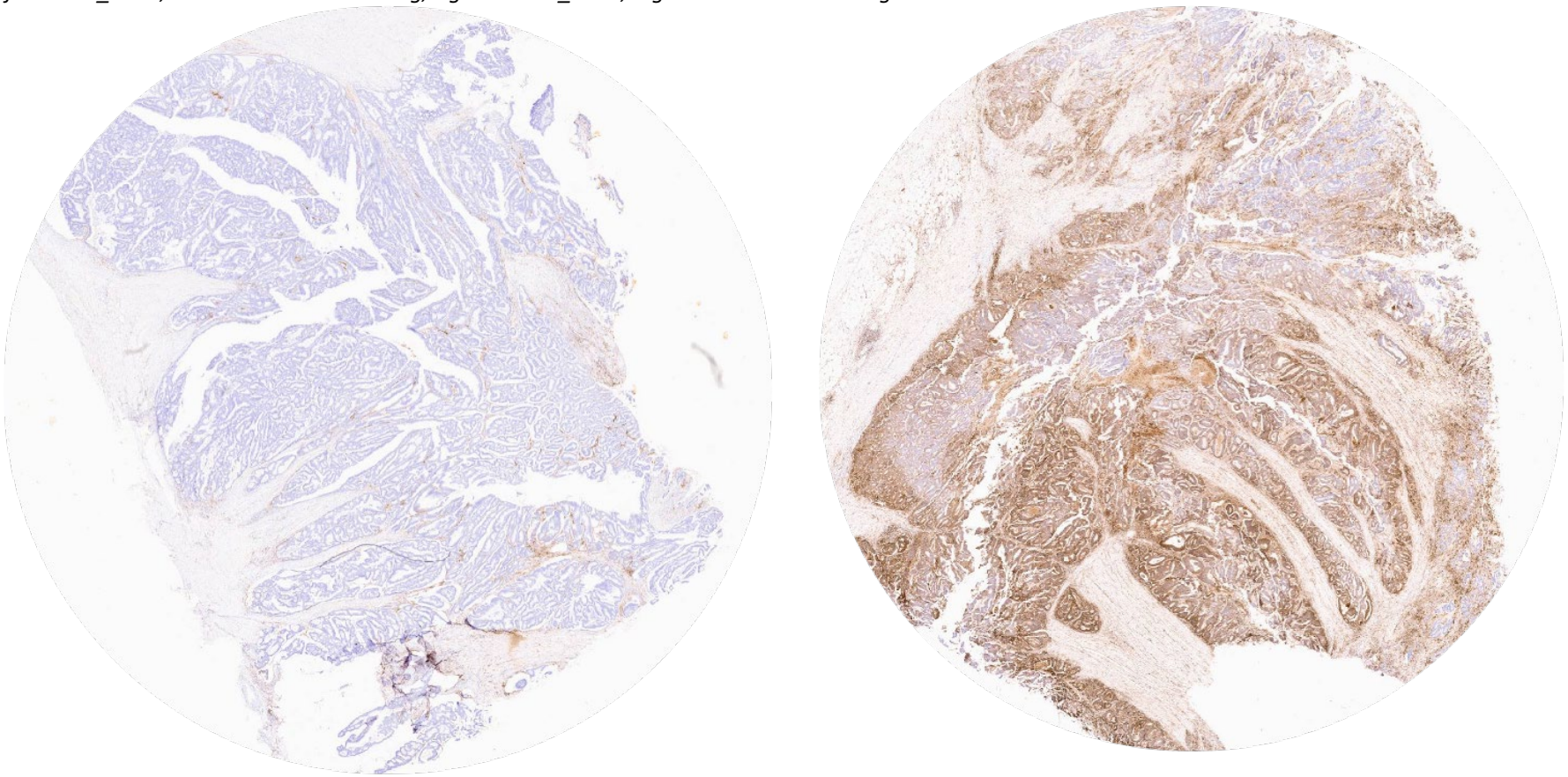


Figure 3.18 – The range of MHC Class II staining in MSS colorectal cancer in the Wales Cancer Bank cohort

FFPE slides were stained for HLA-Class II. Two images from a cohort of 30 MSS patients were chosen for representation of the variability of MHC Class II expression within the MSS subgroup.

Left – CRAC_0129, low MHC Class II staining, right – CRAC_0015, high MHC Class II staining.



3.2.12 MSS-hi-CIRC patients have a comparable immune infiltrate to MSI-H patients

To ensure that MHC Class II could stratify MSS patients who had a high immune infiltrate. I took MSS-hi-CIRC patients (high MHC Class II IHC score) from a previously defined cohort and combined these with MSS and MSI-H patients from a separate cohort (Goussous, 2020, Lal et al., 2015). Figure 3.19 displays representative images of the three subtypes and demonstrates that MSS-hi-CIRC resemble the mutation-rich MSI-H patients rather than their native MSS grouping. This can be quantified by taking the average density of the phenotypes in either the stroma or epithelium. I reduced this data onto two PCA plots of the top 3 principal components (Figure 3.20). These three components were chosen for visualisation due to them having eigenvalues above 1.5. Whilst MSS-lo-CIRC patients seemed to hover around a negative point for PC1 and zero for both PC2 and PC3, the other two subtypes moved along the axes. These axes, being a combination of all the parameters, are essentially measures of overall immunity and demonstrate that MSS-hi-CIRC and MSI-H have an increased immune presence.

Figure 3.19 – Representative multispectral immunofluorescence images of CRC subtypes

Composite images for each subtype were chosen as a representation of 10 MSI-H patients, 8 MSS-lo-CIRC patients, and 12 MSS-hi-CIRC patients. Images demonstrate the range of immune infiltration in the different subgroups. Colours represent the markers outlined in Table 2.4: Red – MHC Class II, Yellow – CD8, Green – CD4, Cyan – T-bet, Orange – FOXP3, RORγt – Magenta, DAPI – Blue.

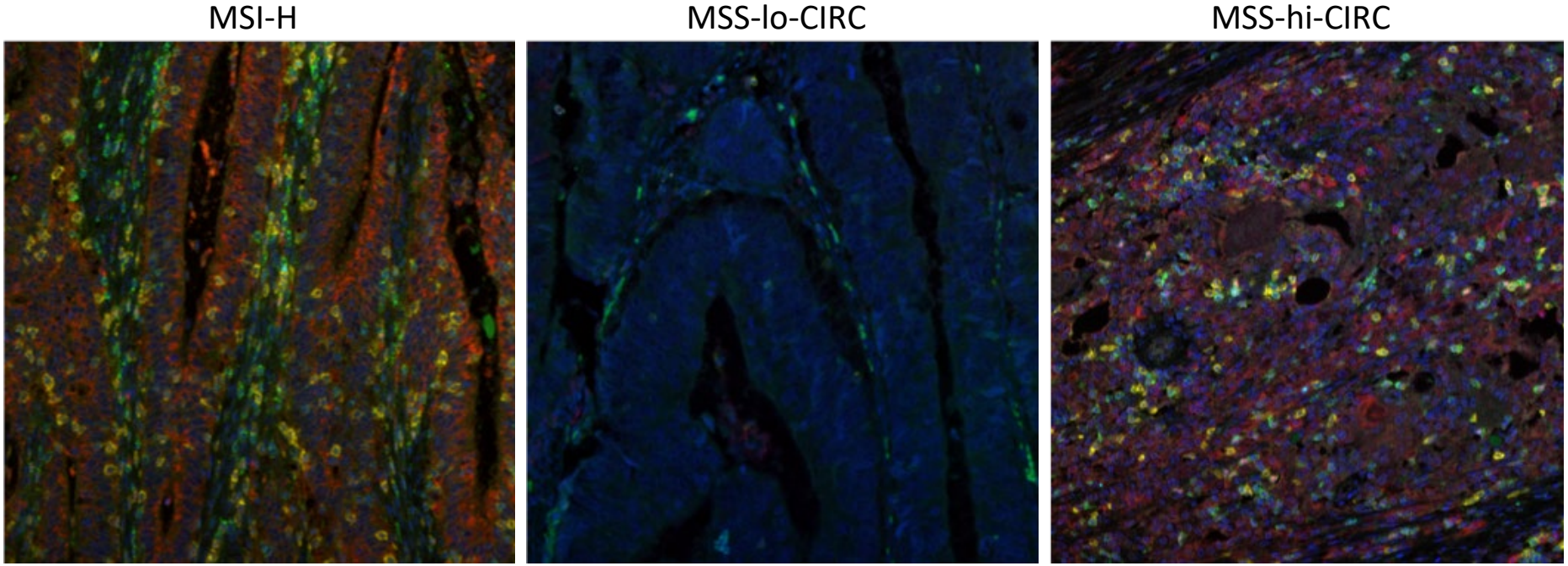
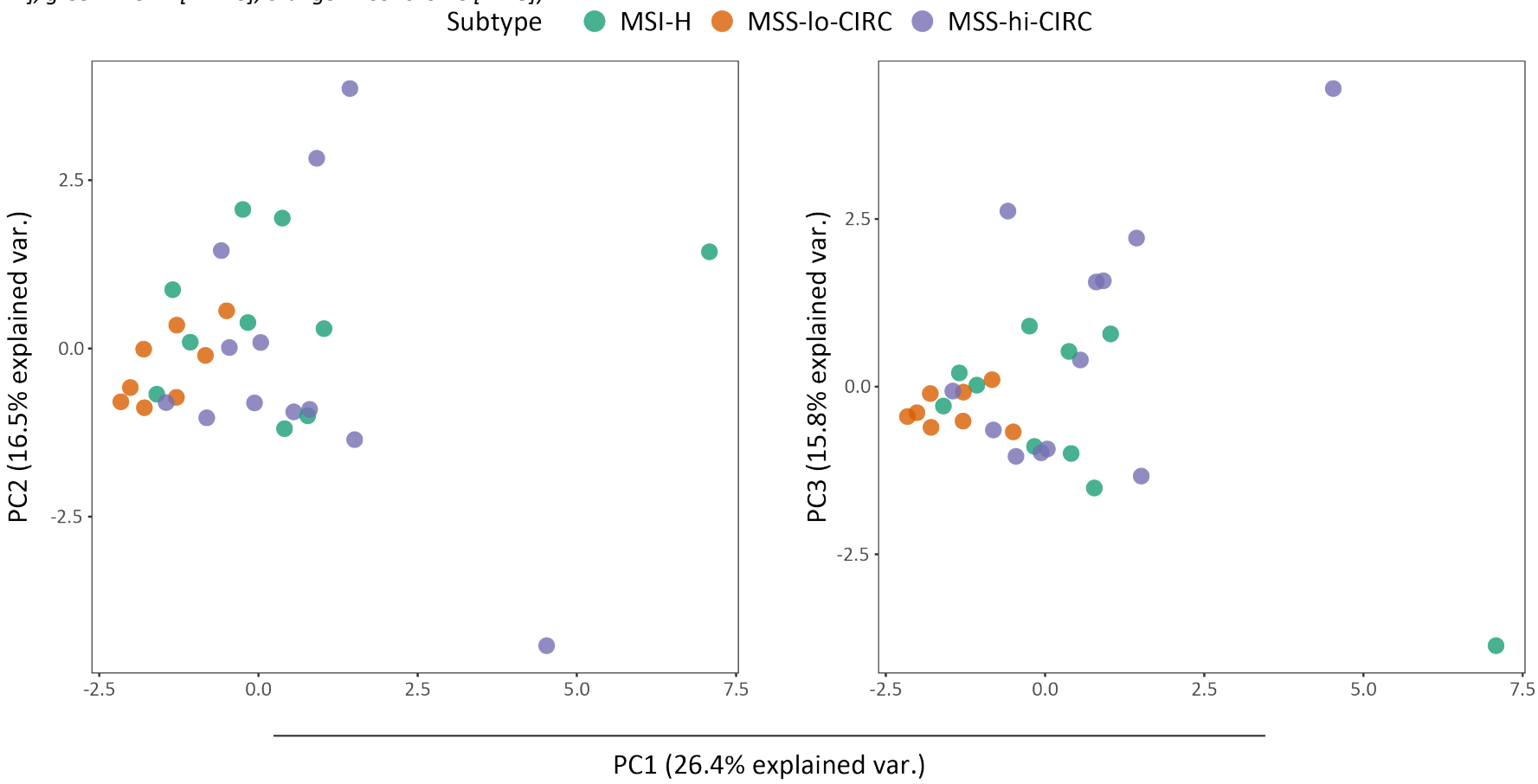


Figure 3.20 – Principal component analysis of immune cell densities characterised by multispectral immunofluorescence

The average density per mm^2 of cells phenotyped as Class II, CD4 only, CD8 only, T_H1 (CD4 + T-bet), T_H17 (CD4 & RORγt), T_C1 (CD8 & T-bet), and T_{REGS} (CD4 & FOXP3) in both the stroma and epithelium were used as input for principal component analysis. The first three principal components had an eigenvalue above 1.5 and were used as axes (PC1 – X axis; PC2 and PC3 – Y axes). Points represent patients and were coloured for subtypes (purple: MSS-hi-CIRC [n = 12], green: MSI-H [n = 10], orange: MSS-lo-CIRC [n = 8]).



3.3 Discussion

Here I utilised multiple clustering mechanisms and multi-platform RNA-based data to define a group of patients whose tumours are MSS but have an increased immune infiltration. These analyses pay homage to, but refine, the threshold-based stratification undertaken by Lal (2016). Furthermore, whilst they concluded by utilising a threshold-based stratification itself, they raise the possibility of combining multiple platforms to determine a consensus set of genes that are common between MSS-hi-CIRC patients on different platforms and across multiple cancers.

The CIRC signature contains nine MHC Class II genes, along with other genes related to a T_H1 and the cytolytic response (Lal et al., 2015), and here I show that MHC Class II alone can accurately predict MSS-hi-CIRC from MSS-lo-CIRC patients using both RNA data and demonstrates high variability across the MSS cohort using IHC. The fundamental role of MHC Class II in anti-cancer immunity has been highlighted by two separate studies investigating *CIITA*, the master regulator of MHC Class II expression (Mortara et al., 2006, Meazza et al., 2003). The transfection of *CIITA* into highly metastatic MHC-Class-II-negative, but MHC-Class-I-positive, tumour cell lines rendered these cells MHC Class II positive, highly immunogenic *in vivo*, and capable of establishing immunological memory. Collectively these studies suggest that MHC Class II, arguably even more than MHC Class I, is pivotal in establishing a robust immune response, and may act as a potential prognostic biomarker. Clinical studies of melanoma have also provided support for MHC Class II as a prognostic biomarker, in which MHC Class II-positive tumours had a higher response rate, higher progression-free survival, and higher overall survival when treated with checkpoint blockade in comparison to MHC Class II negative tumours (Johnson et al., 2016). Although a difference in overall survival wasn't observed here between the MHC Class II-positive MSS-hi-CIRC patients and the MHC Class II-negative MSS-lo-CIRC patients, it is important to note that none of these patients have been treated with checkpoint

blockade, or any treatment. Furthermore, over 70% of overall survival data was missing from these analyses.

Lal (2016) previously showed that MHC Class II expression and the CIRC signature were highly correlated and here I improve on this by demonstrating the high accuracy of MHC Class II RNA expression to predict MSS-hi-CIRC patients with regards to their native, MSS grouping. Furthermore, we demonstrate a highly variable MHC Class II protein expression across our MSS cohort. Despite analyses demonstrating weak positive correlations of MHC Class II protein with the CIRC signature here, the CIRC signature is highly correlated with MHC Class II transcript levels (Lal, 2016) and stronger relationships between MHC Class II expression and its protein have been demonstrated elsewhere (Sillo, 2021, pre-viva). Furthermore, there are many reports demonstrating that MHC Class II expression is regulated at the level of transcription (Handunnetthi et al., 2010). Therefore, despite these correlations not being reflected in this dataset, wider findings reinforce transcriptomic findings in the use of an IHC stain for MHC Class II to stratify this patient cohort. These analyses provided the basis for the ANICCA Class II clinical trial, conducted in Birmingham over the past few years. This trial aimed to test the efficacy of Nivolumab in high class II MHC cancers. An MHC Class II threshold of 50% or more was suggested to capture MSS-hi-CIRC patients. This was subsequently refined to being a tumour-membrane positivity threshold due to studies demonstrating cancer cell MHC Class II expression as a biomarker for success of PD-1/PD-L1 blockade in melanoma (Johnson et al., 2016). Unfortunately, there was difficulty in recruiting patients with tumours that met this threshold of cancer cell Class II MHC positivity. At the time of writing, no patients were found in the advanced CRC setting above this cancer cell MHC Class II threshold.

It is key that the analyses presented here drastically differ from the setting in which ANICCA attempted to recruit. Here, I use primary tumours whereas ANICCA recruited patients who had

metastatic disease which had relapsed post-surgery. The lack of patients recruited in the ANICCA trial suggests multiple possibilities. Firstly, this may suggest that having a high cancer cell MHC Class II expression leads to much better outcome even after conventional treatment modalities, accounting for the lack of Class II-high tumours found in the metastatic setting. Secondly, the transition from a non-spatially resolved signature (CIRC) to one which is spatially resolved (MHC Class II expression on the epithelium – ANICCA) may explain the discrepancy. It is conceivable that MSS-hi-CIRC patients are those with elevated stromal expression of MHC Class II, rather than epithelial-linked. This latter suggestion would be in keeping with the finding that MSS-hi-CIRC are predominantly of a CMS4 subtype, in addition to the Kaplan Meier plot demonstrating the poorer prognosis of the MSS-hi-CIRC cohort. Lastly, it is possible that the MSS-hi-CIRC cohort represents a mixture of patients expressing MHC Class II on either their epithelial or stromal compartments. This latter suggestion is perhaps, the most likely, with further interrogation of the Vectra dataset (Section 5.2.1) demonstrating equivalent MHC Class II expression in both epithelial and stromal compartments.

Although the MHC Class II threshold for ANICCA has subsequently been reduced to aid trial recruitment, these findings support the investigation into the underlying biology of how MSS-hi-CIRC patients gain their increased MHC Class II expression, and by extension their increased immune infiltrate.

Chapter 4: Investigating the Genetic Basis of MSS-hi-CIRC CRC

4.1 Introduction

Cancer, at its core, is a genetic disease. The transition from a normally functioning cell to one that is devoid of proliferative control is encoded in its genetics via mutations. This has meant that, since the advent of sequencing methods, the genetic basis of cancer has been extensively characterised. As highlighted by Fearon and Vogelstein, as briefly covered in the thesis introduction, it was initially thought that the transition from a normally functioning cell to a carcinoma was linear. It wasn't until a landmark study completed in 2012, in which a single tumour was sequenced multiple times, that the idea of cancer heterogeneity was accepted, and that cancer develops analogous to Darwin's evolutionary theory (Gerlinger et al., 2012). Cancer initially develops with driver mutations in genes that safeguard or regulate key processes, such as proliferation, apoptosis, and immune visibility. These initiating mutations provide the basis for more, secondary mutations to occur which increase a cancer cell's likelihood of survival (Hanahan and Weinberg, 2011). Whilst these secondary mutations may act to decrease immune visibility, such as inactivating mutations in the MHC Class I antigen presentation machinery, they may alternatively increase the likelihood of the presentation of antigens with altered sequence. Cancer arises from healthy tissue. Healthy tissue presents self-antigen via MHC Class I, which also serves to inhibit immune attack by natural killer cells, a component of the innate immune system. Since autoreactive T-cells are deleted in the thymus during tolerance, there is generally a low potential for immune activation against the otherwise 'healthy' cancer cell. Mutations, however, present a novel opportunity for altered self to be presented. Mutations, having been processed by the intracellular antigen processing machinery, can be presented as novel MHC-associated peptide sequences called neoantigens. These neoantigens are immensely immunogenic, and therefore, the degree of immunogenicity of a tumour is directly correlated to the number of neoantigens and by extension, the degree of mutational burden (Giannakis et al., 2016, Brown et al., 2014, Rooney et al., 2015, Blass and Ott, 2021). For these

reasons, relative to tumours with proficient repair pathways, MSI cancers or those which have defective DNA damage repair pathways have a much higher immune infiltrate, reflective of substantially greater immunogenicity. This phenomenon likely underpins the highly successful clinical intervention of ICB, as covered in the thesis introduction. However, for those cancers which display much lower levels of mutation burden, these therapies have not been so revolutionary. Importantly, however, low-mutation burden cancers harbour core, oncogenic, driver mutations that in principle could act as targets for immune recognition and that are fixed across the tumour (McGranahan et al., 2015). Moreover, studies have demonstrated that these driver mutations may therapeutically be exploitable. The Rosenberg laboratory proposed that RAS mutations common in CRC may be immunotherapeutic tractable. They identified $CD8^{POS}$ T-cells with reactivity against the KRAS G12D neoepitope in an advanced CRC patient who had multiple lung metastases. After administering nonmyeloablative, lymphodepleting chemotherapy, the patient was administered a single dose of these expanded $CD8^{POS}$ T-cells, and five doses of IL-2. 6 weeks after the cell therapy, all seven metastatic lung metastases had significantly reduced in burden (Tran et al., 2016). Interestingly, similar effects have also been observed following administration of $CD4^{POS}$ T-cells recognising a mutation in *ERRB2IP*, suggesting that these effects are not only MHC Class I dependent (Tran et al., 2014). These studies suggest that, even despite limited neoantigens, in principle only one may be required for immunogenicity. I demonstrate a highly variable intratumoural immunity within the low-mutation burden MSS CRC subgroup, which could be indicative of a difference in recurrently mutated loci, giving rise to differential immunogenic neoepitopes between the MSS cohorts.

Genetic characterisation of cancer has largely focussed on the somatic changes that occur during cancer evolution. However, genetic variability that distinguishes us as individuals, also distinguishes our susceptibility/response to cancer. The human genome differs vastly between two individuals, with up 3.5 - 4.3 million areas in the genome having been noted as sites of inter-individual variability

(Agarwal et al., 2017). These areas of inter-individual variability are termed germline single nucleotide polymorphisms (gSNPs), and can be non-synonymous, synonymous, or in non-coding regions. Non-synonymous gSNPs are, perhaps, the easiest to comprehend in terms of their potential to increase susceptibility to cancer development. Knudson's famous two-hit model of cancer development allows mutations to be inherited. These mutations will not affect gene function due to the presence of a fully functional second allele. Only when a second, inactivating mutation occurs in the redundant allele does cancer develop. Synonymous gSNPs, on the other hand, do not produce a change in the protein sequence, however, their action is likely to occur earlier in the translation process. Synonymous alterations to coding regions have been shown to affect translation rates by interfering with ribosome recruitment, or by reducing mRNA half-life (Mendell and Dietz, 2001), thus reducing gene expression. Intronic gSNPs, also referred to as expression quantitative trait loci (eQTLs) can act locally i.e., variants lie within 1 megabase on either side of a gene's transcriptional start site (TSS), or distantly to the TSS. Locally acting eQTLs are referred to as cis-eQTLs, whereas distant eQTLs (whereby the variant is located at least 5 megabases upstream/downstream, or on a different chromosome to the gene it affects expression of) are considered trans-eQTLs (Nica and Dermitzakis, 2013). The action of eQTLs is perhaps the most understudied of gSNPs, due to the requirement of whole genome sequencing data, and the significant expertise that is required to analyse it.

In addition to contributing to the direct development of cancer, gSNPs influence host immunity. gSNPs occurring in various immune-related genes, or regions which influence gene expression of immune related genes, may in principle modulate risk of cancer development due to the altered capacity for immunoediting, but also by modulating the immune response towards cancer. To understand the differing levels of intratumoural immunity between our MSS cohorts, I characterised the genetic landscape of both settings. Here I hypothesise that MSS-hi-CIRC CRC may share processes

with the highly immunogenic MSI-H CRC cancer grouping or display a distinct genetic process that differentiates it from the MSS-lo-CIRC CRC cohort.

4.2 Results

4.2.1 MSS-hi-CIRC and MSS-lo-CIRC cancers display similar levels of mutational burden

Although MSS-hi-CIRC patients are microsatellite stable, I hypothesised that they may have a large enough mutation burden to cause their increased immune infiltrate. Using 33 different tumours, Panda and colleagues (2017) determined various different mutation burden thresholds (iCAM threshold) that could predict response to immune checkpoint therapy, including in CRC. To distinguish whether MSS-hi-CIRC patients had a mutation burden above this threshold, or larger than the MSS-lo-CIRC patients I compared the number of non-synonymous mutations across the cohorts. Mutation burdens between the MSS subgroups didn't alter, with both MSS-hi-CIRC and MSS-lo-CIRC displaying iCAM thresholds significantly below the 2.5 cut-off determined by Panda and colleagues. In fact, only 3.90% of MSS-hi-CIRC and 3.98% of MSS-lo-CIRC exceeded this threshold. In contrast, 96.7% of the ICB-susceptible, MSI-H passed this criterion (Figure 4.1).

Other papers have also suggested that the main reason for an increased immune infiltrate lies in the number of frameshift mutations, rather than overall non-synonymous mutations. For example, Mlecnik and colleagues (2016) demonstrated that frameshift mutations were much more likely to result in epitopes than missense mutations. However, in concordance to overall non-synonymous mutations, no differences between MSS-hi-CIRC and MSS-lo-CIRC existed with regards to frameshift mutations (Figure 4.2) or any other forms of mutation (data not shown).

Figure 4.1 – Non-synonymous mutation burdens of TCGA-COADREAD patient subgroups
Total non-synonymous mutation burdens were compared across TCGA-COADREAD subtypes. A red dashed line was placed at 2.5 to indicate the identified iCAM threshold for CRC (Panda et al., 2017). The number of patients per subtype are indicated on the X axis.
Statistical analysis used a Kruskal-Wallis test, p-value = 1.64e-29. $CI_{95\%}$ [0.25, 0.32]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

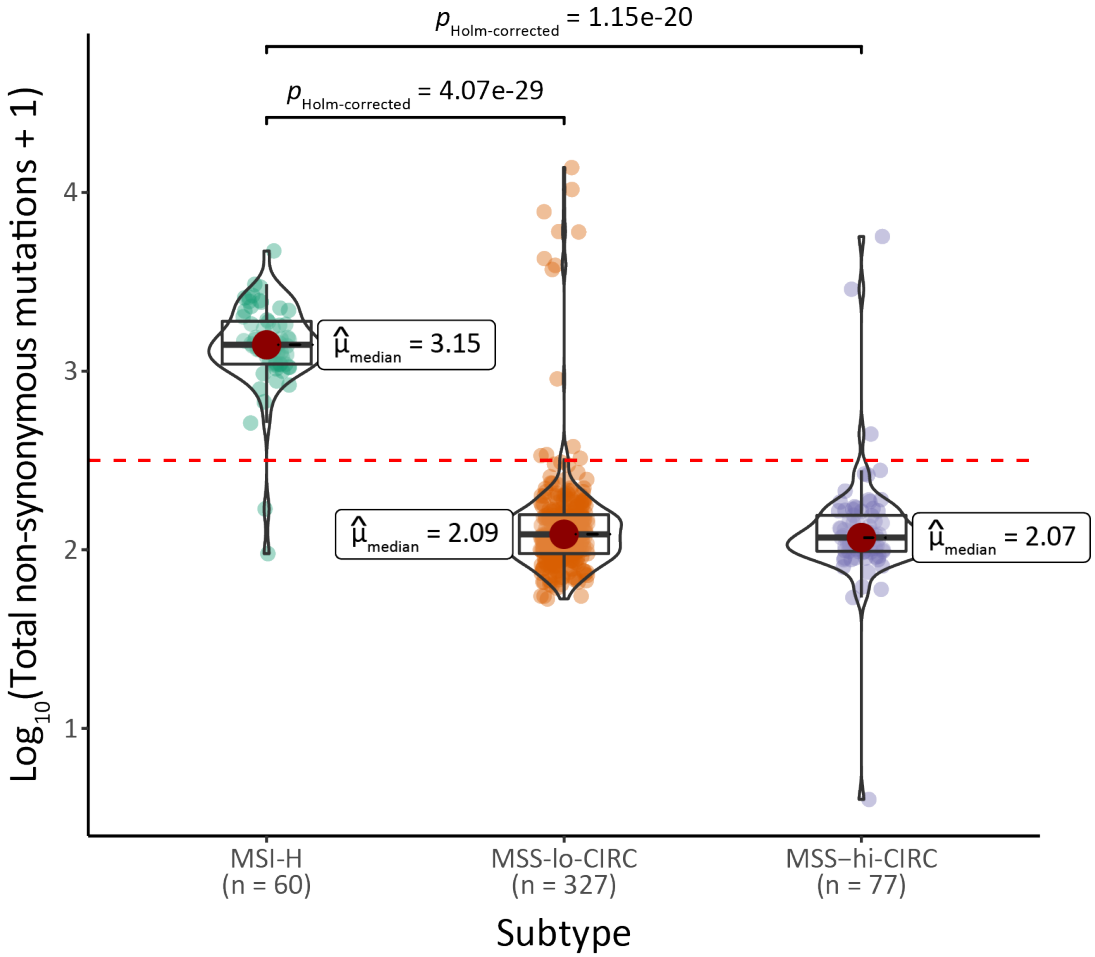
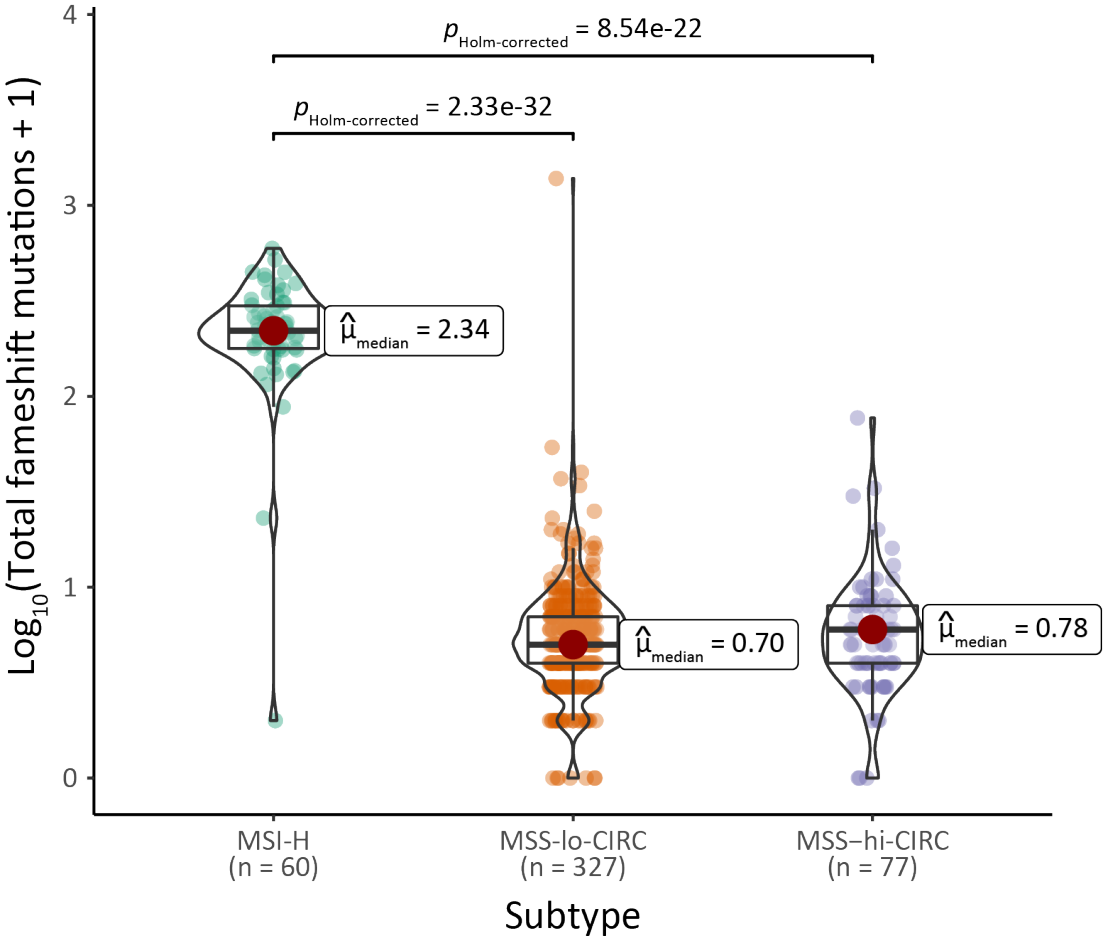


Figure 4.2 – Frameshift mutation burdens of TCGA-COADREAD patient subgroups
Total number of single nucleotide variants causing a frameshift mutation were compared across TCGA-COADREAD subtypes. The number of patients per subtype are indicated on the X axis. Statistical analysis used a Kruskal-Wallis test, p-value = 1.62e-32. *$CI_{95\%}$[*0.28, 0.34*]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.*



4.2.2 MSS-hi-CIRC patients display no specific enrichments of mutation signatures

Due to the lack of quantitative mutation burden difference between MSS-hi-CIRC and MSS-lo-CIRC patients, I next hypothesised that this might arise from a qualitative difference. Using signatures derived from Alexandrov et al (2013), I distinguished five signatures that existed within our data (see Figure 2.2, Appendix 1, and Figure 4.3). Using cosine similarity and published signatures from Alexandrov et al, I determined that our signatures resembled certain signatures with known aetiology. For example, our Signature 3 displayed striking similarity to Alexandrov Signature 6 (Table 4.1). Alexandrov Signature 6 is known to be enriched in cancers with a mismatch repair deficiency. Here, I see that this signature is selectively enriched in our MSI-H patients, demonstrating the power of these analyses (Figure 4.3). Two of the five signatures were enriched in the MSS-hi-CIRC and MSS-lo-CIRC patients, albeit of equal proportions (Appendix 2). Our Signature 1 mapped to Alexandrov Signature 1A, seen to be enriched pan-cancer. The increased C>T transitions found in this signature was concluded to likely be related to spontaneous mutation associated with age although I found no association of MSS-hi-CIRC and MSS-lo-CIRC patients residing in older generations (Table 3.6).The second signature enriched in MSS-hi-CIRC and MSS-lo-CIRC patients was that of SU2. Unfortunately, this signature was unable to be validated by Alexandrov et al, and therefore, no probable association. These analyses highlight, however, that mutation signatures seem to occur in equal proportions across both the MSS-hi-CIRC and MSS-lo-CIRC cohorts, and therefore, are unlikely to share a mutational process responsible for the increased intratumoural immunity seen in MSS-hi-CIRC (Figure 4.3, Appendix 2).

Figure 4.3 – Hierarchical clustering of the five mutation signatures found in the TCGA-COADREAD subtypes

Five mutation signatures were determined in the TCGA-COADREAD cohort using NMF decomposition and the frequency of 96 single nucleotide changes. The frequency of each signature was averaged across patient subtypes and Z score transformed within each subtype. Two-dimensional hierarchical clustering was applied to determine the order of signatures and patient subtypes. Increased red intensity signifies an increased strength of signature in that subtype.

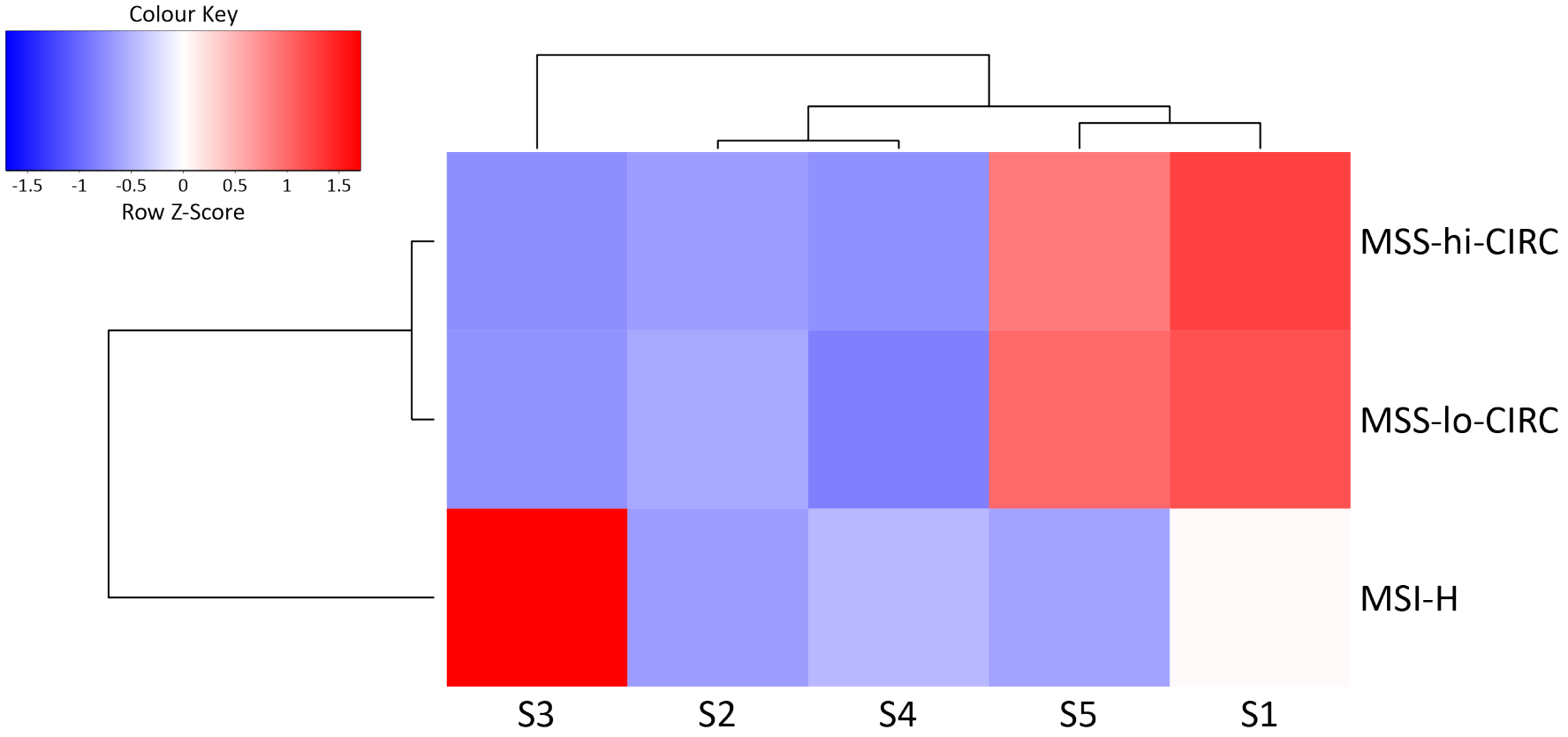


Table 4.1 – Cosine similarity values between TCGA-COADREAD mutation signatures and Alexandrov signatures

Similarity between the mutation signatures found by NMF decomposition in the TCGA-COADREAD cohort (Somatic Signature) and the Alexandrov signatures was evaluated using cosine similarity on the frequency of base changes within each signature. Alexandrov signatures demonstrating the highest cosine similarity value were chosen to represent each Somatic Signature in the TCGA-COADREAD dataset. Cosine similarity values are rounded to 3 significant figures. Alexandrov signatures were derived from the `SomaticSignatures` *R package.*

Somatic Signature	***Alexandrov et al., Signature***	***Cosine Similarity***
S1	S1A	0.948
S2	S10	0.992
S3	S6	0.950
S4	S17	0.765
S5	SU2	0.853

4.2.3 MSS-hi-CIRC and MSS-lo-CIRC share recurrently mutated genes

Due to the lack of a mutation signature arising specifically in the MSS-hi-CIRC cohort, I next suggested that it may only take one, particularly immunogenic mutation to induce an immune response, as was seen in Rosenberg's study with *KRAS* targeting (Tran et al., 2016). Therefore, I investigated the presence of shared mutations between the MSS-hi-CIRC patients which were absent in the MSS-lo-CIRC cohort. Only two genes were recurrently mutated in at least 50% of the MSS-hi-CIRC cohort, *APC* (86%) and *TP53* (73%). Both genes, however, were also recurrently mutated in the MSS-lo-CIRC cohort (84% and 63% for *APC* and *TP53*, respectively). In contrast, the MSI-H cohort had 10 genes which were recurrently mutated in more than 50% of patients (*TTN*; 90%, *MUC16*; 70%, *FAT4*; 68%, *OBSCN*; 63%, *PCLO*; 62%, *BRAF*; 60%, *SYNE1*; 60%, *ZFHX4*; 60%, *KMT2D*; 58%, *RYR2*; 55%, *DNAH5*; 53%, *SETD1B*; 53%, *MUC5B*; 52%, *RYR1*; 52%, *ZFHX3*; 52%), however, none were shared with the MSS-hi-CIRC subgroup. Due to the lack of a specific recurrently mutated gene in MSS-hi-CIRC patients compared to MSS-lo-CIRC, I hypothesised that particular types of mutations occurring on these genes might give rise to an immunogenic peptide specific to MSS-hi-CIRC patients. I plotted the frequency of various mutation types: missense, nonsense, frame shift and splice sites, onto the gene schematics for both *APC* and *TP53*. No real changes were observed between MSS-hi-CIRC and MSS-lo-CIRC with relatively few patients sharing areas where mutations occurred and the changes largely mirroring one another (Figure 4.4 & Figure 4.5). Whilst there were a small cluster of *APC* missense mutations in the MSS-lo-CIRC grouping, occurring in the APC basic and EB1-binding domain, these represented very few patients.

Figure 4.4 – Schematic of the *APC* gene demonstrating the frequency of mutations in TCGA-COADREAD MSS subgroups

The number of patients demonstrating various mutation types were plotted onto a gene schematic of APC for both MSS-hi-CIRC (top) and MSS-lo-CIRC (bottom) in a lollipop plot. The longest transcript ID of APC - NM_001127510 was chosen for visualisation. Lollipops represent various mutation types as depicted by the colours in the key. The Y axis highlights the frequency of patients harbouring the mutations at areas on the gene. Percentages indicate frequency in which APC was mutated in the respective cohorts.

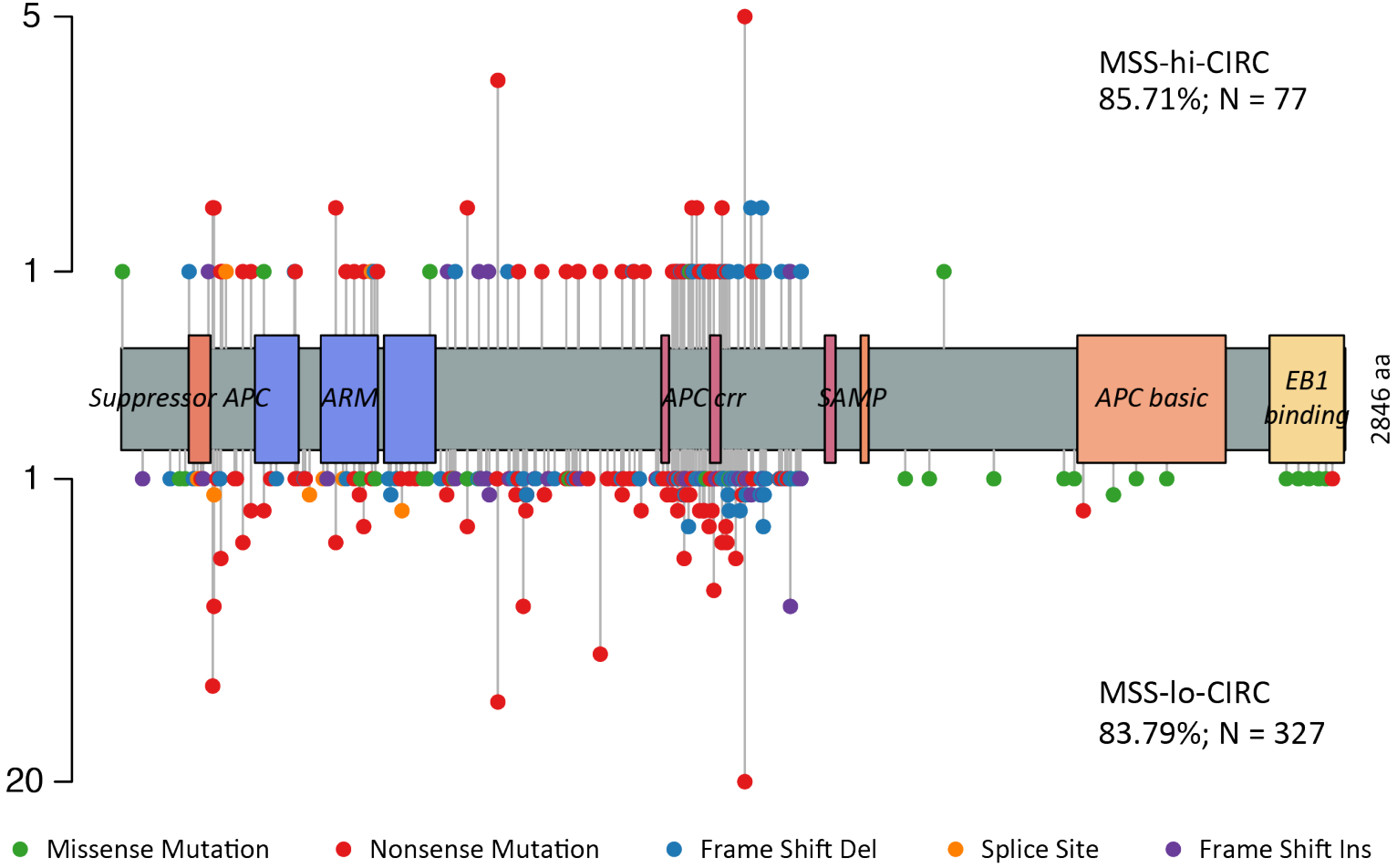
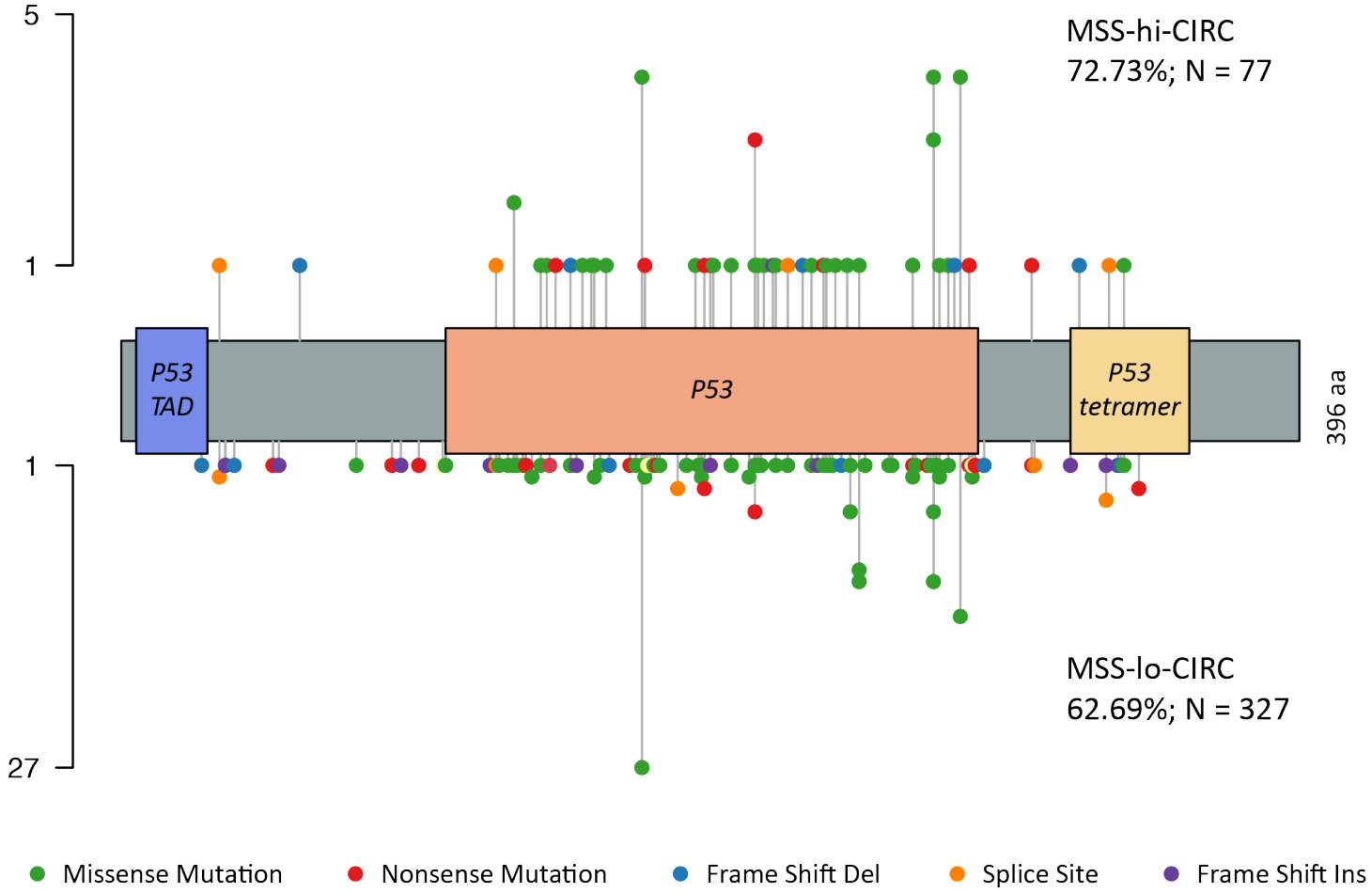


Figure 4.5 – Schematic of the *TP53* gene demonstrating the frequency of mutations in TCGA-COADREAD MSS subgroups

The number of patients demonstrating various mutation types were plotted onto a gene schematic of TP53 for both MSS-hi-CIRC (top) and MSS-lo-CIRC (bottom) in a lollipop plot. The longest transcript ID: NM_000546 was chosen for visualisation. Lollipops represent various mutation types as depicted by the colours in the key. The Y axis highlights the frequency of patients harbouring the mutations at areas on the gene. Percentages indicate frequency in which TP53 was mutated in the respective cohorts.



