

Biomarkers of Mismatch Repair Deficiency in Colorectal Cancer and Cancer Predisposition Syndromes

Richard John Gallon

A thesis submitted to Newcastle University
for the degree of Doctor of Philosophy

Institute of Genetic Medicine
Faculty of Medical Sciences
Newcastle University

September 2018

Abstract

Colorectal cancer (CRC) is the third most common cancer in Western societies and approximately 15% are mismatch repair deficient (MMRd). MMRd CRCs have a distinct prognosis, respond to immunotherapy, and occur at a high rate in patients with Lynch syndrome or constitutional mismatch repair deficiency (CMMRD). Detection of MMR deficiency, therefore, guides treatment and identification of associated cancer-predisposition syndromes. However, there is a need for novel biomarkers to detect MMRd CRC, and innovative assays to improve Lynch syndrome and CMMRD diagnosis.

I assessed autoantibodies generated against MMRd CRCs as a liquid-biopsy biomarker for cancer detection, by analysing the sera of 464 Lynch syndrome gene carriers using a recently published, multiplex method. Although autoantibodies correlated with a history of CRC, a lack of signal from patients who developed CRC shortly after sampling suggests the method has poor sensitivity. Microsatellite instability (MSI) is an established biomarker of MMR deficiency. I used single molecule molecular inversion probes to develop a sequencing-based MSI assay with an automated results analysis, suitable as a companion diagnostic for immunotherapy, and for streamlined Lynch syndrome screening. The assay achieved 100% accuracy in 197 CRCs, and was robust to sample variables, including quantity, quality, and tumour cell content. Subsequently, I adapted the MSI assay to detect low-level MSI in non-neoplastic tissues of CMMRD patients. The assay separated all 32 CMMRD patients from 94 controls. For both CRC and CMMRD diagnostics, the MSI assay is cheaper and faster than current methods, and is scalable to large cohorts.

These results suggest that the humoral immune response to MMRd CRCs cannot readily be used as a biomarker to detect disease, and that alternatives should be sought. However, the MSI assay could be deployed into clinical practice to meet the high demand for MMR deficiency testing of CRCs and to improve CMMRD diagnostics.

Acknowledgements

I would like to thank my supervisors Prof Sir John Burn, Dr Mike Jackson, and Dr Mauro Santibanez-Koref for maintaining the direction of my work, providing inspiration, and facilitating my growth as a scientist. In particular, I thank John for his clinical insight and for connecting me with key collaborators. I thank Mike and Mauro for their day-to-day supervision, their patience, and their invaluable feedback; they have been diligent tutors. I also thank the Barbour Foundation for funding this PhD.

I thank Dr Harsh Sheth, who has been a valued colleague to exchange ideas with, and with whom I worked closely on the MSI assay and related projects. I thank Dr Lisa Redford and Dr Ghanim Alhilal for their work on the MSI assay in previous years. I thank Ottie O'Brien and Amanda Waltham from the Northern Genetics Service, Newcastle, for their provision of clinical samples and data.

I would like to thank all members of the CaPP3 clinical trial, locally and at genetics centres across the UK, for the data and samples that they have provided. From the CaPP3 Central Study Team, Newcastle, I thank Dr Gill Borthwick for organising research collaborations, ethics, and budget. I thank Christine Hayes for innumerable contributions, she has been a constant support in all aspects of this work. I thank Lynn Reed, Lynne Longstaff, Donna Job, and Jackie Greenwood for answering my CaPP3 queries and welcoming me into the team. I thank Amy McAllister for her assistance organising meetings and conference attendance.

Collaboration with other research institutions was critical to this PhD. I thank Dr Matthias Kloor and Prof Magnus von Knebel Döberitz for being fantastic hosts during my placement at the Department of Applied Tumour Biology, Heidelberg University Hospital. I thank Matthias, Dr Miriam Reuschenbach, Dr Aysel Ahadova and Jonathan Dörre for generating antibody titre data, and their collaboration on related projects not presented in this thesis. I thank Dr Katharina Wimmer, Barbara Mühlegger, and others at the Division of Human Genetics, Medical University of Innsbruck, for their collaboration adapting the MSI assay to detect CMMRD, particularly Katharina for sourcing the CMMRD samples.

Finally, I thank any individual not named here who has contributed to my PhD in any way, such as members of the Institute of Genetic Medicine and Newcastle University administrative teams, and friends and family for their support.

Table of Contents

Abstract	i
Acknowledgements	iii
Table of Contents	v
List of Figures	xi
List of Tables	xiii
List of Abbreviations	xv
Chapter 1. Introduction	1
1.1. Maintaining Genomic Stability through Mismatch Repair	1
1.2. Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Cancer and Implications for Prognosis and Response to Chemotherapy	3
1.3. Microsatellite Instability generates Frameshift Mutations that drive Tumorigenesis	6
1.4. Frameshift Mutations stimulate an Anti-Tumour Immune Response	10
1.5. Mismatch Repair Deficient Cancers respond to Immunotherapy	11
1.6. Cancer-predisposition Syndromes associated with Mismatch Repair Deficiency	14
1.6.1. Lynch syndrome biology and clinical management	15
1.6.2. Constitutional mismatch repair deficiency biology and clinical management	21
1.7. The Utility of Biomarkers in Colorectal Cancer and Cancer-predisposition Syndromes	25
1.8. Biomarkers for the Early Detection of Colorectal Cancer	28
1.9. Early Detection of Mismatch Repair Deficient Colorectal Cancer using Autoantibodies	30
1.10. Mismatch Repair Deficiency testing to identify Cancer-predisposition Syndromes	33
1.10.1. Diagnosing Lynch syndrome	33
1.10.2. Diagnosing constitutional mismatch repair deficiency	36
1.11. The Inadequacies of Current Biomarker Tests for Mismatch Repair Deficiency	39
1.12. Detection of Microsatellite Instability using Next Generation Sequencing	43

1.13. Summary and Aims	48
Chapter 2. Materials and Methods	51
2.1. Ethical Approval for Research Conducted	51
2.2. Human Tissue and DNA Samples	51
2.2.1. Samples for the assessment of immunological biomarkers	51
2.2.2. Samples for development of a smMIP-based MSI assay for cancer diagnostics	51
2.2.3. Samples for development of a smMIP-based MSI assay for constitutional mismatch repair deficiency	52
2.3. Cell Line Samples and Cell Culture Protocols	52
2.4. DNA Extraction	53
2.5. DNA Quantification and Dilution	53
2.6. Generation of Samples containing Known Proportions of MSI-high DNA	54
2.7. Detection of Frameshift Peptide Serum Reactivity	54
2.7.1. Generation of median fluorescence intensity data for frameshift peptides	54
2.7.2. Analysis of serum reactivity against frameshift peptides	56
2.8. Design of Single Molecule Molecular Inversion Probes	56
2.8.1. Selection of marker loci for the smMIP-based MSI assay	56
2.8.2. Design of smMIPs using MIPgen	56
2.8.3. Validation of smMIP designs and amplicons	57
2.9. Single Molecule Molecular Inversion Probe Amplification Protocol	58
2.9.1. Probe phosphorylation	58
2.9.2. Target capture and amplification	58
2.10. Library Preparation for Amplicon Sequencing	59
2.11. smMIP Amplicon Sequencing on the Illumina MiSeq	59
2.12. Sequencing Read Analysis	60
2.12.1. Generation of Marker Result tables	60
2.12.2. MSI classification using a naïve Bayes approach	61
2.12.3. Read and variant counting	61
2.12.4. Hotspot mutation calling	61

2.12.5. CMMRD classification	61
2.13. Germline Confirmation of MSH6 c.3557-1G>C Mutation	62
2.14. Statistical Analyses and Graphics	62
Chapter 3. The Utility of Anti-Frameshift Peptide Antibodies in the Serum to detect Mismatch Repair Deficient Colorectal Cancer	63
3.1. Introduction	63
3.2. Aims	65
3.3. Cohort Description and Justification of Method	65
3.4. Subtraction-based Normalisation does not equalise Baseline FSP Serum Reactivity	68
3.5. Regression-based Normalisation equalises Baseline FSP Serum Reactivity	72
3.6. Frameshift Peptides cluster by Serum Reactivity	76
3.7. A History of Colorectal Cancer is associated with Frameshift Peptide Serum Reactivity	77
3.8. Discussion	81
3.9. Conclusions and Future Work	85
Chapter 4. Development of a Short Mononucleotide Repeat Sequencing Assay to Detect Microsatellite Instability in Colorectal Cancer	87
4.1. Introduction	87
4.2. Aims	90
4.3. Multiplex Amplification of Microsatellites using Molecular Inversion Probes	91
4.4. Amplicon Sequencing identifies Variants in Control Samples	91
4.5. Training the MSI Classifier	101
4.6. MSI Classification is Accurate and Reproducible	106
4.7. MSI Classification is Robust to Low MSI-high Content	107
4.8. MSI Classification is Reliable from sequencing 75 Molecules per Marker	110
4.9. <i>BRAF</i> and <i>KRAS</i> Mutations are Reproducibly Detected	112
4.10. Assay Cost and Turnaround Time are Superior to Established Methods	115
4.11. Discussion	116
4.12. Conclusions and Future Work	121