4.2.4 MSS-hi-CIRC and MSS-lo-CIRC are indiscernible in terms of copy number alterations

It seems that in terms of mutation status, MSS-hi-CIRC and MSS-lo-CIRC were almost indiscernible. However, MSS cancers may not be predicted to have particularly immunogenic mutations when one considers the process in which the majority arise. MSS cancers largely arise through chromosomal instability, and therefore, the differences in immunity may lie in differences in the chromosomal aberrations of each cohort. Whilst on average, the frequency of genomic changes altered between the subsets with MSS-lo-CIRC demonstrating a higher percentage (Figure 4.6), I couldn't attribute these differences to any one chromosome (Figure 4.7).

Figure 4.6 – Degree of chromosomal instability in the genomes of TCGA-COADREAD subgroups
The level of chromosomal instability was measured as the fraction of the genome which had a copy number different from 0. These fractions were compared across the TCGA-COADREAD subtypes. The number of patients belonging to each subtype are indicated on the X axis.
Statistical analysis used a Kruskal-Wallis test, p-value = 3.22e-25*. $CI_{95\%}$[*0.17, 0.26*]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.*

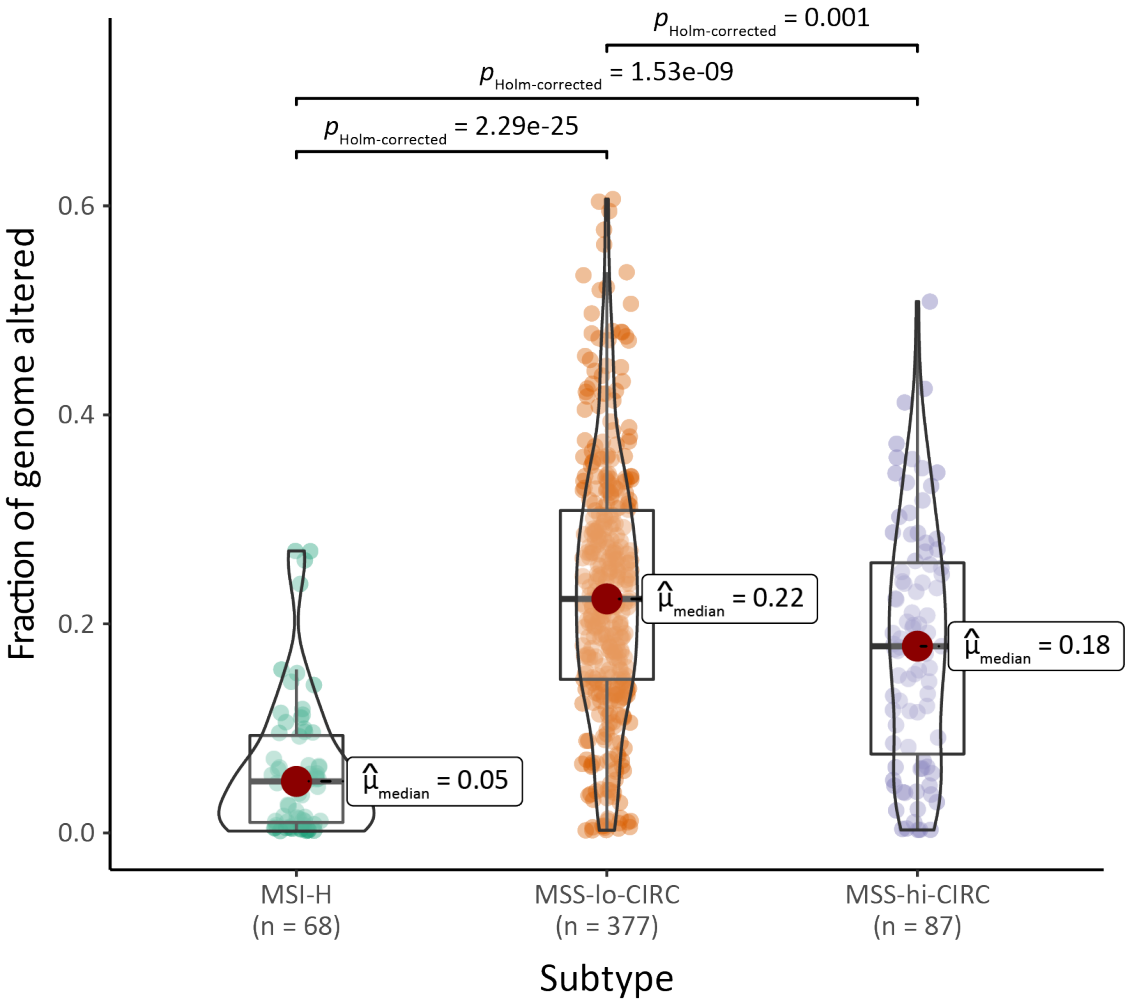
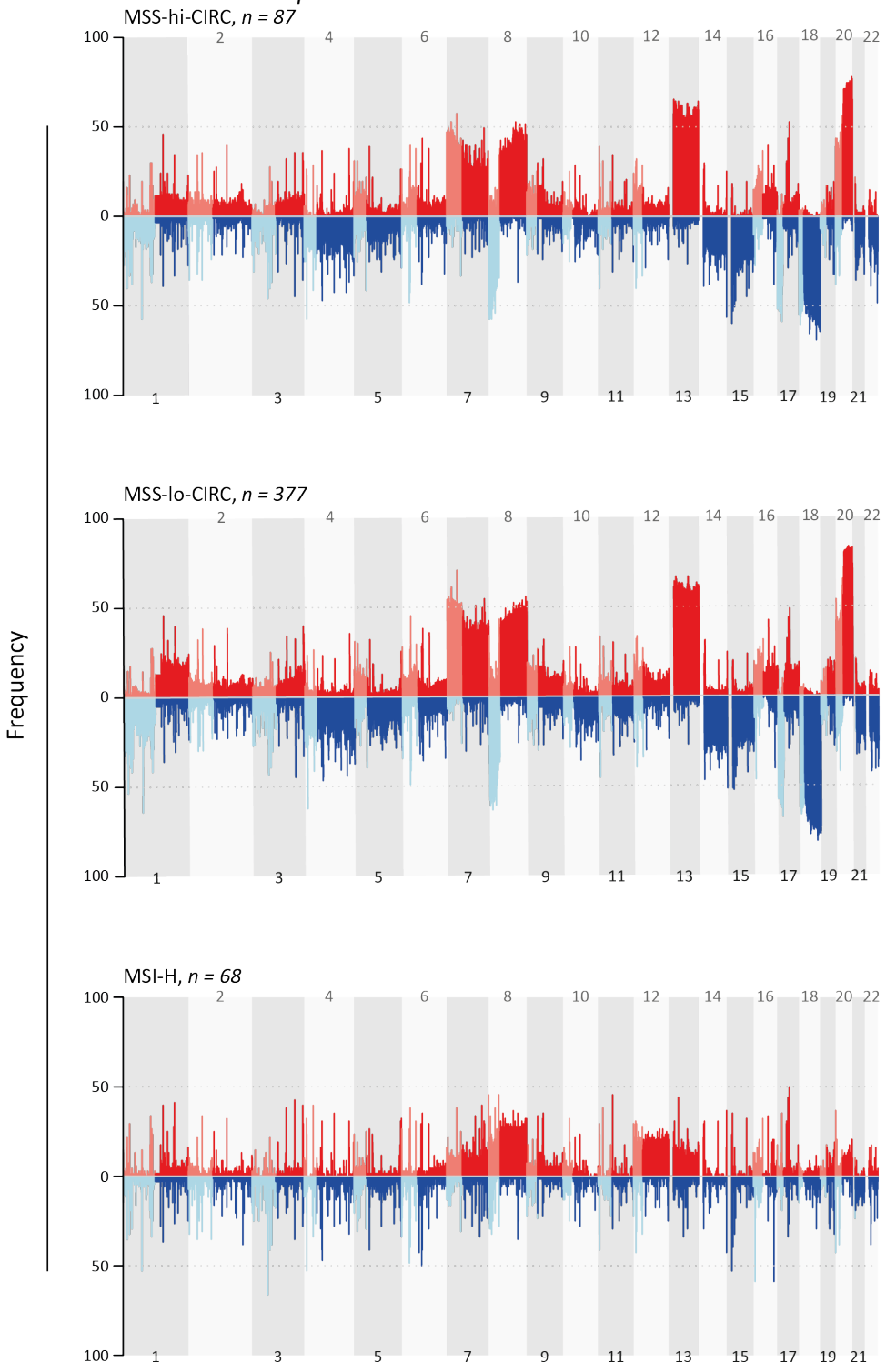


Figure 4.7 – The frequency of chromosomal copy number alterations across TCGA-COADREAD subtypes

Copy number alterations were aggregated as a frequency (Y axis) for individual subtypes and demonstrated as gains (red) and losses (blue) across the chromosomes of the genome (X axis). The number of patients belonging to each subtype is indicated on individual panes (order of panels from top to bottom – MSS-hi-CIRC, MSS-lo-CIRC, MSI-H). Chromosomes are labelled with odd numbers at the base and even chromosomes at the top.



4.2.5 Three mitochondrial SNPs were differentially enriched between the MSS-hi-CIRC and MSS-lo-CIRC cohorts

Until here, I have considered the autosomal changes imposed by the cancer to each of the patient cohorts. However, the importance of germline variants in the development of disease are increasingly important. SNPs which reduce/increase the propensity of detection of antigenic peptides, or those which reduce or increase the threshold for immune activation may be particularly important for the MSS-hi-CIRC and MSS-lo-CIRC cohorts. Here I test the frequency of the ~1,000,000 most common SNPs across these cohorts. Due to the haploid nature of the mitochondrial genome, autosomal and mitochondrial SNPs were subjected to different statistical tests (allele and genotype tests, respectively). Within the SNP arrays, there were 868,263 autosomal SNPs and 107 mitochondrial SNPs for testing. After multiple testing, none of the autosomal SNPs were significant, however, three passed the threshold of 0.01 at an unadjusted P-value (Figure 4.8). However, whilst these represent differences between the cohorts, after investigating the frequency of patients which had these in each of the cohorts, along with that they are no longer significant after FDR correction, I was not convinced they were true findings.

In contrast, three mitochondrial SNPs were significant at an FDR of 0.0164 (Figure 4.9). These SNPs occurred in *ND4*, *ND4L*, and *CO3*. *ND4* and *ND4L* code for subunits of the NADH dehydrogenase enzyme which is a component of complex I of the respiratory chain. *CO3*, on the other hand, codes for a component of the cytochrome c oxidase complex in complex IV of the respiratory chain. The frequency of these SNPs is of particular interest. For complex I SNPs, 100% of MSS-hi-CIRC patients carried the SNP. In contrast, for the CO3 SNP, 100% of MSS-hi-CIRC patients did not carry the SNP. Importantly, only three MSS patients who had one of the SNPs didn't carry all three, suggesting that these SNPs were linked.

Figure 4.8 – Autosomal SNP enrichment analysis between TCGA-COADREAD MSS-hi-CIRC and MSS-lo-CIRC patients

868,263 Autosomal SNPs were tested for an association with MSS-hi-CIRC or MSS-lo-CIRC status using an allele test. Before multiple correction, three SNPs passed the P-value cut-off to be considered significant (shown in orange and labelled for rs number). SNPs with an unadjusted P-value greater than 0.05 were removed from the plot. Chromosomes are shown on the x axis, and the -log$_{10}$(unadjusted P-value) is shown on the y axis. A dotted line is placed at -log$_{10}$(3.16e-06) to indicate the threshold for unadjusted significance [3 significant figures].

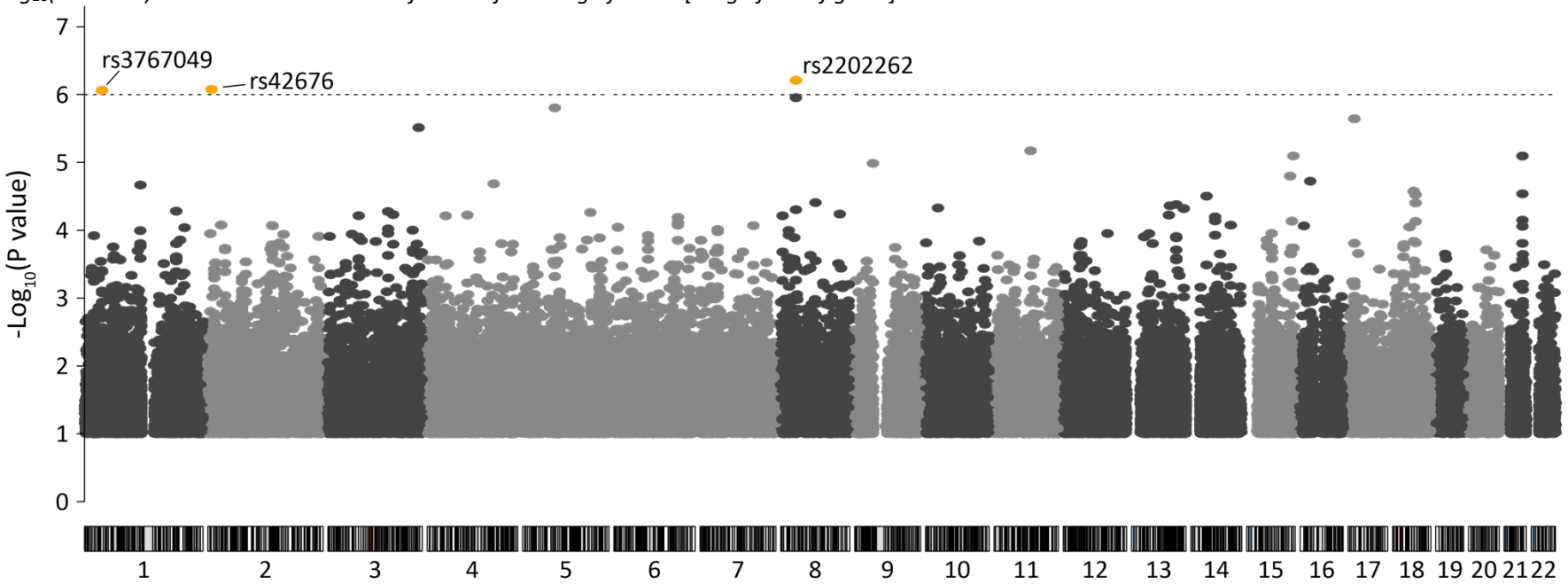


Figure 4.9 – Mitochondrial SNP enrichment analysis between TCGA-COADREAD MSS-hi-CIRC and MSS-lo-CIRC patients

107 mitochondrial SNPs were tested for an association with MSS-hi-CIRC or MSS-lo-CIRC status using a genotype test. After correction for multiple testing, three SNPs were considered significant at -log$_{10}$(0.01) (shown in orange and labelled). Genes in the mitochondrial genome are shown on the x axis, and the -log$_{10}$(unadjusted P-value) is shown on the y axis. A dotted line is placed at -log$_{10}$(0.01) to indicate the threshold for significance.

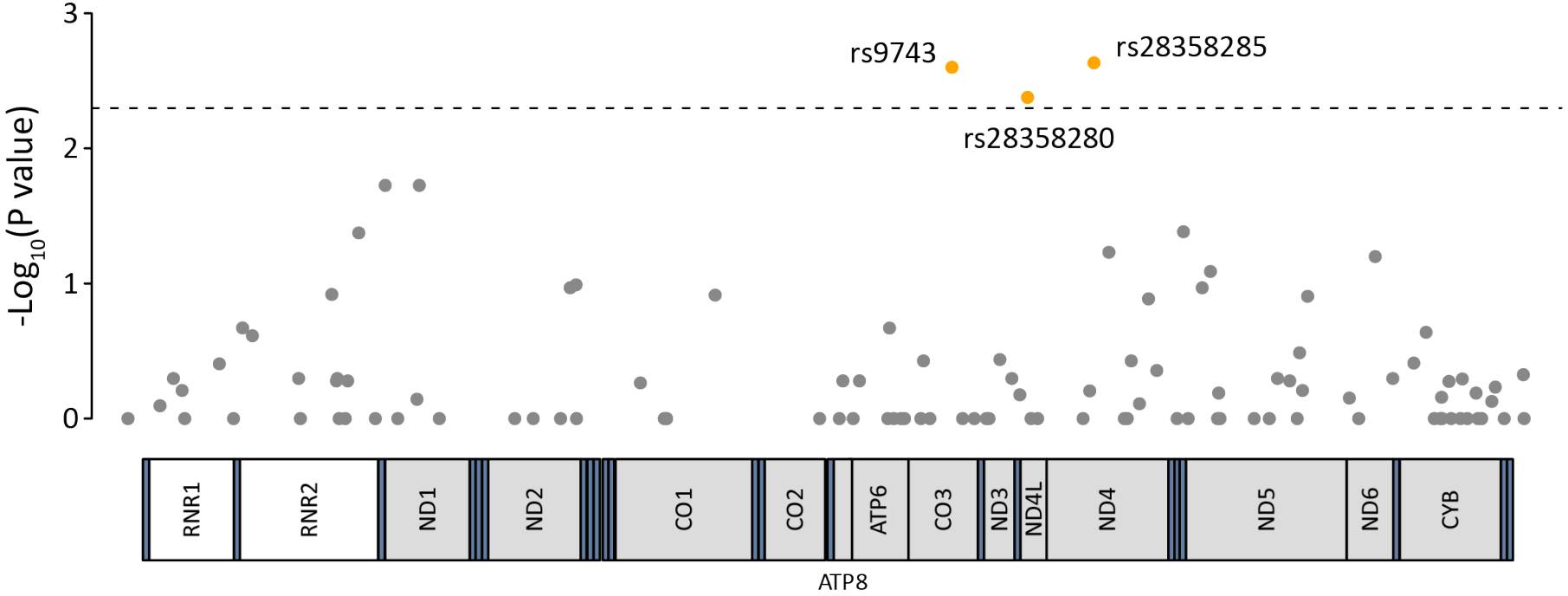


Table 4.2 – Percentage of patients demonstrating homo- or heterozygosity at significant SNPs

Autosomal and mitochondrial SNPs were tested for enrichment between MSS-hi-CIRC and MSS-lo-CIRC patients in the TCGA-COADREAD cohort. For autosomal SNPs, the Phred score represents the allele test, whereas for mitochondrial SNPs the Phred score represents the results of the genotype test. SNPs passing thresholds in Figure 4.8 & Figure 4.9 are shown. No autosomal SNPs were significant after FDR correction. Information about associated genes, the chromosome and frequency in the population was gained from the R package; `rsnps` (Gustavsen et al., 2020)*. NA where no data was available. SNPs are ordered by Phred score. Total number of patients; MSS-lo-CIRC, n = 304; MSS-hi-CIRC, n = 80.*

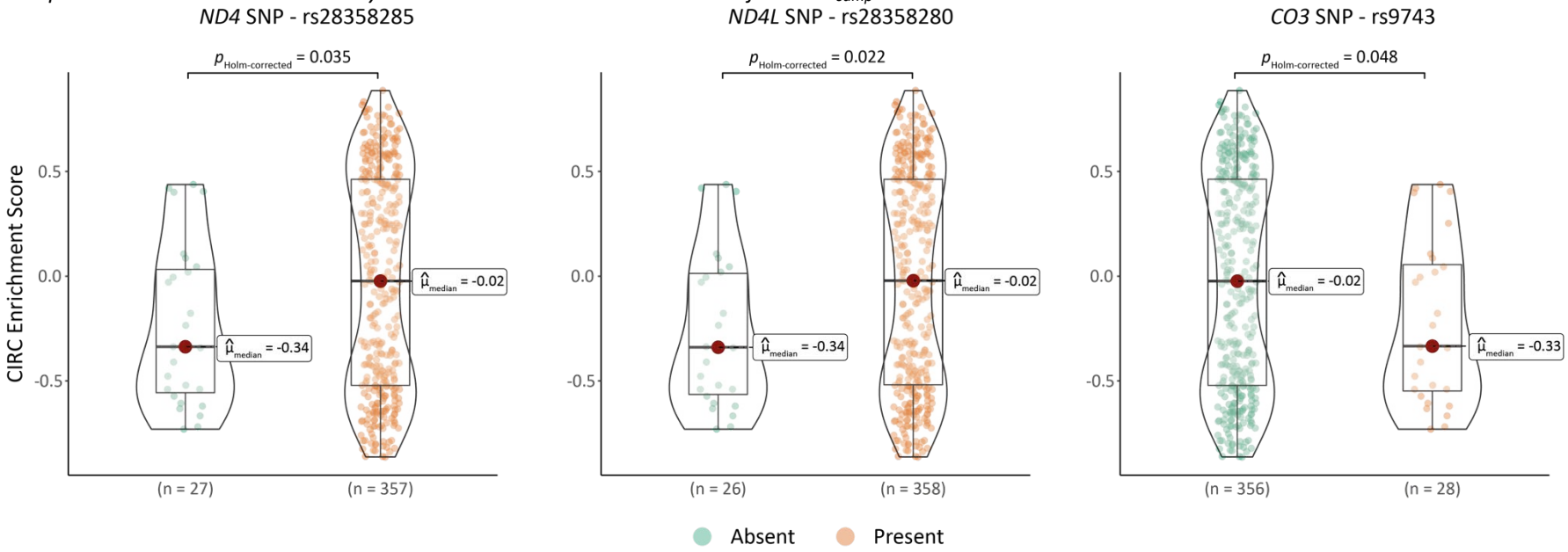
SNPID	Population Frequency (%)	Associated Gene	Chr	Nucleotide change	Phred Score	MSS-lo-CIRC			MSS-hi-CIRC		
						-/-	-/+	+/+	-/-	-/+	+/+
rs2202262	23.6	*NRG1*	8	A > C	62.1	198	95	11	29	40	11
rs42676	87.5	*LINC00298*	2	G > C G > T	60.8	8	54	242	0	1	79
rs3767049	81.9	*GRIK3*	1	C > A C > G C > T	60.6	246	45	13	79	1	0
rs28358285	NA	*ND4*	MT	T > C	26.3	27	NA	277	0	NA	80
rs9743	NA	*CO3*	MT	T > C	26.0	276	NA	28	80	NA	0
rs28358280	NA	*ND4L*	MT	A > G	23.8	26	NA	278	0	NA	80

Whilst the association tests of these SNPs were significant, alongside the fact that 100% of MSS-hi-CIRC patients either carried or didn't carry the respective SNP, a significant proportion of MSS-lo-CIRC patients followed suit. In fact, only ~10% of MSS-lo-CIRC patients didn't follow the same direction as MSS-hi-CIRC patients (Table 4.2). To investigate whether the likelihood of having the presence of these SNPs affected the CIRC score, I compared CIRC scores across presence/absence of SNP regardless of MSS-hi-CIRC status (Figure 4.10). In all three scenarios, the presence of the SNP significantly increased CIRC score. However, this comparison was heavily one-sided with ~92% of patients either having the SNP present (ND4/ND4L) or absent (CO3) in comparisons. Furthermore, despite being significant, the distribution of patients in all three comparisons did not suggest that the presence/absence of these SNPs had any link with CIRC score.

Figure 4.10 – CIRC enrichment scores across mitochondrial SNP homozygosity status in TCGA-COADREAD MSS patients

MSS patients were split by the presence (homozygous SNP) or absence (homozygous reference) of the three mitochondrial SNPs significantly enriched in either MSS-hi-CIRC or MSS-lo-CIRC patients from Table 4.2. The number of patients per group for each of the SNPs are shown on the X axis. Order of panels from left to right; ND4 (rs28458285), ND4L (rs28358280), CO3 (rs9743).

All comparisons used the Mann-Whitney statistical test. P-values are Holm adjusted. N_{samp} = 384 in all scenarios.



4.3 Discussion

In this chapter, I investigated the genetic basis of the MSS-hi-CIRC CRC subgroup in comparison to MSS-lo-CIRC and MSI-H groupings. An interesting finding was the lack of a shared level of mutation burden between MSI-H and MSS-hi-CIRC, especially considering their comparable immune infiltrate. Furthermore, MSS-hi-CIRC patients demonstrated drastically lower mutation burden thresholds than have been suggested by Panda and colleagues to be immunogenically relevant (Panda et al., 2017). Parallel to these findings, MSS-hi-CIRC tumours did not share dysregulated mutational processes with MSI-H tumours, nor did they appear to have an individual genetic process that could explain the differences from their native (MSS) grouping in terms of immune infiltrate. Thirdly, the MSS subgroups lacked distinguishing recurrently expressed mutations with only *APC* and *TP53* being mutated in around 70% of both MSS-hi-CIRC and MSS-lo-CIRC cohorts. Finally, albeit demonstrating a reduced fraction of genome altered in comparison to MSS-lo-CIRC, no individual chromosome demonstrated significant differences in this variable.

Together, these analyses failed to detect any somatic genetic differences between MSS-hi-CIRC and MSS-lo-CIRC as cohorts. However, these results are inevitably still inconclusive. I did not investigate neoantigen burden, and although others have and showed no discernible differences in overall neoantigen burden between the cohorts (Lal, 2016), these analyses investigated a quantitative neoantigen load, rather than a qualitative difference in immunogenicity of individual neoantigens. It is also conceivable that mutations may not be recurrent across the MSS-hi-CIRC cohort, but individualised mutations that are immunogenic in various settings. Such settings may include the expression of differing MHC haplotypes. Recent evidence has shown that the unique combination of MHC alleles that an individual carries will define the available sub-peptidome that can be effectively presented (Marty Pyke et al., 2018, Marty et al., 2017). Here, I didn't consider the MHC haplotype of the patient cohorts, and whilst it is unlikely that MSS-hi-CIRC patients harbour a common MHC

haplotype (Lal, 2016), they may contain MHC haplotypes that complement their individualised mutational spectrums. Similarly, spatial intratumoural heterogeneity (sITH) was also not considered. Referring to the differences within tumours with regards to genetic subclones rather than differences between primary and secondary locations, there are two leading hypotheses on how sITH influences intratumoural immunity. The first argues that a reduced sITH produces a more coordinated immune attack against recurrent mutations 'fixed' across the tumour. In contrast, the second suggests increased sITH increases the likelihood of immunogenic neoepitope presentation through presentation of multiple different immunogenic mutations. Evidence is far stronger for the latter, whereby a higher sITH is linked to favourable outcome in a variety of different cancer types (Andor et al., 2016). However, whilst this may be favourable in the primary disease setting as investigated by Andor and colleagues using TCGA data, upon relapse these tumours will undoubtedly produce a highly immunoedited tumour which demonstrates significantly worse overall and progression free survival – analogous to relapsed MSI-H patients (Kim et al., 2016b). Despite similar mutation burdens, MSS-hi-CIRC tumours could display dramatically increased sITH thereby increasing the presentation of various mutations across the entirety of the tumour, accounting for its increased immune infiltrate. There now exists a multitude of computational programs which can deconvolute the tumour meta-genomes gathered from a single exome sequencing experiment into the abundance of various clones, and thus a surrogate reading for sITH, the use of which should be prioritised in the immediate future of this project (Andor et al., 2014, Roth et al., 2014).

In addition to not considering key genetic factors influencing intratumoural immunity, the approach I employed was reductionist. Here, I considered each element individually that could influence immunogenicity of a tumour, whereas the combination of such factors is likely much more telling. The combination of these factors is, perhaps, easiest to visualise in terms of copy number alterations and mutations. The overlap of copy number alterations occurring in regions of mutations would

increase, or decrease, the probability of mutations to be processed and presented by increasing, or decreasing, their expression. The combination of mutation data, copy number alterations, RNA sequencing, and reverse protein phase arrays would provide a much clearer picture of whether somatic alterations are likely to lead to generation of presentable neoepitopes. Finally, my mutation signature analyses utilised only single base changes. Alexandrov and colleagues have since extended their analyses to incorporate 49 single-base-substitution signatures, 11 doublet-base-substitution signatures, 4 clustered-base-substitution signatures, and 17 small insertion-and-deletion signatures (Alexandrov et al., 2020), and therefore the presence of an individualised mutation signature within this patient cohort remains a possibility.

In terms of germline alterations, no autosomal SNPs were found to be significantly associated with MSS-hi-CIRC or MSS-lo-CIRC subgroups, although these analyses likely lack the statistical power to discern associations in this cohort. To gain the statistical power to find associations, the number of patients included would have to be increased. Alternatively, a more directed approach could be taken as adopted in a paper by the Kirchhoff group (Vogelsang et al., 2016). Utilising the Gene Ontology (GO) and Kyoto Encyclopaedia of Genes and Genomes (KEGG) databases, Vogelsang and colleagues collated a list of 382 genes associated with adaptive immune signalling, differentiation, and cytokine production. eQTLs from the MuTHER project shown to affect these 382 genes were ranked by significance, and the top 50 were used for future investigation. These analyses demonstrated significant associations of variants controlling expression of *IL19* and *BATF3* with better overall survival of melanoma patients. Furthermore, previous students in our extended research group recently completed similar analyses using the Genomics England cohort. Here, they demonstrated a significant relationship of high Immunoscore patients with a variant in *TCF7*. Unfortunately, these patients were not first delineated by MMR status, and therefore, the contribution of this eQTL to the MSS-hi-CIRC setting could not be delineated (Sillo, 2021, pre-viva).

Although the TCGA database has whole genome sequencing data available, unfortunately this amounts to just 28 samples and therefore is unlikely to have sufficient power to validate this association (The Cancer Genome Atlas, 2021).

In contrast to autosomal SNPs, our analyses demonstrated significant association of two mitochondrial SNPs with MSS-hi-CIRC patients, and one with MSS-lo-CIRC patients. These SNPs lay in the genes *ND4*, and *ND4L* (MSS-hi-CIRC), and *CO3* (MSS-lo-CIRC). Whilst the SNP in *ND4* has never been associated with any other disease states, those in *ND4L* and *CO3* have recently been demonstrated as associated with an increased BMI (National Library of Medicine, 2021, Flaquer et al., 2014). This association is interesting, considering that obesity is a leading risk factor for cancer. Furthermore, obesity has recently been linked with leaky barrier function, and colonic inflammation (Ahmad et al., 2017). Using leptin-deficient mice (ob/ob mice), which over-eat due to lack of the hunger satiety signal leptin, the authors detected increased gut permeability as measured by the fluorescence intensity of orally fed FITC-dextran in the blood after 4 hours of ingestion. These findings, independent of diet, were in parallel with reduced expression of the tight junction proteins, Occludin, and ZO-1, as well as the mucin proteins Muc2 and Muc6 in ob/ob mice (Nagpal et al., 2018). Conceivably, a similarly reduced barrier function, resulting from these SNPs driving increased BMI, could account for the increased immune infiltrate observed in MSS-hi-CIRC these patients. A dysfunctional gut barrier could permit translocation of intestinal microbes, causing an elevated immune response directed to these microbes, rather than the cancer. Furthermore, the elevated expression of MHC Class II protein in this patient cohort and its centrality to the CIRC signature may also support the theory of microbially-derived antigenic signals as, classically, MHC Class II is associated with the presentation of exogenous antigen.

Despite this interesting speculation, the distribution of these mitochondrial SNPs across the subgroups leaves considerable uncertainty as to whether they represent true associations. Whilst these SNPs were wholly present, or absent, within the MSS-hi-CIRC cohort, ~90% of MSS-lo-CIRC patients followed the same pattern. Hence, their presence or absence was merely marginally enriched in the MSS-hi-CIRC subgroup relative to MSS-lo-CIRC.

Although the analyses described here are not exhaustive, they strongly imply that the reason for discrepant intratumoural immunity between MSS-hi-CIRC and MSS-lo-CIRC is not a genetic determinant and suggests other drivers may be critical. Moreover, although tentative, the potential link highlighted by mitochondrial SNPs between obesity, a reduced barrier function, and an increase in intratumoural immunity is particularly intriguing, and warrants further investigation into a microenvironmental or microbial source of antigen.

Chapter 5: Investigating Microbial Adjuvantisation of Intratumoural Immunity in MSS-hi-CIRC CRC

5.1 Introduction

The lack of a definitive genetic reason for the increased immune infiltration in MSS-hi-CIRC patients led me to explore other potential drivers of increased immune infiltration. The colon is home to an estimated 100 trillion bacteria, the majority of which belong to the *Bacteroidetes* and *Firmicutes* phyla. These commensals play vital roles in digestion of various polysaccharides to SCFAs, of which constitutes ~70% of the energy used by colonocytes. The crucial role of the gut microbiome is reflected by the numerous mechanisms designed to segregate bacteria to the colonic lumen. The gut maintains a fine balance between tolerance and inflammation. At homeostasis, the gut is naturally enriched for immunosuppressive cell types, such as T_{REGS}, which are replaced by inflammatory cell types, such as T_H17 cells, upon barrier disruption and microbial translocation. Colonic barrier dysregulation leads to a co-ordinated inflammatory response to prevent infection, however, is tightly regulated by a multitude of negative feedback mechanisms to limit tissue damage. Therefore, perhaps most applicable in the colon, the gut microbiome has become a 'hot topic' in cancer biology. Here, it has strong links to the development of multiple other cancers (oesophageal (Baba et al., 2017), liver (Yu and Schwabe, 2017), and pancreatic cancer (Wei et al., 2019)), as well as influencing response to cancer therapies, including ICB (discussed in further detail later) (Routy et al., 2018). Lastly, the successful clinical translation of Faecal Microbiota Transplantation (FMT) in colitis provides evidence that manipulating the gut microbiome in humans could be a feasible route to influencing gut immunity for therapeutic benefit.

The gut microbiome has been shown to play a vital role in the pathogenesis of CRC. Murine studies on an ApcMin background have shown that colonisation of the mouse colon with enterotoxigenic *Bacteroides fragilis* (ETBF), a strain of *B. fragilis* which secretes a metalloprotease toxin termed BFT, dramatically reduced the time frame in which tumours developed (Wu et al., 2009). In contrast, the

non-producing BFT strain caused no increase in tumour formation, or the increased mortality as observed in ETBF-colonised ApcMin mice. Furthermore, those colonised by ETBF showed marked increases in the infiltration of T_H17-polarised T-cells of which blockade of IL-17, or IL-23R, abrogated the associated colonic mucosal proliferation, and neoplasia. Furthermore, Rakoff-Nahoum and Medzhitov demonstrate that the increased mortality from ApcMin mice can be independent of colonisation with a toxin-producing bacterium. ApcMin mice crossed onto a $MyD88^{NEG/NEG}$ background were shown to have reduced mortality, and reduced tumour formation as seen in ApcMin mice (Rakoff-Nahoum and Medzhitov, 2007). These results demonstrate the necessity for TLR signalling in malignant transformation of APC-mutant mice, and thus both studies provide a role for microbial stimulation in cancer pathogenesis.

The link between colitis and cancer is clear. A 2001 meta-analysis of 116 studies demonstrated that the risk of developing CRC after suffering with colitis for 10 years was 2%; this quadrupled by 20 years to 8% and finally, after 30 years of disease the likelihood of developing CRC was 18% (Eaden, 2001). Although the causality of the colitis that drives an inflammation-mediated CRC is the topic of many reviews, other inflammation-driven cancers have clear aetiology. Infection with *Helicobacter pylori* causes chronic gastric inflammation and ulcers, and significantly increases the likelihood of developing gastric cancer. Around half of the world's population is infected with *H. pylori*, which remains a lifelong infection – suggesting the immune system is ineffective at clearing this bacterium (Wroblewski et al., 2010). This unresolved inflammation creates chronic tissue damage, initiating wound healing responses and with them the release of growth and anti-apoptotic factors, and proangiogenic cytokines – propagating a microenvironment suited to malignant transformation. This pro-tumorigenic microenvironment undoubtedly contains T-lymphocytes, which otherwise would indicate a favourable prognosis. However, in a situation in which chronic inflammation perpetuates and drives malignant transformation, it is unclear as to whether activation of the immune system

through ICB therapies would help or hinder anti-tumour efforts (Amedei et al., 2012). The upregulation of PD-L1 upon *H. pylori* infection is a well-documented phenomenon in both *in vitro* and *in vivo* settings (Figueroa-Protti et al., 2019); similarly PD-L1 upregulation has been linked to a favourable outcome after ICB, therefore one would hypothesise efficacy of ICB in *H. pylori*-infected cancer. However, to my knowledge, the stratification of gastric cancer in *H. pylori*-positive and *H. pylori*-negative cancers and their sensitivity to ICB is yet to formally tested in a clinical trial.

In contrast to playing a role in the development of CRC, the enrichment of certain microbial species has been linked to increased intratumoural infiltration, and better prognoses. Riquelme and colleagues demonstrated that long-term pancreatic ductal adenocarcinoma survivors had an enrichment of *Proteobacteria* and *Actinobacteria*. Furthermore, the enrichment of these bacteria, and increased bacterial diversity, were associated with an increased abundance of $CD3^{POS}$ and $CD8^{POS}$ intratumoural T-cells (Riquelme et al., 2019). This phenomenon has been demonstrated elsewhere, in which specific bacterial communities (*Ruminococcaceae* and *Lachnospiraceae*) were shown to recruit T-cells by inducing the production of various chemokines by cancer cells. *In* vitro application of TLR agonists upregulated the expression of CCL20, CXCL9, and CXCL10 as well as inducing the *de novo* expression of CCL3, CCL4, CCL5, and CCL22 by cancer cells. Furthermore, the transfer of CRC cell lines either intraperitoneally or intercaecally drastically altered chemokine expression profiles. Those injected directly into the colon had up to 70-fold higher expression of chemokines, which were lost with antibiotic treatment. These results demonstrated the importance of the microbiota in attracting T-cells (Cremonesi et al., 2018), and helped to establish the presence of a 'favourable microbiome composition' linked with increased immune infiltration and could be used as rationale for using FMT in the cancer setting. This therapy, currently being successfully utilised in the U.K. to treat colitis, aims to replace 'unfavourable microbiomes' with those designed to ameliorate pathology. In contrast to the colitic setting in which these therapies aim to decrease gut inflammation, in the cancer setting

one would aim to transplant microbiota which increased intratumoural immunity. Murine studies have supported such a use, whereby FMT approaches have demonstrated success in conferring susceptibility to ICB in germ-free mice (Routy et al., 2018, Matson et al., 2018, Gopalakrishnan et al., 2018), as well as reducing adverse events associated with ICB (Vétizou et al., 2015). Moreover, the Zitvogel laboratory also demonstrated that FMT from responding patients, or oral supplementation of *Akkermansia muciniphila* after FMT from non-responding patients, into germ-free mice increased infiltration of $CCR9^{POS}CXCR3^{POS}CD4^{POS}$ and cytotoxic T-cells, directly contributing to the observed responses to PD-1 blockade (Routy et al., 2018). Although the authors suggested these effects were IL-12 dependent based on this recruitment, subsequent comments on the narrative by Routy and colleagues, have placed emphasis on *A. muciniphila* as a mucus-degrading bacterium, and its production of SCFAs acting on GPR41 and GPR43 – two orphan G-protein coupled receptors suggested to limit tumour growth through cell cycle arrest and caspase activation, leading to apoptosis and a second mechanism that could restore responsiveness to ICB (Tang et al., 2011, Santoni et al., 2018).

Whilst the pro-tumour or anti-tumour effects that individual microbial species may elicit is unclear, it is evident that the microbiome plays a vital part in both the development and the outcome of various cancers. Here, I investigated the contribution of the gut microbiome to the discrepancies in tumour immune infiltrate evident between MSS-hi-CIRC and MSS-lo-CIRC patients. As investigations in Chapter 4 did not highlight any genetic reason for the discrepancy in tumour immune infiltrate observed between MSS-hi-CIRC and MSS-lo-CIRC patients, I combined multispectral immunofluorescence, RNA sequencing, RNAscope analyses, immunohistochemistry, and metagenomic sequencing data analyses to provide substantial support for a role for the microbiome in acting as an underpinning driver of high immunity in the MSS-hi-CIRC setting.

5.2 Results

5.2.1 T_H17 cell abundance differs between MSS-hi-CIRC, MSS-lo-CIRC, and MSI-H

To determine differences in immune cell abundance, I gathered the median density of each of the subtypes in both the stromal and epithelial compartment using data collected on the Vectra 3.0. Comparing these densities across subtype, I demonstrated that whilst MSS-hi-CIRC and MSI-H were largely comparable across all the subtypes tested they differed in that the majority of the MSS-hi-CIRC immunity was confined to the epithelial layer (Figure 5.1). This tendency to fall in stroma or epithelium was perhaps most striking between T_H17 cells where they almost exclusively fell in the epithelium in MSS-hi-CIRC and in the stroma in MSI-H. Although a significant difference did not exist between the density of epithelial T_H17 cells between the subgroups, there was a general trend of these cells being increased in MSS-hi-CIRC (Figure 5.2; first column, second row). Furthermore, this difference was reinforced when comparing the enrichment of a T_H17 specific geneset in TCGA-COADREAD data (Figure 5.3). In contrast to other areas of the body, T_H17 cells play a vital role at barrier sites. Here, they control the spread of microbes, which barrier sites are coated in. Therefore, to investigate the abundances of microbial species I utilised WGS data from Genomics England. After mapping reads to the human genome, I took the remaining data and mapped these to genome references for bacteria and viruses, allowing comparisons of microbial species across MSS-hi-CIRC and MSS-lo-CIRC settings.

Figure 5.1 – Two-dimensional hierarchical clustering of the cellular density of phenotypes using multispectral immunofluorescence

The median of the average density per mm^2 of cells phenotyped as CD4 only, CD8 only, T_H1 (CD4 + T-bet), T_H17 (CD4 + RORγt), T_{REGS} (CD4 + FOXP3), T_C1 (CD8 + T-bet), and Class II in each compartment (epithelium, stroma) were Z score transformed across subtype and plotted as a heatmap. Two-dimensional hierarchical clustering was applied to determine order of cell phenotypes and patient subtypes. Increased yellow intensity reflects an increased presence whereas increased purple intensity reflects a lower presence.

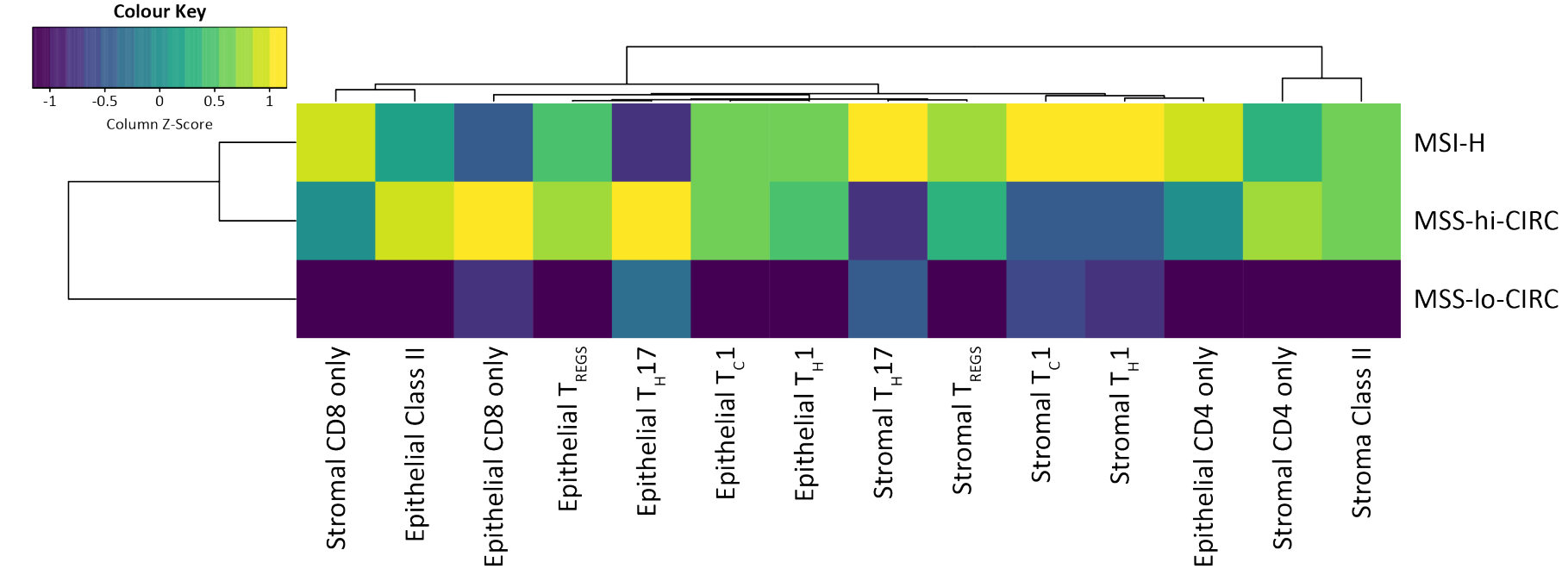


Figure 5.2 – Cellular densities of T-cell subsets across subgroups using multispectral immunofluorescence

The average density of cells phenotyped as 'CD4 only', 'CD8 only', 'T_C1', 'T_H1', 'T_H17', and 'T_{REGS}' in each compartment (epithelium, stroma) were compared across subtypes (MSI-H, n = 10; MSS-lo-CIRC, n = 8; MSS-hi-CIRC, n = 12) using multispectral immunofluorescence. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

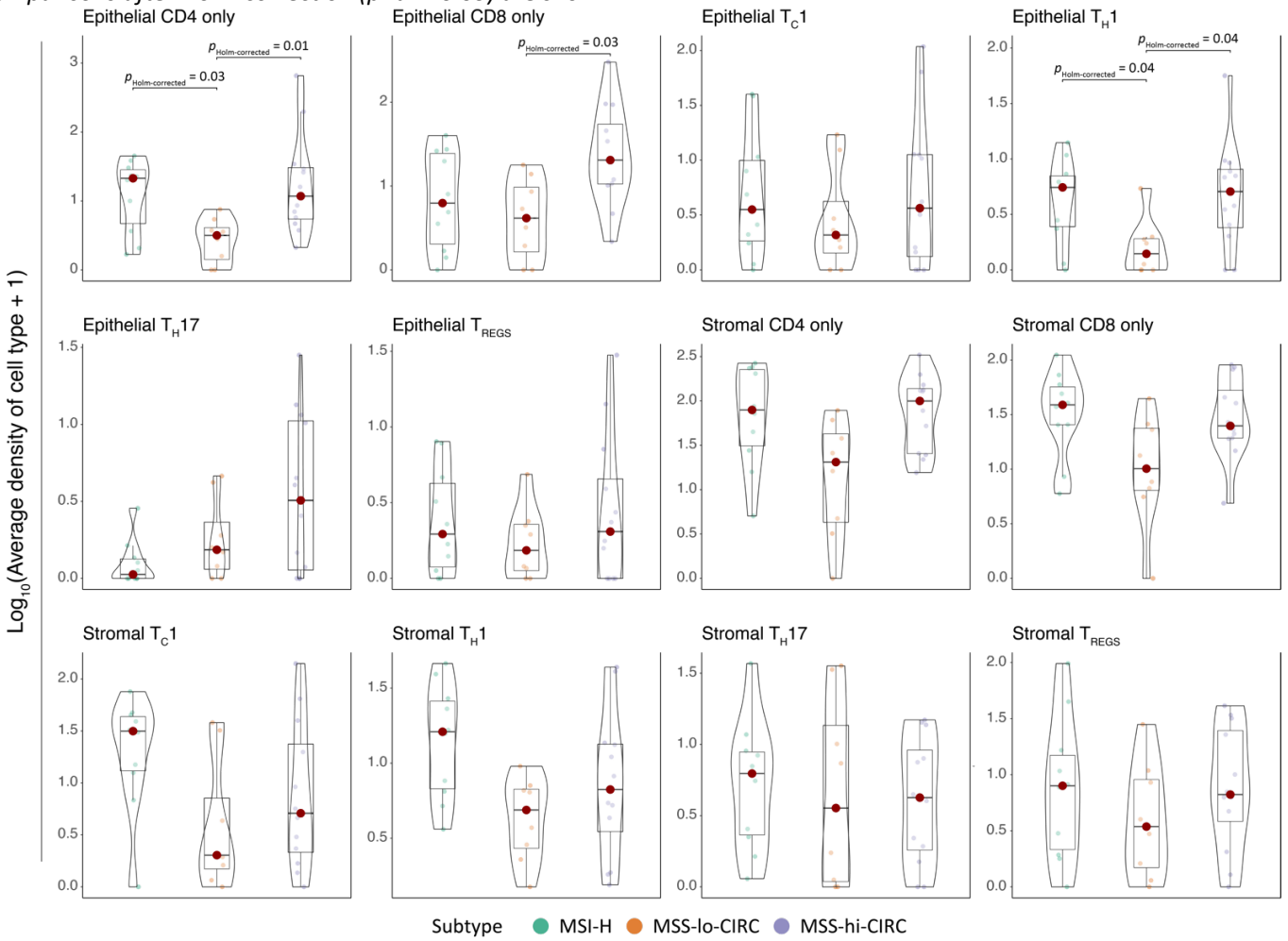
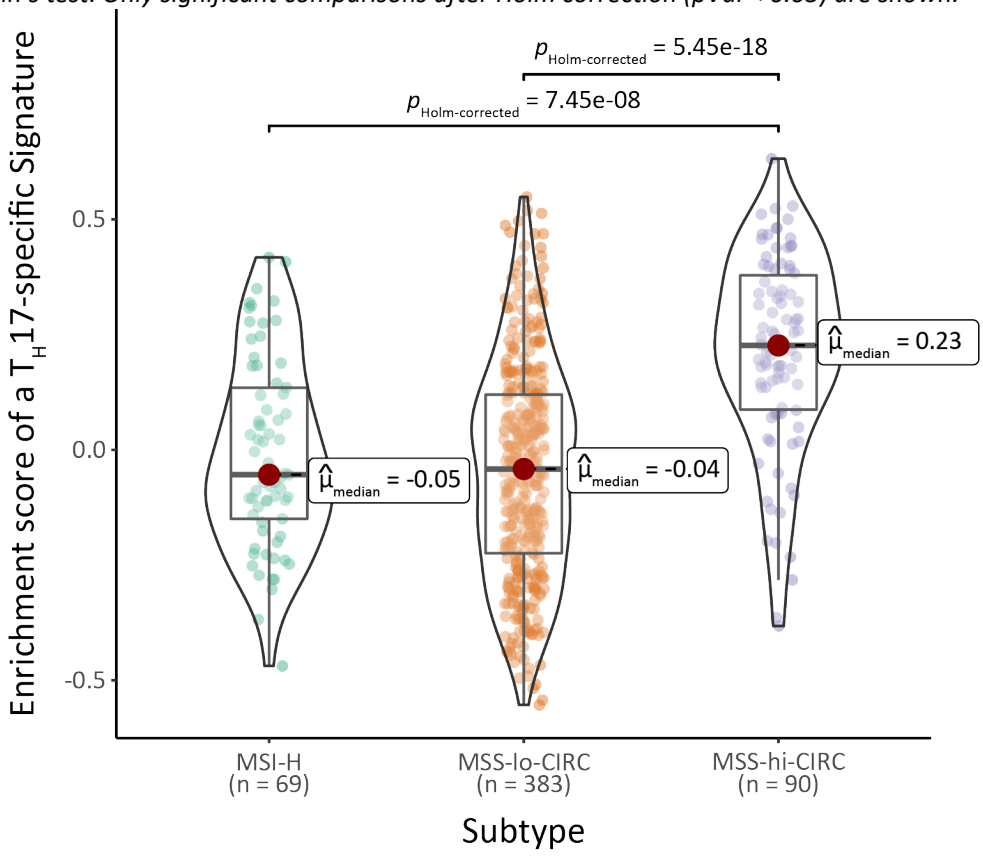


Figure 5.3 – A comparison of T_H17 geneset enrichment scores across TCGA-COADREAD subgroups

Enrichment scores were calculated of a T_H17 cell signature devoid of T_H1-related genes taken from (Castro et al., 2017) and compared across TCGA-COADREAD subgroups. The number of patients per subgroup are indicated on the X axis.

Statistical analysis used a Kruskal-Wallis test, p-value = 1.88e-17. $CI_{95\%}$[0.1, 0.2]. Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



5.2.1 MSI-H cancers have a slight increase in overall bacterial abundance

To investigate whether an abundance of bacteria, or any other microbial species, were enriched in the MSS-hi-CIRC subgroup in comparison to MSS-lo-CIRC and MSI-H, I mapped the unmapped reads from the whole genome sequencing data from the Genomics England cohort to genomes originating from Bacteria, Viruses, and Archaea. When considering the percentage mapped reads to each of the Superkingdoms – Bacteria and Viruses, there were no significant differences between MSS-hi-CIRC and MSS-lo-CIRC cohorts, however, MSI-H displayed a slight enrichment of bacterial reads and a lower fraction of reads mapping to viruses (Figure 5.4). However, after performing RNAscope staining for 16s rRNA, designed to stain all bacterial species present, I found no differences in bacterial abundance, although positive staining was minimal across the cohort (Figure 5.5, Appendix 4).

Using the Genomics England data, I also investigated diversity of the microbes, as it has been shown that diversity of bacteria can significantly impact the degree of inflammation in the gut, with increased diversity causing increased inflammation (Riquelme et al., 2019) However, there were no differences in diversity in either bacteria or viruses (Figure 5.6). To further probe whether there was a difference in response to bacteria between the datasets, I performed differential gene expression on terms associated with microbes.

Figure 5.4 – A comparison of the percentage abundance of microbes across Genomics England subgroups

Whole genome sequencing data from the 100K genomes project was mapped to the human genome. Unmapped reads were subsequently mapped to bacterial and viral genomes. The proportion of unmapped sequences mapping to bacterial (left - Kruskal-Wallis test, p-value = 0.003. $CI_{95\%}$[0.05, 1]) or viral genomes (right - Kruskal-Wallis test, p-value = 0.003. $CI_{95\%}$[0.05, 1].) were compared across subtypes. The number of patients per subtype are shown in the figure key.

Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

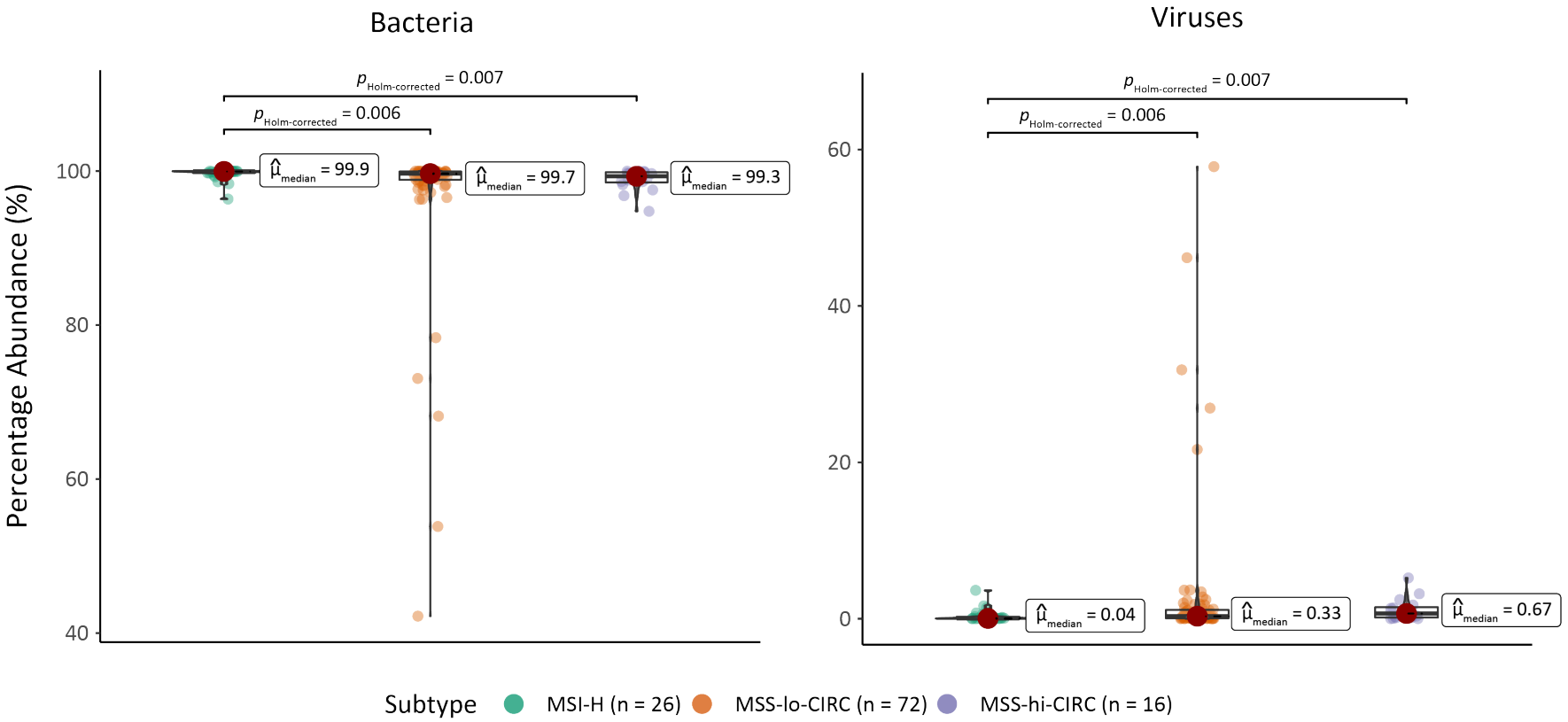


Figure 5.5 – A comparison of bacterial presence between CRC subgroups

FFPE slides were stained for 16s rRNA using an RNAscope probe as a surrogate quantification of bacterial presence. The intensity of 16s rRNA staining was quantified by digital pathology and intensity values were averaged per pixel. Plots are separated by location of staining (left – whole slide, right – epithelium only). The number of patients per subgroup are demonstrated in the figure key.

Statistical analysis used Kruskal-Wallis tests (left; p-value = 0.455. $CI_{95\%}$[0.01, 1], right; p-value = 0.981. $CI_{95\%}$[3.05e-03, 1]). Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.

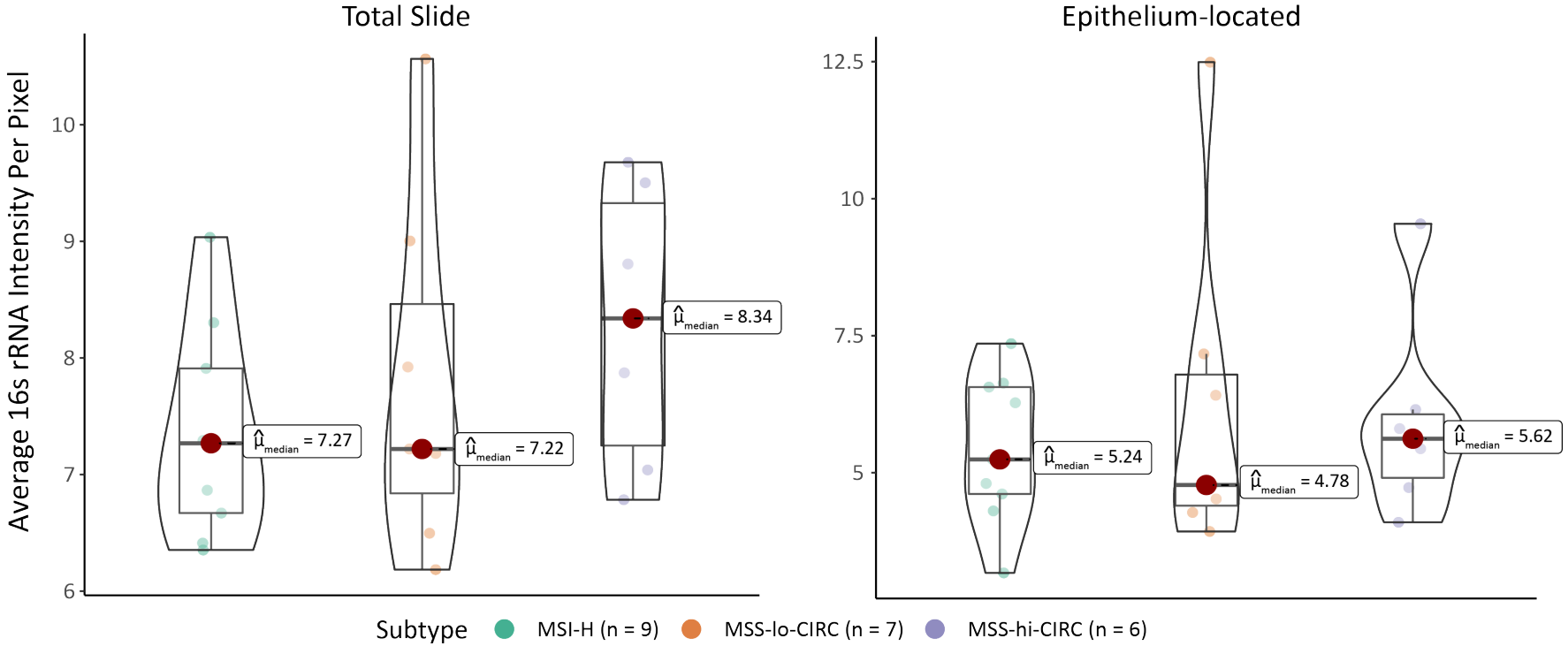
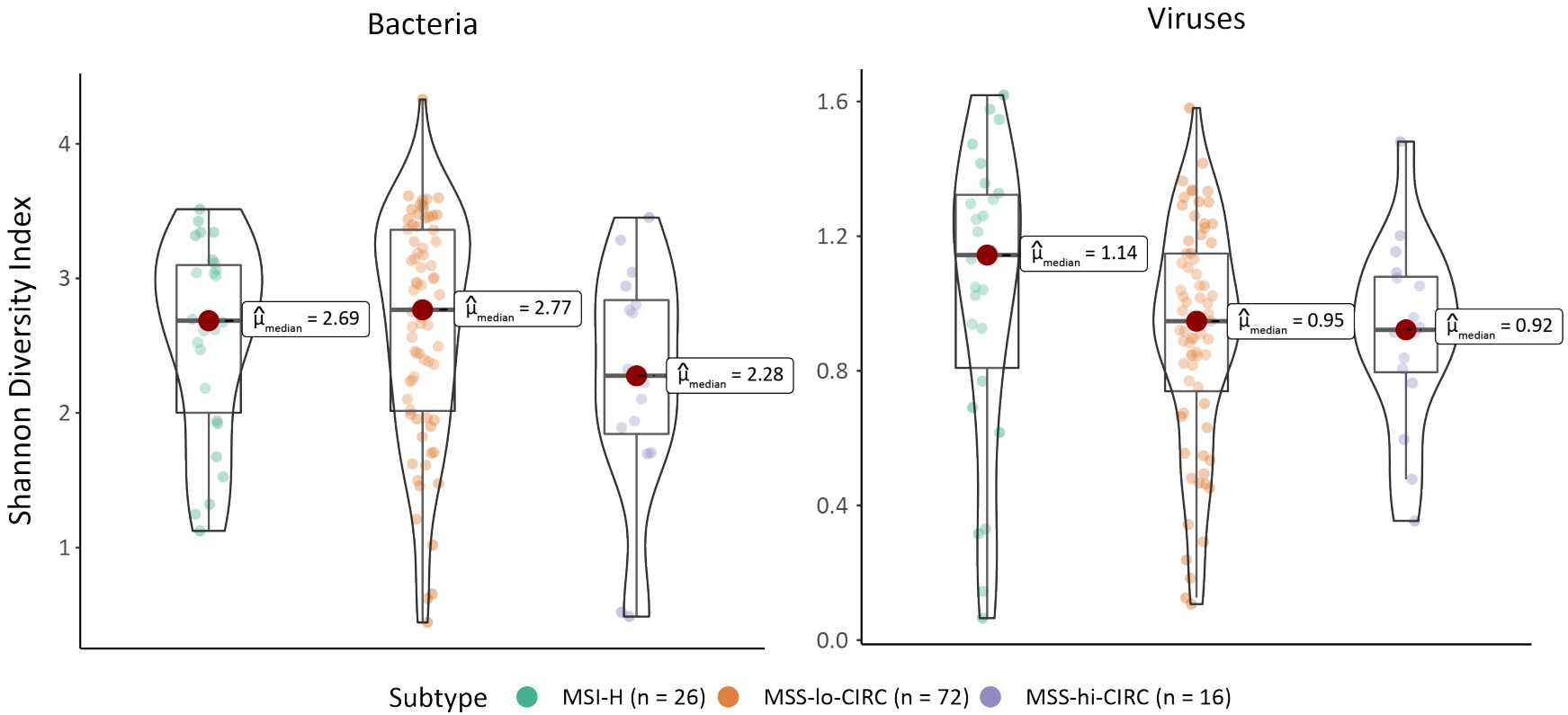


Figure 5.6 – A comparison of bacterial and viral diversity indexes across the subgroups in the Genomics England Dataset

The Shannon diversity index was calculated individually for species of bacteria (left) and viruses (right) and compared across subtypes. The number of patients per subgroup is highlighted in the figure key. Statistical analyses used Kruskal-Wallis tests, (Bacterial diversity; p-value = 0.211. $CI_{95\%}$[2.77e-03, 1], Viral diversity; Kruskal-Wallis test, p-value = 0.123. $CI_{95\%}$[5.58e-03, 1]).

Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



5.2.2 Pathway analysis indicates increased responses to Gram-negative bacteria in the MSS-hi-CIRC cohort

To distinguish whether subgroups were activating pathways associated with microbes, I performed differential pathway expression using search terms associated with microbes: “bacteria”, “bacterium”, “virus”, “viral”, “microbe”, “microbial”, and “archaea”. After removal of genesets consisting of less than 5 genes, I had 95 available genesets for differential pathway analysis between MSS-lo-CIRC and MSS-hi-CIRC (Figure 5.7, Appendix 3). 34 of 38 significant genesets were enriched in the MSS-hi-CIRC direction, indicating an increased response to microbes in these patients. MSI-H also shared this increased enrichment when compared to MSS-lo-CIRC. Although no genesets demonstrated FDR-corrected significance between MSI-H and MSS-hi-CIRC in these patients, 14 genesets could be considered significant if not FDR corrected (Table 5.1).

When considering an unadjusted P-value, only “GO:0016045: detection of bacterium”, and “GO:0050829: defence response to Gram-negative bacteria”, were enriched in the MSS-hi-CIRC setting with respect to MSI-H. More interestingly, whilst “GO: 0016045: detection of bacteria” was enriched in MSI-H compared to the low-immune MSS-lo-CIRC patients, GO:0050829: defence response to Gram-negative bacteria was highly insignificant between MSI-H and MSS-lo-CIRC with an unadjusted P-value of 0.449. The specificity of this geneset in the MSS-hi-CIRC setting is highly suggestive that these patients have an enrichment of Gram-negative bacteria.

Figure 5.7 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC CRC patients using GO database search terms related to microbes

Genes differentially expressed between TCGA-COADREAD MSS-hi-CIRC and MSS-lo-CIRC patients were used as input. 34 pathways relating to "regulation of viral entry/activity", "response to bacteria", "response to virus", and "detection of bacterial product" were found upregulated in MSS-hi-CIRC patients. In contrast, 4 pathways relating to "viral gene expression" were found upregulated in MSS-lo-CIRC patients. Cytoscape was used to produce the enrichment map whereby the size of the node indicates the size of the geneset, edge thickness represents Jaccard similarity scores between genesets, and node colour is based on FDR value (intensity is proportional to significance). Nodes are surrounded by red if they were enriched in the MSS-hi-CIRC setting and blue if they were enriched in the MSS-lo-CIRC setting. Only genesets with an FDR below 0.05 are shown. Nodes were collated into larger groups depending on similarity. Full descriptions are shown in Appendix 3.

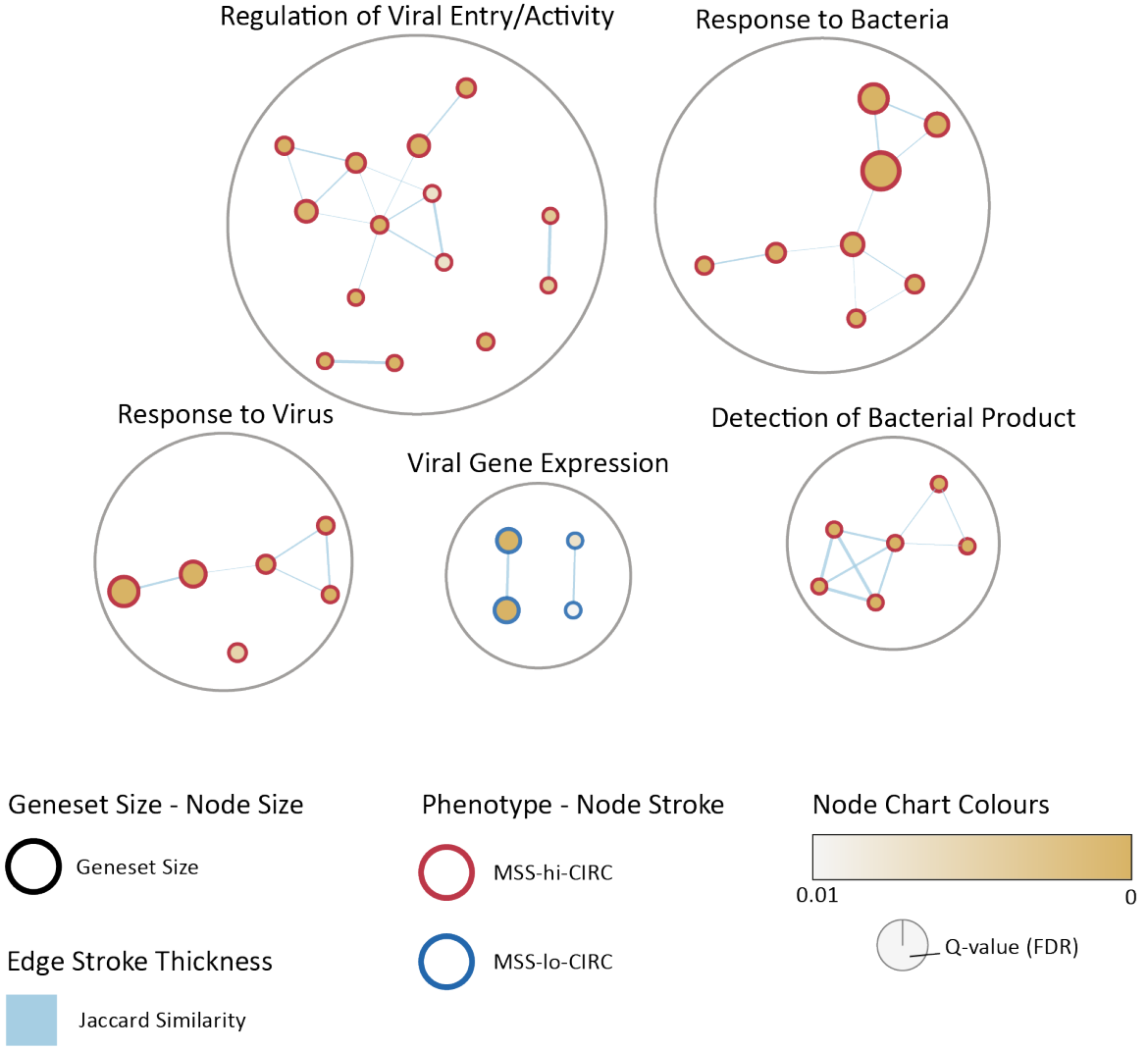


Table 5.1 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H CRC patients using GO terms relating to microbes

Genes differentially expressed between MSS-hi-CIRC and MSI-H patients in the TCGA-COADREAD cohort and GO terms relating to microbes were used as input. 14 pathways were found differentially enriched between MSS-hi-CIRC and MSI-H. Pathways are ordered according to unadjusted P-value.
Genesets are displayed if unadjusted P-value < 0.05. Unadjusted P-values were rounded to 3 significant figures.
NGenes = Number of Genes in the Geneset.

Geneset Name	***NGenes***	***Direction***	***Unadjusted P-value***
GO:0019080: viral gene expression	197	MSI-H	0.00325
GO:0019083: viral transcription	179	MSI-H	0.00498
GO:0039532: negative regulation of viral-induced cytoplasmic pattern recognition receptor signalling pathway	11	MSI-H	0.00543
GO:0039528: cytoplasmic pattern recognition receptor signalling pathway in response to virus	30	MSI-H	0.00644
GO:0039531: regulation of viral-induced cytoplasmic pattern recognition receptor signalling pathway	21	MSI-H	0.00666
GO:0016045: detection of bacterium	12	MSS-hi-CIRC	0.0107
GO:0075733: intracellular transport of virus	56	MSI-H	0.0117
GO:0050829: defence response to Gram-negative bacterium	42	MSS-hi-CIRC	0.0168
GO:0046794: transport of virus	63	MSI-H	0.0193
GO:1902187: negative regulation of viral release from host cell	14	MSI-H	0.0217
GO:0050687: negative regulation of defence response to virus	23	MSI-H	0.0341
GO:0046782: regulation of viral transcription	41	MSI-H	0.0389
GO:0044829: positive regulation by host of viral genome replication	7	MSI-H	0.0447
GO:0050434: positive regulation of viral transcription	25	MSI-H	0.0458

5.2.3 Response to reactive oxygen species is different between MSS-hi-CIRC and MSS-lo-CIRC

To distinguish whether Gram-negative bacteria were enriched in the MSS-hi-CIRC patients, I looked at the genes which were different between "GO:0050829: defence response to Gram-negative bacterium" and GO0050830: defence response to Gram-positive bacterium" to gain a geneset that was Gram-negative specific. Of the 40 genes which differed between the genesets 9 were differentially expressed in the MSS-hi-CIRC cohort with respect to the MSS-lo-CIRC setting (adjusted P-value < 0.01). Of these genes, there were notable members associated with a T_H17 response – *CD4*, *IL6R*, and *CXCL6*. Subsequently, I interrogated the REACTOME database to distinguish other pathways that the remaining 6 genes were members of. A common theme was the production of reactive oxygen species (ROS) and reactive nitrogen species (RNS) either directly (*NOS2* and *SLC11A1*) or indirectly through the metabolism of selenoamino acids (*SELP*, *RPS19*) or interaction with the respiratory transport chain (*OPTN*). Due to the association with the production of ROS and RNS, I interrogated the GO database for terms containing "reactive oxygen" and "reactive nitrogen". 16 genesets were available for interrogation after removal of genesets with less than 5 genes present. Of these genesets, 7 were significantly enriched in the MSS-hi-CIRC setting (FDR < 0.01). The top 4 genesets are shown in Figure 5.8.

I wanted to test if an increased ROS presence could reflect changes in tissue structure, and therefore, I stained for 4-HNE, a product derived from ROS-mediated lipid peroxidation. MHC Class II was positively correlated with 4-HNE staining across the entire slide within the MSS cohort (Figure 5.9). Furthermore, this significant association extended to the epithelium layer.

Figure 5.8 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-COADREAD cohort using pathways relating to the production of ROS (GO:0072593, GO:2000379, GO:1903428) or RNS (GO:2001057)

Plots show the top 4 most differentially enriched pathways from an available 15 pathways of GO terms relating to ROS or RNS. Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-COADREAD cohort were used as input. Barcode plots rank all genes in each geneset in order of enrichment in either MSS-lo-CIRC (left) or MSS-hi-CIRC (right). The blue and red vertical bars in the central section of the plots represent genes from the given geneset. The red and blue lines indicate running enrichment scores for the geneset as the analysis progresses. A higher peak on the right-hand side indicates that the geneset is enriched in the MSS-hi-CIRC setting.

All pathways demonstrate significant enrichment in the MSS-hi-CIRC setting (GO:0072593, FDR = 0.000490; GO:2001057, FDR = 0.000490; GO:2000379, FDR = 0.000113; GO:1903428, FDR = 0.000394).

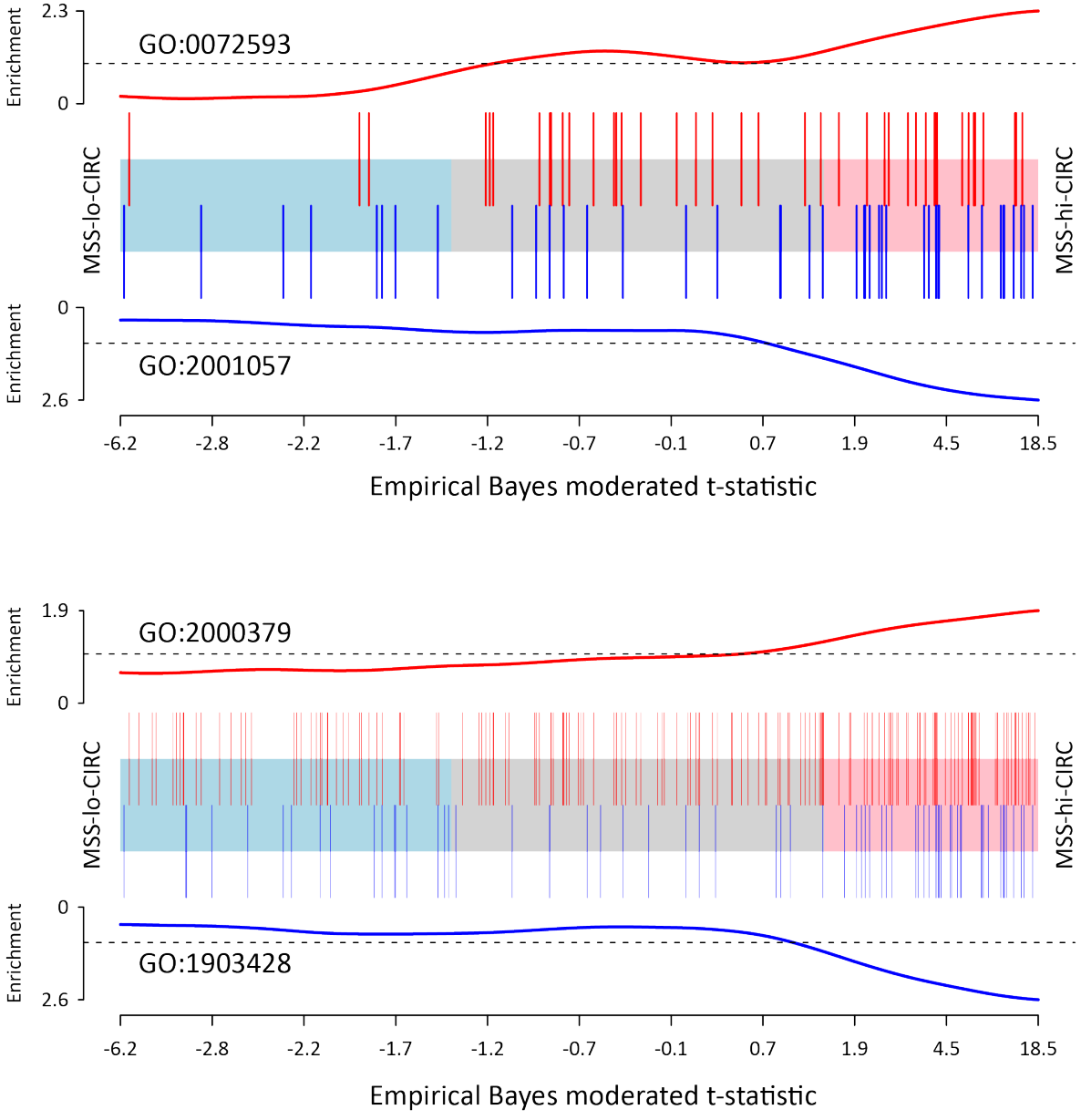
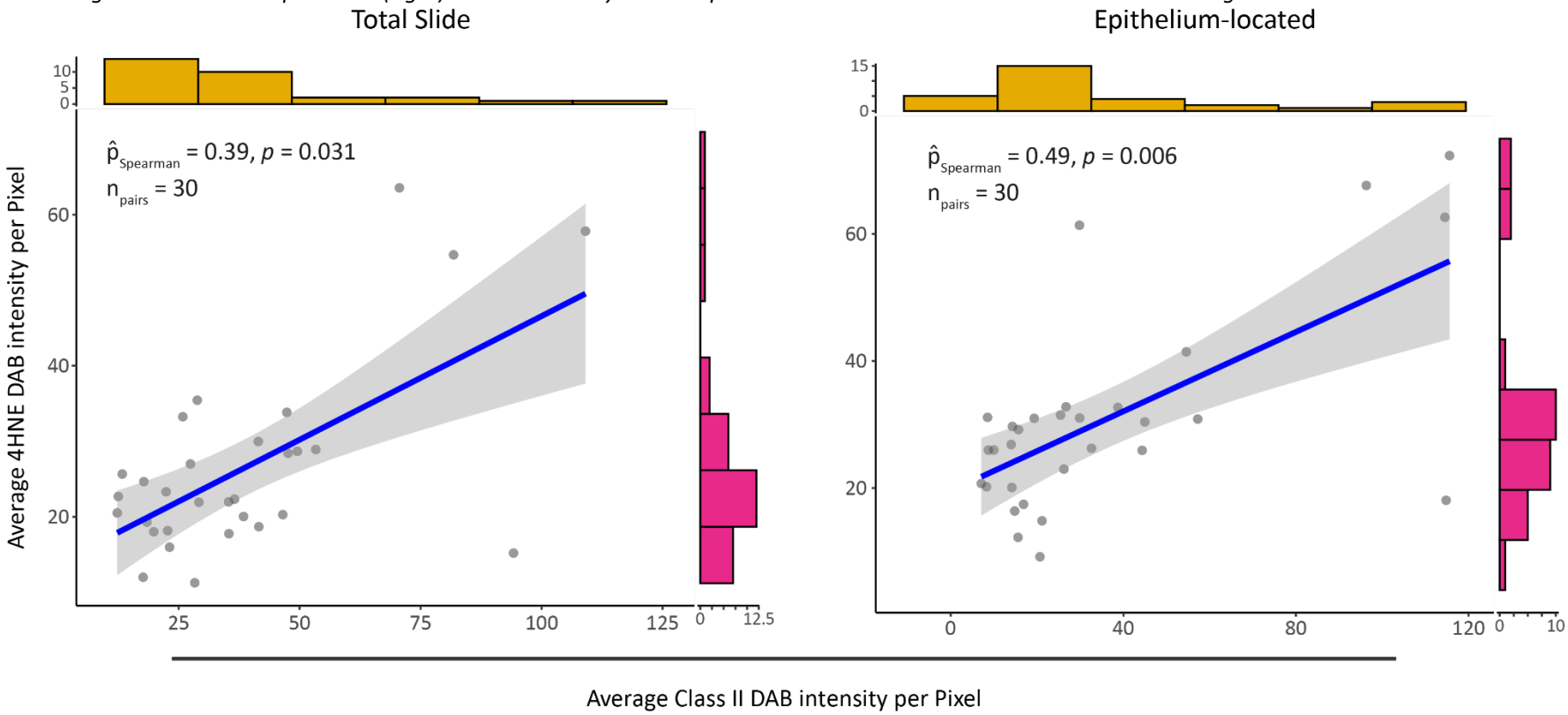


Figure 5.9 – Relationships between the intensity of HLA Class II and 4HNE IHC staining

FFPE slides from MSS patients were stained for HLA Class II in addition to 4HNE, a surrogate of ROS presence. Positive trends were visualised between the average intensity of marker per pixel as quantified by digital pathology software. Patterns were consistent across all tissue located on the slide (left) and within regions detected as epithelium (right). Statistical analysis used Spearman correlations. Bin widths were set according to the Freedman-Diaconis rule.



As MSS-hi-CIRC patients lacked a specific mutation signature that could relate to ROS-induced damage (Chapter 4), I hypothesised that MSS-hi-CIRC patients were successfully ameliorating effects of the increased ROS production, or at least to a degree to prevent genetic damage. ROS are counteracted with deployment of antioxidants, and notably the GO signature "GO:0016209 – antioxidant activity" was significantly enriched in the MSS-hi-CIRC patients in comparison to MSS-lo-CIRC patients (FDR = 0.0005). These data demonstrate that MSS-hi-CIRC patients have greater antioxidant ability to limit effects of the increased ROS they produce.

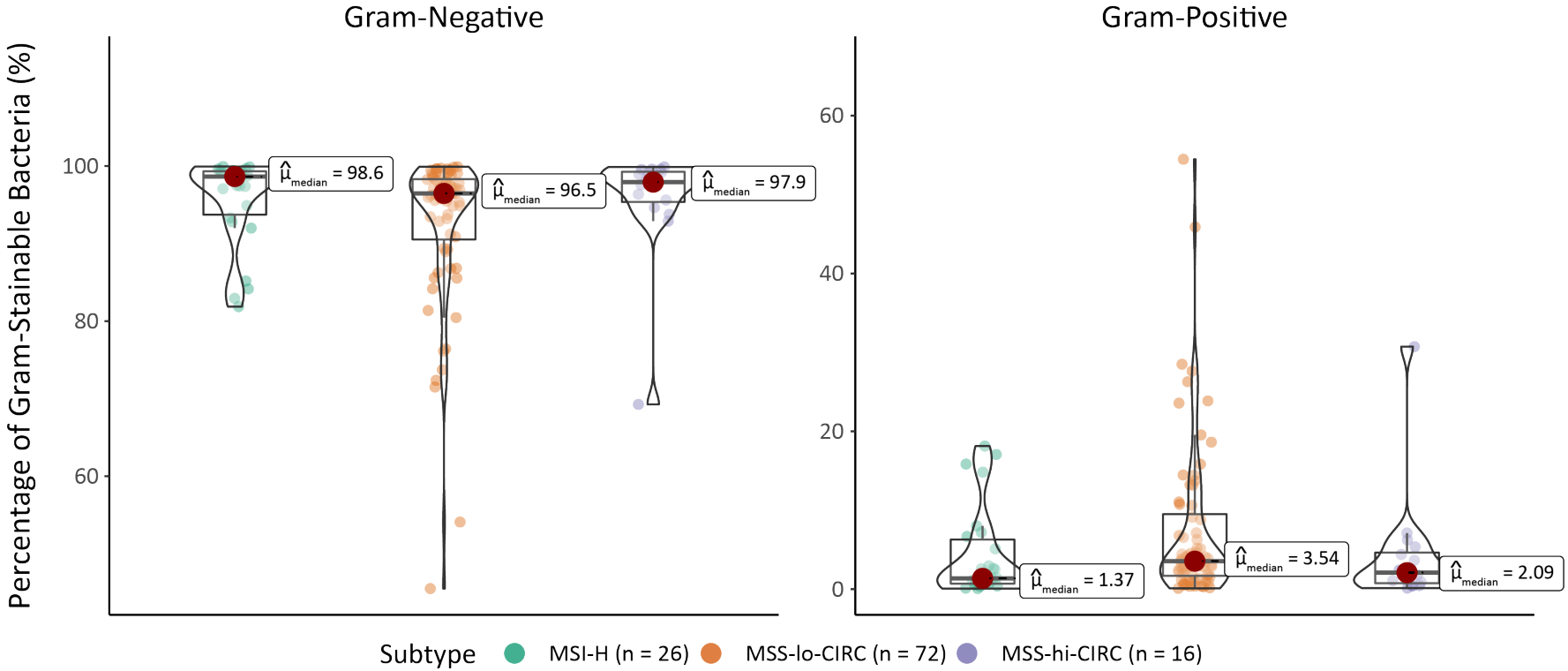
5.2.4 MSS-hi-CIRC does not exhibit increased overall abundance of Gram-negative bacteria

After collating data for gram positivity status, I returned to the Genomics England dataset to determine whether there was an increased abundance of gram-negative bacteria in the MSS-hi-CIRC patients (Figure 5.10), however, demonstrated no significant differences between the subtypes.

Figure 5.10 – A comparison of the percentage abundance of bacteria according to Gram status between the subgroups in the Genomics England Dataset

Bacterial species delineated from unmapped reads in the Genomics England cohort were annotated for Gram status according to the Gram staining rules (Chamberlain, 2019). The percentage abundance of Gram-Negative (Left) and Gram-Positive bacteria were calculated as a percentage of stainable bacteria and compared across subtypes. The number of patients belonging to each subtype is shown in the figure key.

Statistical analysis used Kruskal-Wallis tests (Gram-Negative; p-value = 0.062. $CI_{95\%}$[0.02, 1], Gram-positive; Kruskal-Wallis test, p-value = 0.062. $CI_{95\%}$[0.01, 1]). Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



5.2.5 MSS-hi-CIRC patients exhibit a distinct microbial profile enriched in *Gammaproteobacteria*

Although I detected no discernible differences between patient cohorts in terms of overall abundance of microbes when split by Superkingdom, or within bacteria by Gram-status, an enrichment of particular species which belong to the Gram-negative cohort may explain the difference in intratumoural immunity. I performed differential microbe expression between the three subgroups and found a high number of microbes were differentially expressed in the MSS-hi-CIRC versus MSI-H setting, although this was to be expected due to the increased likelihood of MSI-H developing on the right side (Zhong et al., 2020) . However, interestingly, 126 microbial species were differentially enriched in the MSS-hi-CIRC versus MSS-lo-CIRC setting. Of these 126 species, 92 were also differentially expressed in the MSS-hi-CIRC versus MSI-H setting (Figure 5.11), the top 20 of which are shown in Table 5.2.

Interestingly, 87% of the differentially expressed microbes between MSS-hi-CIRC and MSS-lo-CIRC were enriched in the MSS-hi-CIRC direction indicating increased bacterial presence in this set of patients. Of these 109 microbial species, the majority (85 species) belonged to the *Proteobacteria* phylum, a gram-negative phylum. Furthermore, over three quarters of these microbes were of the *Gammaproteobacteria* class. Finally, microbes of the *Gammaproteobacteria* class comprised the majority of the MSS-hi-CIRC-enriched microbes which were differentially enriched in both comparisons with MSS-lo-CIRC and MSI-H (Figure 5.12 and Figure 5.13, respectively). These findings strongly suggest that MSS-hi-CIRC have an enrichment of *Gammaproteobacteria* which could explain the increased immunity observed in these patients.

Figure 5.11 – Frequency of overlapping differentially enriched microbes between MSS-hi-CIRC vs MSI-H and MSS-hi-CIRC vs MSS-lo-CIRC comparisons in the Genomics England dataset
Differential microbe enrichment analysis performed between MSS-hi-CIRC and MSI-H patients revealed 1,113 microbial species (white). In contrast, MSS-hi-CIRC and MSS-lo-CIRC demonstrated only 126 differentially enriched microbes, the majority of which (92) were shared by the MSS-hi-CIRC vs MSI-H comparison.

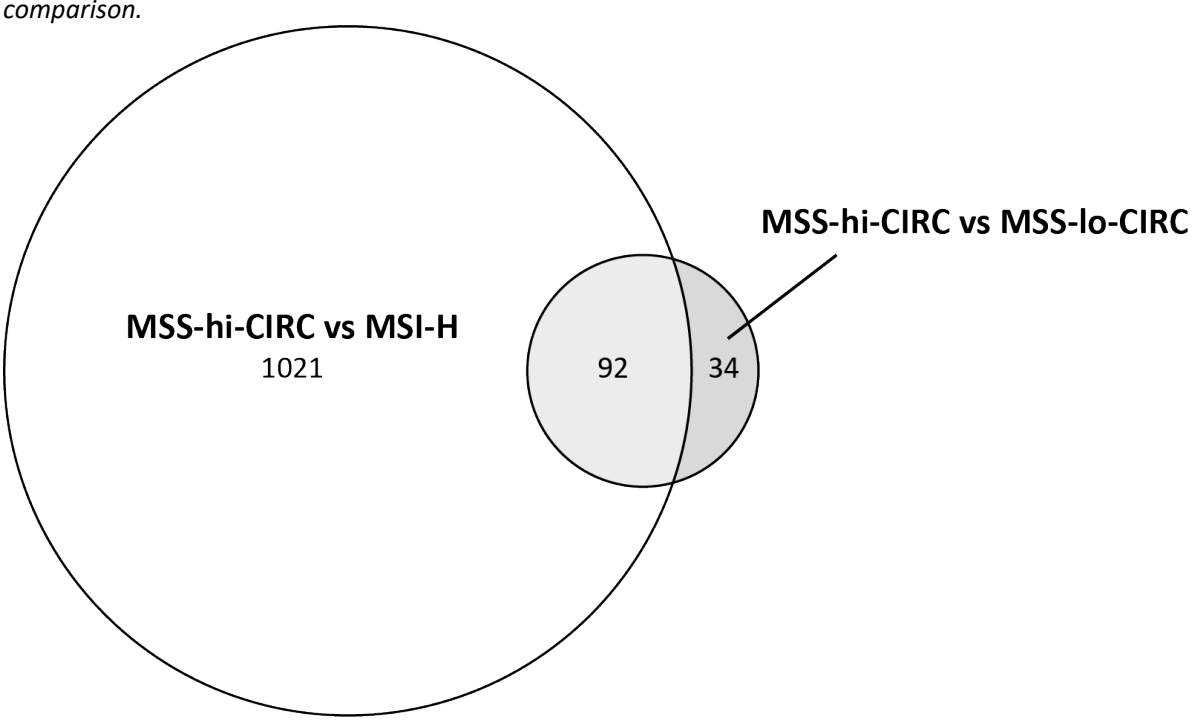


Table 5.2 – List of the top 20 differentially enriched microbes common to MSS-hi-CIRC vs MSS-lo-CIRC and MSS-hi-CIRC vs MSI-H comparisons in the Genomics England dataset

The top 20 of 92 microbes are shown. Microbes are ordered by the average adjusted P-value across the two comparisons. A positive Log(Fold Change) (FC) indicates enrichment in the MSS-hi-CIRC group whereas a negative Log(FC) indicates an enrichment in MSS-lo-CIRC or MSI-H setting. Values are rounded to 3 significant figures.