Chapter 5. Accurate Detection of Constitutional Mismatch Repair Deficiency by a Sequencing-based Microsatellite Instability Assay	123
5.1. Introduction	123
5.2. Aims	124
5.3. Study Samples and Method	124
5.4. Single Molecule Reads reduce Error in Microsatellite Length Variant Detection	125
5.5. CMMRD Samples are identifiable by Deviation in Microsatellite Lengths from Controls	130
5.6. CMMRD Samples are Identifiable with High Accuracy	133
5.7. Identification of Contamination in a Control Sample	137
5.8. Discussion	140
5.9. Conclusions and Future Work	143
Chapter 6. General Discussion and Future Work	145
6.1. The Clinical Utility and Analytical Validity of Mismatch Repair Deficiency Biomarkers and Tests in Cancer Diagnostics	145
6.1.1. Anti-frameshift peptide antibodies as a liquid biopsy biomarker of colorectal cancer	145
6.1.2. A sequencing-based microsatellite instability assay for colorectal cancer diagnostics	146
6.1.3. A sequencing-based microsatellite instability assay to detect constitutional mismatch repair deficiency	147
6.2. The Future Direction of Biomarker Tests for Mismatch Repair Deficiency	149
6.2.1. Surveillance for mismatch repair deficient cancers in high risk populations	149
6.2.2. Microsatellite instability testing in cancer diagnostics	151
6.2.3. Microsatellite instability testing of non-neoplastic tissues	155
6.3. Concluding Remarks	159
Chapter 7. Appendices	161
7.1. Appendix A: Colorectal Cancer Sample Data and Source	161
7.2. Appendix B: Constitutional Mismatch Repair Deficiency and Control Patient Samples	163
7.3. Appendix C: Frameshift Peptides analysed for Serum Reactivity	169

7.4. Appendix D: Marker Loci for the smMIP and Sequencing-based MSI Assay	170
7.5. Appendix E: Sequences of Molecular Inversion Probes, PCR Primers, and Sequencing Primers	171
7.6. Appendix F: Example Sample Sheet for MiSeq Loading	175
7.7. Appendix G: PCR Primer Sequences for MSH6 c.3557-1G>C	175
7.8. Appendix H: Distribution of Frameshift Peptide versus FLAG-only Control Median Fluorescence Intensity	176
7.9. Appendix I: Read-balancing the Multiplex Pool of Molecular Inversion Probes	183
7.10. Appendix J: Reagent Costs of the smMIP and Sequencing-based MSI Assay	185
7.11. Appendix K: Modelling the Proportion of smSequences with WT Microsatellite Length by the Beta Distribution	189
7.12. Appendix L: Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. Ahadova, Gallon et al, 2018.	191

Note: The respective manuscript is appended to Section 7.12 as a distinct publication, and hence it follows its own page numbering. Chapter 8 continues the page numbering of the rest of this document.

Chapter 8. References	193
------------------------------	------------

List of Figures

Figure 1.1. Simple schematic of the mismatch repair system.	2
Figure 1.2. The classical pathway of genetic changes in colorectal tumorigenesis.	7
Figure 1.3. Frameshift mutations identified in TGF β R2.	9
Figure 1.4. The mechanism of action of immune checkpoint blockade by pembrolizumab in mismatch repair deficient cancers.	13
Figure 1.5. The colorectal cancer syndromes within hereditary non-polyposis colorectal cancer.	15
Figure 1.6. The cumulative cancer risks in Lynch syndrome gene carriers by age 75 years.	16
Figure 1.7. Mismatch repair deficient adenoma associated with mismatch repair deficient colorectal crypt focus.	18
Figure 1.8. Comparative histology of colorectal adenomas.	19
Figure 1.9. The ages of 197 constitutional mismatch repair deficiency patients at cancer diagnosis.	23
Figure 1.10. Microsatellite instability (MSI) detection by fragment length analysis.	41
Figure 1.11. Utilising molecular barcodes in next generation sequencing to discriminate true variants from PCR and sequencing errors.	47
Figure 2.1: Design of PCR primers to verify sequence content of smMIP amplicons.	58
Figure 2.2: Example Marker Result table.	60
Figure 3.1: Detection of antibody signal against a panel of 32 frameshift peptides.	64
Figure 3.2: A multiplex method of detecting anti-frameshift peptide antibodies in serum (Reuschenbach et al, 2014).	67
Figure 3.3: Data structure for median fluorescence intensity.	70
Figure 3.4: Distribution of serum reactivity between frameshift peptides and between patients.	71
Figure 3.5: Distribution of serum reactivity between frameshift peptides and between patients.	73
Figure 3.6: Distribution of serum reactivity between frameshift peptides.	74
Figure 3.7: Distribution of serum reactivity between frameshift peptides in positive control samples.	75
Figure 3.8: Frameshift peptide clustering by serum reactivity.	76
Figure 3.9: Correlation of patient variables.	78