** secondary endosymbiont of*

Class	Order	Family	Genus	Species	MSS-hi-CIRC vs MSS-lo-CIRC		MSS-hi-CIRC vs MSI-H	
					Adj. Pval	*Log(FC)*	*Adj. Pval*	*Log(FC)*
Gammaproteobacteria	*Pasteurellales*	*Pasteurellaceae*	*Actinobacillus*	*A. succinogenes*	6.53E-13	1.95	8.00E-11	2.74
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*Enterobacter*	*E. sp. LU1*	2.75E-09	2.14	1.57E-09	3.88
Gammaproteobacteria	*Enterobacterales*	*Erwiniaceae*	*Erwinia*	*Candidatus E. haradaeae*	3.21E-08	1.67	1.39E-09	3.39
Gammaproteobacteria	*Enterobacterales*	*Morganellaceae*	*Photorhabdus*	*P. asymbiotica*	8.70E-10	2.35	4.35E-08	3.33
Betaproteobacteria	*Neisseriales*	*Neisseriaceae*	*Simonsiella*	*S. muelleri*	2.87E-09	3.55	8.99E-08	4.70
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*NA*	*E. bacterium ENNIH3*	1.91E-07	1.71	5.58E-07	3.01
Gammaproteobacteria	*Orbales*	*Orbaceae*	*Gilliamella*	*G. apicola*	1.06E-06	1.22	3.35E-10	3.29
Gammaproteobacteria	*Enterobacterales*	*Erwiniaceae*	*Erwinia*	*E. tasmaniensis*	8.19E-07	1.27	9.39E-07	2.22
Gammaproteobacteria	*Enterobacterales*	*Yersiniaceae*	*Rahnella*	*R. aquatilis*	3.62E-06	0.886	1.29E-09	2.77
Actinomycetia	*Streptomycetales*	*Streptomycetaceae*	*Kitasatospora*	*K. albolonga*	7.76E-06	0.786	1.43E-08	2.09
Gammaproteobacteria	*Pseudomonadales*	*Moraxellaceae*	*Acinetobacter*	*A ursingii*	1.86E-05	-1.15	2.28E-07	-2.92
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*Buttiauxella*	*B. sp. 3AFRM03*	3.62E-06	1.45	1.68E-05	2.50
Gammaproteobacteria	*Enterobacterales*	*Erwiniaceae*	*Erwinia*	*E. amylovora*	3.88E-10	1.92	2.20E-05	1.94
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*NA*	*Trabutina mannipara**	2.46E-05	0.822	8.64E-07	1.78
Gammaproteobacteria	*NA*	*NA*	*Gallaecimonas*	*G. mangrovi*	2.73E-05	0.762	1.77E-06	1.61
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*Kosakonia*	*K. sp. CCTCC M2018092*	1.50E-05	1.04	1.76E-05	1.93
Actinomycetia	*Corynebacteriales*	*Mycobacteriaceae*	*Mycobacteroides*	*M. saopaulense*	1.86E-05	0.817	1.91E-05	1.40
Gammaproteobacteria	*Enterobacterales*	*Enterobacteriaceae*	*Kosakonia*	*K. arachidis*	4.20E-05	1.35	1.05E-05	2.56

Gammaproteobacteria	*Enterobacterales*	*Pectobacteriaceae*	*Dickeya*	*D. dadantii*	8.68E-05	0.985	5.68E-06	2.10
Gammaproteobacteria	*Cellvibrionales*	*Microbulbiferaceae*	*Microbulbifer*	*M. sp. A4B17*	2.37E-10	1.83	1.19E-04	1.38

Figure 5.12 – Sankey plot demonstrating the phyla of differentially enriched microbes in MSS-hi-CIRC compared to MSS-lo-CIRC

Differential microbe analysis was performed between MSS-hi-CIRC and MSS-lo-CIRC at the species level in the Genomics England cohort. Classes shown represent only those enriched in the MSS-hi-CIRC direction. For visualisation purposes, data were restricted to Class of microbes. Flow size correlates to frequency of differentially enriched species. Only Bacterial Classes with > 2 species were annotated.

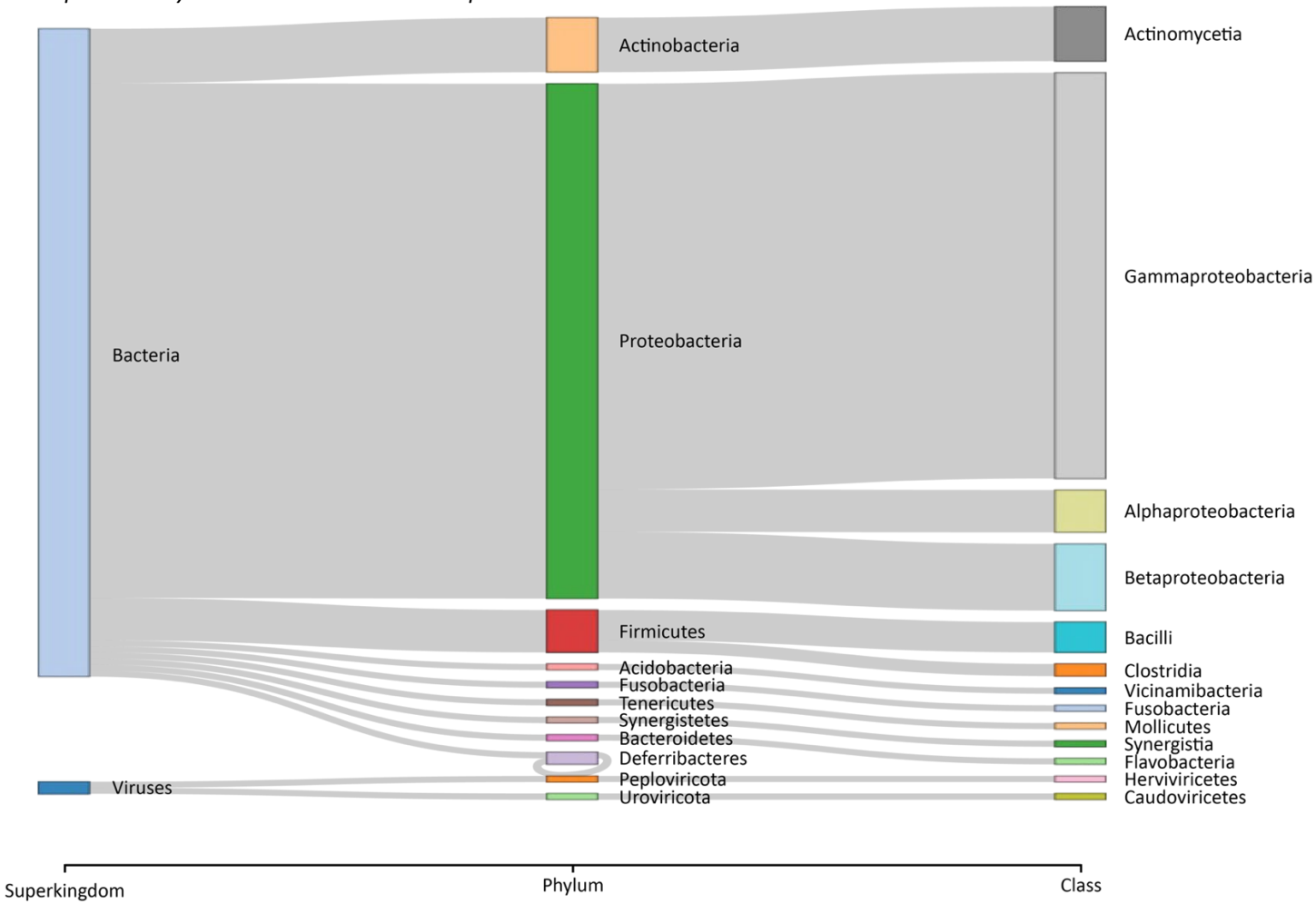
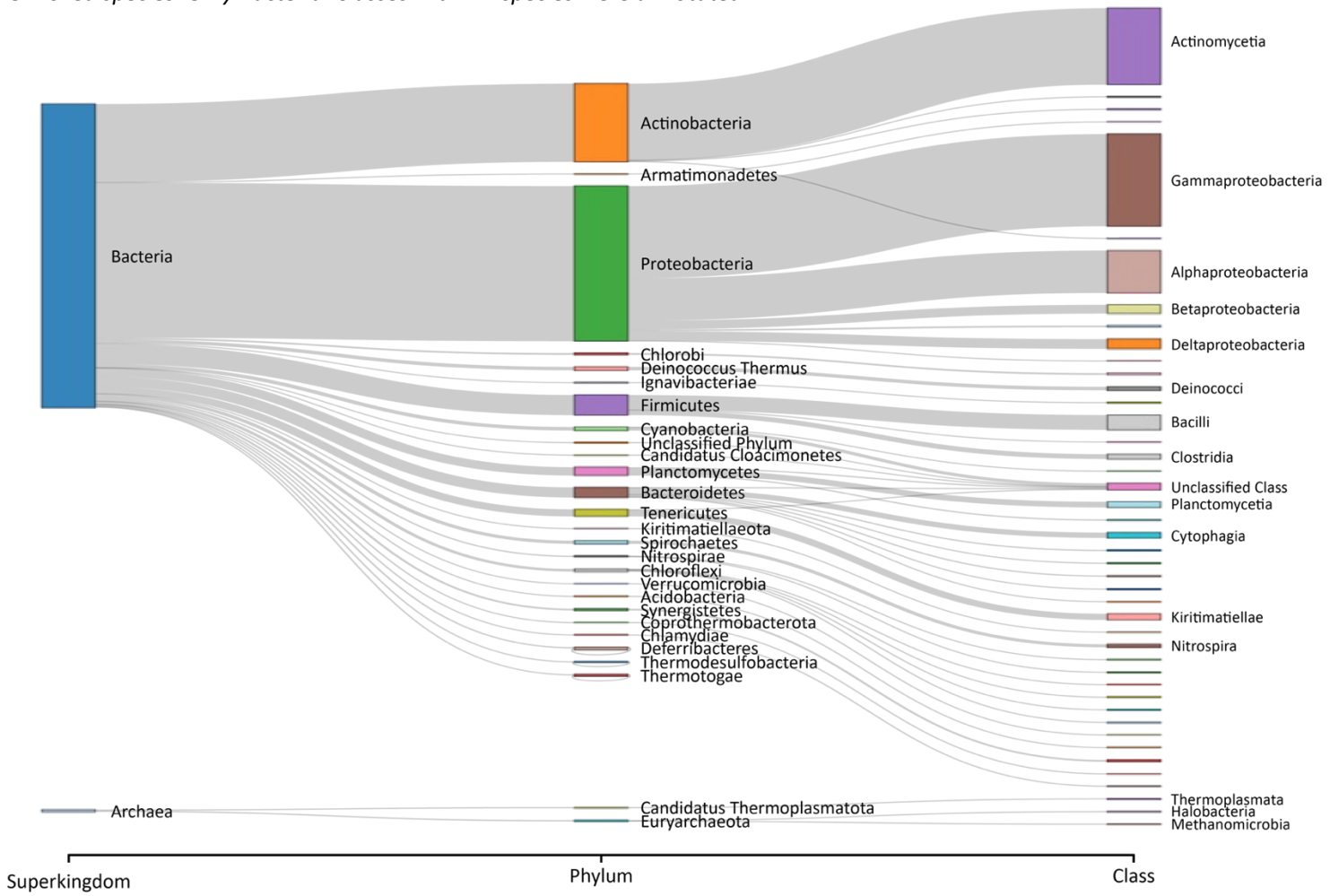


Figure 5.13 – Sankey plot demonstrating the phyla of differentially enriched microbes in MSS-hi-CIRC compared to MSI-H

Differential microbe analysis was performed between MSS-hi-CIRC and MSI-H at the species level in the Genomics England cohort. Classes shown represent only those enriched in the MSS-hi-CIRC direction. For visualisation purposes, data were restricted to classes of microbes. Flow size correlates to frequency of differentially enriched species. Only Bacterial Classes with > 2 species were annotated.



5.3 Discussion

Here, I investigated whether the source of the increased immune infiltrate demonstrated between MSS-hi-CIRC and MSS-lo-CIRC could be environmental/microbial. Utilising multispectral immunofluorescence data, I demonstrated an almost identical enrichment of T-cell subtypes between MSS-hi-CIRC and MSI-H cancers. There were, however, discrepancies in the distribution of these cell types with MSS-hi-CIRC being more epithelial-distributed than MSI-H. This was surprising, especially considering the wide reports of high intraepithelial immunity demonstrated in MSI-H (Buckowitz et al., 2005) and the tendency for MSS-hi-CIRC to include CMS4 (stromal) subtype cancers (Table 3.5 and Table 3.6).

A particularly interesting discrepancy was the differential distribution of T_H17 cells between MSS-hi-CIRC and the other subtypes. Whereas MSS-lo-CIRC demonstrated equivalently low epithelial and stromal distributions of T_H17 cells and MSI-H demonstrated a stromal T_H17 localisation, T_H17 cells in MSS-hi-CIRC were found to reside within the epithelium. In contrast to other parts of the body, T_H17 cells are enriched at barrier sites in preference to T_H1 and T_H2 where they protect the host from microorganisms, suggesting MSS-hi-CIRC may have an increased microbial presence. Despite this, metagenomic sequencing demonstrated a slight increase in bacterial burden in MSI-H patients, rather than MSS-hi-CIRC, although this difference didn't translate to a difference in staining for 16s rRNA by RNA scope. The data gathered from 16s rRNA staining requires further exploration, however. Positive and negative controls demonstrated no significant difference, and only one positive control showed strong staining (Appendix 4). Whilst this likely demonstrates an inherent technical issue with the protocol, the reason for reduced staining quality of 16s rRNA could be further exacerbated by the lack of sufficient pre-treatment. Gram-positive bacteria have both an inner membrane and a thick cell wall, and Gram-negative bacteria display a multi-layered structure consisting of a complex network of polysaccharides, peptides, phospholipids, and lipopolysaccharides

which renders them largely impermeable (Gravel et al., 2017), meaning they may require permeabilization prior to application of the RNA scope probe. Subsequent analyses by others have demonstrated the requirement for pre-treatment, whereby bacterial staining in FFPE slides of gastric cancer were dramatically increased upon pre-treatment with lysozyme (Rossiter, 2021) These analyses, therefore, should be repeated with lysozyme pre-treatment, to ensure all bacteria are evenly stained. Furthermore, this study relied heavily on the usage of FFPE tissue, which limited certain analyses. Tissue is severely dehydrated during this process, collapsing the mucus layer, and eliminating the distribution of microbes in the inner and outer mucus layers of the colon. Should these analyses be repeated, the use of Carnoy-fixed tissue would preserve the architecture of the mucus layer and allow for segregation of microbes in the inner and outer mucus layers (Johansson et al., 2008), although this would require fresh tissue as Carnoy fixing isn't a procedure routinely used in pathology departments.

I demonstrate using RNA sequencing, IHC, and WGS-derived microbial data that MSS-hi-CIRC CRC patients have an increased abundance of certain microbial species, and increased expression of pathways associated with their clearance – including structural changes distinguished by IHC. Microbes enriched in the MSS-hi-CIRC setting were overwhelmingly Gram-negative. 94.4% of those differentially enriched in MSS-hi-CIRC in comparison to MSS-lo-CIRC were Gram-negative, whereas 89.4% of those enriched in MSS-hi-CIRC in comparison to MSI-H were Gram-negative. Out of the 92 microbes which were enriched in MSS-hi-CIRC in both comparisons, 93.5% were Gram-negative. These data support the RNA sequencing findings which demonstrate an increased response to Gram-negative bacteria in MSS-hi-CIRC. Perhaps the most interesting output included the selective enrichment of the *Proteobacteria* phyla. Whilst *Proteobacteria* are present in the gut, they are usually present in relatively low quantities in comparison to the dominating presence of bacteria from the *Bacteroidetes* and *Firmicutes* phyla. In contrast, *Proteobacteria* are naturally enriched at

other barrier sites, namely the oral cavity (The Human Microbiome Consortium, 2012b). This phenomenon of oral bacteria appearing in CRC is not a new concept. Perhaps the most widely researched, *Fusobacterium nucleatum,* is heavily linked with the development of CRC (Kostic et al., 2012, Thomas et al., 2019). Interestingly, I demonstrate a significant enrichment of *F. nucleatum* in MSI-H patients with respect to both MSS-hi-CIRC (FDR = 4.35e-08) and MSS-lo-CIRC (FDR = 3.52e-09). These findings correlate with literature, which have demonstrated the strong association of *F. nucleatum* infection and a deficient MMR system (Ito et al., 2015). Here, however, I demonstrate that MSS-hi-CIRC have enrichments of the *Gammaproteobacteria* class. *Gammaproteobacteria* represent one of the most diverse classes, with over 250 genera. It contains well-known human pathogens such as those from the *Salmonella* genus, and those which can be symbiotic or pathogenic such as particular strains of *Escherichia coli* (Williams et al., 2020). Whilst *Gammaproteobacteria* such as *E.coli* have been linked with the development of CRC (Hernández-Luna et al., 2019), they are perhaps most famously linked with colitis-based diseases, in particular Crohn's disease and Ulcerative colitis. Various species of the *Gammaproteobacteria* class are enriched in inflammatory bowel diseases, however, the mechanisms in which they are enriched in these settings of inflammation are not well-understood (Rizzatti et al., 2017). A dominating theory is the oxygen hypothesis whereby particular species of bacteria thrive under various levels of oxygen (Rigottier-Gois, 2013). The colonic inflammatory reaction releases high levels of ROS, as mentioned in the Thesis introduction and demonstrated here by increased 4HNE staining in MSS-hi-CIRC patients. Interestingly, Winter et al., showed that a species of *Gammaproteobacteria*, *Salmonella enterica*, could combine ROS with lumenal sulfuric compounds to form a respiratory electron acceptor, providing it with a survival advantage over other microbes. The abundance of microbes with similar capabilities in MSS-hi-CIRC could explain the lack of a ROS-mediated mutation signature in Chapter 4 (Winter et al., 2010).

These analyses suggest future studies investigating the abundance of *Gammaproteobacteria* in these samples should be prioritised. Whilst the use of WGS data is advantageous compared to 16s rRNA approaches in gaining a greater genomic diversity and a lower risk of PCR-related artefacts, 16s rRNA amplicon sequencing is routinely performed to determine the relative abundance of microbial species in both stool and mucosal samples (Ranjan et al., 2016) and represents a cheaper alternative. The use of 16s rRNA sequencing, however, would require the collection of fresh tissue rather than the FFPE that was available to me. Whilst studies have been completed using FFPE material as input 16s rRNA sequencing, the validity of such studies is questionable (Pinto-Ribeiro et al., 2020). The processing of FFPE slides is not a sterile process, and therefore, 16s rRNA sequencing results using FFPE is likely to include readouts from contaminants alongside colonic bacterial abundance. The use of FFPE, however, is suited to targeted work. The design of qPCR primers towards the *Proteobacteria* phyla or the *Gammaproteobacteria* class could provide a cheap and quick method to determining abundance of these bacteria. The relative ease of performing this means that it could be performed on a relatively large scale and provide enough power to establish a difference in abundance. Furthermore, one could design a probe for the 16s rRNA gene which selectively highlighted *Proteobacteria* or *Gammaproteobacteria*, potentially providing information about spatial localisation with regards to immune subset and/or MHC Class II positive IECs.

Another interesting finding shown here was that MSS-hi-CIRC patients demonstrated epithelial staining of MHC Class II. IECs have been shown to have ubiquitous low-level expression of the Class II genes, which heightens on inflammation (Mayer et al., 1991). Furthermore, Koyama and colleagues have recently shown, in mice, the necessity for an intact microbiome in MHC Class II induction in IECs, and subsequent priming of $CD4^{POS}$ T-cells in driving Graft versus Host Disease (GvHD) (Koyama et al., 2019). This study, in highlighting the MHC Class II^{POS} IECs expressed the co-stimulatory molecule CD80 which, upon application of CD80 blocking antibodies abrogated $CD4^{POS}$ T-cell

activation, demonstrates the potential for IECs to act as non-professional antigen-presenting cells (APCs) in times of barrier breakdown. It's easy to speculate, considering the findings demonstrated here, that IECs in MSS-hi-CIRC CRC may present bacterial antigen to infiltrating T-cells. In support of this, Kalaora and colleagues demonstrated that tumour-colonising commensals can initiate immune responses, directed at the bacterial antigen. Utilising immunopeptidomic techniques, the authors isolated antigens from both MHC Class I and MHC Class II molecules using fresh human melanoma samples and subjected them to analysis by mass spectrometry. By cross-referencing eluted peptides to commensal proteome databases, the authors demonstrated that certain commensals could be presented via both MHC Class I and MHC Class II and that TILs isolated from fresh melanoma samples produced IFNγ and upregulated CD69 in response to MHC Class I-restricted peptides when loaded onto B-cells for presentation (Kalaora et al., 2021). Although, critically, the authors didn't demonstrate the ability for TILs to recognise antigen presented on MHC Class II, these analyses provide a rationale for a similar mechanism to be occurring in MSS-hi-CIRC patients.

Here, I demonstrate enrichments of epithelial-localisation T_H17 cells in MSS-hi-CIRC cancers. This finding has translated to enrichments of bacterial classes within the MSS-hi-CIRC cohort. Whilst determination of bacteria-restricted responses requires further investigation in the context of MSS-hi-CIRC, findings from melanoma and data presented here support the suggestion that a bacterial-derived antigen may increase the intratumoural heat of low-mutation burden cancers. To distinguish whether these findings might extend to other low-mutation burden cancers, I interrogated other datasets from the TCGA. The stomach adenocarcinoma (STAD) and the uterine corpus endometrial carcinoma (UCEC) datasets were chosen for analysis based on the routine delineation of MMR status, in addition to both cancers representing microbiome-rich, barrier-based malignancies.

Chapter 6: Bioinformatic Investigation of MSS-hi-CIRC Biology in other Low-Mutation Burden Cancers

6.1 Introduction

To distinguish whether the effects I noted in CRC were applicable to other low-mutation burden cancers, I downloaded datasets of both gastric cancer (TCGA-STAD) and uterine corpus endometrial carcinoma (TCGA-UCEC). Of note, both gastric and endometrial cancer are routinely checked for MMR status, analogous to CRC, whereby MSI-H cancers account for around 10% of cases. Importantly, MSI-H cancers in both settings have demonstrated good responses to ICB and both can be treated with ICB under current FDA regulations. In 2017, the FDA approved the use of pembrolizumab in a tissue-agnostic manner provided the cancer was unresectable or metastatic, demonstrated MSI, and had progressed following prior treatment (The U.S Department of Heath, 2017). Finally, both gastric and endometrial cancer display a comparably low mutation burden relative to CRC, and like CRC they both originate from mucosal sites, each providing a habitat for a variety of commensal bacteria (Alexandrov et al., 2013). These characteristics position gastric cancer and endometrial cancer as suitable candidates to determine whether the findings from CRC can extend to other areas of the body.

Gastric cancer refers to cancers which arise in the stomach, most of which develop in the gland cells of the inner stomach lining (adenocarcinomas). Whilst the incidence of gastric cancer is decreasing, it represents the third mostly deadly cancer in the world (Bray et al., 2018). For localised gastric cancer, the only curative treatment is surgery in which parts, or all, of the stomach are removed in a gastrectomy. For advanced gastric cancer, systemic chemotherapy is administered for palliative care. However, over the past decade, targeted treatments have become available for gastric cancer in the palliative setting such as trastuzumab for HER2-overexpressing gastric cancer (Charalampakis et al., 2018). Anatomically, the stomach is split into three main compartments – the cardia, body, and fundus. Unlike the colon, which segregates its various physiological tasks in distinct regions, the

stomach compartments work in unison to store food, and mechanically (via co-ordinated contractions) and chemically (through the production of various acids) digest food into chyme ready for further digestion. Although the stomach represents an intensely acidic environment which is largely inhabitable by microbes, the discovery of *H. pylori* infection as a major initiating factor in gastric cancer disproved the hypothesis that the stomach was a sterile area. Subsequent studies have demonstrated that the healthy stomach hosts a microbiome predominantly consisting of bacteria belonging to the *Prevotella*, *Streptococcus*, *Veillonella*, *Rothia*, and *Haemophilus* genera (Nardone and Compare, 2015).

Endometrial cancer, also referred to as uterine corpus endometrial carcinoma, describe cancers which develop in the inner lining of the uterus, termed the endometrium. The majority of endometrial cancer are diagnosed at an early stage due to the abnormal vaginal bleeding associated with its presence, and therefore, reflect a relatively high 5-year survival rate of 75.6% after diagnosis (Cancer Research UK, 2019). Initial treatment involves a surgical procedure called a hysterectomy in which parts, or all, of the uterus is removed. However, for those patients whose cancer recurs due to incomplete removal, or for those 20% of patients who present at an advanced stage, a combination of chemotherapy, and radiotherapy is often used. Host genetics play a role in the aetiology of endometrial cancer, with patients with Lynch syndrome demonstrating an increased tendency to develop MSI-H endometrial cancer (Hampel et al., 2006). However, recently, the action of the uterine microbiome has been considered a risk factor (Walther-António et al., 2016). Previously thought of as a sterile environment, the uterus is now accepted as having an abundant microbiota (Benner et al., 2018).

The prognostic influence of TILs in cancer extends to both gastric cancer and endometrial cancer, whereby studies have demonstrated that cancers with a dense immune infiltrate demonstrate better

prognoses (Jiang et al., 2017, Cermakova et al., 2014). Importantly, analogous to CRC, reports have also described a heterogeneous immune response in MMR proficient cancers from both gastric cancer and endometrial cancer cohorts (Jin et al., 2021, Liu, 2019). Here, I use the stomach adenocarcinoma (STAD) and uterine corpus endometrial carcinoma (UCEC) datasets from TCGA to investigate whether a set of MSS-hi-CIRC patients could be delineated, and whether these patients share characteristics that I previously defined in the CRC MSS-hi-CIRC cohort.

6.2 Results

6.2.1 TCGA-STAD and TCGA-UCEC demonstrate similar percentages of MSS-hi-CIRC patients to CRC patients

I subjected TCGA-STAD and TCGA-UCEC to the same pipeline as denoted in Chapter 3, using the initial geneset described by Lal (2016) for input to Phenograph clustering. The number of phenograph clusters for TCGA-STAD and TCGA-UCEC were three and five, respectively (Modularity values, 0.290 and 0.488). Cluster 3 demonstrated the highest CIRC enrichment score for TCGA-STAD, whereas both Cluster 1 and Cluster 5 demonstrated the highest for TCGA-UCEC (Figure 6.1). The percentage of MSI-H patients falling into selected clusters amounted to 46.2% and 38.0% for TCGA-STAD and TCGA-UCEC, respectively. This meant a CIRC enrichment score of above 0.398 (TCGA-STAD) and 0.397 (TCGA-UCEC) were used to denote MSS-hi-CIRC patients. One *POLE*/*POLD1* mutant existed in this description of TCGA-STAD MSS-hi-CIRC patients and was removed, whereas nine *POLE*/*POLD1* mutants existed in the TCGA-UCEC description of MSS-hi-CIRC patients and were removed. These delineations resulted in the detection of 82 TCGA-STAD MSS-hi-CIRC patients (28.1% of MSS patients) and 68 TCGA-UCEC MSS-hi-CIRC patients (19.9% of MSS patients) (Figure 6.2). I next sought to clarify the mutational landscape of these patients, analogous to analyses completed in Chapter 4 for TCGA-COADREAD patients.

Figure 6.1 – CIRC enrichment scores across phenograph clusters of MSI-H patients from TCGA-STAD and TCGA-UCEC cohorts

Phenograph clustering of MSI-H patients using the initial geneset in TCGA-STAD (left) and TCGA-UCEC (right) cohorts revealed three and five clusters, respectively. CIRC enrichment scores were calculated using RNA sequencing data and compared across the clusters. The number of patients per cluster are shown on the X axes.

Statistical analyses used Kruskal-Wallis tests (TCGA-STAD; p-value = 5.79e-12. $CI_{95\%}$[0.75, 1], TCGA-UCEC; p-value = 1.56e-25. $CI_{95\%}$[0.76, 1]). Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown. Colours represent the different clusters found in the data.

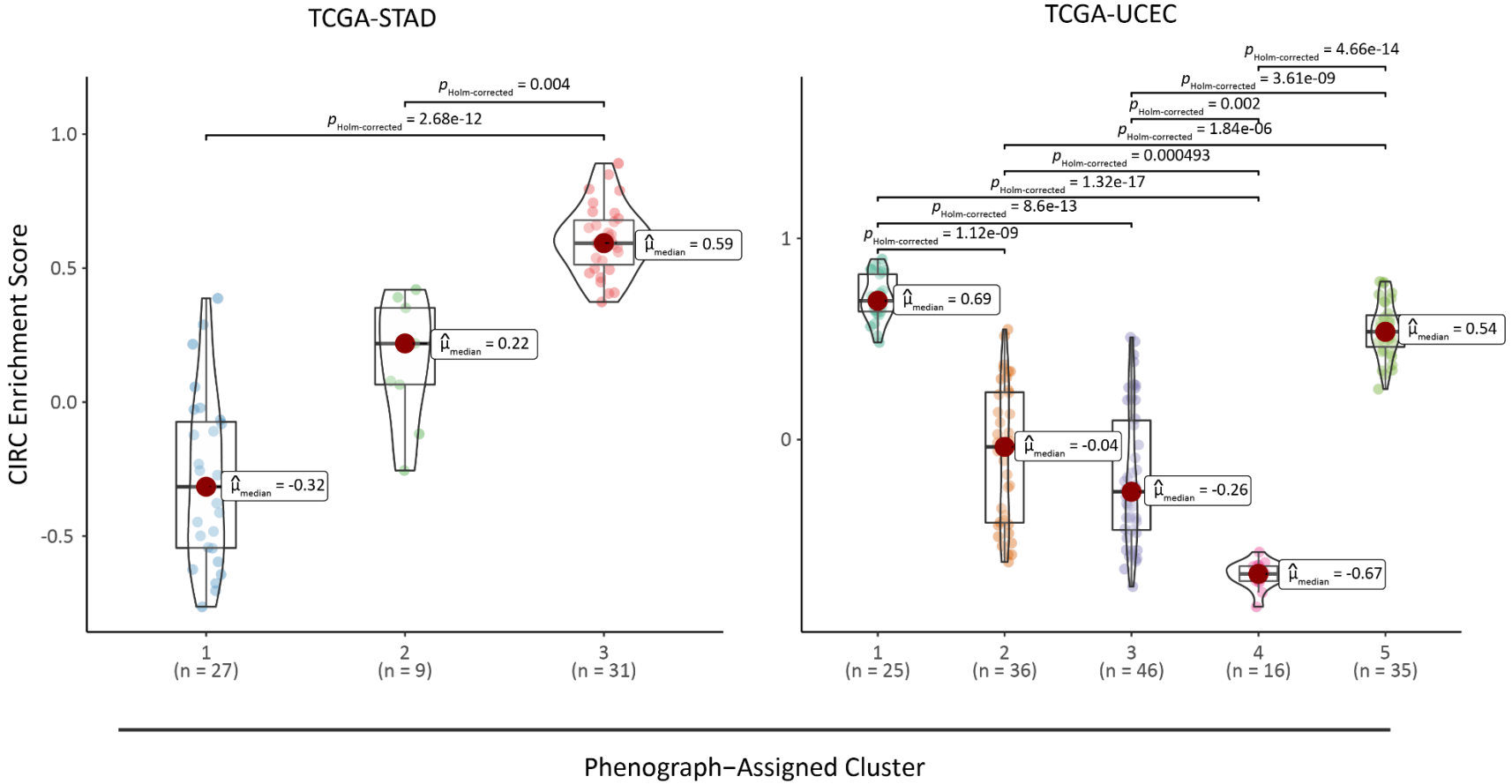
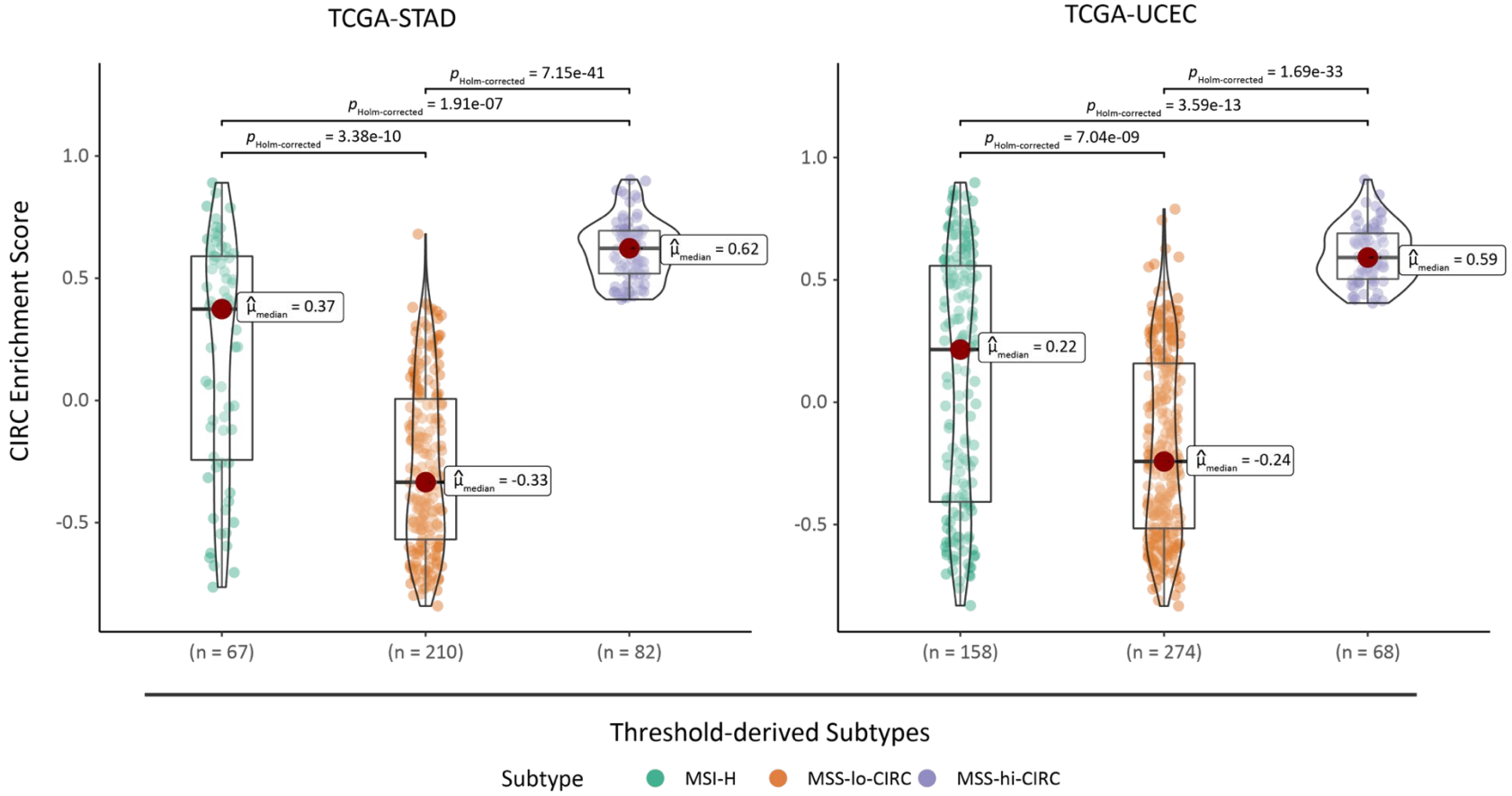


Figure 6.2 – CIRC enrichment scores of TCGA-STAD and TCGA-UCEC subtypes as defined by MMR status and CIRC enrichment score thresholds
MSS patients in TCGA-STAD (left) and TCGA-UCEC (right) datasets were split into MSS-hi-CIRC and MSS-lo-CIRC by an CIRC enrichment score cut-off as determined by Phenograph clustering on MSI-H patients from the separate cohorts. CIRC enrichment scores were calculated using RNA sequencing and compared across subtypes. The number of patients per subgroup are shown on the X axes.
Statistical analysis used Kruskal-Wallis tests (TCGA-STAD; p-value = 4.77e-42. $CI_{95\%}$[0.49, 1], TCGA-UCEC; p-value = 2.87e-34. $CI_{95\%}$[0.28, 1]). Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



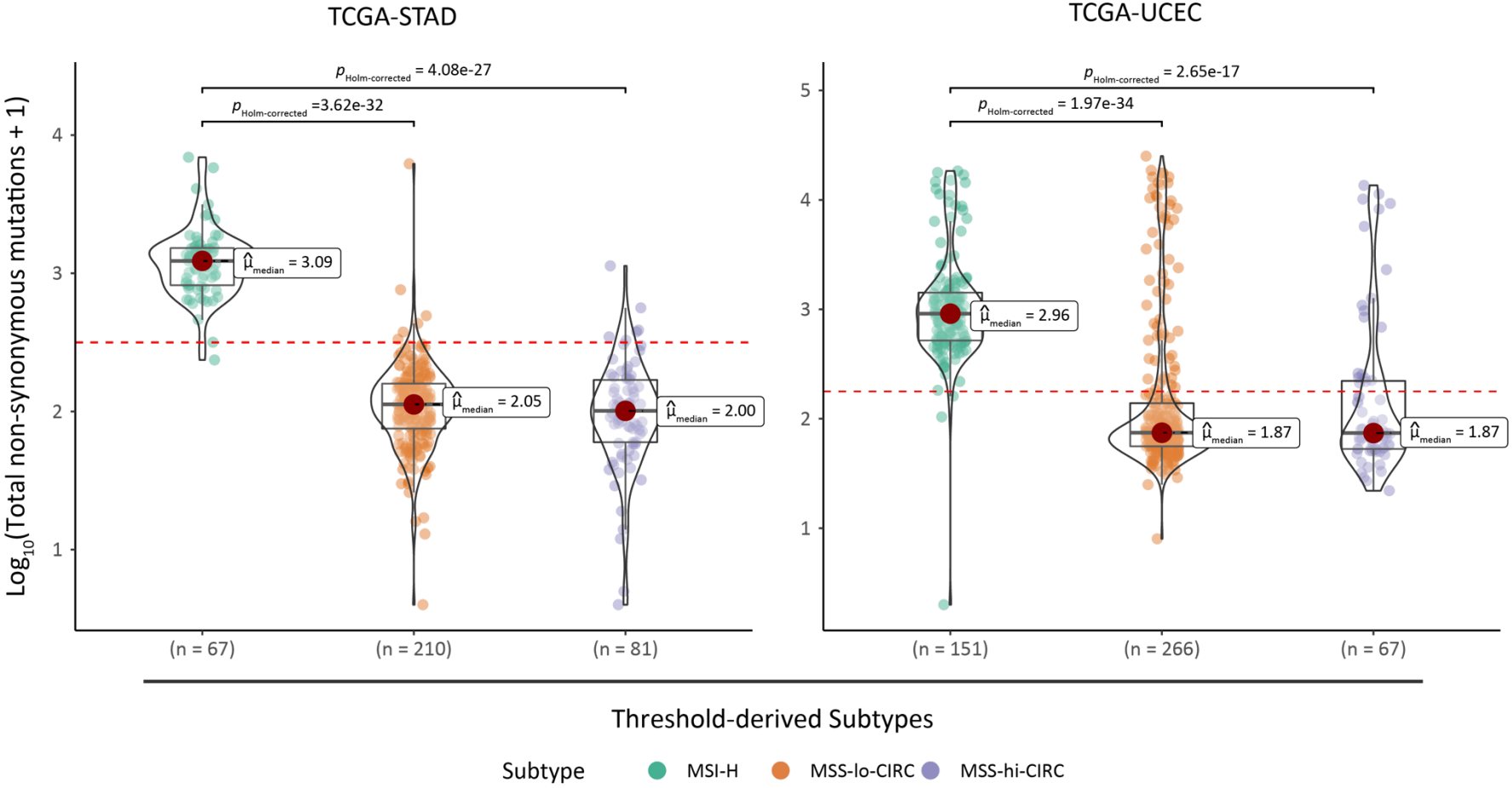
6.2.2 No differences in mutational burden exist between MSS-hi-CIRC and MSS-lo-CIRC in both TCGA-STAD and TCGA-UCEC cohorts

I next investigated mutational burden between the various subgroups. Analogous to findings in the TCGA-COADREAD dataset, MSS-hi-CIRC patients demonstrated no increase in mutation burden, whereas MSI-H patients demonstrated a greater number of mutations in both TCGA-STAD and TCGA-UCEC cohorts (Figure 6.3). Furthermore, the number of TCGA-STAD MSS-hi-CIRC patients passing the iCAM mutation burden thresholds delineated by Panda and colleagues represented just 7.4% of patients (98.5% of MSI-H, and 2.86% of MSS-lo-CIRC patients passed this threshold). These data are in accordance with the COADREAD data set whereby 3.90% of MSS-hi-CIRC and 3.98% of MSS-lo-CIRC exceeded the CRC threshold (96.7% of MSI-H COADREAD patients exceeded threshold). Panda et al demonstrated that the threshold for UCEC was lower, which led to an increased number of MSS-hi-CIRC patients passing this criterion (28.4%), however, this was much lower than MSI-H patients (98.0%), and like that of from MSS-lo-CIRC (21.4%). Finally, I could not detect any recurrently mutated genes occurring in more that 50% of each of the cohorts in MSS-hi-CIRC and MSS-lo-CIRC, whereas MSI-H mutants from both cohorts demonstrated at least six genes commonly mutated in at least 50% of the cohort (data not shown).

Figure 6.3 – Non-synonymous mutation burdens of patient subgroups in TCGA-STAD and TCGA-UCEC cohorts

The number of non-synonymous mutations for each sample were totalled and compared across subtypes. A red dashed line was placed at 2.5 (TCGA-STAD – Left) and at 2.25 (TCGA-UCEC - Right) to indicate the identified iCAM thresholds for gastric and endometrial cancer, respectively (Panda et al., 2017). The number of patients per subtype are indicated on the X axes.

Statistical analysis used Kruskal-Wallis tests (TCGA-STAD; p-value = 3.14e-35. $CI_{95\%}$[0.42, 0.47], TCGA-UCEC; p-value = 2.75e-36. $CI_{95\%}$[0.27, 0.41]). Pairwise investigations used Dunn's test. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



6.2.3 Mutation signatures did not alter between MSS-hi-CIRC and MSS-lo-CIRC cohorts

In accordance with methods used to delineated mutation signatures within the TCGA-COADREAD cohort, I performed these analyses using TCGA-STAD and TCGA-UCEC cohorts. Analyses revealed that both TCGA-STAD and TCGA-UCEC had three mutation signatures present. Comparison of these signatures across subtypes revealed no differences between MSS-hi-CIRC and MSS-lo-CIRC (Figure 6.4 and Figure 6.5), however, analogous to TCGA-COADREAD analyses, MSI-H patients exhibited the enrichment of signatures which were highly concordant with Alexandrov signature 6 (S1 in TCGA-STAD and S1 in TCGA-UCEC), the signature enriched in tumours with MMR deficiencies (92.6% similar – TCGA-STAD, 96.5% similar – TCGA-UCEC).

Figure 6.4 – Hierarchical clustering of the three mutation signatures found in the TCGA-STAD subtypes

Three mutation signatures were determined in the TCGA-STAD cohort using NMF decomposition and the frequency of 96 single nucleotide changes. The frequency of each signature was averaged across the patient subtypes and Z score transformed within each subtype. Two-dimensional hierarchical clustering was applied for ordering of signatures and patient subgroups. Increased red intensity signifies an increased strength of signature in that subtype.

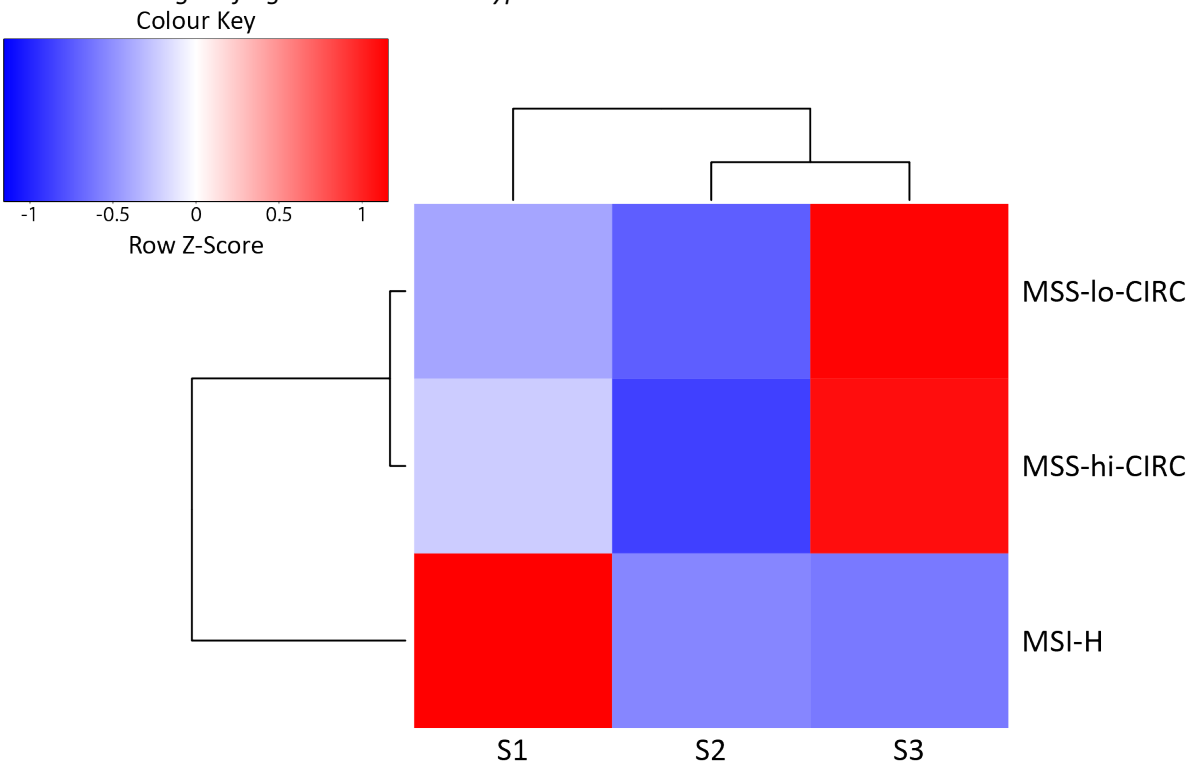
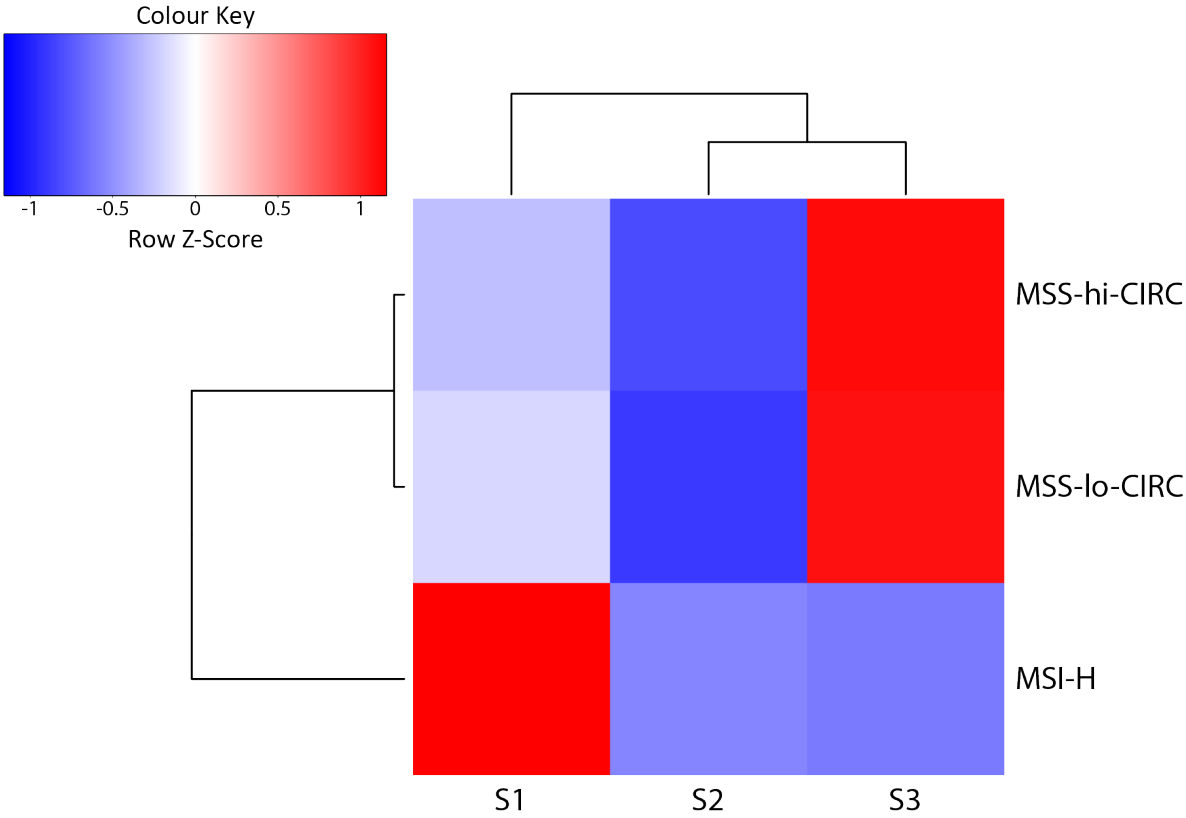


Figure 6.5 – Hierarchical clustering of the three mutation signatures found in the TCGA-UCEC subtypes

Three mutation signatures were determined in the TCGA-UCEC cohort using NMF decomposition and the frequency of 96 single nucleotide changes. The frequency of each signature was averaged across the patient subtypes and Z score transformed within each subtype. Two-dimensional hierarchical clustering was applied for ordering of signatures and patient subgroups. Increased red intensity signifies an increased strength of signature in that subtype.



6.2.4 MSS-hi-CIRC patients in the TCGA-STAD cohort demonstrate significant enrichment of microbes

In accordance with methods in Chapter 5, I compared the enrichment of pathways associated to microbes in the TCGA-STAD dataset by searching for pathways with the terms: “bacteria”, “bacterium”, “virus”, “viral”, “microbe”, “microbial”, and “archaea” in to GO database. Comparison of genesets demonstrated 40 genesets were significantly differentially enriched comparing MSS-hi-CIRC and MSS-lo-CIRC (Table 6.1), and 17 significant genesets between MSS-hi-CIRC and MSI-H (FDR < 0.05) (Table 6.2). Whereas all genesets differentially enriched between MSS-hi-CIRC and MSS-lo-CIRC were significant in the MSS-hi-CIRC direction, only those which were associated with bacterial presence were significant in MSS-hi-CIRC patients in the comparison with MSI-H patients. Interestingly, both a defence response to Gram-negative and Gram-positive bacteria were heightened in TCGA-STAD MSS-hi-CIRC patients.

Table 6.1 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC gastric cancer patients using GO terms relating to microbes
Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC patients in the TCGA-STAD cohort and GO terms relating to microbes were used as input. 40 pathways were found differentially enriched between MSS-hi-CIRC and MSS-lo-CIRC, all of which were enriched in the MSS-hi-CIRC direction. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.
NGenes = Number of Genes in the Geneset.

Geneset	NGenes	FDR	Direction
GO:0071219 cellular response to molecule of bacterial origin	177	6.42E-16	MSS-hi-CIRC
GO:0051607 defense response to virus	223	1.94E-11	MSS-hi-CIRC
GO:0002237 response to molecule of bacterial origin	300	1.94E-11	MSS-hi-CIRC
GO:0009615 response to virus	313	3.17E-11	MSS-hi-CIRC
GO:0009617 response to bacterium	500	1.57E-10	MSS-hi-CIRC
GO:0016045 detection of bacterium	16	9.65E-10	MSS-hi-CIRC
GO:0045071 negative regulation of viral genome replication	50	6.98E-09	MSS-hi-CIRC
GO:0050691 regulation of defense response to virus by host	37	8.39E-09	MSS-hi-CIRC
GO:0002230 positive regulation of defense response to virus by host	28	2.39E-08	MSS-hi-CIRC
GO:0048525 negative regulation of viral process	81	2.45E-08	MSS-hi-CIRC
GO:0050688 regulation of defense response to virus	63	2.11E-07	MSS-hi-CIRC
GO:0050830 defense response to Gram-positive bacterium	68	1.11E-06	MSS-hi-CIRC
GO:0050690 regulation of defense response to virus by virus	29	1.68E-06	MSS-hi-CIRC
GO:0042742 defense response to bacterium	178	1.84E-06	MSS-hi-CIRC
GO:0046596 regulation of viral entry into host cell	37	4.87E-06	MSS-hi-CIRC
GO:0046598 positive regulation of viral entry into host cell	11	7.06E-06	MSS-hi-CIRC
GO:1903900 regulation of viral life cycle	137	1.58E-05	MSS-hi-CIRC
GO:0032490 detection of molecule of bacterial origin	12	2.11E-05	MSS-hi-CIRC
GO:0045869 negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	16	2.39E-05	MSS-hi-CIRC
GO:0019064 fusion of virus membrane with host plasma membrane	8	5.00E-05	MSS-hi-CIRC
GO:0039663 membrane fusion involved in viral entry into host cell	8	5.00E-05	MSS-hi-CIRC
GO:0070339 response to bacterial lipopeptide	6	5.00E-05	MSS-hi-CIRC
GO:0071220 cellular response to bacterial lipoprotein	6	5.00E-05	MSS-hi-CIRC

GO:0071221 cellular response to bacterial lipopeptide	6	5.00E-05	MSS-hi-CIRC
GO:0045069 regulation of viral genome replication	78	5.33E-05	MSS-hi-CIRC
GO:0046718 viral entry into host cell	132	0.000156	MSS-hi-CIRC
GO:0050792 regulation of viral process	173	0.000333	MSS-hi-CIRC
GO:0032493 response to bacterial lipoprotein	8	0.000462	MSS-hi-CIRC
GO:0045091 regulation of single stranded viral RNA replication via double stranded DNA intermediate	19	0.000478	MSS-hi-CIRC
GO:0039692 single stranded viral RNA replication via double stranded DNA intermediate	20	0.000488	MSS-hi-CIRC
GO:0019079 viral genome replication	121	0.000723	MSS-hi-CIRC
GO:1903901 negative regulation of viral life cycle	21	0.004177	MSS-hi-CIRC
GO:0050829 defense response to Gram-negative bacterium	48	0.004575	MSS-hi-CIRC
GO:0001618 virus receptor activity	72	0.005297	MSS-hi-CIRC
GO:0046597 negative regulation of viral entry into host cell	18	0.006023	MSS-hi-CIRC
GO:0019058 viral life cycle	331	0.008826	MSS-hi-CIRC
GO:0098586 cellular response to virus	60	0.010846	MSS-hi-CIRC
GO:0016032 viral process	916	0.020638	MSS-hi-CIRC
GO:0039694 viral RNA genome replication	35	0.026802	MSS-hi-CIRC
GO:0043922 negative regulation by host of viral transcription	14	0.035776	MSS-hi-CIRC

Table 6.2 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H gastric cancer patients using GO terms relating to microbes

Genes differentially expressed between MSS-hi-CIRC and MSI-H patients in the TCGA-STAD cohort and GO terms relating to microbes were used as input. 17 pathways were found differentially enriched between MSS-hi-CIRC and MSI-H, 9 of which were enriched in the MSS-hi-CIRC direction. All significant pathways related to the detection of/response to bacteria. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.

NGenes = Number of Genes in the Geneset.

Geneset	NGenes	FDR	Direction
GO:0019080 viral gene expression	197	0.00111	MSI-H
GO:0019083 viral transcription	179	0.00138	MSI-H
GO:0075733 intracellular transport of virus	56	0.00155	MSI-H
GO:0016045 detection of bacterium	16	0.00155	MSS-hi-CIRC
GO:0046794 transport of virus	63	0.0107	MSI-H
GO:0009617 response to bacterium	500	0.0122	MSS-hi-CIRC
GO:0071219 cellular response to molecule of bacterial origin	177	0.0227	MSS-hi-CIRC
GO:0042742 defence response to bacterium	178	0.0232	MSS-hi-CIRC
GO:0032490 detection of molecule of bacterial origin	12	0.0232	MSS-hi-CIRC
GO:0019081 viral translation	16	0.0283	MSI-H
GO:0050829 defence response to Gram-negative bacterium	48	0.0283	MSS-hi-CIRC
GO:0046755 viral budding	26	0.0283	MSI-H
GO:0002237 response to molecule of bacterial origin	300	0.0283	MSS-hi-CIRC
GO:0050830 defence response to Gram-positive bacterium	68	0.0315	MSS-hi-CIRC
GO:0045869 negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	16	0.0315	MSS-hi-CIRC
GO:0075522 IRES-dependent viral translational initiation	11	0.0338	MSI-H
GO:0039702 viral budding via host ESCRT complex	21	0.0442	MSI-H

6.2.1 MSS-hi-CIRC patients in the TCGA-UCEC cohort demonstrate significant enrichment of microbes

Identical to analyses in the TCGA-STAD cohort, I considered the microbe-associated pathways from the GO database between MSS-hi-CIRC and MSS-lo-CIRC, and between MSS-hi-CIRC and MSI-H, in the TCGA-UCEC cohort. Comparison of genesets demonstrated 50 genesets were significant comparing MSS-hi-CIRC and MSS-lo-CIRC (Table 6.3), and 42 significant genesets between MSS-hi-CIRC and MSI-H (FDR < 0.05) (Table 6.4). In both instances, all genesets were significant in the MSS-hi-CIRC direction, demonstrating a significant microbial presence in MSS-hi-CIRC patients. Interestingly, only two pathways were significant in the MSI-H versus MSS-lo-CIRC comparison, "GO:0071219 cellular response to molecule of bacterial origin" (FDR = 0.00427) and "GO:0002237 response to molecule of bacterial origin" (FDR = 0.0267), both of which were enriched in the MSI-H setting.

Table 6.3 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC endometrial cancer patients using GO terms relating to microbes
Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC patients in the TCGA-UCEC cohort and GO terms relating to microbes were used as input. 50 pathways were found differentially enriched between MSS-hi-CIRC and MSS-lo-CIRC, all of which were enriched in the MSS-hi-CIRC direction. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.
NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	FDR	Direction
GO:0071219 cellular response to molecule of bacterial origin	159	1.19E-27	MSS-hi-CIRC
GO:0002237 response to molecule of bacterial origin	271	1.71E-21	MSS-hi-CIRC
GO:0009617 response to bacterium	436	2.90E-21	MSS-hi-CIRC
GO:0042742 defense response to bacterium	142	2.18E-17	MSS-hi-CIRC
GO:0050830 defense response to Gram-positive bacterium	52	2.85E-17	MSS-hi-CIRC
GO:0009615 response to virus	303	5.52E-15	MSS-hi-CIRC
GO:0051607 defense response to virus	217	1.73E-14	MSS-hi-CIRC
GO:0016045 detection of bacterium	12	4.24E-12	MSS-hi-CIRC
GO:0050691 regulation of defense response to virus by host	36	2.62E-10	MSS-hi-CIRC
GO:0048525 negative regulation of viral process	79	2.62E-10	MSS-hi-CIRC
GO:0002230 positive regulation of defense response to virus by host	27	4.20E-10	MSS-hi-CIRC
GO:0032490 detection of molecule of bacterial origin	12	9.77E-10	MSS-hi-CIRC
GO:0045071 negative regulation of viral genome replication	49	1.08E-09	MSS-hi-CIRC
GO:0050690 regulation of defense response to virus by virus	29	4.66E-09	MSS-hi-CIRC
GO:0046596 regulation of viral entry into host cell	34	6.85E-09	MSS-hi-CIRC
GO:1903900 regulation of viral life cycle	132	8.67E-09	MSS-hi-CIRC
GO:0046718 viral entry into host cell	124	8.67E-09	MSS-hi-CIRC
GO:0050688 regulation of defense response to virus	62	1.14E-08	MSS-hi-CIRC
GO:0050829 defense response to Gram-negative bacterium	38	1.83E-08	MSS-hi-CIRC
GO:0070339 response to bacterial lipopeptide	6	9.42E-08	MSS-hi-CIRC
GO:0071220 cellular response to bacterial lipoprotein	6	9.42E-08	MSS-hi-CIRC
GO:0071221 cellular response to bacterial lipopeptide	6	9.42E-08	MSS-hi-CIRC
GO:0046598 positive regulation of viral entry into host cell	11	5.88E-07	MSS-hi-CIRC
GO:0019730 antimicrobial humoral response	69	6.63E-07	MSS-hi-CIRC

GO:0050792 regulation of viral process	168	6.63E-07	MSS-hi-CIRC
GO:0061844 antimicrobial humoral immune response mediated by antimicrobial peptide	36	6.63E-07	MSS-hi-CIRC
GO:0019064 fusion of virus membrane with host plasma membrane	8	1.63E-06	MSS-hi-CIRC
GO:0039663 membrane fusion involved in viral entry into host cell	8	1.63E-06	MSS-hi-CIRC
GO:0001618 virus receptor activity	67	3.20E-06	MSS-hi-CIRC
GO:0045069 regulation of viral genome replication	77	3.55E-06	MSS-hi-CIRC
GO:0032493 response to bacterial lipoprotein	8	7.35E-06	MSS-hi-CIRC
GO:0019058 viral life cycle	322	1.93E-05	MSS-hi-CIRC
GO:1900426 positive regulation of defense response to bacterium	7	5.66E-05	MSS-hi-CIRC
GO:0019079 viral genome replication	120	7.95E-05	MSS-hi-CIRC
GO:1903901 negative regulation of viral life cycle	20	0.000137	MSS-hi-CIRC
GO:0046597 negative regulation of viral entry into host cell	17	0.000242	MSS-hi-CIRC
GO:1903902 positive regulation of viral life cycle	31	0.000420	MSS-hi-CIRC
GO:0098586 cellular response to virus	59	0.000442	MSS-hi-CIRC
GO:0045869 negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	15	0.000802	MSS-hi-CIRC
GO:0016032 viral process	897	0.00198	MSS-hi-CIRC
GO:0045091 regulation of single stranded viral RNA replication via double stranded DNA intermediate	18	0.00502	MSS-hi-CIRC
GO:0019731 antibacterial humoral response	28	0.00502	MSS-hi-CIRC
GO:0039692 single stranded viral RNA replication via double stranded DNA intermediate	19	0.00651	MSS-hi-CIRC
GO:0019065 receptor-mediated endocytosis of virus by host cell	7	0.0106	MSS-hi-CIRC
GO:0075509 endocytosis involved in viral entry into host cell	7	0.0106	MSS-hi-CIRC
GO:1900424 regulation of defense response to bacterium	13	0.0106	MSS-hi-CIRC
GO:0043922 negative regulation by host of viral transcription	14	0.0321	MSS-hi-CIRC
GO:0048524 positive regulation of viral process	88	0.0343	MSS-hi-CIRC
GO:0044788 modulation by host of viral process	27	0.0356	MSS-hi-CIRC
GO:0044794 positive regulation by host of viral process	14	0.0424	MSS-hi-CIRC

Table 6.4 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H endometrial cancer patients using GO terms relating to microbes

Genes differentially expressed between MSS-hi-CIRC and MSI-H patients in the TCGA-UCEC cohort and GO terms relating to microbes were used as input. 42 pathways were found differentially enriched between MSS-hi-CIRC and MSI-H, all of which were enriched in the MSS-hi-CIRC direction. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.

NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	FDR	Direction
GO:0071219 cellular response to molecule of bacterial origin	159	7.51E-10	MSS-hi-CIRC
GO:0009617 response to bacterium	436	3.50E-09	MSS-hi-CIRC
GO:0042742 defense response to bacterium	142	5.45E-09	MSS-hi-CIRC
GO:0009615 response to virus	303	1.66E-08	MSS-hi-CIRC
GO:0050830 defense response to Gram-positive bacterium	52	1.66E-08	MSS-hi-CIRC
GO:0002237 response to molecule of bacterial origin	271	1.66E-08	MSS-hi-CIRC
GO:0051607 defense response to virus	217	1.76E-08	MSS-hi-CIRC
GO:0045071 negative regulation of viral genome replication	49	1.44E-07	MSS-hi-CIRC
GO:0016045 detection of bacterium	12	6.03E-07	MSS-hi-CIRC
GO:0050691 regulation of defense response to virus by host	36	8.06E-07	MSS-hi-CIRC
GO:0002230 positive regulation of defense response to virus by host	27	8.06E-07	MSS-hi-CIRC
GO:0048525 negative regulation of viral process	79	9.20E-07	MSS-hi-CIRC
GO:0046596 regulation of viral entry into host cell	34	2.24E-06	MSS-hi-CIRC
GO:0050829 defense response to Gram-negative bacterium	38	5.87E-06	MSS-hi-CIRC
GO:0050688 regulation of defense response to virus	62	2.23E-05	MSS-hi-CIRC
GO:1903900 regulation of viral life cycle	132	3.31E-05	MSS-hi-CIRC
GO:0046718 viral entry into host cell	124	4.69E-05	MSS-hi-CIRC
GO:0032490 detection of molecule of bacterial origin	12	6.52E-05	MSS-hi-CIRC
GO:0019064 fusion of virus membrane with host plasma membrane	8	0.000143	MSS-hi-CIRC
GO:0039663 membrane fusion involved in viral entry into host cell	8	0.000143	MSS-hi-CIRC
GO:1903901 negative regulation of viral life cycle	20	0.000264	MSS-hi-CIRC
GO:0045069 regulation of viral genome replication	77	0.000308	MSS-hi-CIRC
GO:0046597 negative regulation of viral entry into host cell	17	0.000400	MSS-hi-CIRC
GO:0061844 antimicrobial humoral immune response mediated by antimicrobial peptide	36	0.000446	MSS-hi-CIRC

GO:0019730 antimicrobial humoral response	69	0.000488	MSS-hi-CIRC
GO:0050792 regulation of viral process	168	0.00100	MSS-hi-CIRC
GO:0050690 regulation of defense response to virus by virus	29	0.00180	MSS-hi-CIRC
GO:0046598 positive regulation of viral entry into host cell	11	0.00180	MSS-hi-CIRC
GO:0001618 virus receptor activity	67	0.00214	MSS-hi-CIRC
GO:1900426 positive regulation of defense response to bacterium	7	0.00291	MSS-hi-CIRC
GO:0070339 response to bacterial lipopeptide	6	0.00563	MSS-hi-CIRC
GO:0071220 cellular response to bacterial lipoprotein	6	0.00563	MSS-hi-CIRC
GO:0071221 cellular response to bacterial lipopeptide	6	0.00563	MSS-hi-CIRC
GO:0045869 negative regulation of single stranded viral RNA replication via double stranded DNA intermediate	15	0.00715	MSS-hi-CIRC
GO:0019079 viral genome replication	120	0.00758	MSS-hi-CIRC
GO:0032493 response to bacterial lipoprotein	8	0.0157	MSS-hi-CIRC
GO:0019058 viral life cycle	322	0.0172	MSS-hi-CIRC
GO:0019731 antibacterial humoral response	28	0.0196	MSS-hi-CIRC
GO:0045091 regulation of single stranded viral RNA replication via double stranded DNA intermediate	18	0.0454	MSS-hi-CIRC
GO:1900424 regulation of defense response to bacterium	13	0.0459	MSS-hi-CIRC
GO:0019065 receptor-mediated endocytosis of virus by host cell	7	0.0463	MSS-hi-CIRC
GO:0075509 endocytosis involved in viral entry into host cell	7	0.0463	MSS-hi-CIRC

6.2.2 TCGA-UCEC MSS-hi-CIRC patients demonstrate increased responses to reactive oxygen species

In addition to the analyses investigating microbial pathways, I decided to investigate whether TCGA-STAD and TCGA-UCEC MSS-hi-CIRC patients shared the increased response to ROS and RNS that was observed in CRC MSS-hi-CIRC patients. Here, analogous to analyses performed on the TCGA-COADREAD cohort, I performed pathway analysis on GO terms with "reactive oxygen" or "reactive nitrogen" in their name. TCGA-STAD MSS-hi-CIRC patients showed no significant differences in any of the tested pathways in either comparisons with MSS-lo-CIRC or MSI-H (Table 6.5 and Table 6.6). However, notably there was a clear trend for ROS being elevated in MSS-hi-CIRC in comparison to MSS-lo-CIRC, marginally missing significance after FDR correction. In contrast, MSS-hi-CIRC patients from the TCGA-UCEC dataset demonstrated significant differences in pathways relating to biosynthetic or metabolic production of ROS in comparison to MSS-lo-CIRC patients (Table 6.7), but not in comparison to MSI-H patients (Table 6.8).

Table 6.5 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC gastric cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS)

Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-STAD cohort and GO terms relating to ROS or RNS were used as input. 5 out of 15 available pathways demonstrated a trend to be increased in the MSS-hi-CIRC patients. All 5 pathways related to the production of ROS. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.
NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	P-value	FDR	Direction
GO:2001057 reactive nitrogen species metabolic process	69	0.00371	0.0587	MSS-hi-CIRC
GO:2000379 positive regulation of reactive oxygen species metabolic process	90	0.00782	0.0587	MSS-hi-CIRC
GO:1903428 positive regulation of reactive oxygen species biosynthetic process	46	0.0132	0.0587	MSS-hi-CIRC
GO:0072593 reactive oxygen species metabolic process	244	0.0158	0.0587	MSS-hi-CIRC
GO:1903426 regulation of reactive oxygen species biosynthetic process	82	0.0184	0.0587	MSS-hi-CIRC
GO:2000377 regulation of reactive oxygen species metabolic process	169	0.0317	0.0865	MSS-hi-CIRC
GO:1903409 reactive oxygen species biosynthetic process	104	0.0820	0.187	MSS-hi-CIRC
GO:0097468 programmed cell death in response to reactive oxygen species	11	0.196	0.391	MSS-lo-CIRC
GO:1903427 negative regulation of reactive oxygen species biosynthetic process	27	0.225	0.400	MSS-hi-CIRC
GO:0000302 response to reactive oxygen species	205	0.358	0.521	MSS-hi-CIRC
GO:1902170 cellular response to reactive nitrogen species	15	0.430	0.574	MSS-hi-CIRC
GO:2000378 negative regulation of reactive oxygen species metabolic process	53	0.588	0.724	MSS-hi-CIRC
GO:0034614 cellular response to reactive oxygen species	143	0.681	0.779	MSS-hi-CIRC
GO:1901032 negative regulation of response to reactive oxygen species	16	0.763	0.814	MSS-lo-CIRC
GO:1901031 regulation of response to reactive oxygen species	29	0.831	0.831	MSS-lo-CIRC

Table 6.6 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H gastric cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS)

Genes differentially expressed between MSS-hi-CIRC and MSI-H in the TCGA-STAD cohort and GO terms relating to ROS or RNS were used as input. No pathways were deemed significant after FDR correction. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.

NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	P-value	FDR	Direction
GO:0034614 cellular response to reactive oxygen species	143	0.0423	0.377	MS-H
GO:0097468 programmed cell death in response to reactive oxygen species	11	0.0706	0.377	MSI-H
GO:0000302 response to reactive oxygen species	205	0.130	0.519	MSI-H
GO:1903427 negative regulation of reactive oxygen species biosynthetic process	27	0.234	0.712	MSS-hi-CIRC
GO:1902170 cellular response to reactive nitrogen species	15	0.267	0.712	MSI-H
GO:1901032 negative regulation of response to reactive oxygen species	16	0.352	0.754	MSI-H
GO:1901031 regulation of response to reactive oxygen species	29	0.377	0.754	MSI-H
GO:1903428 positive regulation of reactive oxygen species biosynthetic process	46	0.665	0.950	MSI-H
GO:1903426 regulation of reactive oxygen species biosynthetic process	82	0.690	0.950	MSS-hi-CIRC
GO:2001057 reactive nitrogen species metabolic process	69	0.743	0.950	MSS-hi-CIRC
GO:0072593 reactive oxygen species metabolic process	244	0.769	0.950	MSS-hi-CIRC
GO:2000378 negative regulation of reactive oxygen species metabolic process	53	0.820	0.950	MSS-hi-CIRC
GO:2000377 regulation of reactive oxygen species metabolic process	169	0.831	0.950	MSS-hi-CIRC
GO:1903409 reactive oxygen species biosynthetic process	104	0.929	0.974	MSI-H
GO:2000379 positive regulation of reactive oxygen species metabolic process	90	0.974	0.974	MSS-hi-CIRC

Table 6.7 – Geneset enrichment analysis between MSS-hi-CIRC and MSS-lo-CIRC endometrial cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS)

Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-UCEC cohort and GO terms relating to ROS or RNS were used as input. 7 out of 15 available pathways were significantly enriched in the MSS-hi-CIRC patients. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.

NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	P-value	FDR	Direction
GO:2000379 positive regulation of reactive oxygen species metabolic process	83	9.77E-07	1.56E-05	MSS-hi-CIRC
GO:2001057 reactive nitrogen species metabolic process	66	1.75E-05	0.000140	MSS-hi-CIRC
GO:0072593 reactive oxygen species metabolic process	226	3.78E-05	0.000169	MSS-hi-CIRC
GO:1903428 positive regulation of reactive oxygen species biosynthetic process	44	4.24E-05	0.000169	MSS-hi-CIRC
GO:2000377 regulation of reactive oxygen species metabolic process	159	0.000163	0.000521	MSS-hi-CIRC
GO:1903426 regulation of reactive oxygen species biosynthetic process	77	0.000232	0.000619	MSS-hi-CIRC
GO:1903409 reactive oxygen species biosynthetic process	97	0.00151	0.00346	MSS-hi-CIRC
GO:0000302 response to reactive oxygen species	198	0.0349	0.0698	MSS-hi-CIRC
GO:0034614 cellular response to reactive oxygen species	138	0.142	0.252	MSS-hi-CIRC
GO:1903427 negative regulation of reactive oxygen species biosynthetic process	24	0.242	0.387	MSS-hi-CIRC
GO:1901032 negative regulation of response to reactive oxygen species	15	0.411	0.596	MSS-hi-CIRC
GO:2000378 negative regulation of reactive oxygen species metabolic process	50	0.458	0.596	MSS-hi-CIRC
GO:1902170 cellular response to reactive nitrogen species	16	0.484	0.596	MSS-hi-CIRC
GO:1901031 regulation of response to reactive oxygen species	28	0.604	0.690	MSS-hi-CIRC
GO:0097468 programmed cell death in response to reactive oxygen species	11	0.730	0.779	MSS-lo-CIRC

Table 6.8 – Geneset enrichment analysis between MSS-hi-CIRC and MSI-H endometrial cancer patients using terms relating to reaction oxygen species (ROS) or reactive nitrogen species (RNS)

Genes differentially expressed between MSS-hi-CIRC and MSI-H in the TCGA-UCEC cohort and GO terms relating to ROS or RNS were used as input. No pathways were deemed significant after FDR correction. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Pathways are ordered according to FDR.

NGenes = Number of Genes in the Geneset.

Geneset Name	NGenes	P-value	FDR	Direction
GO:1903426 regulation of reactive oxygen species biosynthetic process	77	0.0114	0.136	MSS-hi-CIRC
GO:2000379 positive regulation of reactive oxygen species metabolic process	83	0.0229	0.136	MSS-hi-CIRC
GO:2001057 reactive nitrogen species metabolic process	66	0.0255	0.136	MSS-hi-CIRC
GO:2000377 regulation of reactive oxygen species metabolic process	159	0.0475	0.149	MSS-hi-CIRC
GO:1903409 reactive oxygen species biosynthetic process	97	0.0557	0.149	MSS-hi-CIRC
GO:0072593 reactive oxygen species metabolic process	226	0.0577	0.149	MSS-hi-CIRC
GO:1903428 positive regulation of reactive oxygen species biosynthetic process	44	0.0651	0.149	MSS-hi-CIRC
GO:1903427 negative regulation of reactive oxygen species biosynthetic process	24	0.197	0.393	MSS-hi-CIRC
GO:0034614 cellular response to reactive oxygen species	138	0.413	0.715	MSI-H
GO:1901031 regulation of response to reactive oxygen species	28	0.567	0.766	MSI-H
GO:0000302 response to reactive oxygen species	198	0.575	0.766	MSI-H
GO:0097468 programmed cell death in response to reactive oxygen species	11	0.665	0.818	MSI-H
GO:1901032 negative regulation of response to reactive oxygen species	15	0.723	0.826	MSI-H
GO:1902170 cellular response to reactive nitrogen species	16	0.843	0.844	MSS-hi-CIRC
GO:2000378 negative regulation of reactive oxygen species metabolic process	50	0.844	0.844	MSI-H

6.3 Discussion

Although based purely on TCGA data, here I demonstrate that the findings from Chapters 3, 4, and 5 are largely applicable in both gastric and endometrial cancer. A significant proportion of proficient MMR cancers from both settings were predicted to be MSS-hi-CIRC, although this was notably higher in gastric cancer (28.1% - TCGA-STAD, and 19%- TCGA-UCEC). The increased abundance of MSS-hi-CIRC patients in gastric cancer might be explained by the high penetrance of *H. pylori* infections amongst gastric cancer patients. Around 70% of gastric cancer has associated *H. pylori* infection, and although not evaluated here, one might hypothesise that bacterial load may correlate strongly with the CIRC signature (Lee et al., 2016). Whilst the lack of differences in genesets relating to ROS or RNS between the subtypes may also be explained by the predominance of *H. pylori* infection within all subtypes (Drake et al., 1998), there was a strong trend for these to be enriched in the MSS-hi-CIRC setting. These findings suggest that MSS-hi-CIRC TCGA-STAD patients may represent those with ongoing or an increased *H. pylori* bacterial load. These questions could easily be answered should data for the bacterial load of *H. pylori* be available in TCGA clinical data, or access given to WGS data for stomach cancers in the Genomics England cohort, paralleling analyses performed on the CRC cohort. Of interest, the increased likelihood of MSI-H cancers to be enriched for viral presence was shared between TCGA-STAD and TCGA-COADREAD cohorts. Whilst this may reveal an inherent mechanism of viral-mediated development for MSI-H tumours in these settings, they may also provide favourable conditions for viruses to thrive.

Analogous to the findings presented in Chapter 5, MSS-hi-CIRC patients from both TCGA-STAD and TCGA-UCEC cohorts demonstrated increased significance of GO pathways related to bacterial presence. Moreover, here the significance also extended to FDR-corrected comparisons with MSI-H cohorts. These differences may reflect ‘sidedness’ of MSI-H CRC. Despite the tendency for gastric

cancer MSI-H cancers to be distally located, this localisation is unlikely to reflect a difference in species between the two locations due to the physical 'churning' during digestion. Although the significant differences demonstrated using pathways analysis supports this former conclusion, it is conceivable that cancers located near the oesophageal or pyloric sphincters could have very different microbial compositions, due to the differences in microbiota between the oesophagus and small intestine (The Human Microbiome Consortium, 2012b). Interestingly, TCGA-UCEC MSS-hi-CIRC cancers demonstrated an enrichment of both presence of bacterial and viral species with respect to both MSS-lo-CIRC and MSI-H patients. This is particularly interesting as enrichment of microbes have previously been demonstrated to play a vital role in development of endometrial cancer. Whilst this phenomenon of microbial adjuvantisation seems common across COADREAD, STAD, and UCEC sites, the nature of these microbes likely differs. Utilisation of WGS data from Genomics England could answer such questions, however, we hypothesise that MSS-hi-CIRC cancers provide a similar microenvironment that permits outgrowth of anaerobic commensals over naturally enriched aerobic inhabitants.

Although these analyses indicate that my earlier findings from CRC broadly apply to two other low-mutation burden cancers, it should be noted that the cancers analysed (STAD and UCEC) were also those with a dense microbiota present. It would be particularly interesting to determine whether these findings applied to other, low-mutation burden cancers which don't act as a barrier to the outside world such as in pancreatic cancer, or renal cell carcinoma (RCC). Pancreatic cancer demonstrates one of the worst prognoses as it presents with very little indication until late-stage disease. Although reports demonstrating that MSI-H pancreatic cancer patients are rare (Luchini et al., 2021), and generally immune-desolate, a high immune environment has been reported in some patients (Liu et al., 2020). In contrast, RCC is considered a highly immunogenic tumour. Utilising TCGA data, Şenbabaoğlu and colleagues demonstrated that RCC had the highest T-cell infiltration

scores, despite their low mutation burden. Furthermore, the authors demonstrated no genetic reasons for the increased T-cell infiltrate seen in these cancers, in addition to an increased T_H17 response (Şenbabaoğlu et al., 2016). These results are analogous to my analyses presented in this thesis. Whilst MSI-H is deemed a rare phenomenon amongst RCC patients, it is conceivable that there exists a significant proportion of MSS-hi-CIRC cancers within this malignancy.

Chapter 7: Overall Discussion

Immune checkpoint blockade therapies have been adopted as frontline treatments in cancers such as melanoma and lung cancer (predominantly NSCLC) and have begun to emerge as options in subgroups of other cancers that display MSI. One of the most pressing issues in cancer immunology is how to extend the success observed with these therapies to the low-mutational burden setting. There are now wide reports demonstrating the high variability in immune tone across the MSS CRC cohort (De Vries et al., 2020, Li et al., 2020, Mlecnik et al., 2016, Lal et al., 2015, Nosho et al., 2010, Ogino et al., 2009). Perhaps the most widely cited, Mlecnik and colleagues demonstrated a significant subset of MSS patients which had elevated immune signatures as well as an increased intratumoural density of $CD3^{POS}$ and $CD8^{POS}$ lymphocytes. They reported high immune infiltrates in around 50% of MSS patients, significantly more than we report in Chapter 3 (Mlecnik et al., 2016). Whilst this now seems to be an accepted phenomenon in MSS cancer, to my knowledge there have been no reports dedicated to investigating the mechanistic drivers of immune presence in these cancers. This thesis attempted to ameliorate this by systematically investigating potential mechanistic drivers.

7.1.1 The lack of MSS-hi-CIRC in the recurrence setting

A phase II clinical trial assessing the efficacy of Nivolumab in MSS MHC Class II-high cancers (ANICCA) was launched in 2019. Whereas recruitment was due to end in September 2021, a complete lack of MHC Class II MSS CRC patients was found in the recurrence setting. Subsequently, the threshold for MHC Class II positivity has been lowered to 1%, with a new recruitment end date of 2023. Therefore, no patients have been treated in this trial to date. The lack of MSS-hi-CIRC patients recruited in the ANICCA clinical trial was surprising especially considering we, and others, have demonstrated a significant proportion of MSS CRC predicted to be high immune CRC (Mlecnik et al., 2016). There are multiple possibilities to explain this discrepancy.

Firstly, MHC Class II expression and high immune infiltration in general may not translate to the recurrence setting. Our analyses of MSS-hi-CIRC were restricted to primary tumours (as has been the case for many other analyses of high immune CRC), whereas ANICCA recruited patients following disease relapse; it is therefore conceivable that MHC Class II expression (and potent intratumoural immunity) is lost/reduced upon cancer recurrence. Numerous reports demonstrate the ability of tumours to lose MHC Class I expression upon progression (Jhunjhunwala et al., 2021), however, there are also reports demonstrating that defects in the MHC Class II processing pathway are important for progression of MSI-H cancer [reviewed in (Seliger et al., 2017)](Michel et al., 2010, Surmann et al., 2015). Speaking to variability in MHC Class II expression on tumour cells, a student in our wider research group demonstrated the differing inducibility of certain MSS CRC organoids to upregulate MHC Class II in response to IFNγ (Pickles, 2021). A minority of these organoids were found to be incapable of upregulating MHC Class II irrespective of the timeframe of IFNγ treatment, due to *JAK* mutations that affected the responsiveness to IFNγ through *IRF1*/*STAT1* deficiencies – a well-documented mechanism of immune escape in MSI-H cancers (Albacker et al., 2017, Pickles, 2021). Furthermore, given that IFNγ responsiveness could be tumour cell intrinsic we cannot exclude the possibility that surgery itself could affect the balance between tumour clones, allowing for outgrowth of those clones that have lost sensitivity to IFNγ, and by extension their ability to upregulate MHC Class II. Whilst no patients in our MSS subgroups displayed these mutations based on analysis of TCGA data relating to their primary tumour, it is conceivable that cancer subclones bearing mutations in the IFNγ responsiveness pathways could be selected for over time allowing for outgrowth post-surgery, especially considering the large $CD4^{POS}$ infiltrate in MSS-hi-CIRC cancer. In contrast, Pickles demonstrated the ability of most organoids to express MHC Class II on continued exposure to IFNγ, despite not expressing MHC Class II *in situ* (Pickles, 2021). These results suggest MHC Class II expression is to a significant degree governed by environmental exposure. Therefore, an alternative and perhaps a more likely route of surgical perturbation might be that upon surgery in

which tumours are removed, along with surrounding tissue, an infectious agent(s) inhabiting the cancer cells or surrounding microenvironment may be removed with it. This could conceivably allow for clearance of the remaining microbes but not necessarily all cancer cells, allowing for regrowth of cancer in a non-infected environment.

A second, and arguably more likely possibility, based on the established benefits of increased intratumoural immunity on prognosis, is that MSS-hi-CIRC patients don't progress to disease recurrence, and that the intratumoural immunity observed in MSS-hi-CIRC effectively contains the cancer, enabling surgery that is effectively curative. Whilst not being able to test the efficacy of ICB in this patient cohort, this mechanism predicts that the inflammation observed in MSS-hi-CIRC patients is beneficial, and warrants further investigation regarding its environmental drivers, with a view to developing therapies that recapitulate it in the MSS-lo-CIRC setting. Based on the data presented in Chapter 5, one could foresee a therapeutic strategy conceptually analogous to Bacillus Calmette-Guerin (BCG) treatment for bladder cancer. BCG, which comprises attenuated *Mycobacterium bovis*, is administered intravesically, meaning directly into the bladder, inducing a highly inflammatory immune response in the surrounding urothelium and tumour (Ratliff, 1992). Importantly for results presented in Chapter 5 regarding the increased abundance of *Gammaproteobacteria* in immune-high MSS cancers, studies have shown that treatment with other *Gammaproteobacteria* recapitulate the immune responses observed with BCG (Kim et al., 2015). After intravenous injection with attenuated *Salmonella typhimurium*, the authors showed that *S. typhimurium* selectively targeted murine subcutaneous tumours, induced an increase in adaptive immune cell infiltration, and a reduction in tumour burden (Kim et al., 2015). Furthermore, whilst these experiments relied on the colonisation of tumours with a pathogen, other data suggest that the microbe that colonises tumours does not need to be pathogenic to aid anti-tumour immunity. Sivan and colleagues demonstrated that oral administration of the commensal bacteria *Bifidobacterium* reduced tumour volume in subcutaneous

melanoma murine models. Importantly, in CD8-depleted mice and in cases whereby a heat-inactivated version of *Bifidobacterium* was used no differences in tumour control were reported. This highlights the importance of CD8 T-cells in the control of tumour growth as well as the necessity for live bacteria (Sivan et al., 2015). These results, whilst striking, do not exclude the possibility that other commensals are able to regulate anti-tumour immunity, nor do they exclude synergistic interactions with other bacteria or indeed viruses. Of note, administration of *Bifidobacterium* slightly altered the composition of the microbiota, with a reduction noted in members of the *Clostridia* order, as well as other butyrate-producing species – suggestive of a microbiotal reprogramming to a more inflammatory state. In the era of Faecal Microbiota Transplantation (FMT), the potential to reprogram patients' microbiota to a 'favourable state' is now a possibility. FMT is now routinely used to treat colitis, and in mouse models of cancer has demonstrated great promise in conferring a sensitivity to ICB (Routy et al., 2018, Matson et al., 2018, Gopalakrishnan et al., 2018).

Whether MSS-hi-CIRC represents a phenomenon restricted to primary tumours but that is lost at some point post-surgery during the progression of residual tumours towards disease relapse, or whether it represents a cohort of patients in which the immune infiltration enables effective containment of the tumour, resulting in surgical resection that is effectively curative, remains both undetermined, and a key future question. Importantly, these two respective scenarios could be delineated with the access to primary ANICCA samples. Measurement of Class II expression on patient-paired primary and post-recurrence tumour samples should clarify whether MSS-hi-CIRC patients have subsequently lost Class II expression by the point of disease recurrence, or alternatively whether MSS disease recurrence exclusively occurs in patients with MSS-lo-CIRC primary tumours. However, regardless of which scenario applies, understanding the mechanism of MSS cancer with an increased immune tone is vital.

7.1.2 Mechanistic drivers of the MSS-hi-CIRC grouping

Some researchers might explain this MSS-hi-CIRC grouping by hypothesising such patients possess unusual combinations of MHC haplotype with specific tumour mutations, permitting presentation of particularly critical/immunogenic neoantigens that drive strong intratumoural immunity. While I do not directly refute such a hypothesis, the genetic data in this thesis, combined with other mutation-based findings elsewhere (Lal, 2016, Mlecnik et al., 2016, Scheper et al., 2019), and the data and arguments in Chapter 5, suggest the influence of an equally powerful tumour cell-extrinsic driver of high immunity (i.e., the microbiome) which should be included in mechanistic models of intratumoural immunity. Here, I demonstrate an increased abundance of the *Proteobacteria* phyla, as well as other members of the oral microbiome (*Simonsiella muelleri*) that are enriched in MSS-hi-CIRC patients in comparison to MSS-lo-CIRC and MSI-H, which could be the driver of increased immunity in these patients. Increases in oral microbiome members have been described elsewhere, however, have previously only been associated with inhibition of an immune response (Mima et al., 2015). Whilst we cannot rule out bacterial species acting to inhibit immunity in MSS-lo-CIRC patients, we find only 17 microbial species enriched in these MSS-lo-CIRC tumours in comparison to MSS-hi-CIRC tumours. In contrast, 109 microbial species were significantly enriched in the MSS-hi-CIRC setting, suggesting a stimulatory effect on T-cell responses against MSS-hi-CIRC cancers. The method in which bacteria drive T-cell responses could be both direct, indirect, or, more likely, a mixture of both. However, it is likely that MHC Class II expression by IECs is central to these mechanisms. As covered briefly in the Chapter 3 discussion, the transfection of *CIITA* into MHC Class II negative tumour cell lines rendered these cells MHC Class II positive and rejectable by 92% of mice. Importantly, authors showed that the tumour microenvironment of these mice had increased infiltrates of $CD4^{POS}$ and $CD8^{POS}$ T-cells, and that IFNγ production was vital in the production of an anti-tumour response (Mortara et al., 2006). Increased IEC MHC Class II expression is dependent on environmental IFNγ. The source of this IFNγ can be many folds. $CD8^{POS}$ T-cells, $CD4^{POS}$ T-cells, natural

killer (NK) cells, NK T-cells, ILC1 cells, and γδ T-cells are all present within and around the IEC layer and have the capacity to produce IFNγ. IFNγ production by these cells, however, can be activated in both antigen-driven mechanisms, such as via innate immune receptors, or by exposure to inflammatory cytokines such as IL-18 and IL-12. Innate pathways are almost certainly at play, however, whether infiltrating antigen-specific T-cells are directed at any of the microbes enriched in the MSS-hi-CIRC setting is unclear.

Findings indicating that *Salmonella*-mediated tumour clearance occurs in nude mice provided the first indications that tumour rejection could occur in a T-cell independent manner, although MHC Class II expression was still induced (Luo et al., 2001, Kaimala et al., 2014). *Salmonella* infection increases expression of vital inflammasome components, such as NLRP3, which increases supernatant concentrations of IL-1β, and mediates MC38 tumour clearance in murine models, providing a means of inflammasome-mediated clearance (Phan et al., 2015). Interestingly, we demonstrated that *GBP5* was the most upregulated gene between MSS-hi-CIRC and MSS-lo-CIRC patients. Increased expression of *GBP5*, a suggested co-factor in promoting NLRP3-dependent inflammasome assembly in by IFNγ, supports the role of the inflammasome-mediated anti-bacterial responses in MSS-hi-CIRC tumours (Kim et al., 2016a), however, the role of *GBP5* in this system requires further verification by IHC. Whilst these murine studies demonstrate the ability of innate-mediated clearance, it is important to note that when T-cells are present in the host, they do induce a T-cell response upon the microbial-colonisation of tumours, providing the potential for a specific, anti-microbial, and intratumoural response (Koyama et al., 2019).

Other studies have demonstrated a potential for anti-commensal T-cell responses in melanoma. As overviewed in the Chapter 5 discussion, the Samuels laboratory were able to isolate peptides from bacteria and demonstrate their ability to be presented as antigen, activate $CD8^{POS}$ responses, and

reduce tumour burden (Kalaora et al., 2021). In similarity to this study, and should fresh resection tissue be available, a similar methodology could be adopted here, however, the colon is prone to high contamination. Should this work be recapitulated in this setting, the need for various controls (no template PCR controls etc.) ensuring low contaminations are a priority. Given the sheer abundance of bacteria that the colon encounters it would not be surprising if a similar microbial-antigen specific response is not occurring in this context. Furthermore, bacterial antigens may be able to elicit a response that is cross-reactive to otherwise lowly-expressing or low-efficacy tumour antigens. The concept of molecular mimicry, which is hypothesised to underpin the pathogenesis of various autoimmune diseases, suggests that microbial antigens share homology with weaker stimuli present on IECs. Multiple instances of potential molecular mimicry have now been described (Cusick et al., 2012), however, perhaps the most relevant here are those which are hypothesised to occur in gastritis, or in Crohn's disease, although the self-antigen in the latter is undetermined. By isolating T-cells from biopsies of *H. pylori* infected patients with autoimmune chronic gastritis (AIG), T-cells were shown to proliferate in response to both *H. pylori* lysate and peptides of the H^+K^+-ATPase abundant in the parietal cell canaliculi of the stomach. Interestingly, by cross-referencing peptides shown to induce responses in both settings, 20 H^+K^+ ATPase peptide sequences were shown to have homology to proteins from *H. pylori*, including LPS – a widely expressed protein by gram-negative bacteria (Amedei et al., 2003). These results demonstrate the potential for sequences derived from bacteria to have homology and even elicit responses to self-antigen. Interestingly, this concept has also been demonstrated in tumour models, and exploited to induce targeting of tumour associated antigens (Fluckiger et al., 2020, James and Bond, 2012, Boesch et al., 2021). It is entirely conceivable that tumour mutations may increase the significance of cross-reactivity from molecular mimic peptides, due to their altered antigenic sequence circumventing self-tolerance.

The issue of whether T-cells in MSS-hi-CIRC are directed at microbes and/or tumour antigens remains to be resolved. It is likely that an interplay of both mechanisms is occurring, and it is possible to envisage how a positive feedback loop might be established. Initial bacterial translocation could drive an IL-12 dependent activation of IFNγ production, and IFNγ then increases MHC Class II expression on the colonic epithelium. Increased MHC Class II expression allows for the presentation of cognate ligand to effector-memory T-cells that reside as an on-site population (Sathaliyawala et al., 2013). Whether this scenario represents microbial adjuntivisation of tumour antigen-specific responses purely through innate pathway activation, or alternatively via molecular mimicry involving T-cell responses induced to microbial antigen that then cross-react to tumour antigens, is important to understand. These questions could potentially be answered with the acquisition of fresh resection tissue and performing analyses analogous to those outlined in Kalaora et al., (2021).

7.1.3 The link with obesity

Whilst highly speculative, the link between SNPs found enriched in MSS-hi-CIRC patients in Chapter 4 and obesity is an interesting area for further investigation. Obesity, defined as having a body mass index above 30, is an epidemic in the Western world. It imposes significant health tolls on an individual with a recent meta-analysis demonstrating that overweight individuals were twice as likely to developing cardiometabolic diseases than healthy weight counterparts. Moreover, individuals with class II (BMI between 35 and 39.9) or class III (BMI equal to or greater than 40) obesity were 15 times more at risk of developing cardiometabolic disease (Kivimäki et al., 2017). This association extends to cancer. Current evidence names obesity as a major preventable risk factor for multiple different cancers, including CRC (Ma et al., 2013), gastric cancer (Chen et al., 2013), and endometrial cancer (Setiawan et al., 2013). Multiple possible mechanisms exist that could explain the increased risk of cancer with increasing obesity, however, perhaps the most relevant here are increased hormone production, and chronic low-level inflammation.

Adipose tissue generates high levels of various hormones. The sex hormone oestrogen is elevated in obese individuals, due to the adipocyte expression of aromatase – the enzyme responsible for conversion of androgens to oestrogen. An increase in the concentration of oestrogen, when not inhibited by the concomitant increased production of progesterone, will provide growth signals to the endometrium and lay the groundwork for cancer to develop (Rodriguez et al., 2019). Other hormones produced by adipocytes, such as leptin, have been shown to have a proliferative effect on multiple different cancer cell lines *in vitro*, such as gastric cancer and CRC cells (Schneider et al., 2001, Hardwick et al., 2001). However, such proliferative action did not extend to an increase in tumour burden in ApcMin mice, despite the 2.4-fold increase in plasma leptin levels compared to non-leptin fed controls (Aparicio et al., 2005). Despite the lack of an effect in mice, a report from 2004 demonstrated that human colon cancer increased in likelihood with increasing serum leptin concentrations. Furthermore, these risks were most pronounced for those cancers developing on the left side of the colon, and therefore, may increase the risk of MSS-based colon cancer (Stattin et al., 2004). Finally, and perhaps most relevant here, obese individuals present with chronic low-level inflammation. Multiple reports have demonstrated a correlation between obesity and systemic increases in concentrations of pro-inflammatory cytokines, most notably IL-6 (Fontana et al., 2007, Illán-Gómez et al., 2012). By evaluating IL-6 plasma levels in obese patients after bariatric surgery, Illan-Gomez and colleagues demonstrated a substantial decrease in circulating IL-6 plasma levels in the 12-months post-surgery. Furthermore, they demonstrated a significant association of IL-6 levels and BMI. The increased expression of circulating IL-6 is implicated in the pathogenesis of a variety of diseases, however, notably in the development of gastric and colorectal cancers (Ellulu et al., 2017).

The increased concentration of circulating inflammatory cytokines were not the only alterations associated with being obese, with numerous studies demonstrating organ-specific effects. High-fat

diets have been shown to have significant effects on gut transit time, but also the propensity for a 'leaky gut'. Attempting to recapitulate findings demonstrating increases in gut permeability in ob/ob mice, Teixeira and colleagues compared concentrations of lactulose and mannitol in the urine of lean and obese women (Cani et al., 2009, Teixeira et al., 2012). The study demonstrated a trend for an increased lactulose:mannitol ratio in the urine, concurrent with increased intestinal permeability in the obese group. An increase in gut permeability would not only allow paracellular passage of sugars as shown in the aforementioned study but would also be expected to enable increased bacterial translocation and changes in microbiota compositions, providing a means for initiation of intestinal inflammation. Consistent intestinal inflammation, as observed in obesity, in linked with the preferential bloom of certain phyla of bacteria. Under steady state, *Proteobacteria* makeup ~4.5% of bacteria in the gut. However, this phylum was seen to be expanded in children who ate high-fat diets (De Filippo et al., 2010). Interestingly, *Proteobacteria* seem to be one of the commensals most sensitive to environmental factors, such as diet. The Knight group, utilising PCR for the variable region of the 16s rRNA genes and multiplex sequencing, took snapshots of the microbiota over 396 timepoints and demonstrated that *Proteobacteria* could increase (albeit transiently) to comprise up to 45% of the gut microbiome (Caporaso et al., 2011).

Whereas one might conclude that *Proteobacteria* can cause inflammation in the gut, it is perhaps more likely that they thrive in the inflamed niche. *Enterobacteriaceae*, which are benign commensals in steady state, and members of the *Gammaproteobacteria* phylum, can use nitrates generated from host inflammatory responses as alternative electron acceptors, allowing them to prosper during inflammation, where they outcompete microbes reliant on fermentation for growth (e.g., *Firmicutes* and *Bacteroidetes*) (Shin et al., 2015). Interestingly, my data demonstrate an increased abundance of *Gammaproteobacteria* with a particular enrichment of *Enterobacteriaceae* in the MSS-hi-CIRC setting. As literature demonstrates the enrichment of these bacteria in that of both obese, and colitis

patients, I propose a mechanism, albeit highly speculative, about the aetiology of MSS-hi-CIRC cancers (Figure 7.1).

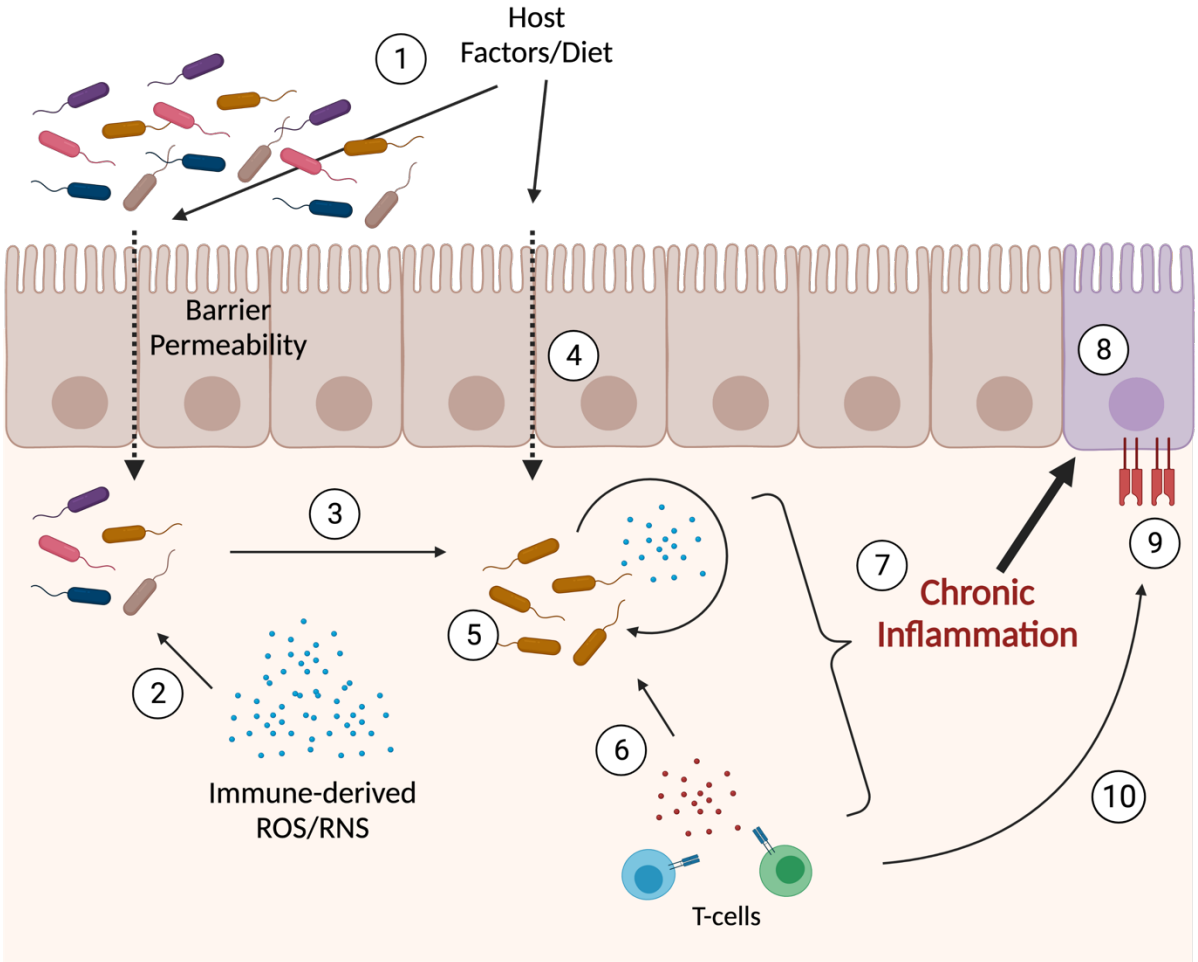
7.1.1 The MSS-hi-CIRC hypothesis

My proposed mechanism for aetiology of MSS-hi-CIRC cancers is outlined in Figure 7.1. The disruption of the colonic intestinal barrier, favoured by host factors (e.g., SNPs linked to leaky barrier function or obesity) and/or environmental factors (e.g., diet, and potentially downstream obesity), permits increased translocation of bacteria through the gut mucosa. Such increased bacterial translocation induces inflammatory responses, including the production of ROS and RNS. An environment high in ROS and RNS provides conditions that permit proliferation of anaerobic microbes, such as *Proteobacteria*. The inability to clear *Proteobacteria* because of ongoing inflammation, due to continued insult to the intestinal barrier (driven by host/environmental factors as above), provides a positive feedback mechanism which further exacerbates inflammation, leading to inflammation-induced cancer, and allowing for a *Proteobacteria* bloom. Cancer development and progression is accompanied by immunosuppressive mechanisms that limit immune cell mediated responses, impeding clearance of *Proteobacteria*. ROS and RNS produced by the immune system fail to induce genetic damage to the cancer due to the antioxidant effects of *Proteobacteria* in utilising these as fuel for respiration. This theory provides a mechanism in which initial bacterial translocation leads to a symbiosis between anaerobic microorganisms and cancer.

Figure 7.1 – Schematic denoting the theory of MSS-hi-CIRC

1) *Disruption of barrier permeability via host or dietary factors allows for increased translocation of commensals*
2) *On site immune system reacts to microbes, producing an environment rich in reactive oxygen species (ROS) and reactive nitrogen species (RNS)*
3) *A ROS/RNS high environment selects for species best equipped to deal with these environments (Proteobacteria)*
4) *Continued insult means inflammation is not resolved*
5) *Increased ROS/RNS allows for the species best equipped to deal with ROS/RNS to bloom*
6) *Adaptive immune system attacks bacteria, producing a highly inflammatory environment*
7) *Chronic inflammation ensues from unresolved insult*
8) *Chronic inflammation allows for malignant transformation of enterocytes (purple enterocyte)*
9) *In response to highly inflammatory environment, the colonic epithelium increases MHC Class II expression (red molecule)*
10) *Increased MHC Class II presentation provides a means for T-cell recognition. Potential MHC Class II antigen sources are:*
    a. *Microbial origin (species demonstrated in point 5)*
    b. *Tumour origin (not shown)*

Created with BioRender.com



Importantly, this theory is substantially testable. Central to this theory is the barrier disruption which acts as an initiating factor. H&E stains for a large cohort of CRC patients could be used to analyse barrier integrity, either directly by training pathologists, or alternatively by training AI algorithms detecting barrier breakdown. In the latter approach, by utilising annotated regions for crypt-like structures, we could measure the divergence from trained regions by utilising scores such as inbuilt probability scores when classifying objects (Visiopharm). Importantly, however, without the use of 'healthy' controls, it would be difficult to distinguish cause and effect. Intestinal inflammation almost undoubtedly produces severe barrier disruption, and therefore, differences are unlikely to be visualised between MSS-hi-CIRC and MSI-H. In contrast, the use of 'healthy' samples may indicate these differences, albeit subtly. Should an increased barrier permeability be an initiating factor in MSS-hi-CIRC, we may expect an increase in barrier breakdown along the length of the colon, including in healthy regions. Whilst this may require high sample numbers to distinguish a difference, there are other factors which could lead to increased immune activation in the absence of complete barrier disruption. For example, a reduced mucus layer or a reduced distance between microbial commensals and the underlying epithelium would also increase immune activation and may provide surrogate readouts for increased commensal-activation of the immune system. The comparison of normal samples from donors would provide insight into whether a general increase in bacterial translocation is important, or whether it is the preferential bloom. Although investigated using the 16s rRNA RNA scope data, which had inherent technical issues associated with it, we show no differences in bacterial abundance across the sample type (tumour or surrounding normal). Furthermore, it is conceivable that the increased permeability is not enough to cause a widespread increase in the abundance of bacteria, but preferential blooms of certain species that are widespread. These questions could be answered by utilising larger cohorts and 16s rRNA RNA scope probes directed to the *Gammaproteobacteria* phylum.

Whilst I suggest links to an increased BMI and high fat diets as common initiating factors in establishing MSS-hi-CIRC immunobiology, the involvement of an either an increased BMI or a high fat diet are not absolute requirements of this theory. The mechanism suggested here is somewhat analogous to that observed in smoking-induced cancer. Although a history of smoking is associated with the development of lung cancer, an individual does not need to be an active smoker for this to be linked. Interestingly, tobacco-use is linked with tumours with an increased mutation burden, and by extension an increased immune infiltrate. Somewhat perversely, in the era of ICB, smoking-induced cancer has a better prognosis than sporadically developed lung cancer. A similar mechanism may be at play here, whereby a history of being overweight, or an increased consumption of high-fat foods cause chronic damage leading to CRC (or gastric cancer), but conversely may ultimately trigger relatively strong intratumoural immunity, which subsequently links with better prognosis.

Central to this theory is the microbial adjuvantisation of T-cell responses in the MSS-hi-CIRC setting. Whilst the CIRC signature can delineate an increase in immune temperature in certain patients, it is not a spatially resolved signature and therefore cannot discriminate specific areas of high immune infiltration, of which we would expect with bacterial adjuvantisation. I would expect that areas of tumour with a dense immune infiltrate would have an increase in pathways relating to bacteria-driven responses (i.e., inflammasome activation) relative to those areas with lower immune infiltrates. Our data using the GeoMx digital spatial profiling platform could elucidate such differences, however, unfortunately, due to time constraints and data troubleshooting issues, this experiment couldn't be analysed in time for thesis submission. In brief, FFPE sections from the tumours of four cases each from the MSS-lo-CIRC, MSS-hi-CIRC, and MSI-H patient groupings were selected for spatial transcriptomics analysis. These sections were subjected to region of interest (ROI) selection in which we selected representative areas demonstrating varying characteristics. Of interest, ROIs included: (i) epithelial areas which contained high immune infiltrates, as defined as

$CD3^{HIGH}CD45^{HIGH}$ and $PanCK^{HIGH}$, (ii) epithelial areas with low immune infiltrates, as defined by $CD3^{LOW}$ $CD45^{LOW}$ and $PanCK^{HIGH}$, as well as (iii) stromal regions with differing immune infiltrates ($PanCK^{LOW}$). These areas were subjected to the Whole Transcriptome Atlas panel; as expected, high immune areas were notably rarer in MSS-lo-CIRC patient samples. These analyses, considering my hypothesis, could reveal pathways/transcripts that are upregulated in high immune epithelial areas compared with low immune epithelial areas within MSS-hi-CIRC patients. These could delineate potential pathways key to microbial-adjuvantisation of T-cell responses or, alternatively, those that prevent T-cell action.

7.1.2 Future directions

Overall, this work has helped define factors that fuel an increased intratumoural infiltrate in MSS CRC. I suggest, albeit somewhat speculatively, a mechanism in which diet and possibly genetic factors could combine to influence barrier permeability and provide a niche for both tumour development, and adjuvantisation of intratumoural immunity. Importantly, I highlight key areas which should be prioritised for future experiments. The samples acquired by ANICCA collectively represent a cohort with a wealth of data attached. These comprise patients whose CRC has recurred after surgery, and assuming primary samples are accessible, would allow the answer for one of the most important questions to this work – why are MSS-hi-CIRC tumours not represented in recurrent CRC? Should the primary samples of ANICCA patients all demonstrate little-to-no MHC Class II expression then it is highly likely that the immunity in MSS-hi-CIRC is effective at containing the tumour, such that surgery is effectively curative. This has major implications for subsequent findings demonstrated in Chapter 5, perhaps opening therapeutic interventions in MSS-lo-CIRC patients that aim to mimic this natural bacterial adjuntivisation of anti-tumour immunity. Secondly, the collection of fresh resection samples should arguably be prioritised with the aim of determining the antigen-specificity of T-cells. Should these analyses reveal a basis for microbial adjuvantisation then therapeutic interventions analogous

to the BCG vaccine, or those utilising FMT manipulation could be beneficial for MSS-lo-CIRC patients. A pipeline which delivered fresh resection tissue could also be modified to ensure Carnoy fixing of samples, allowing the investigation of microbial distribution in mucus layers. Finally, and perhaps most easily, the abundance of *Gammaproteobacteria* in MSS-hi-CIRC samples should be quantified either by qPCR for sheer abundance or via designing a 16s rRNA RNA scope probe to specifically target *Gammaproteobacteria* for use in either singleplex or multiplex IHC.

Chapter 8: Collaborative Statement

All work in thesis was my own, other than:

IHC staining of MHC Class II on the Genomics England cohort and RNA sequencing data (Toritseju Silo)

Vectra 3.0 data from Ghaleb Goussous (originally performed by the BTA and the HBRC)

Alignment of Genomics England whole genome sequencing data to genomes of bacteria, viruses, and archaea (Boris Noyvert)

16s rRNA RNA scope staining (BTA and the HBRC)

Chapter 9: Reference List

AGARWAL, D., NOWAK, C., ZHANG, N. R., PUSZTAI, L. & HATZIS, C. 2017. Functional germline variants as potential co-oncogenes. *npj Breast Cancer,* 3.

AHMAD, R., RAH, B., BASTOLA, D., DHAWAN, P. & SINGH, A. B. 2017. Obesity-induces Organ and Tissue Specific Tight Junction Restructuring and Barrier Deregulation by Claudin Switching. *Scientific Reports,* 7.

AL NABHANI, Z., DIETRICH, G., HUGOT, J.-P. & BARREAU, F. 2017. Nod2: The intestinal gate keeper. *PLOS Pathogens,* 13**,** e1006177.

ALBACKER, L. A., WU, J., SMITH, P., WARMUTH, M., STEPHENS, P. J., ZHU, P., YU, L. & CHMIELECKI, J. 2017. Loss of function JAK1 mutations occur at high frequency in cancers with microsatellite instability and are suggestive of immune evasion. *PLOS ONE,* 12**,** e0176181.

ALEXANDROV, L. B., JONES, P. H., WEDGE, D. C., SALE, J. E., CAMPBELL, P. J., NIK-ZAINAL, S. & STRATTON, M. R. 2015. Clock-like mutational processes in human somatic cells. *Nat Genet,* 47**,** 1402-7.

ALEXANDROV, L. B., KIM, J., HARADHVALA, N. J., HUANG, M. N., TIAN NG, A. W., WU, Y., BOOT, A., COVINGTON, K. R., GORDENIN, D. A., BERGSTROM, E. N., ISLAM, S. M. A., LOPEZ-BIGAS, N., KLIMCZAK, L. J., MCPHERSON, J. R., MORGANELLA, S., SABARINATHAN, R., WHEELER, D. A., MUSTONEN, V., GETZ, G., ROZEN, S. G. & STRATTON, M. R. 2020. The repertoire of mutational signatures in human cancer. *Nature,* 578**,** 94-101.

ALEXANDROV, L. B., NIK-ZAINAL, S., WEDGE, D. C., APARICIO, S. A., BEHJATI, S., BIANKIN, A. V., BIGNELL, G. R., BOLLI, N., BORG, A., BORRESEN-DALE, A. L., BOYAULT, S., BURKHARDT, B., BUTLER, A. P., CALDAS, C., DAVIES, H. R., DESMEDT, C., EILS, R., EYFJORD, J. E., FOEKENS, J. A., GREAVES, M., HOSODA, F., HUTTER, B., ILICIC, T., IMBEAUD, S., IMIELINSKI, M., JAGER, N., JONES, D. T., JONES, D., KNAPPSKOG, S., KOOL, M., LAKHANI, S. R., LOPEZ-OTIN, C., MARTIN, S., MUNSHI, N. C., NAKAMURA, H., NORTHCOTT, P. A., PAJIC, M., PAPAEMMANUIL, E., PARADISO, A., PEARSON, J. V., PUENTE, X. S., RAINE, K., RAMAKRISHNA, M., RICHARDSON, A. L., RICHTER, J., ROSENSTIEL, P., SCHLESNER, M., SCHUMACHER, T. N., SPAN, P. N., TEAGUE, J. W., TOTOKI, Y., TUTT, A. N., VALDES-MAS, R., VAN BUUREN, M. M., VAN 'T VEER, L., VINCENT-SALOMON, A., WADDELL, N., YATES, L. R., AUSTRALIAN PANCREATIC CANCER GENOME, I., CONSORTIUM, I. B. C., CONSORTIUM, I. M.-S., PEDBRAIN, I., ZUCMAN-ROSSI, J., FUTREAL, P. A., MCDERMOTT, U., LICHTER, P., MEYERSON, M., GRIMMOND, S. M., SIEBERT, R., CAMPO, E., SHIBATA, T., PFISTER, S. M., CAMPBELL, P. J. & STRATTON, M. R. 2013. Signatures of mutational processes in human cancer. *Nature,* 500**,** 415-21.

ALLAIRE, J. J., GANDRUD, C., RUSSELL, K. & YETMAN, C. 2017. networkD3: D3 JavaScript Network Graphs from R.

ALSAAB, H. O., SAU, S., ALZHRANI, R., TATIPARTI, K., BHISE, K., KASHAW, S. K. & IYER, A. K. 2017. PD-1 and PD-L1 Checkpoint Signaling Inhibition for Cancer Immunotherapy: Mechanism, Combinations, and Clinical Outcome. *Front Pharmacol,* 8**,** 561.

AMEDEI, A., BERGMAN, M. P., APPELMELK, B. J., AZZURRI, A., BENAGIANO, M., TAMBURINI, C., VAN DER ZEE, R., TELFORD, J. L., VANDENBROUCKE-GRAULS, C. M. J. E., D'ELIOS, M. M. & DEL PRETE, G. 2003. Molecular Mimicry between Helicobacter pylori Antigens and H+,K+–Adenosine Triphosphatase in Human Gastric Autoimmunity. *Journal of Experimental Medicine,* 198**,** 1147-1156.

AMEDEI, A., DELLA BELLA, C., SILVESTRI, E., PRISCO, D. & D'ELIOS, M. M. 2012. T Cells in Gastric Cancer: Friends or Foes. *Clinical and Developmental Immunology,* 2012**,** 1-10.

ANDOR, N., GRAHAM, T. A., JANSEN, M., XIA, L. C., AKTIPIS, C. A., PETRITSCH, C., JI, H. P. & MALEY, C. C. 2016. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine,* 22**,** 105-113.

ANDOR, N., HARNESS, J. V., MÜLLER, S., MEWES, H. W. & PETRITSCH, C. 2014. EXPANDS: expanding ploidy and allele frequency on nested subpopulations. *Bioinformatics,* 30**,** 50-60.

APARICIO, T., KOTELEVETS, L., TSOCAS, A., LAIGNEAU, J.-P., SOBHANI, I., CHASTRE, E. & LEHY, T. 2005. Leptin stimulates the proliferation of human colon cancer cells in vitro but does not promote the growth of colon cancer xenografts in nude mice or intestinal tumorigenesis in ApcMin/+ mice. *Gut,* 54**,** 1136-1145.

ARCE VARGAS, F., FURNESS, A. J. S., LITCHFIELD, K., JOSHI, K., ROSENTHAL, R., GHORANI, E., SOLOMON, I., LESKO, M. H., RUEF, N., RODDIE, C., HENRY, J. Y., SPAIN, L., BEN AISSA, A., GEORGIOU, A., WONG, Y. N. S., SMITH, M., STRAUSS, D., HAYES, A., NICOL, D., O'BRIEN, T., MARTENSSON, L., LJUNGARS, A., TEIGE, I., FRENDEUS, B., MELANOMA, T. R., RENAL, T. R., CONSORTIA, T. R. L., PULE, M., MARAFIOTI, T., GORE, M., LARKIN, J., TURAJLIC, S., SWANTON, C., PEGGS, K. S. & QUEZADA, S. A. 2018. Fc Effector Function Contributes to the Activity of Human Anti-CTLA-4 Antibodies. *Cancer Cell,* 33**,** 649-663 e4.

ARUMUGAM, M., RAES, J., PELLETIER, E., LE PASLIER, D., YAMADA, T., MENDE, D. R., FERNANDES, G. R., TAP, J., BRULS, T., BATTO, J.-M., BERTALAN, M., BORRUEL, N., CASELLAS, F., FERNANDEZ, L., GAUTIER, L., HANSEN, T., HATTORI, M., HAYASHI, T., KLEEREBEZEM, M., KUROKAWA, K., LECLERC, M., LEVENEZ, F., MANICHANH, C., NIELSEN, H. B., NIELSEN, T., PONS, N., POULAIN, J., QIN, J., SICHERITZ-PONTEN, T., TIMS, S., TORRENTS, D., UGARTE, E., ZOETENDAL, E. G., WANG, J., GUARNER, F., PEDERSEN, O., DE VOS, W. M., BRUNAK, S., DORÉ, J., WEISSENBACH, J., EHRLICH, S. D. & BORK, P. 2011. Enterotypes of the human gut microbiome. *Nature,* 473**,** 174-180.

AVIELLO, G. & KNAUS, U. G. 2018. NADPH oxidases and ROS signaling in the gastrointestinal tract. *Mucosal Immunology,* 11**,** 1011-1023.

BABA, Y., IWATSUKI, M., YOSHIDA, N., WATANABE, M. & BABA, H. 2017. Review of the gut microbiome and esophageal cancer: Pathogenesis and potential clinical implications. *Annals of Gastroenterological Surgery,* 1**,** 99-104.

BENNER, M., FERWERDA, G., JOOSTEN, I. & VAN DER MOLEN, R. G. 2018. How uterine microbiota might be responsible for a receptive, fertile endometrium. *Human Reproduction Update,* 24**,** 393-415.

BIANCONI, E., PIOVESAN, A., FACCHIN, F., BERAUDI, A., CASADEI, R., FRABETTI, F., VITALE, L., PELLERI, M. C., TASSANI, S., PIVA, F., PEREZ-AMODIO, S., STRIPPOLI, P. & CANAIDER, S. 2013. An estimation of the number of cells in the human body. *Ann Hum Biol,* 40**,** 463-71.

BINDEA, G., MLECNIK, B., TOSOLINI, M., KIRILOVSKY, A., WALDNER, M., OBENAUF, A. C., ANGELL, H., FREDRIKSEN, T., LAFONTAINE, L., BERGER, A., BRUNEVAL, P., FRIDMAN, W. H., BECKER, C., PAGES, F., SPEICHER, M. R., TRAJANOSKI, Z. & GALON, J. 2013. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity,* 39**,** 782-95.

BLASS, E. & OTT, P. A. 2021. Advances in the development of personalized neoantigen-based therapeutic cancer vaccines. *Nature Reviews Clinical Oncology,* 18**,** 215-229.

BLUM, J. S., WEARSCH, P. A. & CRESSWELL, P. 2013. Pathways of antigen processing. *Annu Rev Immunol,* 31**,** 443-73.

BOESCH, M., BATY, F., ROTHSCHILD, S. I., TAMM, M., JOERGER, M., FRÜH, M. & BRUTSCHE, M. H. 2021. Tumour neoantigen mimicry by microbial species in cancer immunotherapy. *British Journal of Cancer,* 125**,** 313-323.

BOLAND, C. R. & GOEL, A. 2010. Microsatellite instability in colorectal cancer. *Gastroenterology,* 138**,** 2073-2087 e3.

BORGHAEI, H., PAZ-ARES, L., HORN, L., SPIGEL, D. R., STEINS, M., READY, N. E., CHOW, L. Q., VOKES, E. E., FELIP, E., HOLGADO, E., BARLESI, F., KOHLHAUFL, M., ARRIETA, O., BURGIO, M. A., FAYETTE, J., LENA, H., PODDUBSKAYA, E., GERBER, D. E., GETTINGER, S. N., RUDIN, C. M., RIZVI, N., CRINO, L., BLUMENSCHEIN, G. R., JR., ANTONIA, S. J., DORANGE, C., HARBISON, C. T., GRAF FINCKENSTEIN, F. & BRAHMER, J. R. 2015. Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer. *N Engl J Med,* 373**,** 1627-39.

BRAY, F., FERLAY, J., SOERJOMATARAM, I., SIEGEL, R. L., TORRE, L. A. & JEMAL, A. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians,* 68**,** 394-424.

BROWN, S. D., WARREN, R. L., GIBB, E. A., MARTIN, S. D., SPINELLI, J. J., NELSON, B. H. & HOLT, R. A. 2014. Neo-antigens predicted by tumor genome meta-analysis correlate with increased patient survival. *Genome Research,* 24**,** 743-750.

BUCKOWITZ, A., KNAEBEL, H. P., BENNER, A., BLAKER, H., GEBERT, J., KIENLE, P., VON KNEBEL DOEBERITZ, M. & KLOOR, M. 2005. Microsatellite instability in colorectal cancer is associated with local lymphocyte infiltration and low frequency of distant metastases. *Br J Cancer,* 92**,** 1746-53.

BUHARD, O., SURAWEERA, N., LECTARD, A., DUVAL, A. & HAMELIN, R. 2004. Quasimonomorphic Mononucleotide Repeats for High-Level Microsatellite Instability Analysis. *Disease Markers,* 20**,** 251-257.

BUNKER, J. J. & BENDELAC, A. 2018. IgA Responses to Microbiota. *Immunity,* 49**,** 211-224.

CAI, C. W., BLASE, J. R., ZHANG, X., EICKHOFF, C. S. & HOFT, D. F. 2016. Th17 Cells Are More Protective Than Th1 Cells Against the Intracellular Parasite Trypanosoma cruzi. *PLOS Pathogens,* 12**,** e1005902.

CANCER RESEARCH UK, C. 2019. *Uterine Cancer Survival Statistics* [Online]. CRUK. Available: https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/uterine-cancer/survival#ref- [Accessed 16/11/2021 2021].

CANCER.NET EDITORIAL BOARD, N. 2012. *Colorectal Cancer – Treatment Options* [Online]. American Society of Clinical Oncology (ASCO). Available: https://www.cancer.net/cancer-types/colorectal-cancer/treatment-options [Accessed 17/04/2018 2018].

CANI, P. D., POSSEMIERS, S., VAN DE WIELE, T., GUIOT, Y., EVERARD, A., ROTTIER, O., GEURTS, L., NASLAIN, D., NEYRINCK, A., LAMBERT, D. M., MUCCIOLI, G. G. & DELZENNE, N. M. 2009. Changes in gut microbiota control inflammation in obese mice through a mechanism involving GLP-2-driven improvement of gut permeability. *Gut,* 58**,** 1091-1103.

CAPORASO, J. G., LAUBER, C. L., COSTELLO, E. K., BERG-LYONS, D., GONZALEZ, A., STOMBAUGH, J., KNIGHTS, D., GAJER, P., RAVEL, J., FIERER, N., GORDON, J. I. & KNIGHT, R. 2011. Moving pictures of the human microbiome. *Genome Biology,* 12**,** R50.

CARLSON, M. 2016. hgu133plus2.db: Affymetrix Human Genome U133 Plus 2.0 Array annotation data (chip
hgu133plus2). R package version 3.2.3 ed.

CARLSON, M. 2020. org.Hs.eg.db: Genome wide annotation for Human. R package version 3.11.4 ed.

CASTRO, G., LIU, X., NGO, K., DE LEON-TABALDO, A., ZHAO, S., LUNA-ROMAN, R., YU, J., CAO, T., KUHN, R., WILKINSON, P., HERMAN, K., NELEN, M. I., BLEVITT, J., XUE, X., FOURIE, A. & FUNG-LEUNG, W. P. 2017. RORgammat and RORalpha signature genes in human Th17 cells. *PLoS One,* 12**,** e0181868.

CERMAKOVA, P., MELICHAR, B., TOMSOVA, M., ZOUL, Z., KALABOVA, H., SPACEK, J. & DOLEZEL, M. 2014. Prognostic Significance of CD3+ Tumor-Infiltraing Lymphocytes in Patients with Endometrial Carcinoma. *Anticancer Research,* 34**,** 5555-5561.

CHAE, Y. K., ANKER, J. F., BAIS, P., NAMBURI, S., GILES, F. J. & CHUANG, J. H. 2018. Mutations in DNA repair genes are associated with increased neo-antigen load and activated T cell infiltration in lung adenocarcinoma. *Oncotarget,* 9**,** 7949-7960.

CHAMBERLAIN, N. R. 2019. *Differential Staining of Bacteria; Knowing Your Gram Stain Reactions Using Three Simple Rules* [Online]. University of Health Sciences/Kirksville College of Osteopathic Medicine. Available: https://www.atsu.edu/faculty/chamberlain/mosdoh/gramstainingrules.htm [Accessed 26/10/2021 2021].

CHARALAMPAKIS, N., ECONOMOPOULOU, P., KOTSANTIS, I., TOLIA, M., SCHIZAS, D., LIAKAKOS, T., ELIMOVA, E., AJANI, J. A. & PSYRRI, A. 2018. Medical management of gastric cancer: a 2017 update. *Cancer Medicine,* 7**,** 123-133.

CHEN, H. 2015. Rphenograph: R implementation of the phenograph algorithm. version: 0.99.1. ed.

CHEN, Y., LIU, L., WANG, X., WANG, J., YAN, Z., CHENG, J., GONG, G. & LI, G. 2013. Body Mass Index and Risk of Gastric Cancer: A Meta-analysis of a Population with More Than Ten Million from 24 Prospective Studies. *Cancer Epidemiology Biomarkers & Prevention,* 22**,** 1395-1408.

CHEROUTRE, H., LAMBOLEZ, F. & MUCIDA, D. 2011. The light and dark sides of intestinal intraepithelial lymphocytes. *Nature Reviews Immunology,* 11**,** 445-456.

CLEMENTE, C. G., MIHM, M. C., BUFALINO, R., ZURRIDA, S., COLLINI, P. & CASCINELLI, N. 1996. Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma. *Cancer,* 77**,** 1303-1310.

CLINICALTRIALS.GOV 2019. A Phase II Trial Assessing Nivolumab in Class II Expressing Microsatellite Stable Colorectal Cancer. https://ClinicalTrials.gov/show/NCT03981146.

CLINICALTRIALS.GOV 2020. Study of Pembrolizumab (MK-3475) vs Standard Therapy in Participants With Microsatellite Instability-High (MSI-H) or Mismatch Repair Deficient (dMMR) Stage IV Colorectal Carcinoma (MK-3475-177/KEYNOTE-177). https://ClinicalTrials.gov/show/NCT02563002.

COBO, E. & CHADEE, K. 2013. Antimicrobial Human β-Defensins in the Colon and Their Role in Infectious and Non-Infectious Diseases. *Pathogens,* 2**,** 177-192.

COLAPRICO, A., SILVA, T. C., OLSEN, C., GAROFANO, L., CAVA, C., GAROLINI, D., SABEDOT, T. S., MALTA, T. M., PAGNOTTA, S. M., CASTIGLIONI, I., CECCARELLI, M., BONTEMPI, G. & NOUSHMEHR, H. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research,* 44**,** e71-e71.

COUTURE-BEIL, A. 2018. rjson: JSON for R. R package version 0.2.20 ed. CRAN.

CREMONESI, E., GOVERNA, V., GARZON, J. F. G., MELE, V., AMICARELLA, F., MURARO, M. G., TRELLA, E., GALATI-FOURNIER, V., OERTLI, D., DÄSTER, S. R., DROESER, R. A., WEIXLER, B., BOLLI, M., ROSSO, R., NITSCHE, U., KHANNA, N., EGLI, A., KECK, S., SLOTTA-HUSPENINA, J., TERRACCIANO, L. M., ZAJAC, P., SPAGNOLI, G. C., EPPENBERGER-CASTORI, S., JANSSEN, K.-P., BORSIG, L. & IEZZI, G. 2018. Gut microbiota modulate T cell trafficking into human colorectal cancer. *Gut,* 67**,** 1984-1994.

CRUK, C. 2015. *Bowel Cancer - Screening* [Online]. Available: http://about-cancer.cancerresearchuk.org/about-cancer/bowel-cancer/getting-diagnosed/screening [Accessed 22/05/2018 2018].

CURTIS, M. M. & WAY, S. S. 2009. Interleukin-17 in host defence against bacterial, mycobacterial and fungal pathogens. *Immunology,* 126**,** 177-85.

CUSICK, M. F., LIBBEY, J. E. & FUJINAMI, R. S. 2012. Molecular Mimicry as a Mechanism of Autoimmune Disease. *Clinical Reviews in Allergy & Immunology,* 42**,** 102-111.

DA SILVA, V., RAMOS, M., GROENEN, M., CROOIJMANS, R., JOHANSSON, A., REGITANO, L., COUTINHO, L., ZIMMER, R., WALDRON, L. & GEISTLINGER, L. 2019. CNVRanger: association analysis of CNVs with gene expression and quantitative phenotypes. *Bioinformatics*.

DALERBA, P., MACCALLI, C., CASATI, C., CASTELLI, C. & PARMIANI, G. 2003. Immunology and immunotherapy of colorectal cancer. *Critical Reviews in Oncology/Hematology,* 46**,** 33-57.

DAVIS, S. & MELTZER, P. S. 2007. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics,* 23**,** 1846-1847.

DAVIS-RICHARDSON, A. G. & TRIPLETT, E. W. 2015. A model for the role of gut bacteria in the development of autoimmunity for type 1 diabetes. *Diabetologia,* 58**,** 1386-1393.
DE FILIPPO, C., CAVALIERI, D., DI PAOLA, M., RAMAZZOTTI, M., POULLET, J. B., MASSART, S., COLLINI, S., PIERACCINI, G. & LIONETTI, P. 2010. Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. *Proceedings of the National Academy of Sciences,* 107**,** 14691-14696.
DE KIVIT, S., TOBIN, M. C., FORSYTH, C. B., KESHAVARZIAN, A. & LANDAY, A. L. 2014. Regulation of intestinal immune responses through TLR activation: implications for pro- and prebiotics. *Frontiers in Immunology,* 5.
DE SMEDT, L., LEMAHIEU, J., PALMANS, S., GOVAERE, O., TOUSSEYN, T., VAN CUTSEM, E., PRENEN, H., TEJPAR, S., SPAEPEN, M., MATTHIJS, G., DECAESTECKER, C., MOLES LOPEZ, X., DEMETTER, P., SALMON, I. & SAGAERT, X. 2015. Microsatellite instable vs stable colon carcinomas: analysis of tumour heterogeneity, inflammation and angiogenesis. *British Journal of Cancer,* 113**,** 500-509.
DE SOUSA E MELO, F., KURTOVA, A. V., HARNOSS, J. M., KLJAVIN, N., HOECK, J. D., HUNG, J., ANDERSON, J. E., STORM, E. E., MODRUSAN, Z., KOEPPEN, H., DIJKGRAAF, G. J., PISKOL, R. & DE SAUVAGE, F. J. 2017. A distinct role for Lgr5(+) stem cells in primary and metastatic colon cancer. *Nature,* 543**,** 676-680.
DE VRIES, N. L., VAN UNEN, V., IJSSELSTEIJN, M. E., ABDELAAL, T., VAN DER BREGGEN, R., FARINA SARASQUETA, A., MAHFOUZ, A., PEETERS, K. C. M. J., HÖLLT, T., LELIEVELDT, B. P. F., KONING, F. & DE MIRANDA, N. F. C. C. 2020. High-dimensional cytometric analysis of colorectal cancer reveals novel mediators of antitumour immunity. *Gut,* 69**,** 691-703.
DENOIX, P. F. 1946. Enquete permanent dans les centre anticancereaux. *Nat. Hyg,* 1**,** 70-75.
DESAI, M. S., SEEKATZ, A. M., KOROPATKIN, N. M., KAMADA, N., HICKEY, C. A., WOLTER, M., PUDLO, N. A., KITAMOTO, S., TERRAPON, N., MULLER, A., YOUNG, V. B., HENRISSAT, B., WILMES, P., STAPPENBECK, T. S., NÚÑEZ, G. & MARTENS, E. C. 2016. A Dietary Fiber-Deprived Gut Microbiota Degrades the Colonic Mucus Barrier and Enhances Pathogen Susceptibility. *Cell,* 167**,** 1339-1353.e21.
DING, T. & SCHLOSS, P. D. 2014. Dynamics and associations of microbial community types across the human body. *Nature,* 509**,** 357-360.
DONNEM, T., KILVAER, T. K., ANDERSEN, S., RICHARDSEN, E., PAULSEN, E. E., HALD, S. M., AL-SAAD, S., BRUSTUGUN, O. T., HELLAND, A., LUND-IVERSEN, M., SOLBERG, S., GRONBERG, B. H., WAHL, S. G., HELGELAND, L., FLOTTEN, O., POHL, M., AL-SHIBLI, K., SANDANGER, T. M., PEZZELLA, F., BUSUND, L. T. & BREMNES, R. M. 2016. Strategies for clinical implementation of TNM-Immunoscore in resected nonsmall-cell lung cancer. *Ann Oncol,* 27**,** 225-32.
DONOHOE, D., GARGE, N., ZHANG, X., SUN, W., O'CONNELL, T., BUNGER, M. & BULTMAN, S. 2011. The Microbiome and Butyrate Regulate Energy Metabolism and Autophagy in the Mammalian Colon. *Cell Metabolism,* 13**,** 517-526.
DRAKE, I. M., MAPSTONE, N. P., SCHORAH, C. J., WHITE, K. L. M., CHALMERS, D. M., DIXON, M. F. & AXON, A. T. R. 1998. Reactive oxygen species activity and lipid peroxidation inHelicobacter pylori associated gastritis: relation to gastric mucosal ascorbic acid concentrations and effect of H pylori eradication. *Gut,* 42**,** 768-771.
DUNN, G. P., OLD, L. J. & SCHREIBER, R. D. 2004. The Three Es of Cancer Immunoediting. *Annual Review of Immunology,* 22**,** 329-360.
DURINCK, S., MOREAU, Y., KASPRZYK, A., DAVIS, S., DE MOOR, B., BRAZMA, A. & HUBER, W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics,* 21**,** 3439-3440.
DURINCK, S., SPELLMAN, P. T., BIRNEY, E. & HUBER, W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nature Protocols,* 4**,** 1184-1191.

DZUTSEV, A., HOGG, A., SUI, Y., SOLAYMANI-MOHAMMADI, S., YU, H., FREY, B., WANG, Y. & BERZOFSKY, J. A. 2017. Differential T cell homing to colon vs. small intestine is imprinted by local CD11c + APCs that determine homing receptors. *Journal of Leukocyte Biology,* 102**,** 1381-1388.

EADEN, J. A. 2001. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut,* 48**,** 526-535.

EIDE, P. W., BRUUN, J., LOTHE, R. A. & SVEEN, A. 2017. CMScaller: an R package for consensus molecular subtyping of colorectal cancer pre-clinical models. *Scientific Reports,* 7.

ELLULU, M. S., PATIMAH, I., KHAZA'AI, H., RAHMAT, A. & ABED, Y. 2017. Obesity and inflammation: the linking mechanism and the complications. *Archives of Medical Science,* 4**,** 851-863.

FELLNER, C. 2017. Promising Drugs in Clinical Development to Treat Advanced Colorectal Cancer. *Pharmacy and Therapeutics,* 42**,** 262-265.

FIGUEROA-PROTTI, L., SOTO-MOLINARI, R., CALDERÓN-OSORNO, M., MORA, J. & ALPÍZAR-ALPÍZAR, W. 2019. Gastric Cancer in the Era of Immune Checkpoint Blockade. *Journal of Oncology,* 2019**,** 1-11.

FLAQUER, A., BAUMBACH, C., KRIEBEL, J., MEITINGER, T., PETERES, A., WALDENBERGER, M., GRALLERT, H. & STRAUCH, K. 2014. Mitochondrial Genetic Variants Identified to Be Associated with BMI in Adults. *PLoS One,* 9**,** e105116.

FLUCKIGER, A., DAILLÈRE, R., SASSI, M., SIXT, B. S., LIU, P., LOOS, F., RICHARD, C., RABU, C., ALOU, M. T., GOUBET, A.-G., LEMAITRE, F., FERRERE, G., DEROSA, L., DUONG, C. P. M., MESSAOUDENE, M., GAGNÉ, A., JOUBERT, P., DE SORDI, L., DEBARBIEUX, L., SIMON, S., SCARLATA, C.-M., AYYOUB, M., PALERMO, B., FACCIOLO, F., BOIDOT, R., WHEELER, R., BONECA, I. G., SZTUPINSZKI, Z., PAPP, K., CSABAI, I., PASOLLI, E., SEGATA, N., LOPEZ-OTIN, C., SZALLASI, Z., ANDRE, F., IEBBA, V., QUINIOU, V., KLATZMANN, D., BOUKHALIL, J., KHELAIFIA, S., RAOULT, D., ALBIGES, L., ESCUDIER, B., EGGERMONT, A., MAMI-CHOUAIB, F., NISTICO, P., GHIRINGHELLI, F., ROUTY, B., LABARRIÈRE, N., CATTOIR, V., KROEMER, G. & ZITVOGEL, L. 2020. Cross-reactivity between tumor MHC class I–restricted antigens and an enterococcal bacteriophage. *Science,* 369**,** 936-942.

FONTANA, L., EAGON, J. C., TRUJILLO, M. E., SCHERER, P. E. & KLEIN, S. 2007. Visceral Fat Adipokine Secretion Is Associated With Systemic Inflammation in Obese Humans. *Diabetes,* 56**,** 1010-1013.

GALON, J., COSTES, A., SANCHEZ-CABO, F., KIRILOVSKY, A., MLECNIK, B., LAGORCE-PAGÉS, C., TOSOLINI, M., CAMUS, M., BERGER, A., WIND, P., ZINZINDOHOUÉ, F., BRUNEVAL, P., CUGNENC, P. H., TRAJANOSKI, Z., FRIDMAN, W. H. & PAGÉS, F. 2006. Type, Density, and Location of Immune Cells Within Human Colorectal Tumors Predict Clinical Outcome. *Science,* 313**,** 1960-1964.

GALON, J., MLECNIK, B., BINDEA, G., ANGELL, H. K., BERGER, A., LAGORCE, C., LUGLI, A., ZLOBEC, I., HARTMANN, A., BIFULCO, C., NAGTEGAAL, I. D., PALMQVIST, R., MASUCCI, G. V., BOTTI, G., TATANGELO, F., DELRIO, P., MAIO, M., LAGHI, L., GRIZZI, F., ASSLABER, M., D'ARRIGO, C., VIDAL-VANACLOCHA, F., ZAVADOVA, E., CHOUCHANE, L., OHASHI, P. S., HAFEZI-BAKHTIARI, S., WOUTERS, B. G., ROEHRL, M., NGUYEN, L., KAWAKAMI, Y., HAZAMA, S., OKUNO, K., OGINO, S., GIBBS, P., WARING, P., SATO, N., TORIGOE, T., ITOH, K., PATEL, P. S., SHUKLA, S. N., WANG, Y., KOPETZ, S., SINICROPE, F. A., SCRIPCARIU, V., ASCIERTO, P. A., MARINCOLA, F. M., FOX, B. A. & PAGÈS, F. 2014. Towards the introduction of the 'Immunoscore' in the classification of malignant tumours. *The Journal of Pathology,* 232**,** 199-209.

GALON, J., PAGÈS, F., MARINCOLA, F. M., ANGELL, H. K., THURIN, M., LUGLI, A., ZLOBEC, I., BERGER, A., BIFULCO, C., BOTTI, G., TATANGELO, F., BRITTEN, C. M., KREITER, S., CHOUCHANE, L., DELRIO, P., ARNDT, H., ASSLABER, M., MAIO, M., MASUCCI, G. V., MIHM, M., VIDAL-VANACLOCHA, F., ALLISON, J. P., GNJATIC, S., HAKANSSON, L., HUBER, C., SINGH-JASUJA, H., OTTENSMEIER, C., ZWIERZINA, H., LAGHI, L., GRIZZI, F., OHASHI, P. S., SHAW, P. A., CLARKE, B. A., WOUTERS, B. G., KAWAKAMI, Y., HAZAMA, S., OKUNO, K., WANG, E., O'DONNELL-

TORMEY, J., LAGORCE, C., PAWELEC, G., NISHIMURA, M. I., HAWKINS, R., LAPOINTE, R., LUNDQVIST, A., KHLEIF, S. N., OGINO, S., GIBBS, P., WARING, P., SATO, N., TORIGOE, T., ITOH, K., PATEL, P. S., SHUKLA, S. N., PALMQVIST, R., NAGTEGAAL, I. D., WANG, Y., D'ARRIGO, C., KOPETZ, S., SINICROPE, F. A., TRINCHIERI, G., GAJEWSKI, T. F., ASCIERTO, P. A. & FOX, B. A. 2012. Cancer classification using the Immunoscore: a worldwide task force. *Journal of Translational Medicine,* 10**,** 205.

GARCIA, G., GIORGI, F. 2021. svpluscnv: Integrative analyses of CNV segmentation and SV variant calls.

GAUTIER, L., COPE, L., BOLSTAD, B. M. & IRIZARRY, R. A. 2004. affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics,* 20**,** 307-315.

GEHRING, J., FISCHER, B., LAWRENCE, M. & HUBER, W. 2015. SomaticSignatures: Inferring Mutational Signatures from Single Nucleotide Variants. btv408 ed. Bioinformatics.

GEL, B. & SERRA, E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics,* 33**,** 3088-3090.

GERLINGER, M., ROWAN, A. J., HORSWELL, S., LARKIN, J., ENDESFELDER, D., GRONROOS, E., MARTINEZ, P., MATTHEWS, N., STEWART, A., TARPEY, P., VARELA, I., PHILLIMORE, B., BEGUM, S., MCDONALD, N. Q., BUTLER, A., JONES, D., RAINE, K., LATIMER, C., SANTOS, C. R., NOHADANI, M., EKLUND, A. C., SPENCER-DENE, B., CLARK, G., PICKERING, L., STAMP, G., GORE, M., SZALLASI, Z., DOWNWARD, J., FUTREAL, P. A. & SWANTON, C. 2012. Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *New England Journal of Medicine,* 366**,** 883-892.

GHORESCHI, K., LAURENCE, A., YANG, X.-P., TATO, C. M., MCGEACHY, M. J., KONKEL, J. E., RAMOS, H. L., WEI, L., DAVIDSON, T. S., BOULADOUX, N., GRAINGER, J. R., CHEN, Q., KANNO, Y., WATFORD, W. T., SUN, H.-W., EBERL, G., SHEVACH, E. M., BELKAID, Y., CUA, D. J., CHEN, W. & O'SHEA, J. J. 2010. Generation of pathogenic TH17 cells in the absence of TGF-β signalling. *Nature,* 467**,** 967-971.

GIANNAKIS, M., XINMENG, SACHET, ZHI, COHEN, O., NISHIHARA, R., BAHL, S., CAO, Y., AMIN-MANSOUR, A., YAMAUCHI, M., SUKAWA, Y., STEWART, C., ROSENBERG, M., MIMA, K., INAMURA, K., NOSHO, K., JONATHAN, MICHAEL, EDWARD, ANDREW, NG, K., JEFFREY, ELIEZER, GETZ, G., STACEY, ERIC, CATHERINE, CHARLES, OGINO, S. & LEVI 2016. Genomic Correlates of Immune-Cell Infiltrates in Colorectal Carcinoma. *Cell Reports,* 15**,** 857-865.

GOPALAKRISHNAN, V., SPENCER, C. N., NEZI, L., REUBEN, A., ANDREWS, M. C., KARPINETS, T. V., PRIETO, P. A., VICENTE, D., HOFFMAN, K., WEI, S. C., COGDILL, A. P., ZHAO, L., HUDGENS, C. W., HUTCHINSON, D. S., MANZO, T., PETACCIA DE MACEDO, M., COTECHINI, T., KUMAR, T., CHEN, W. S., REDDY, S. M., SZCZEPANIAK SLOANE, R., GALLOWAY-PENA, J., JIANG, H., CHEN, P. L., SHPALL, E. J., REZVANI, K., ALOUSI, A. M., CHEMALY, R. F., SHELBURNE, S., VENCE, L. M., OKHUYSEN, P. C., JENSEN, V. B., SWENNES, A. G., MCALLISTER, F., MARCELO RIQUELME SANCHEZ, E., ZHANG, Y., LE CHATELIER, E., ZITVOGEL, L., PONS, N., AUSTIN-BRENEMAN, J. L., HAYDU, L. E., BURTON, E. M., GARDNER, J. M., SIRMANS, E., HU, J., LAZAR, A. J., TSUJIKAWA, T., DIAB, A., TAWBI, H., GLITZA, I. C., HWU, W. J., PATEL, S. P., WOODMAN, S. E., AMARIA, R. N., DAVIES, M. A., GERSHENWALD, J. E., HWU, P., LEE, J. E., ZHANG, J., COUSSENS, L. M., COOPER, Z. A., FUTREAL, P. A., DANIEL, C. R., AJAMI, N. J., PETROSINO, J. F., TETZLAFF, M. T., SHARMA, P., ALLISON, J. P., JENQ, R. R. & WARGO, J. A. 2018. Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science,* 359**,** 97-103.

GOUSSOUS, G. 2020. *Exploroing the Immune Microenvironment of Primary and Metastatic Colorectal Cancer.* Doctor of Philosophy PhD Thesis, University of Birmingham.

GRAVEL, J., PARADIS-BLEAU, C. & SCHMITZER, A. R. 2017. Adaptation of a bacterial membrane permeabilization assay for quantitative evaluation of benzalkonium chloride as a membrane-disrupting agent. *MedChemComm,* 8**,** 1408-1413.

GROSSMAN, R. L., HEATH, A. P., FERRETTI, V., VARMUS, H. E., LOWY, D. R., KIBBE, W. A. & STAUDT, L. M. 2016. Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine,* 375**,** 1109-1112.

GUINNEY, J., DIENSTMANN, R., WANG, X., DE REYNIÈS, A., SCHLICKER, A., SONESON, C., MARISA, L., ROEPMAN, P., NYAMUNDANDA, G., ANGELINO, P., BOT, B. M., MORRIS, J. S., SIMON, I. M., GERSTER, S., FESSLER, E., DE SOUSA E MELO, F., MISSIAGLIA, E., RAMAY, H., BARRAS, D., HOMICSKO, K., MARU, D., MANYAM, G. C., BROOM, B., BOIGE, V., PEREZ-VILLAMIL, B., LADERAS, T., SALAZAR, R., GRAY, J. W., HANAHAN, D., TABERNERO, J., BERNARDS, R., FRIEND, S. H., LAURENT-PUIG, P., MEDEMA, J. P., SADANANDAM, A., WESSELS, L., DELORENZI, M., KOPETZ, S., VERMEULEN, L. & TEJPAR, S. 2015. The consensus molecular subtypes of colorectal cancer. *Nature Medicine,* 21**,** 1350-1356.

GUSTAVSEN, J., RÜEGER, S., CHAMBERLAIN, S., USHEY, K. & ZHU, H. 2020. rsnps: Get 'SNP' ('Single-Nucleotide' 'Polymorphism') Data on the Web. R package version 0.4.0 ed. CRAN.

GUSTAVSEN, J. A., PAI, S., ISSERLIN, R., DEMCHAK, B. & PICO, A. R. 2019. RCy3: Network biology using Cytoscape from within R. *F1000Research,* 8**,** 1774.

HAMPEL, H., FRANKEL, W., PANESCU, J., LOCKMAN, J., SOTAMAA, K., FIX, D., COMERAS, I., LA JEUNESSE, J., NAKAGAWA, H., WESTMAN, J. A., PRIOR, T. W., CLENDENNING, M., PENZONE, P., LOMBARDI, J., DUNN, P., COHN, D. E., COPELAND, L., EATON, L., FOWLER, J., LEWANDOWSKI, G., VACCARELLO, L., BELL, J., REID, G. & DE LA CHAPELLE, A. 2006. Screening for Lynch Syndrome (Hereditary Nonpolyposis Colorectal Cancer) among Endometrial Cancer Patients. *Cancer Research,* 66**,** 7810-7817.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *Cell,* 144**,** 646-74.

HANDUNNETTHI, L., RAMAGOPALAN, S. V., EBERS, G. C. & KNIGHT, J. C. 2010. Regulation of major histocompatibility complex class II gene expression, genetic variation and disease. *Genes & Immunity,* 11**,** 99-112.

HANSEN, K. D., IRIZARRY, R. A. & WU, Z. 2012. Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics,* 13**,** 204-216.

HÄNZELMANN, S., CASTELO, R. & GUINNEY, J. 2013. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics,* 14**,** 7.

HARDWICK, J. C., VAN DEN BRINK, G. R., OFFERHAUS, G. J., VAN DEVENTER, S. J. & PEPPELENBOSCH, M. P. 2001. Leptin is a growth factor for colonic epithelial cells. *Gastroenterology,* 121**,** 79-90.

HASTIE, T., TIBSHIRANI, R., NARASIMHAN, B. & CHU, G. 2020. impute: Imputation for microarray data. R package version 1.62.0 ed.

HELLMANN, M. D., RIZVI, N. A., GOLDMAN, J. W., GETTINGER, S. N., BORGHAEI, H., BRAHMER, J. R., READY, N. E., GERBER, D. E., CHOW, L. Q., JUERGENS, R. A., SHEPHERD, F. A., LAURIE, S. A., GEESE, W. J., AGRAWAL, S., YOUNG, T. C., LI, X. & ANTONIA, S. J. 2017. Nivolumab plus ipilimumab as first-line treatment for advanced non-small-cell lung cancer (CheckMate 012): results of an open-label, phase 1, multicohort study. *The Lancet Oncology,* 18**,** 31-41.

HERNÁNDEZ-LUNA, M. A., LÓPEZ-BRIONES, S & LURIA-PÉREZ, R. 2019. The Four Horsemen in Colon Cancer. *Journal of Oncology,* 2019.

HODI, F. S., O'DAY, S. J., MCDERMOTT, D. F., WEBER, R. W., SOSMAN, J. A., HAANEN, J. B., GONZALEZ, R., ROBERT, C., SCHADENDORF, D., HASSEL, J. C., AKERLEY, W., VAN DEN EERTWEGH, A. J., LUTZKY, J., LORIGAN, P., VAUBEL, J. M., LINETTE, G. P., HOGG, D., OTTENSMEIER, C. H., LEBBE, C., PESCHEL, C., QUIRT, I., CLARK, J. I., WOLCHOK, J. D., WEBER, J. S., TIAN, J., YELLIN, M. J., NICHOL, G. M., HOOS, A. & URBA, W. J. 2010. Improved survival with ipilimumab in patients with metastatic melanoma. *N Engl J Med,* 363**,** 711-23.

HUGOT, J.-P., CHAMAILLARD, M., ZOUALI, H., LESAGE, S., CÉZARD, J.-P., BELAICHE, J., ALMER, S., TYSK, C., O'MORAIN, C. A., GASSULL, M., BINDER, V., FINKEL, Y., CORTOT, A., MODIGLIANI, R., LAURENT-PUIG, P., GOWER-ROUSSEAU, C., MACRY, J., COLOMBEL, J.-F., SAHBATOU, M. &

THOMAS, G. 2001. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature,* 411**,** 599-603.
HUMAN MICROBIOME PROJECT DATA PORTAL, H. 2021. Available: https://portal.hmpdacc.org/ [Accessed 27/01/21 2021].
ILLÁN-GÓMEZ, F., GONZÁLVEZ-ORTEGA, M., OREA-SOLER, I., ALCARAZ-TAFALLA, S., ARAGÓN-ALONSO, A., PASCUAL-DÍAZ, M., PÉREZ-PAREDES, M. & LOZANO-ALMELA, L. 2012. Obesity and Inflammation: Change in Adiponextin, C-Reactive Protein, Tumour Necrosis Factor-Alpha and Interleukin-6 After Bariatric Surgery. *Obesity Surgery,* 22**,** 950-955.
INGHAM-DEMPSTER, T., WALKER, D. C. & CORFE, B. M. 2017. An agent-based model of anoikis in the colon crypt displays novel emergent behaviour consistent with biological observations. *R Soc Open Sci,* 4**,** 160858.
ITO, M., KANNO, S., NOSHO, K., SUKAWA, Y., MITSUHASHI, K., KURIHARA, H., IGARASHI, H., TAKAHASHI, T., TACHIBANA, M., TAKAHASHI, H., YOSHII, S., TAKENOUCHI, T., HASEGAWA, T., OKITA, K., HIRATA, K., MARUYAMA, R., SUZUKI, H., IMAI, K., YAMAMOTO, H. & SHINOMURA, Y. 2015. Association ofFusobacterium nucleatumwith clinical and molecular features in colorectal serrated pathway. *International Journal of Cancer,* 137**,** 1258-1268.
JAMES, T. C. & BOND, U. 2012. Molecular Mimics of the Tumour Antigen MUC1. *PLoS ONE,* 7**,** e49728.
JHUNJHUNWALA, S., HAMMER, C. & DELAMARRE, L. 2021. Antigen presentation in cancer: insights into tumour immunogenicity and immune evasion. *Nature Reviews Cancer,* 21**,** 298-312.
JIANG, W., LIU, K., GUO, Q., CHENG, J., SHEN, L., CAO, Y., WU, J., SHI, J., CAO, H., LIU, B., TAO, K., WANG, G. & CAI, K. 2017. Tumor-infiltrating immune cells and prognosis in gastric cancer: a systematic review and meta-analysis. *Oncotarget,* 8**,** 62312 - 62329.
JIMÉNEZ-SÁNCHEZ, A., CAST, O. & MILLER, M. L. 2019. Comprehensive Benchmarking and Integration of Tumor Microenvironment Cell Estimation Methods. *Cancer Research,* 79**,** 6238-6246.
JIN, Y., TAO, L., JIN, S. & CAI, W. 2021. Patterns of immune infiltration in gastric cancer and their clinical significance. *Japanese Journal of Clinical Oncology,* 51**,** 1067-1079.
JOHANSSON, M. E. V., PHILLIPSON, M., PETERSSON, J., VELCICH, A., HOLM, L. & HANSSON, G. C. 2008. The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences,* 105**,** 15064-15069.
JOHNSON, D. B., ESTRADA, M. V., SALGADO, R., SANCHEZ, V., DOXIE, D. B., OPALENIK, S. R., VILGELM, A. E., FELD, E., JOHNSON, A. S., GREENPLATE, A. R., SANDERS, M. E., LOVLY, C. M., FREDERICK, D. T., KELLEY, M. C., RICHMOND, A., IRISH, J. M., SHYR, Y., SULLIVAN, R. J., PUZANOV, I., SOSMAN, J. A. & BALKO, J. M. 2016. Melanoma-specific MHC-II expression represents a tumour-autonomous phenotype and predicts response to anti-PD-1/PD-L1 therapy. *Nat Commun,* 7**,** 10582.
KAIMALA, S., MOHAMED, Y. A., NADER, N., ISSAC, J., ELKORD, E., CHOUAIB, S., FERNANDEZ-CABEZUDO, M. J. & AL-RAMADI, B. K. 2014. Salmonella-mediated tumor regression involves targeting of tumor myeloid suppressor cells causing a shift to M1-like phenotype and reduction in suppressive capacity. *Cancer Immunology, Immunotherapy,* 63**,** 587-599.
KALAORA, S., NAGLER, A., NEJMAN, D., ALON, M., BARBOLIN, C., BARNEA, E., KETELAARS, S. L. C., CHENG, K., VERVIER, K., SHENTAL, N., BUSSI, Y., ROTKOPF, R., LEVY, R., BENEDEK, G., TRABISH, S., DADOSH, T., LEVIN-ZAIDMAN, S., GELLER, L. T., WANG, K., GREENBERG, P., YAGEL, G., PERI, A., FUKS, G., BHARDWAJ, N., REUBEN, A., HERMIDA, L., JOHNSON, S. B., GALLOWAY-PEÑA, J. R., SHROPSHIRE, W. C., BERNATCHEZ, C., HAYMAKER, C., ARORA, R., ROITMAN, L., EILAM, R., WEINBERGER, A., LOTAN-POMPAN, M., LOTEM, M., ADMON, A., LEVIN, Y., LAWLEY, T. D., ADAMS, D. J., LEVESQUE, M. P., BESSER, M. J., SCHACHTER, J., GOLANI, O., SEGAL, E., GEVA-ZATORSKY, N., RUPPIN, E., KVISTBORG, P., PETERSON, S. N., WARGO, J. A., STRAUSSMAN, R. & SAMUELS, Y. 2021. Identification of bacteria-derived HLA-bound peptides in melanoma. *Nature,* 592**,** 138-143.

KAPLAN, D. H., SHANKARAN, V., DIGHE, A. S., STOCKERT, E., AGUET, M., OLD, L. J. & SCHREIBER, R. D. 1998. Demonstration of an interferon-dependent tumor surveillance system in immunocompetent mice. *Proceedings of the National Academy of Sciences,* 95**,** 7556-7561.

KAPLAN, K. B., BURDS, A. A., SWEDLOW, J. R., BEKIR, S. S., SORGER, P. K. & NÄTHKE, I. S. 2001. A role for adenomatous polyposis coli protein in chromosomal segregation. *Nature Cell Biology,* 3**,** 429-432.

KASSAMBARA, A. 2020. ggpubr: 'ggplot2' Based Publication Ready Plots. R package version 0.4.0 ed.

KAUFFMANN, A., GENTLEMAN, R. & HUBER, W. 2009. arrayQualityMetrics--a bioconductor package for quality assessment of microarray data. 25**,** 415-416.

KIM, B.-H., CHEE, J. D., BRADFIELD, C. J., PARK, E.-S., KUMAR, P. & MACMICKING, J. D. 2016a. Interferon-induced guanylate-binding proteins in inflammasome activation and host defense. *Nature Immunology,* 17**,** 481-489.

KIM, C. G., AHN, J. B., JUNG, M., BEOM, S. H., KIM, C., KIM, J. H., HEO, S. J., PARK, H. S., KIM, J. H., KIM, N. K., MIN, B. S., KIM, H., KOOM, W. S. & SHIN, S. J. 2016b. Effects of microsatellite instability on recurrence patterns and outcomes in colorectal cancers. 115**,** 25-33.

KIM, D., PAGGI, J. M., PARK, C., BENNETT, C. & SALZBERG, S. L. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology,* 37**,** 907-915.

KIM, D., SONG, L., BREITWIESER, F. P. & SALZBERG, S. L. 2016c. Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research,* 26**,** 1721-1729.

KIM, E. S. 2017. Avelumab: First Global Approval. *Drugs,* 77**,** 929-937.

KIM, J.-E., PHAN, T. X., NGUYEN, V. H., DINH-VU, H.-V., ZHENG, J. H., YUN, M., PARK, S.-G., HONG, Y., CHOY, H. E., SZARDENINGS, M., HWANG, W., PARK, J.-A., PARK, S., IM, S.-H. & MIN, J.-J. 2015. Salmonella typhimurium Suppresses Tumor Growth via the Pro-Inflammatory Cytokine Interleukin-1β. *Theranostics,* 5**,** 1328-1342.

KIVIMÄKI, M., KUOSMA, E., FERRIE, J. E., LUUKKONEN, R., NYBERG, S. T., ALFREDSSON, L., BATTY, G. D., BRUNNER, E. J., FRANSSON, E., GOLDBERG, M., KNUTSSON, A., KOSKENVUO, M., NORDIN, M., OKSANEN, T., PENTTI, J., RUGULIES, R., SHIPLEY, M. J., SINGH-MANOUX, A., STEPTOE, A., SUOMINEN, S. B., THEORELL, T., VAHTERA, J., VIRTANEN, M., WESTERHOLM, P., WESTERLUND, H., ZINS, M., HAMER, M., BELL, J. A., TABAK, A. G. & JOKELA, M. 2017. Overweight, obesity, and risk of cardiometabolic multimorbidity: pooled analysis of individual-level data for 120 813 adults from 16 cohort studies from the USA and Europe. *The Lancet Public Health,* 2**,** e277-e285.

KOBAYASHI, K., CAHAMAILLARD, M., OGURA, Y., HENEGARIU, O., INOHARA, N., NUNEZ, G. & FLAVELL, R. 2005. Nod2-dependent regulation of innate and adaptive immunity in the intestinal tract. *Science,* 307**,** 731-734.

KOSINSKI, C., LI, V. S., CHAN, A. S., ZHANG, J., HO, C., TSUI, W. Y., CHAN, T. L., MIFFLIN, R. C., POWELL, D. W., YUEN, S. T., LEUNG, S. Y. & CHEN, X. 2007. Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc Natl Acad Sci U S A,* 104**,** 15418-23.

KOSTIC, A. D., GEVERS, D., PEDAMALLU, C. S., MICHAUD, M., DUKE, F., EARL, A. M., OJESINA, A. I., JUNG, J., BASS, A. J., TABERNERO, J., BASELGA, J., LIU, C., SHIVDASANI, R. A., OGINO, S., BIRREN, B. W., HUTTENHOWER, C., GARRETT, W. S. & MEYERSON, M. 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Research,* 22**,** 292-298.

KOYAMA, M., MUKHOPADHYAY, P., SCHUSTER, I. S., HENDEN, A. S., HÜLSDÜNKER, J., VARELIAS, A., VETIZOU, M., KUNS, R. D., ROBB, R. J., ZHANG, P., BLAZAR, B. R., THOMAS, R., BEGUN, J., WADDELL, N., TRINCHIERI, G., ZEISER, R., CLOUSTON, A. D., DEGLI-ESPOSTI, M. A. & HILL, G. R. 2019. MHC Class II Antigen Presentation by the Intestinal Epithelium Initiates Graft-versus-Host Disease and Is Influenced by the Microbiota. *Immunity,* 51**,** 885-898.e7.

LAL, N. 2016. *Exploring Immunogenomic Influences on the Microenvironment of Colorectal Cancer.* Doctor of Philosophy PhD Thesis, University of Birmingham.

LAL, N., BEGGS, A. D., WILLCOX, B. E. & MIDDLETON, G. W. 2015. An immunogenomic stratification of colorectal cancer: Implications for development of targeted immunotherapy. *Oncoimmunology,* 4**,** e976052.

LATCHMAN, Y., WOOD, C. R., CHERNOVA, T., CHAUDHARY, D., BORDE, M., CHERNOVA, I., IWAI, Y., LONG, A. J., BROWN, J. A., NUNES, R., GREENFIELD, E. A., BOURQUE, K., BOUSSIOTIS, V. A., CARTER, L. L., CARRENO, B. M., MALENKOVICH, N., NISHIMURA, H., OKAZAKI, T., HONJO, T., SHARPE, A. H. & FREEMAN, G. J. 2001. PD-L2 is a second ligand for PD-1 and inhibits T cell activation. *Nature Immunology,* 2**,** 261-268.

LE, D. T., DURHAM, J. N., SMITH, K. N., WANG, H., BARTLETT, B. R., AULAKH, L. K., LU, S., KEMBERLING, H., WILT, C., LUBER, B. S., WONG, F., AZAD, N. S., RUCKI, A. A., LAHERU, D., DONEHOWER, R. C., ZAHEER, A., FISHER, G. A., CROCENZI, T. S., LEE, J. J., GRETEN, T. F., DUFFY, A. G., CIOMBOR, K. K., EYRING, A. D., LAM, B. H., JOE, A., KANG, P., HOLDHOFF, M., DANILOVA, L., COPE, L., MEYER, C., ZHOU, S., GOLDBERG, R. M., ARMSTRONG, D. K., BEVER, K. M., FADER, A. N., TAUBE, J. M., PAPADOPOULOS, N., KINZLER, K. W., ESHELMAN, J. R., VOGELSTEIN, B., ANDERS, R. A. & DIAZ, L. A., JR. 2017. Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Cancer Research,* 75**,** 409-413.

LE, D. T., KIM, T. W., VAN CUTSEM, E., GEVA, R., JÄGER, D., HARA, H., BURGE, M., O'NEIL, B., KAVAN, P., YOSHINO, T., GUIMBAUD, R., TANIGUCHI, H., ELEZ, E., AL-BATRAN, S.-E., BOLAND, P. M., CROCENZI, T., ATREYA, C. E., CUI, Y., DAI, T., MARINELLO, P., DIAZ JR, L. A. & ANDRÉ, T. 2020. Phase II Open-Label Study of Pembrolizumab in Treatment-Refractory, Microsatellite Instability–High/Mismatch Repair–Deficient Metastatic Colorectal Cancer: KEYNOTE-164. *Journal of Clinical Oncology,* 38**,** 11-19.

LE, D. T., URAM, J. N., WANG, H., BARTLETT, B. R., KEMBERLING, H., EYRING, A. D., SKORA, A. D., LUBER, B. S., AZAD, N. S., LAHERU, D., BIEDRZYCKI, B., DONEHOWER, R. C., ZAHEER, A., FISHER, G. A., CROCENZI, T. S., LEE, J. J., DUFFY, S. M., GOLDBERG, R. M., DE LA CHAPELLE, A., KOSHIJI, M., BHAIJEE, F., HUEBNER, T., HRUBAN, R. H., WOOD, L. D., CUKA, N., PARDOLL, D. M., PAPADOPOULOS, N., KINZLER, K. W., ZHOU, S., CORNISH, T. C., TAUBE, J. M., ANDERS, R. A., ESHLEMAN, J. R., VOGELSTEIN, B. & DIAZ, L. A., JR. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *N Engl J Med,* 372**,** 2509-20.

LEE, J., MO, J.-H., KATAKURA, K., ALKALAY, I., RUCKER, A. N., LIU, Y.-T., LEE, H.-K., SHEN, C., COJOCARU, G., SHENOUDA, S., KAGNOFF, M., ECKMANN, L., BEN-NERIAH, Y. & RAZ, E. 2006. Maintenance of colonic homeostasis by distinctive apical TLR9 signalling in intestinal epithelial cells. *Nature Cell Biology,* 8**,** 1327-1336.

LEE, Y.-C., CHIANG, T.-H., CHOU, C.-K., TU, Y.-K., LIAO, W.-C., WU, M.-S. & GRAHAM, D. Y. 2016. Association Between Helicobacter pylori Eradication and Gastric Cancer Incidence: A Systematic Review and Meta-analysis. *Gastroenterology,* 150**,** 1113-1124.e5.

LEEK, J. T., JOHNSON, W. E., PARKER, H. S., FERTIG, E. J., JAFFE, A. E., ZHANG, Y., STOREY, J. D. & COLLADO TORRES, L. 2020. sva: Surrogate Variable Analysis. R package version 3.36.0 ed.

LI, D. & WU, M. 2021. Pattern recognition receptors in health and diseases. *Signal Transduction and Targeted Therapy,* 6.

LI, J., JIE, H. B., LEI, Y., GILDENER-LEAPMAN, N., TRIVEDI, S., GREEN, T., KANE, L. P. & FERRIS, R. L. 2015. PD-1/SHP-2 inhibits Tc1/Th1 phenotypic responses and the activation of T cells in the tumor microenvironment. *Cancer Res,* 75**,** 508-518.

LI, X., LING, A., KELLGREN, T. G., LUNDHOLM, M., LÖFGREN-BURSTRÖM, A., ZINGMARK, C., RUTEGÅRD, M., LJUSLINDER, I., PALMQVIST, R. & EDIN, S. 2020. A Detailed Flow Cytometric Analysis of Immune Activity Profiles in Molecular Subtypes of Colorectal Cancer. *Cancers,* 12**,** 3440.

LIAO, Y., SMYTH, G. K. & SHI, W. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. *Nucleic Acids Research,* 47**,** e47-e47.

LINDOR, N. M., BUGART, L. J., LEONTOVICH, O., GOLDBERG, R. M., CUNNINGHAM, J. M., SARGENT, D. J., WALSH-VOCKLEY, C., PETERSEN, G. M., WALSH, M. D., LEGGETT, B. A., YOUNG, J. P., BARKER, M. A., JASS, J. R., HOPPER, J., GALLINGER, S., BAPAT, B., REDSTON, M. & THIBODEAU, S. N. 2002. Immunohistochemistry Versus Microsatellite Instability Testing in Phenotyping Colorectal Tumors. *J Clin Oncol,* 20**,** 1043-1048.

LIU, H., WANG, J., HE, T., BECKER, S., ZHANG, G., LI, D. & MA, X. 2018. Butyrate: A Double-Edged Sword for Health? *Advances in Nutrition,* 9**,** 21-29.

LIU, R., LIAO, Y.-Z., ZHANG, W. & ZHOU, H.-H. 2020. Relevance of Immune Infiltration and Clinical Outcomes in Pancreatic Ductal Adenocarcinoma Subtypes. *Frontiers in Oncology,* 10**,** 575264.

LIU, Y. 2019. Immune response characterization of endometrial cancer. *Oncotarget,* 10**,** 982-992.

LUCHINI, C., BROSENS, L. A. A., WOOD, L. D., CHATTERJEE, D., SHIN, J. I., SCIAMMARELLA, C., FIADONE, G., MALLEO, G., SALVIA, R., KRYKLYVA, V., PIREDDA, M. L., CHENG, L., LAWLOR, R. T., ADSAY, V. & SCARPA, A. 2021. Comprehensive characterisation of pancreatic ductal adenocarcinoma with microsatellite instability: histology, molecular pathology and clinical implications. *Gut,* 70**,** 148-156.

LUO, W., FRIEDMAN, M. S., SHEDDEN, K., HANKENSON, K. D. & WOOLF, P. J. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics,* 10**,** 161.

LUO, X., LI, Z., LIN, S., LE, T., ITTENSOHN, M., BERMUDES, D., RUNYAB, J. D., SHEN, S.-Y., CHEN, J., KING, I. C. & ZHENG, L.-M. 2001. Antitumor Effect of VNP20009, an Attenuated *Salmonella,* in Murine Tumor Models. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics,* 12**,** 501-508.

MA, Y., YANG, Y., WANG, F., ZHANG, P., SHI, C., ZOU, Y. & QIN, H. 2013. Obesity and Risk of Colorectal Cancer: A Systematic Review of Prospective Studies. *PLoS ONE,* 8**,** e53916.

MACDONALD, B. T., TAMAI, K. & HE, X. 2009. Wnt/beta-catenin signaling: components, mechanisms, and diseases. *Dev Cell,* 17**,** 9-26.

MAK, T. W., SAUNDERS, M. E. & JETT, B. D. 2014. Chapter 12 - Mucosal and Cutaneous Immunity. *Primer to the Immune Response.* Second Edition ed. Academic Cell.

MARTY PYKE, R., THOMPSON, W. K., SALEM, R. M., FONT-BURGADA, J., ZANETTI, M. & CARTER, H. 2018. Evolutionary Pressure against MHC Class II Binding Cancer Mutations. *Cell,* 175**,** 416-428.e13.

MARTY, R., KAABINEJADIAN, S., ROSSELL, D., SLIFKER, M. J., VAN DE HAAR, J., ENGIN, H. B., DE PRISCO, N., IDEKER, T., HILDEBRAND, W. H., FONT-BURGADA, J. & CARTER, H. 2017. MHC-I Genotype Restricts the Oncogenic Mutational Landscape. *Cell,* 171**,** 1272-1283.e15.

MATSON, V., FESSLER, J., BAO, R., CHONGSUWAT, T., ZHA, Y., ALEGRE, M.-L., LUKE, J. J. & GAJEWSKI, T. F. 2018. The commensal microbiome is associated with anti–PD-1 efficacy in metastatic melanoma patients. *Science,* 359**,** 104-108.

MAYAKONDA, A., LIN, D.-C., ASSENOV, Y., PLASS, C. & KOEFFLER, H. P. 2018. Maftools: efficient and comprehensive analysis of somatic variants in cancer. *Genome Research,* 28**,** 1747-1756.

MAYER, L., ESISENHARDT, D., SALOMON, P., BAUER, W., PLOUS, R. & PICCININI, L. 1991. Epression of Class II Molecules on Intestinal Epithelial Cells in Humans: Differences between Normal and Inflammatory Bowel Disease. *Gastroenterology,* 100**,** 3-12.

MCCALL, M. N., BOLSTAD, B. M. & IRIZARRY, R. A. 2010. Frozen robust multiarray analysis (fRMA). *Biostatistics,* 11**,** 242-53.

MCGRANAHAN, N., FAVERO, F., DE BRUIN, E. C., BIRKBAK, N. J., SZALLASI, Z. & SWANTON, C. 2015. Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine,* 7**,** 283ra54-283ra54.

MCKENNA, S. M. & DAVIES, K. J. A. 1988. The inhibition of bacterial growth by hypochlorous acid. Possible role in the bactericidal activity of phagocytes. *Biochemical Journal,* 254**,** 685-692.
MCMURRAY, J. L., VON BORSTEL, A., TAHER, T. E., SYRIMI, E., TAYLOR, G. S., SHARIF, M., ROSSJOHN, J., REMMERSWAAL, E. B. M., BEMELMAN, F. J., VIEIRA BRAGA, F. A., CHEN, X., TEICHMANN, S. A., MOHAMMED, F., BERRY, A. A., LYKE, K. E., WILLIAMSON, K. C., STUBBINGTON, M. J. T., DAVEY, M. S., WILLCOX, C. R. & WILLCOX, B. E. 2022. Transcriptional profiling of human Vδ1 T cells reveals a pathogen-driven adaptive differentiation program. *Cell Reports,* 39**,** 110858.
MEAZZA, R., COMES, A., ORENGO, A. M., FERRINI, S. & ACCOLLA, R. S. 2003. Tumor rejection by gene transfer of the MHC class II transactivator in murine mammary adenocarcinoma cells. *Eur J Immunol,* 33**,** 1183-92.
MELLER, S., DI DOMIZIO, J., VOO, K. S., FRIEDRICH, H. C., CHAMILOS, G., GANGULY, D., CONRAD, C., GREGORIO, J., LE ROY, D., ROGER, T., LADBURY, J. E., HOMEY, B., WATOWICH, S., MODLIN, R. L., KONTOYIANNIS, D. P., LIU, Y.-J., AROLD, S. T. & GILLIET, M. 2015. TH17 cells promote microbial killing and innate immune sensing of DNA via interleukin 26. *Nature Immunology,* 16**,** 970-979.
MELLMAN, I., COUKOS, G. & DRANOFF, G. 2011. Cancer immunotherapy comes of age. *Nature,* 480**,** 480-9.
MENDELL, J. T. & DIETZ, H. C. 2001. When the Message Goes Awry. *Cell,* 107**,** 411-414.
MERICO, D., ISSERLIN, R., STUEKER, O., EMILI, A. & BADER, G. D. 2010. Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation. *PLoS ONE,* 5**,** e13984.
MEYER, D., DIMITRIADOU, E., HORNIK, K., WEINGESSEL, A. & LEISCH, F. 2017. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. version 1.6-8. ed.
MICHEL, S., LINNEBACHER, M., ALCANIZ, J., VOSS, M., WAGNER, R., DIPPOLD, W., BECKER, C., VON KNEBEL DOEBERITZ, M., FERRONE, S. & KLOOR, M. 2010. Lack of HLA class II antigen expression in microsatellite unstable colorectal carcinomas is caused by mutations in HLA class II regulatory genes. *International Journal of Cancer***,** NA-NA.
MIK, M., BERUT, M., DZIKI, L., TRZCINSKI, R. & DZIKI, A. 2017. Right- and left-sided colon cancer – clinical and pathological differences of the disease entity in one organ. *Archives of Medical Science,* 1**,** 157-162.
MILLEN, R., HENDRY, S., NARASIMHAN, V., ABBOTT, R., CROXFORD, M., GIBBS, P., TIE, J., WONG, H. L., JONES, I., KOSMIDER, S., BYRNE, D., ZALCBERG, J., FOX, S., DESAI, J., VISVANATHAN, K., RAMSAY, R. G. & TRAN, B. 2020. CD8 + tumor-infiltrating lymphocytes within the primary tumor of patients with synchronous de novo metastatic colorectal carcinoma do not track with survival. *Clinical & Translational Immunology,* 9.
MIMA, K., SUKAWA, Y., NISHIHARA, R., QIAN, Z. R., YAMAUCHI, M., INAMURA, K., KIM, S. A., MASUDA, A., NOWAK, J. A., NOSHO, K., KOSTIC, A. D., GIANNAKIS, M., WATANABE, H., BULLMAN, S., MILNER, D. A., HARRIS, C. C., GIOVANNUCCI, E., GARRAWAY, L. A., FREEMAN, G. J., DRANOFF, G., CHAN, A. T., GARRETT, W. S., HUTTENHOWER, C., FUCHS, C. S. & OGINO, S. 2015. Fusobacterium nucleatum and T Cells in Colorectal Carcinoma. *JAMA Oncology,* 1**,** 653.
MLECNIK, B., BINDEA, G., ANGELL, H. K., MABY, P., ANGELOVA, M., TOUGERON, D., CHURCH, S. E., LAFONTAINE, L., FISCHER, M., FREDRIKSEN, T., SASSO, M., BILOCQ, A. M., KIRILOVSKY, A., OBENAUF, A. C., HAMIEH, M., BERGER, A., BRUNEVAL, P., TUECH, J. J., SABOURIN, J. C., LE PESSOT, F., MAUILLON, J., RAFII, A., LAURENT-PUIG, P., SPEICHER, M. R., TRAJANOSKI, Z., MICHEL, P., SESBOUE, R., FREBOURG, T., PAGES, F., VALGE-ARCHER, V., LATOUCHE, J. B. & GALON, J. 2016. Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability. *Immunity,* 44**,** 698-711.

MOOI, J. K., WIRAPATI, P., ASHER, R., LEE, C. K., SAVAS, P., PRICE, T. J., TOWNSEND, A., HARDINGHAM, J., BUCHANAN, D., WILLIAMS, D., TEJPAR, S., MARIADASON, J. M. & TEBBUTT, N. C. 2018. The prognostic impact of consensus molecular subtypes (CMS) and its predictive effects for bevacizumab benefit in metastatic colorectal cancer: molecular analysis of the AGITG MAX clinical trial. *Annals of Oncology,* 29**,** 2240-2246.

MORGAN, M., OBENCHAIN, V., HESTER, J. & PAGÈS, H. 2020. SummarizedExperiment: SummarizedExperiment container. R package version 1.18.2 ed.

MORRISON, D. J. & PRESTON, T. 2016. Formation of short chain fatty acids by the gut microbiota and their impact on human metabolism. *Gut Microbes,* 7**,** 189-200.

MORTARA, L., CASTELLANI, P., MEAZZA, R., TOSI, G., DE LERMA BARBARO, A., PROCOPIO, F. A., COMES, A., ZARDI, L., FERRINI, S. & ACCOLLA, R. S. 2006. 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,* 12**,** 3435-43.

MOTZER, R. J., TANNIR, N. M., MCDERMOTT, D. F., AREN FRONTERA, O., MELICHAR, B., CHOUEIRI, T. K., PLIMACK, E. R., BARTHELEMY, P., PORTA, C., GEORGE, S., POWLES, T., DONSKOV, F., NEIMAN, V., KOLLMANNSBERGER, C. K., SALMAN, P., GURNEY, H., HAWKINS, R., RAVAUD, A., GRIMM, M. O., BRACARDA, S., BARRIOS, C. H., TOMITA, Y., CASTELLANO, D., RINI, B. I., CHEN, A. C., MEKAN, S., MCHENRY, M. B., WIND-ROTOLO, M., DOAN, J., SHARMA, P., HAMMERS, H. J., ESCUDIER, B. & CHECKMATE, I. 2018. Nivolumab plus Ipilimumab versus Sunitinib in Advanced Renal-Cell Carcinoma. *N Engl J Med,* 378**,** 1277-1290.

MOUGIAKAKOS, D., JOHANSSON, C. C. & KIESSLING, R. 2009. Naturally occurring regulatory T cells show reduced sensitivity toward oxidative stress–induced cell death. *Blood,* 113**,** 3542-3545.

MOUNIR, M., LUCCHETTA, M., SILVA, T. C., OLSEN, C., BONTEMPI, G., CHEN, X., NOUSHMEHR, H., COLAPRICO, A. & PAPALEO, E. 2019. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLOS Computational Biology,* 15**,** e1006701.

MOWAT, A. M. & AGACE, W. W. 2014. Regional specialization within the intestinal immune system. *Nature Reviews Immunology,* 14**,** 667-685.

MURPHY, K. M., ZHANG, S., GEIGER, T., HAFEZ, M. J., BACHER, J., BERG, K. D. & ESHLEMAN, J. R. 2006. Comparison of the Microsatellite Instability Analysis System and the Bethesda Panel for the Determination of Microsatellite Instability in Colorectal Cancers. *The Journal of Molecular Diagnostics,* 8**,** 305-311.

NAGPAL, R., NEWMAN, T., WANG, S., JAIN, S., LOVATO, J. & YADAV, H. 2018. Obesity-linked gut microbiome dybiosis associated with derangements in gut permeability and intestinal cellular homeostasis independent of diet. *Journal of Diabetes Research,* 2018.

NARDONE, G. & COMPARE, D. 2015. The human gastric microbiota: Is it time to rethink the pathogenesis of stomach diseases? *United European Gastroenterology Journal,* 3**,** 255-260.

NATIONAL INSTITUTE FOR CLINICAL EXCELLENCE, N. 2017. *Cetuximab and panitumumab for previously untreated metastatic colorectal cancer* [Online]. Available: https://www.nice.org.uk/guidance/ta439/chapter/1-Recommendations [Accessed 07/11/2021 2021].

NATIONAL INSTITUTE FOR CLINICAL EXCELLENCE, N. 2021. *Nivolumab with ipilimumab for previously treated metastatic colorectal cancer with high microsatellite instability or mismatch repair deficiency* [Online]. Available: https://www.nice.org.uk/guidance/ta716/chapter/1-Recommendations [Accessed 09/11/2021 2021].

NATIONAL LIBRARY OF MEDICINE, N. 2021. *Reference SNP report for rs28358285* [Online]. Available: https://www.ncbi.nlm.nih.gov/snp/rs28358285?vertical_tab=true#submissions [Accessed 07/11/2021 2021].

NEUMANN, P. A., KOCH, S., HILGARTH, R. S., PEREZ-CHANONA, E., DENNING, P., JOBIN, C. & NUSRAT, A. 2014. Gut commensal bacteria and regional Wnt gene expression in the proximal versus distal colon. *Am J Pathol,* 184**,** 592-9.

NICA, A. C. & DERMITZAKIS, E. T. 2013. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 368**,** 20120362.

NICHOLSON, J. K., HOLMES, E., KINROSS, J., BURCELIN, R., GIBSON, G., JIA, W. & PETTERSSON, S. 2012. Host-Gut Microbiota Metabolic Interactions. *Science,* 336**,** 1262-1267.

NOSHO, K., BABA, Y., TANAKA, N., SHIMA, K., HAYASHI, M., MEYERHARDT, J. A., GIOVANNUCCI, E., DRANOFF, G., FUCHS, C. S. & OGINO, S. 2010. Tumour-infiltrating T-cell subsets, molecular changes in colorectal cancer, and prognosis: cohort study and literature review. *The Journal of Pathology,* 222**,** 350-366.

OGINO, S., NOSHO, K., IRAHARA, N., MEYERHARDT, J. A., BABA, Y., SHIMA, K., GLICKMAN, J. N., FERRONE, C. R., MINO-KENUDSON, M., TANAKA, N., DRANOFF, G., GIOVANNUCCI, E. L. & FUCHS, C. S. 2009. Lymphocytic Reaction to Colorectal Cancer Is Associated with Longer Survival, Independent of Lymph Node Count, Microsatellite Instability, and CpG Island Methylator Phenotype. *Clinical Cancer Research,* 15**,** 6412-6420.

OGURA, Y., BONEN, D. K., INOHARA, N., NICOLAE, D. L., CHEN, F. F., RAMOS, R., BRITTON, H., MORAN, T., KARALIUSKAS, R., DUERR, R. H., ACHKAR, J.-P., BRANT, S. R., BAYLESS, T. M., KIRSCHNER, B. S., HANAUER, S. B., NUÑEZ, G. & CHO, J. H. 2001. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature,* 411**,** 603-606.

OKSANEN, J., BLANCHET, G., FRIENDLY, M., KINDT, R., LEGENDRE, P., MCGLINN, D., MINCHIN, P. R., O'HARA, R. B., SIMPSON, G. L., SOLYMOS, P., STEVENS, H. H., SZOECS, E. & WAGNER, H. 2020. vegan: Community Ecology Package. R package version 2.5-7 ed.: CRAN.

PANDA, A., BETIGERI, A., SUBRAMANIAN, K., ROSS, J. S., PAVLICK, D. C., ALI, S., MARKOWSKI, P., SILK, A., KAUFMAN, H. L., LATTIME, E., MEHNERT, J. M., SULLIVAN, R., LOVLY, C. M., SOSMAN, J., JOHNSON, D. B., BHANOT, G. & GANESAN, S. 2017. Identifying a Clinically Applicable Mutational Burden Threshold as a Potential Biomarker of Response to Immune Checkpoint Therapy in Solid Tumors. *JCO Precision Oncology***,** 1-13.

PARDOLL, D. 2003. Does the immune system see tumors as foreign or self? *Annu Rev Immunol,* 21**,** 807-39.

PATIL, I. 2021. Visualisation with statistical details: The 'ggstatsplot' approach. *PsyArxiv*.

PATIL, S., RAO, R. S. & MAJUMDAR, B. 2015. T-cell Exhaustion and Cancer Immunotherapy. *Journal of International Oral Health,* 7**,** i-ii.

PAULSON, J. N., OLSON, N., BRACCIA, D., WAGNER, J., TALUKDER, H., POP, M. & BRAVO, H. C. 2013a. metagenomeSeq: Statistical analysis for sparse high-throughput sequncing.

PAULSON, J. N., STINE, O. C., BRAVO, H. C. & POP, M. 2013b. Differential abundance analysis for microbial marker-gene surveys. *Nature Methods,* 10**,** 1200-1202.

PAWLIK, T. M., RAUT, C. P. & RODRIGUEZ-BIGAS, M. A. 2004. Colorectal carcinogenesis: MSI-H versus MSI-L. *Disease Markers,* 20**,** 199-206.

PÈNE, J., CHEVALIER, S., PREISSER, L., VÉNÉREAU, E., GUILLEUX, M.-H., GHANNAM, S., MOLÈS, J.-P., DANGER, Y., RAVON, E., LESAUX, S., YSSEL, H. & GASCAN, H. 2008. Chronically Inflamed Human Tissues Are Infiltrated by Highly Differentiated Th17 Lymphocytes. *The Journal of Immunology,* 180**,** 7423-7430.

PEREZ-RUIZ, E. & BERRAONDO, P. 2016. Immunological Landscape and Clinical Management of Rectal Cancer. *Front Immunol,* 7**,** 61.

PFEIFFER, P. & QVORTRUP, C. 2020. Does the Consensus Molecular Subtypes classification add to selection of precision medicine in patients with metastatic colorectal cancer? *Digestive Medicine Research,* 3**,** 20-20.

PHAN, T. X., NGUYEN, V. H., DUONG, M. T.-Q., HONG, Y., CHOY, H. E. & MIN, J.-J. 2015. Activation of inflammasome by attenuated Salmonella typhimuriumin bacteria-mediated cancer therapy. *Microbiology and Immunology,* 59**,** 664-675.

PICCOLO, S. R., SUN, Y., CAMPBELL, J. D., LENBURG, M. E., BILD, A. H. & JOHNSON, W. E. 2012. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics,* 100**,** 337-344.

PICCOLO, S. R., WITHERS, M. R., FRANCIS, O. E., BILD, A. H. & JOHNSON, W. E. 2013. Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences,* 110**,** 17778-17783.

PICKLES, O. 2021. *Regulation of MHC Class II in Colorectal Cancer.* PhD, University of Birmingham.

PINTO-RIBEIRO, I., FERREIRA, R. M., PEREIRA-MARQUES, J., PINTO, V., MACEDO, G., CARNEIRO, F. & FIGUEIREDO, C. 2020. Evaluation of the Use of Formalin-Fixed and Paraffin-Embedded Archive Gastric Tissues for Microbiota Characterization Using Next-Generation Sequencing. *International Journal of Molecular Sciences,* 21**,** 1096.

PUBMED. 2021. *microbiome - Search Results - PubMed* [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/ [Accessed].

R CORE TEAM, R. 2018. R: A language and environment for statistical computing. 3.5.0 ed. Vienna, Austria: R Foundation for Statistical Computing.

RAKOFF-NAHOUM, S. & MEDZHITOV, R. 2007. Regulation of spontaneous intestinal tumorigenesis through the adaptor protein MyD88. *Science,* 317**,** 124-127.

RAMANAN, D., SAN TANG, M., BOWCUTT, R., LOKE, P. & CADWELL, K. 2014. Bacterial Sensor Nod2 Prevents Inflammation of the Small Intestine by Restricting the Expansion of the Commensal Bacteroides vulgatus. *Immunity,* 41**,** 311-324.

RANJAN, R., RANI, A., METWALLY, A., MCGEE, H. S. & PERKINS, D. L. 2016. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications,* 469**,** 967-977.

RATLIFF, T. L. 1992. Role of the immune response in BCG for bladder cancer. *European Urology,* 21**,** 17-21.

REICHARDT, N., DUNCAN, S. H., YOUNG, P., BELENGUER, A., MCWILLIAM LEITCH, C., SCOTT, K. P., FLINT, H. J. & LOUIS, P. 2014. Phylogenetic distribution of three pathways for propionate production within the human gut microbiota. *The ISME Journal,* 8**,** 1323-1335.

RIGOTTIER-GOIS, L. 2013. Dysbiosis in inflammatory bowel diseases: the oxygen hypothesis. *The ISME Journal,* 7**,** 1256-1261.

RINNINELLA, E., RAOUL, P., CINTONI, M., FRANCESCHI, F., MIGGIANO, G., GASBARRINI, A. & MELE, M. 2019. What is the Healthy Gut Microbiota Composition? A Changing Ecosystem across Age, Environment, Diet, and Diseases. *Microorganisms,* 7**,** 14.

RIQUELME, E., ZHANG, Y., ZHANG, L., MONTIEL, M., ZOLTAN, M., DONG, W., QUESADA, P., SAHIN, I., CHANDRA, V., SAN LUCAS, A., SCHEET, P., XU, H., HANASH, S. M., FENG, L., BURKS, J. K., DO, K.-A., PETERSON, C. B., NEJMAN, D., TZENG, C.-W. D., KIM, M. P., SEARS, C. L., AJAMI, N., PETROSINO, J., WOOD, L. D., MAITRA, A., STRAUSSMAN, R., KATZ, M., WHITE, J. R., JENQ, R., WARGO, J. & MCALLISTER, F. 2019. Tumor Microbiome Diversity and Composition Influence Pancreatic Cancer Outcomes. *Cell,* 178**,** 795-806.e12.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research,* 43**,** e47-e47.

RIZVI, N. A., HELLMANN, M. D., SNYDER, A., KVISTBORG, P., MAKAROV, V., HAVEL, J. J., LEE, W., YUAN, J., WONG, P., HO, T. S., MILLER, M. L., REKHTMAN, N., MOREIRA, A. L., IBRAHIM, F., BRUGGEMAN, C., GASMI, B., ZAPPASODI, R., MAEDA, Y., SANDER, C., GARON, E. B., MERGHOUB, T., WOLCHOK, J. D., SCHUMACHER, T. N. & CHAN, T. A. 2015. Cancer

immunology. Mutational landscape determines sensitivity to PD-1 blockade in non-small cell lung cancer. *Science,* 348**,** 124-8.

RIZZATTI, G., LOPETUSO, L. R., GIBIINO, G., BINDA, C. & GASBARRINI, A. 2017. Proteobacteria: A Common Factor in Human Diseases. *BioMed Research International,* 2017**,** 1-7.

RODRIGUEZ, A. C., BLANCHARD, Z., MAURER, K. A. & GERTZ, J. 2019. Estrogen Signaling in Endometrial Cancer: a Key Oncogenic Pathway with Several Open Questions. *Hormones and Cancer,* 10**,** 51-63.

ROEDIGER, W. E. 1980. Role of anaerobic bacteria in the metabolic welfare of the colonic mucosa in man. *Gut,* 21**,** 793-798.

ROGIER, E., FRANTZ, A., BRUNO, M. & KAETZEL, C. 2014. Secretory IgA is Concentrated in the Outer Layer of Colonic Mucus along with Gut Bacteria. *Pathogens,* 3**,** 390-403.

ROONEY, M. S., SHUKLA, S. A., WU, C., GETZ, G. & HACOHEN, N. 2015. Molecular and Genetic Properties of Tumors Associated with Local Immune Cytolytic Activity. *Cell,* 160**,** 48-61.

ROSENBERG, S. A., PACKARD, B. S., AEBERSOLD, P. M., SOLOMON, D., TOPALIAN, S. L., TOY, S. T., SIMON, P., LOTZE, M. T., YANG, J. C., SEIPP, C. A., SIMPSON, C., CARTER, C., BOCK, S., SCHWARTZENTRUBER, D., WEI, J. P. & WHITE, D. E. 1988. Use of Tumor-Infiltrating Lymphocytes and Interleukin-2 in the Immunotherapy of Patients with Metastatic Melanoma: A Preliminary Report. *N Engl J Med,* 319**,** 1676-1680.

ROSSITER, A. 10/09/2021 2021. *RE: Permeabilisation of bacteria prior to 16s rRNA RNA scope staining.*

ROTH, A., KHATTRA, J., YAP, D., WAN, A., LAKS, E., BIELE, J., HA, G., APARICIO, S., BOUCHARD-CÔTÉ, A. & SHAH, S. P. 2014. PyClone: statistical inference of clonal population structure in cancer. *Nature Methods,* 11**,** 396-398.

ROUTY, B., LE CHATELIER, E., DEROSA, L., DUONG, C. P. M., TALOU, M. T., DAILLERE, R., FLUCKIGER, A., MESSAOUDENE, M., RAUBER, C., ROBERTI, M. P., FIDELLE, M., FLAMENT, C., POIRIER-COLAME, V., OPOLON, P., KLEIN, C., IRIBARREN, K., MONDRAGON, L., JACQUELOT, N., QU, B., FERRERE, G., CLEMONSON, C., MEZQUITA, L., MASIP, J. R., NALTET, C., BROSSEAU, S., KADERBHAI, C., RICHARD, C., RIZVI, H., LEVENEZ, F., GALLERON, N., QUINQUIS, B., PONS, N., RYFFEL, B., MINARD-COLIN, V., GONIN, P., SORIA, J. C., DEUTSCH, E., LORIOT, Y., GHIRINGHELLI, F., ZALCMAN, G., GOLDWASSER, A., RAOULT, D., ALBIGES, L., KROEMER, G. & ZITVOGEL, L. 2018. Gut microbiome influences efficacy of PD-1-based immunotherapy against epithelial tumors. *Science,* 359**,** 91-97.

SANTONI, M., PIVA, F., CONTI, A., SANTONI, A., CIMADAMORE, A., SCARPELLI, M., BATTELLI, N. & MONTIRONI, R. 2018. Re: Gut Microbiome Influences Efficacy of PD-1-based Immunotherapy Against Epithelial Tumors. *Words of Wisdom,* 74**,** 521-522.

SATHALIYAWALA, T., KUBOTA, M., YUDANIN, N., TURNER, D., CAMP, P., JOSEPH, KARA, LERNER, H., GOLDSTEIN, M., SYKES, M., KATO, T. & DONNA 2013. Distribution and Compartmentalization of Human Circulating and Tissue-Resident Memory T Cell Subsets. *Immunity,* 38**,** 187-197.

SAWAYAMA, H., MIYAMOTO, Y., OGAWA, K., YOSHIDA, N. & BABA, H. 2020. Investigation of colorectal cancer in accordance with consensus molecular subtype classification. *Annals of Gastroenterological Surgery,* 4**,** 528-539.

SCHEPER, W., KELDERMAN, S., FANCHI, L. F., LINNEMANN, C., BENDLE, G., DE ROOIJ, M. A. J., HIRT, C., MEZZADRA, R., SLAGTER, M., DIJKSTRA, K., KLUIN, R. J. C., SNAEBJORNSSON, P., MILNE, K., NELSON, B. H., ZIJLMANS, H., KENTER, G., VOEST, E. E., HAANEN, J. B. A. G. & SCHUMACHER, T. N. 2019. Low and variable tumor reactivity of the intratumoral TCR repertoire in human cancers. *Nature Medicine,* 25**,** 89-94.

SCHNEIDER, R., BORNSTEIN, S. R., CHROUSOS, G. P., BOXBERGER, S., EHNINGER, G. & BREIDERT, M. 2001. Leptin mediates a proliferative response in human gastric mucosa cells with functional receptor. *Hormone and Metabolic Research,* 33**,** 1-6.

SCHUMACHER, K., HAENSCH, W., ROEFZAAD, C. & SCHALG, P. 2001. Prognostic signficance of Activated CD8 T cell infiltrations within eosophageal carcinomas. *Cancer Research,* 61.

SELIGER, B., KLOOR, M. & FERRONE, S. 2017. HLA class II antigen-processing pathway in tumors: Molecular defects and clinical relevance. *OncoImmunology,* 6**,** e1171447.

ŞENBABAOĞLU, Y., GEJMAN, R. S., WINER, A. G., LIU, M., VAN ALLEN, E. M., DE VELASCO, G., MIAO, D., OSTROVNAYA, I., DRILL, E., LUNA, A., WEINHOLD, N., LEE, W., MANLEY, B. J., KHALIL, D. N., KAFFENBERGER, S. D., CHEN, Y., DANILOVA, L., VOSS, M. H., COLEMAN, J. A., RUSSO, P., REUTER, V. E., CHAN, T. A., CHENG, E. H., SCHEINBERG, D. A., LI, M. O., CHOUEIRI, T. K., HSIEH, J. J., SANDER, C. & HAKIMI, A. A. 2016. Tumor immune microenvironment characterization in clear cell renal cell carcinoma identifies prognostic and immunotherapeutically relevant messenger RNA signatures. *Genome Biology,* 17.

SETIAWAN, V. W., YANG, H. P., PIKE, M. C., MCCANN, S. E., YU, H., XIANG, Y.-B., WOLK, A., WENTZENSEN, N., WEISS, N. S., WEBB, P. M., VAN DEN BRANDT, P. A., VAN DE VIJVER, K., THOMPSON, P. J., STROM, B. L., SPURDLE, A. B., SOSLOW, R. A., SHU, X.-O., SCHAIRER, C., SACERDOTE, C., ROHAN, T. E., ROBIEN, K., RISCH, H. A., RICCERI, F., REBBECK, T. R., RASTOGI, R., PRESCOTT, J., POLIDORO, S., PARK, Y., OLSON, S. H., MOYSICH, K. B., MILLER, A. B., MCCULLOUGH, M. L., MATSUNO, R. K., MAGLIOCCO, A. M., LURIE, G., LU, L., LISSOWSKA, J., LIANG, X., LACEY, J. V., KOLONEL, L. N., HENDERSON, B. E., HANKINSON, S. E., HÅKANSSON, N., GOODMAN, M. T., GAUDET, M. M., GARCIA-CLOSAS, M., FRIEDENREICH, C. M., FREUDENHEIM, J. L., DOHERTY, J., DE VIVO, I., COURNEYA, K. S., COOK, L. S., CHEN, C., CERHAN, J. R., CAI, H., BRINTON, L. A., BERNSTEIN, L., ANDERSON, K. E., ANTON-CULVER, H., SCHOUTEN, L. J. & HORN-ROSS, P. L. 2013. Type I and II Endometrial Cancers: Have They Different Risk Factors? *Journal of Clinical Oncology,* 31**,** 2607-2618.

SHANNON, P. 2003. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research,* 13**,** 2498-2504.

SHARMA, A., SUBUDHI, S. K., BLANDO, J., SCUTTI, J., VENCE, L., WARGO, J., ALLISON, J. P., RIBAS, A. & SHARMA, P. 2019. Anti-CTLA-4 Immunotherapy Does Not Deplete FOXP3+ Regulatory T Cells (Tregs) in Human Cancers. *Clinical Cancer Research,* 25**,** 1233-1238.

SHEPPARD, K. A., FITZ, L. J., LEE, J. M., BENANDER, C., GEORGE, J. A., WOOTERS, J., QIU, Y., JUSSIF, J. M., CARTER, L. L., WOOD, C. R. & CHAUDHARY, D. 2004. PD-1 inhibits T-cell receptors phosphorylation of the ZAP70/CD3ζ signalosome and downstream signalling to PDCθ. *FEBS Letters,* 574**,** 37-41.

SHERRILL-MIX, S. 2021. taxonomizr: Functions to Work with NCBI Accessions and Taxonomy. R package version 0.8.0 ed.

SHIN, N.-R., WHON, T. W. & BAE, J.-W. 2015. Proteobacteria: microbial signature of dysbiosis in gut microbiota. *Trends in Biotechnology,* 33**,** 496-503.

SHIU, K.-K., ANDRE, T., KIM, T. W., JENSEN, B. V., JENSEN, L. H., PUNT, C. J. A., SMITH, D. M., GARCIA-CARBONERO, R., BENAVIDES, M., GIBBS, P., FOUCHARDIERE, C. D. L., RIVERA, F., ELEZ, E., BENDELL, J. C., LE, D. T., YOSHINO, T., YANG, P., FAROOQUI, M. Z. H., MARINELLO, P. & DIAZ, L. A. 2021. KEYNOTE-177: Phase III randomized study of pembrolizumab versus chemotherapy for microsatellite instability-high advanced colorectal cancer. *Journal of Clinical Oncology,* 39**,** 6-6.

SIEGEL, R. L., MILLER, K. D., GODING SAUER, A., FEDEWA, S. A., BUTTERLY, L. F., ANDERSON, J. C., CERCEK, A., SMITH, R. A. & JEMAL, A. 2020. Colorectal cancer statistics, 2020. *CA: A Cancer Journal for Clinicians,* 70**,** 145-164.

SILLO, T. 2021, pre-viva. *Germline and somatic determinants of the immune contexture in colorectal cancer.* PhD, University of Birmingham.

SILVA, T. C., COLAPRICO, A., OLSEN, C., D'ANGELO, F., BONTEMPI, G., CECCARELLI, M. & NOUSHMEHR, H. 2016. TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages. *F1000Research,* 5**,** 1542.

SING, T., SANDER, O., BEERENWINKEL, N. & LENGAUER, T. 2005. ROCR: visualizing classifier performance in R. *Bioinformatics,* 21**,** 3940-3941.
SIVAN, A., CORRALES, L., HUBERT, N., WILLIAMS, J. B., AQUINO-MICHAELS, K., EARLEY, Z. M., BENYAMIN, F. W., MAN LEI, Y., JABRI, B., ALEGRE, M.-L., CHANG, E. B. & GAJEWSKI, T. F. 2015. Commensal Bifidobacterium promotes antitumor immunity and facilitates anti–PD-L1 efficacy. *Science,* 350**,** 1084-1089.
STATTIN, P., LUKANOVA, A., BIESSY, C., SÖDERBERG, S., PALMQVIST, R., KAAKS, R., OLSSON, T. & JELLUM, E. 2004. Obesity and colon cancer: Does leptin provide a link? *International Journal of Cancer,* 109**,** 149-152.
SURMANN, E.-M., VOIGT, A. Y., MICHEL, S., BAUER, K., REUSCHENBACH, M., FERRONE, S., VON KNEBEL DOEBERITZ, M. & KLOOR, M. 2015. Association of high CD4-positive T cell infiltration with mutations in HLA class II-regulatory genes in microsatellite-unstable colorectal cancer. *Cancer Immunology, Immunotherapy,* 64**,** 357-366.
SWANSON, P. A., KUMAR, A., SAMARIN, S., VIJAY-KUMAR, M., KUNDU, K., MURTHY, N., HANSEN, J., NUSRAT, A. & NEISH, A. S. 2011. Enteric commensal bacteria potentiate epithelial restitution via reactive oxygen species-mediated inactivation of focal adhesion kinase phosphatases. *Proceedings of the National Academy of Sciences,* 108**,** 8803-8808.
SYED, Y. Y. 2017. Durvalumab: First Global Approval. *Drugs,* 77**,** 1369-1376.
TANG, Y., CHEN, Y., JIANG, H., ROBBINS, G. T. & NIE, D. 2011. G-protein-coupled receptor for short-chain fatty acids suppresses colon cancer. *International Journal of Cancer,* 128**,** 847-856.
TARIQ, K. & GHIAS, K. 2016. Colorectal cancer carcinogenesis: a review of mechanisms. *Cancer Biol Med,* 13**,** 120-35.
TEAM, T. B. D. 2020. BSgenome.Hsapiens.UCSC.hg38: Full genome sequences for Homo sapiens (UCSC version hg38, based on GRCh38.p12). R package version 1.4.3 ed.
TEIXEIRA, T. F. S., SOUZA, N. C. S., CHIARELLO, P. G., FRANCESCHINI, S. C. C., BRESSAN, J., FERREIRA, C. L. L. F. & PELUZIO, M. 2012. Intestinal permeability parameters in obese patients are correlated with metabolic syndrome risk factors. *Clinical Nutrition,* 31**,** 735 - 740.
TEMPLE LANG, D. 2020. RCurl: General Network (HTTP/FTP/...) Client Interface for R. R package version 1.98-1.2 ed.
THALLINGER, C., FUREDER, T., PREUSSER, M., HELLER, G., MULLAUER, L., HOLLER, C., PROSCH, H., FRANK, N., SWIERZEWSKI, R., BERGER, W., JAGER, U. & ZIELINSKI, C. 2018. Review of cancer treatment with immune checkpoint inhibitors : Current concepts, expectations, limitations and pitfalls. *Wien Klin Wochenschr,* 130**,** 85-91.
THE CANCER GENOME ATLAS, T. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-7.
THE CANCER GENOME ATLAS, T. 2021. *HCMI-CMDC Database* [Online]. Available: https://portal.gdc.cancer.gov/projects/HCMI-CMDC [Accessed 09/11/2021 2021].
THE HUMAN MICROBIOME CONSORTIUM, H. 2012a. A framework for human microbiome research. *Nature,* 486**,** 215-221.
THE HUMAN MICROBIOME CONSORTIUM, H. 2012b. Structure, function and diversity of the healthy human microbiome. *Nature,* 486**,** 207-214.
THE U.S DEPARTMENT OF HEATH, F. 2017. *FDA grants accelerated approval to pembrolizumab for first tissue/site agnostic indication* [Online]. Available: https://www.fda.gov/drugs/resources-information-approved-drugs/fda-grants-accelerated-approval-pembrolizumab-first-tissuesite-agnostic-indication [Accessed 16/11/2021].
THOMAS, A. M., MANGHI, P., ASNICAR, F., PASOLLI, E., ARMANINI, F., ZOLFO, M., BEGHINI, F., MANARA, S., KARCHER, N., POZZI, C., GANDINI, S., SERRANO, D., TARALLO, S., FRANCAVILLA, A., GALLO, G., TROMPETTO, M., FERRERO, G., MIZUTANI, S., SHIROMA, H., SHIBA, S., SHIBATA, T., YACHIDA, S., YAMADA, T., WIRBEL, J., SCHROTZ-KING, P., ULRICH, C. M., BRENNER, H., ARUMUGAM, M., BORK, P., ZELLER, G., CORDERO, F., DIAS-NETO, E., SETUBAL,

J. C., TETT, A., PARDINI, B., RESCIGNO, M., WALDRON, L., NACCARATI, A. & SEGATA, N. 2019. Metagenomic analysis of colorectal cancer datasets identifies cross-cohort microbial diagnostic signatures and a link with choline degradation. *Nature Medicine,* 25**,** 667-678.

TRAN, E., ROBBINS, P. F., LU, Y.-C., PRICKETT, T. D., GARTNER, J. J., JIA, L., PASETTO, A., ZHENG, Z., RAY, S., GROH, E. M., KRILEY, I. R. & ROSENBERG, S. A. 2016. T-Cell Transfer Therapy Targeting Mutant KRAS in Cancer. *New England Journal of Medicine,* 375**,** 2255-2262.

TRAN, E., TURCOTTE, S., GROS, A., ROBBINS, P. F., LU, Y. C., DUDLEY, M. E., WUNDERLICH, J. R., SOMERVILLE, R. P., HOGAN, K., HINRICHS, C. S., PARKHURST, M. R., YANG, J. C. & ROSENBERG, S. A. 2014. Cancer Immunotherapy Based on Mutation-Specific CD4+ T Cells in a Patient with Epithelial Cancer. *Science,* 344**,** 641-645.

VAN DER BRUGGEN, P., TRAVERSARI, C., CHOMEZ, P., LURQUIN, C., DE PLAEN, E., VAN DEN EYNDE, B., KNUTH, A. & BOON, T. 1991. A Gene Encoding an Antigen Recognized by Cytolytic T Lymphocytes on a Human Melanoma. *Science,* 254**,** 1643-1647.

VAN DER HEE, B. & WELLS, J. M. 2021. Microbial Regulation of Host Physiology by Short-chain Fatty Acids. *Trends in Microbiology,* 29**,** 700-712.

VAN DER KRAAK, L. A., SCHNEIDER, C., DANG, V., BURR, A. H. P., WEISS, E. S., VARGHESE, J. A., YANG, L., HAND, T. W. & CANNA, S. W. 2021. Genetic and commensal induction of IL-18 drive intestinal epithelial MHCII via IFNγ. *Mucosal Immunology*.

VAN GOOL, I. C., EGGINK, F. A., FREEMAN-MILLS, L., STELLOO, E., MARCHI, E., DE BRUYN, M., PALLES, C., NOUT, R. A., DE KROON, C. D., OSSE, E. M., KLENERMAN, P., CREUTZBERG, C. L., TOMLINSON, I. P. M., SMIT, V. T. H. B. M., NIJMAN, H. W., BOSSE, T. & CHURCH, D. N. 2015. POLE Proofreading Mutations Elicit an Antitumor Immune Response in Endometrial Cancer. *Clinical Cancer Research,* 21**,** 3347-3355.

VÉTIZOU, M., PITT, J. M., DAILLÈRE, R., LEPAGE, P., WALDSCHMITT, N., FLAMENT, C., RUSAKIEWICZ, S., ROUTY, B., ROBERTI, M. P., DUONG, C. P. M., POIRIER-COLAME, V., ROUX, A., BECHAREF, S., FORMENTI, S., GOLDEN, E., CORDING, S., EBERL, G., SCHLITZER, A., GINHOUX, F., MANI, S., YAMAZAKI, T., JACQUELOT, N., ENOT, D. P., BÉRARD, M., NIGOU, J., OPOLON, P., EGGERMONT, A., WOERTHER, P.-L., CHACHATY, E., CHAPUT, N., ROBERT, C., MATEUS, C., KROEMER, G., RAOULT, D., BONECA, I. G., CARBONNEL, F., CHAMAILLARD, M. & ZITVOGEL, L. 2015. Anticancer immunotherapy by CTLA-4 blockade relies on the gut microbiota. *Science,* 350**,** 1079-1084.

VOGELSANG, M., MARTINEZ, C. N., RENDLEMAN, J., BAPODRA, A., MALECEK, K., ROMANCHUK, A., KAZLOW, E., SHAPIRO, R. L., BERMAN, R. S., KROGSGAARD, M., OSMAN, I. & KIRCHHOFF, T. 2016. The Expression Quantitative Trait Loci in Immune Pathways and their Effect on Cutaneous Melanoma Prognosis. *Clinical Cancer Research,* 22**,** 3268-3280.

WALKER, L. S. & SANSOM, D. M. 2015. Confusing Signals: Recent Progress in CTLA-4 Biology. *Trends in Immunology,* 36**,** 63-70.

WALTHER-ANTÓNIO, M. R. S., CHEN, J., MULTINU, F., HOKENSTAD, A., DISTAD, T. J., CHEEK, E. H., KEENEY, G. L., CREEDON, D. J., NELSON, H., MARIANI, A. & CHIA, N. 2016. Potential contribution of the uterine microbiome in the development of endometrial cancer. *Genome Medicine,* 8.

WANG, H.-B., WANG, P.-Y., WANG, X., WAN, Y.-L. & LIU, Y.-C. 2012. Butyrate Enhances Intestinal Epithelial Barrier Function via Up-Regulation of Tight Junction Protein Claudin-1 Transcription. *Digestive Diseases and Sciences,* 57**,** 3126-3135.

WEI, M.-Y., SHI, S., LIANG, C., MENG, Q.-C., HUA, J., ZHANG, Y.-Y., LIU, J., ZHANG, B., XU, J. & YU, X.-J. 2019. The microbiota and microbiome in pancreatic cancer: more influential than expected. *Molecular Cancer,* 18.

WEISENBERG, E. 2017. *TNM staging of colorectcal carcinoma (AJCC 8th edition).* [Online]. AJCC. Available: http://www.pathologyoutlines.com/topic/colontumorstaging8ed.html. [Accessed 18/04/18 2018].

WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., MCGOWAN, L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., PEDERSEN, T., MILLER, E., BACHE, S., MÜLLER, K., OOMS, J., ROBINSON, D., SEIDEL, D., SPINU, V., TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. 2019. Welcome to the Tidyverse. *Journal of Open Source Software,* 4**,** 1686.

WIEMANN, B. & STARNES, C. O. 1994. Coley's toxins, tumor necrosis factor and cancer research: A historical perspective. 64**,** 529-564.

WILLEMSEN, L. E. M. 2003. Short chain fatty acids stimulate epithelial mucin 2 expression through differential effects on prostaglandin E1 and E2 production by intestinal myofibroblasts. *Gut,* 52**,** 1442-1447.

WILLIAMS, K. P., GILLESPIE, J. J., SOBRAL, B. W., NORDBERG, E. K., SNYDER, E. E., SHALLOM, J. M. & DICKERMAN, A. W. 2020. Phylogeny of Gammaproteobacteria. *Journal of Bacteriology,* 192.

WINTER, S. E., THIENNIMITR, P., WINTER, M. G., BUTLER, B. P., HUSEBY, D. L., CRAWFORD, R. W., RUSSELL, J. M., BEVINS, C. L., ADAMS, L. G., TSOLIS, R. M., ROTH, J. R. & BÄUMLER, A. J. 2010. Gut inflammation provides a respiratory electron acceptor for Salmonella. *Nature,* 467**,** 426-429.

WROBLEWSKI, L. E., PEEK, R. M. & WILSON, K. T. 2010. Helicobacter pylori and Gastric Cancer: Factors That Modulate Disease Risk. *Clinical Microbiology Reviews,* 23**,** 713-739.

WU, S., RHEE, K.-J., ALBESIANO, E., RABIZADEH, S., WU, X., YEN, H.-R., HUSO, D. L., BRANCATI, F. L., WICK, E., MCALLISTER, F., HOUSSEAU, F., PARDOLL, D. M. & SEARS, C. L. 2009. A human colonic commensal promotes colon tumorigenesis via activation of T helper type 17 T cell responses. *Nature Medicine,* 15**,** 1016-1022.

YAROSZ, E. L. & CHANG, C.-H. 2018. The Role of Reactive Oxygen Species in Regulating T Cell-mediated Immunity and Disease. *Immune Network,* 18.

YU, L.-X. & SCHWABE, R. F. 2017. The gut microbiome and liver cancer: mechanisms and clinical translation. *Nature Reviews Gastroenterology & Hepatology,* 14**,** 527-539.

ZHANG, Y. & ZHANG, Z. 2020. The history and advances in cancer immunotherapy: understanding the characteristics of tumor-infiltrating immune cells and their therapeutic implications. *Cellular & Molecular Immunology,* 17**,** 807-821.

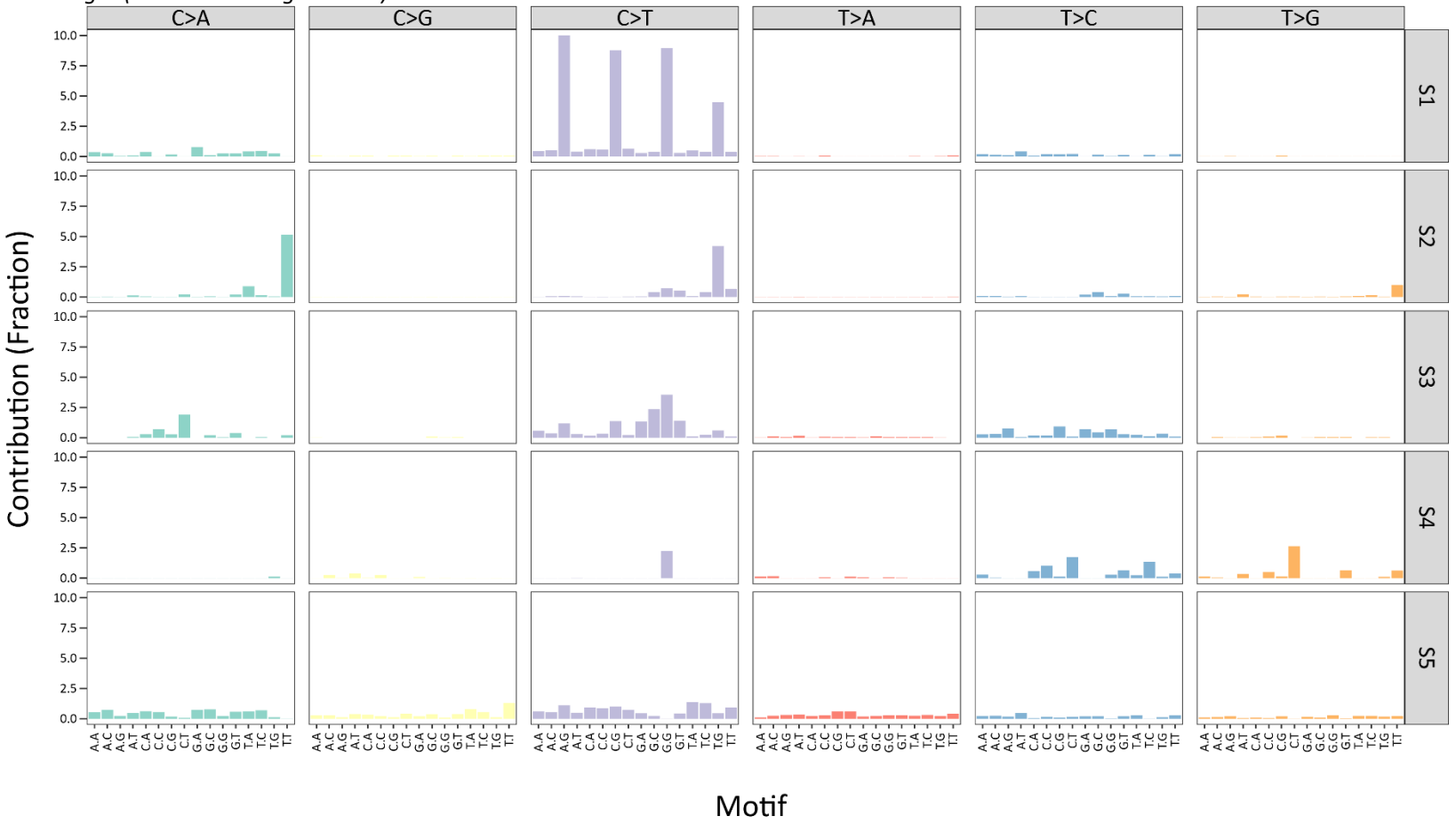
ZHAO, Y., CHEN, F., WU, W., SUN, M., BILOTTA, A. J., YAO, S., XIAO, Y., HUANG, X., EAVES-PYLES, T. D., GOLOVKO, G., FOFANOV, Y., D'SOUZA, W., ZHAO, Q., LIU, Z. & CONG, Y. 2018. GPR43 mediates microbiota metabolite SCFA regulation of antimicrobial peptide expression in intestinal epithelial cells via activation of mTOR and STAT3. *Mucosal Immunology,* 11**,** 752-762.

ZHONG, M., XIONG, Y., YE, Z., ZHAO, J., ZHONG, L., LIU, Y., ZHU, Y., TIAN, L., QIU, X. & HONG, X. 2020. Microbial Community Profiling Distinguishes Left-Sided and Right-Sided Colon Cancer. *Frontiers in Cellular and Infection Microbiology,* 10.

Chapter 10: Appendices

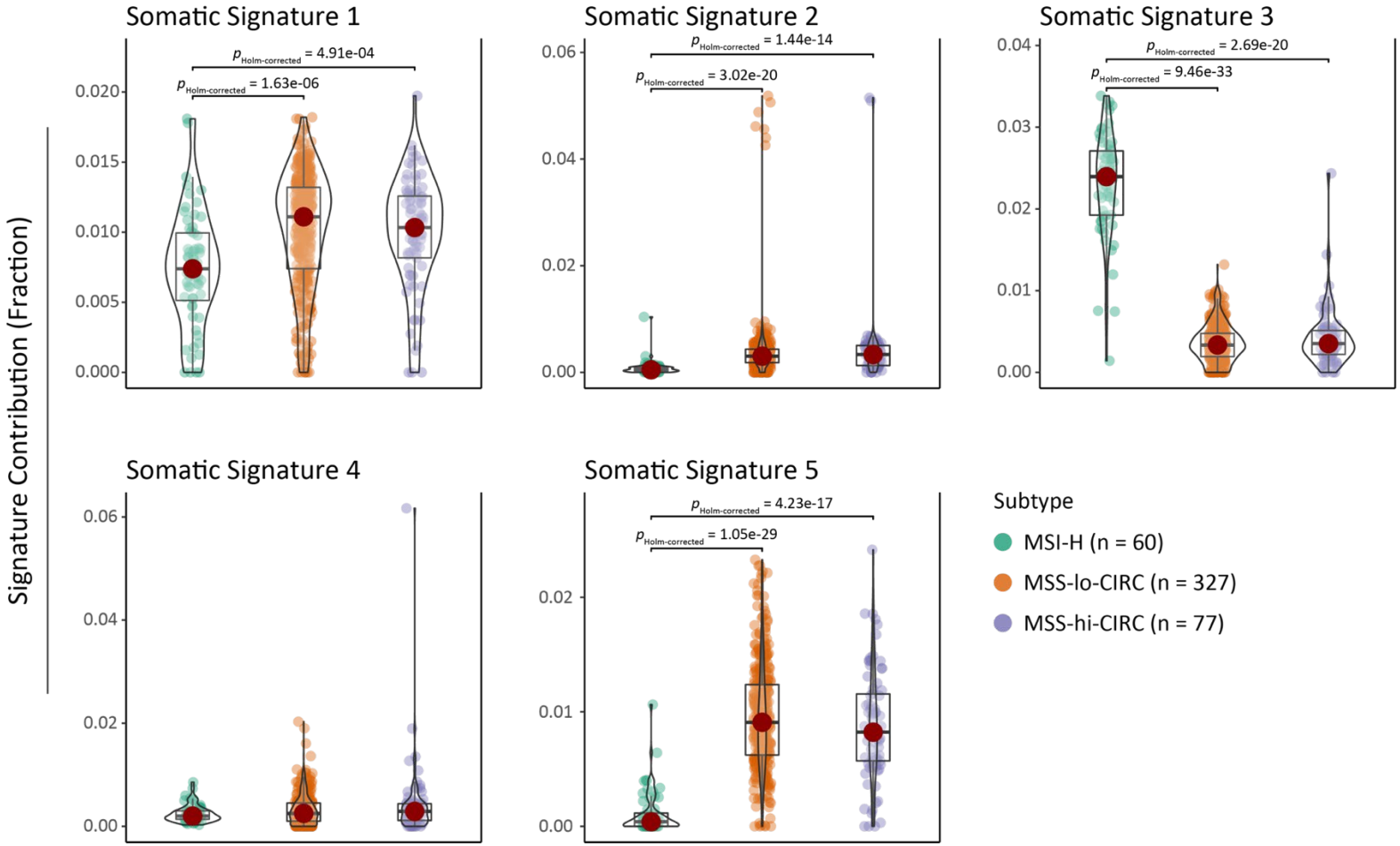
Appendix 1 – The contribution of the 96 individual base changes to each of the five TCGA-COADREAD mutation signatures

*Appendix is in relation to Figure 2.2 & Figure 2.5. The frequency of the 96 individual base changes making up five mutation signatures are demonstrated. The top grid indicates the base change whereby C>A is a change of a cytosine to an adenine. The X axis indicates the flanking nucleotides whereby A.A are two adenine surrounding the base change of interest (for example A**C**A to A**A**A, A**C**C to A**A**C, cont.). The X axis repeats the same pattern (left to right - A.A, A.C, A.G, A.T, C.A, C.C, C.G, C.T, G.A, G.C, G.G, G.T, T.A, T.C, T.G, T.T) for each base change. The Y axis indicates the contribution of each change to the signature shown on the right (S1 = Somatic Signature 1).*



Appendix 2 – Collection of violin plots demonstrating differences in Mutation signature contribution between TCGA-COADREAD subgroups

Appendix is in relation to Figure 4.3. Five mutation signatures were delineated in the TCGA-COADREAD cohort using NMF decomposition. The contribution of each Somatic Signature to each patient was calculated as a fraction and compared across subgroups. Plots are organised by Somatic Signature. The number of patients belonging to each subgroup is shown in the Figure key. Statistical analyses used Kruskal Wallis tests. Pairwise investigations used Dunn's tests. Only significant comparisons after Holm correction (pVal < 0.05) are shown.



Appendix 3 – Microbe genesets significant between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-COADREAD cohort

Appendix relates to Figure 5.7. Genes differentially expressed between MSS-hi-CIRC and MSS-lo-CIRC in the TCGA-COADREAD cohort and GO terms relating to microbes were used as input. 38 pathways were found differentially enriched between MSS-hi-CIRC and MSS-lo-CIRC. Genesets were considered significant at an FDR < 0.05. FDR are rounded to 3 significant figures. Cytoscape group relates to the groups shown in Figure 5.7.
NGenes = Number of Genes in the Geneset.

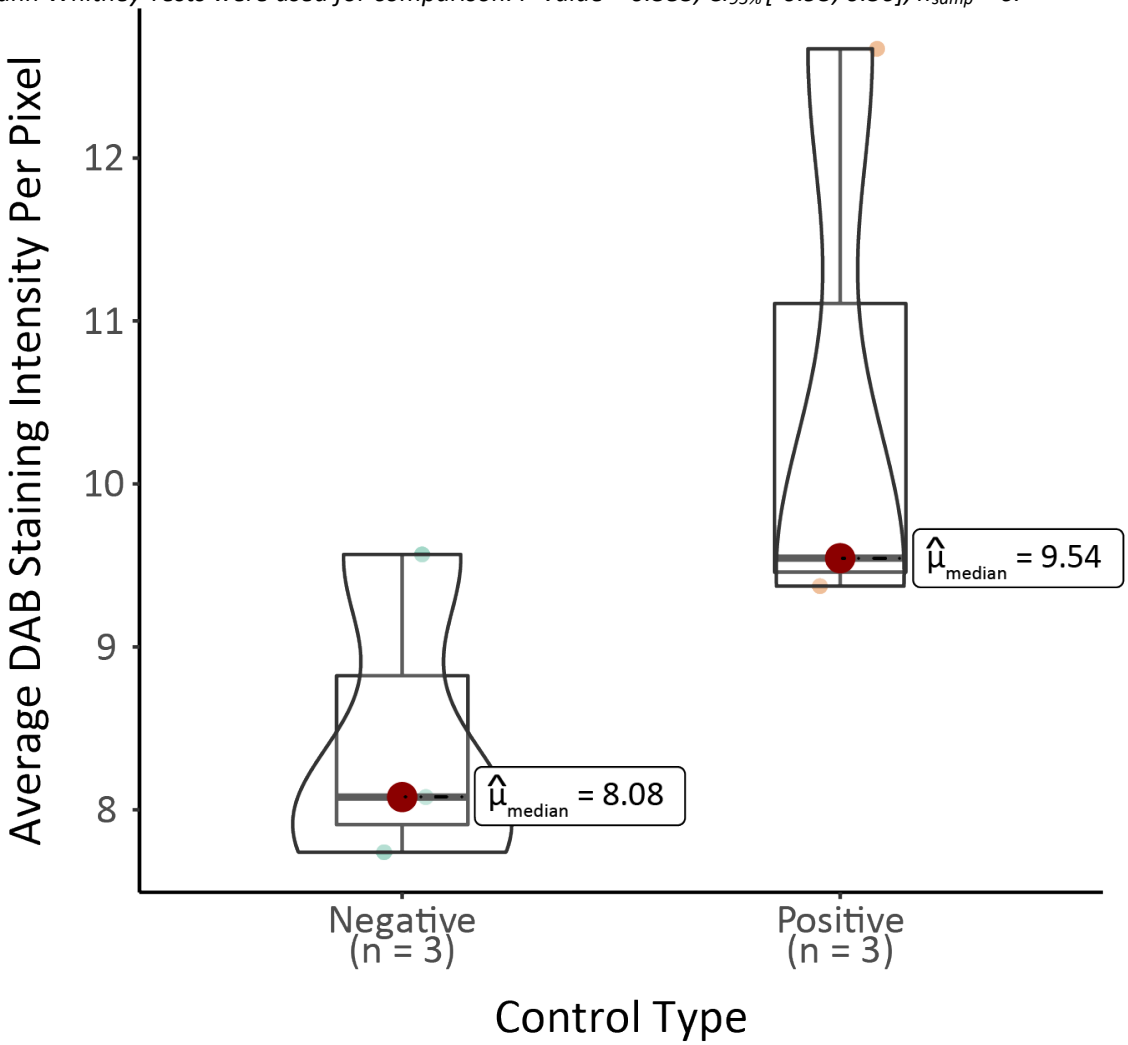
Geneset Name	NGenes	FDR	Direction	Cytoscape Group
GO:0071219 cellular response to molecule of bacterial origin	166	5.48e-20	hi-CIRC	Response to Bacteria
GO:0002237 response to molecule of bacterial origin	281	1.37E-17	hi-CIRC	Response to Bacteria
GO:0009617 response to bacterium	452	2.76E-17	hi-CIRC	Response to Bacteria
GO:0016045 detection of bacterium	12	6.65E-12	hi-CIRC	Detection of Bacterial Product
GO:0042742 defense response to bacterium	150	6.65E-12	hi-CIRC	Response to Bacteria
GO:0050830 defense response to Gram-positive bacterium	56	9.38E-11	hi-CIRC	Response to Bacteria
GO:0009615 response to virus	298	2.81E-09	hi-CIRC	Response to Virus
GO:0051607 defense response to virus	213	1.35E-08	hi-CIRC	Response to Virus
GO:0032490 detection of molecule of bacterial origin	11	1.01E-07	hi-CIRC	Detection of Bacterial Product
GO:0070339 response to bacterial lipopeptide	6	1.11E-07	hi-CIRC	Detection of Bacterial Product
GO:0071220 cellular response to bacterial lipoprotein	6	1.11E-07	hi-CIRC	Detection of Bacterial Product
GO:0071221 cellular response to bacterial lipopeptide	6	1.11E-07	hi-CIRC	Detection of Bacterial Product
GO:0032493 response to bacterial lipoprotein	8	1.71E-07	hi-CIRC	Detection of Bacterial Product
GO:0050691 regulation of defense response to virus by host	37	6.47E-07	hi-CIRC	Response to Virus
GO:0046718 viral entry into host cell	126	1.92E-06	hi-CIRC	Regulation of Viral Entry/Activity
GO:0002230 positive regulation of defense response to virus by host	28	2.35E-06	hi-CIRC	Response to Virus
GO:0050688 regulation of defense response to virus	62	3.18E-06	hi-CIRC	Response to Virus
GO:0046596 regulation of viral entry into host cell	35	6.16E-06	hi-CIRC	Regulation of Viral Entry/Activity
GO:0045071 negative regulation of viral genome replication	47	1.05E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0048525 negative regulation of viral process	78	1.91E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019730 antimicrobial humoral response	78	1.97E-05	hi-CIRC	Response to Bacteria
GO:0050829 defense response to Gram-negative bacterium	42	2.46E-05	hi-CIRC	Response to Bacteria
GO:0001618 virus receptor activity	68	3.34E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019080 viral gene expression	197	4.13E-05	lo-CIRC	Viral Gene Expression

GO:0061844 antimicrobial humoral immune response mediated by antimicrobial peptide	45	4.17E-05	hi-CIRC	Response to Bacteria
GO:0046598 positive regulation of viral entry into host cell	11	6.37E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019064 fusion of virus membrane with host plasma membrane	8	7.35E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0039663 membrane fusion involved in viral entry into host cell	8	7.35E-05	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019083 viral transcription	179	7.81E-05	lo-CIRC	Viral Gene Expression
GO:0050690 regulation of defense response to virus by virus	29	0.000251	hi-CIRC	Response to Virus
GO:1903900 regulation of viral life cycle	131	0.000564	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019065 receptor-mediated endocytosis of virus by host cell	7	0.00272	hi-CIRC	Regulation of Viral Entry/Activity
GO:0075509 endocytosis involved in viral entry into host cell	7	0.00272	hi-CIRC	Regulation of Viral Entry/Activity
GO:0098586 cellular response to virus	58	0.00402	hi-CIRC	Regulation of Viral Entry/Activity
GO:0046597 negative regulation of viral entry into host cell	18	0.00561	hi-CIRC	Regulation of Viral Entry/Activity
GO:0075522 IRES-dependent viral translational initiation	11	0.00561	lo-CIRC	Viral Gene Expression
GO:1903901 negative regulation of viral life cycle	21	0.00561	hi-CIRC	Regulation of Viral Entry/Activity
GO:0019081 viral translation	16	0.00810	lo-CIRC	Viral Gene Expression

Appendix 4 – 16s rRNA staining across positive and negative controls

Appendix relates to Figure 5.5. Comparison of positive and negative controls for 16s rRNA RNAscope probe demonstrated no differences in intensity of staining.

Mann Whitney Tests were used for comparison. P-value = 0.383, $CI_{95\%}$ [-0.93, 0.36], n_{samp} = 6.



Appendix 5 – 16s rRNA staining between normal colon and tumour

Comparison of tumour with adjacent healthy colon tissue demonstrates no differences in intensity of 16s rRNA RNAscope probe staining.

Statistical analysis used a Kruskal-Wallis test, p-value = 0.244. $CI_{95\%}$[0.12, 1]. Pairwise investigations used Dunn's test. All comparisons are shown after Holm correction. Intra-subtype comparisons between normal and tumour slides were performed. Total number of samples = 43. MSI-H Normal – 7.95, MSI-H Tumour – 7.27, MSS-lo-CIRC Normal – 8.90, MSS-lo-CIRC Tumour – 7.22, MSS-hi-CIRC Normal – 8.23, MSS-lo-CIRC Tumour – 7.87.