Figure 3.10: Distribution of serum reactivity in patients with or without a history of colorectal cancer.	80
Figure 4.1: Protocol for multiplex loci capture, molecular barcode tagging, and amplification by single molecule molecular inversion probes.	89
Figure 4.2: Amplification of marker loci using single molecule molecular inversion probes.	92
Figure 4.3: The distribution of reads to different lengths of microsatellite in the 25 markers in 4 samples.	96-97
Figure 4.4: Distribution of reads by microsatellite length and allele.	98
Figure 4.5: Detection of <i>RAS/RAF</i> mutations.	100
Figure 4.6: Distribution of training cohort samples by the relative frequency of microsatellite deletions	102
Figure 4.7: Distribution of samples relative to classifier thresholds.	104
Figure 4.8: Self-classification of the training cohort.	106
Figure 4.9: MSI classifier validation.	107
Figure 4.10: The robustness of the MSI classifier to low MSI-high content.	109
Figure 4.11: Visualisation of smMIP amplicons.	111
Figure 4.12: Sequencing results and classification of low quantity samples.	113
Figure 4.13: Variant allele frequency compared between repeat testing of 32 CRCs.	115
Figure 5.1: The frequency distribution of molecular barcode groups by the number of reads within each group.	125
Figure 5.2: Definition of single molecule sequences.	126
Figure 5.3: Count of reads with different microsatellite lengths, using either all reads irrespective of molecular barcode, or single molecule sequences.	127
Figure 5.4: Using single molecule sequences reduces the error in detection of variants in microsatellite length.	129
Figure 5.5: Modelling the distribution of the proportion of smSequences containing wild type length of microsatellite.	131
Figure 5.6: Confirmation of the identity of sample 99.	134
Figure 5.7: Score distribution of CMMRD and control samples.	135
Figure 5.8: The proportion of reads assigned to different SNPs in sample 7 repeats.	139
Figure 7.1: Distribution of frameshift peptide versus FLAG-only control median fluorescence intensity.	176-182
Figure 7.2: Comparison of Beta and empirical distributions of microsatellite length variants in controls.	189-190

List of Tables

Table 1.1. Surveillance recommendations for constitutional mismatch repair deficiency.	25
Table 1.2. The utility of biomarkers in colorectal cancer-related healthcare.	27
Table 1.3. Clinical criteria for Lynch syndrome screening.	34
Table 1.4. Clinical criteria for constitutional mismatch repair deficiency (CMMRD) screening.	37
Table 2.1: Generation of samples with varying MSI-high DNA content.	54
Table 3.1: Patient details of the Lynch syndrome gene carrier cohort.	66
Table 3.2: Median regression of each frameshift peptide compared to the FLAG-only control.	72
Table 3.3: Multivariate analysis of frameshift peptide serum reactivity and patient variables.	78
Table 3.4: Regression coefficients from multiple logistic regression of patient colorectal cancer history and frameshift peptide serum reactivity.	80
Table 4.1: Statistics from the two sequencing runs.	94
Table 4.2: Mean reads detected per marker per sample from the two sequencing runs.	94
Table 4.3: Detection of allelic bias of deletions in the microsatellite markers.	99
Table 4.4: The proportion of samples in which allelic bias can be assessed for each marker.	103
Table 4.5: Microsatellite instability classification by fragment length analysis of DNA-mixtures of varying MSI-high DNA content.	110
Table 4.6: Frequency of <i>BRAF</i> and <i>KRAS</i> mutations in colorectal cancers.	114
Table 4.7: Summary of cost analysis and turnaround time.	116
Table 4.8: Quality controls for the smMIP-based MSI assay for reliable classification.	122
Table 5.1: Conversion of observed proportion of smSequences containing a wild type microsatellite length to a probability and per sample score.	132
Table 5.2: Genotype, sample scores and gMSI results of the 32 patients with constitutional mismatch repair deficiency.	136
Table 5.3: Repeat testing of three control samples as a quality control.	137
Table 5.4: Summary of read data from sample 7 in two sequencing runs.	138
Table 7.1: CRCs in the classifier training cohort.	161

Table 7.2: CRCs in the assay validation cohort CRCs.	162
Table 7.3: Clinical details and test results of constitutional mismatch repair deficiency and control samples.	163-167
Table 7.4: Synthetic frameshift peptides and their amino acid sequence.	169
Table 7.5: Marker loci for the smMIP and sequencing-based MSI assay.	170
Table 7.6: Oligonucleotide sequences of all probes and primers used in the development of the smMIP and sequencing-based MSI assay.	171-173
Table 7.7: Example sample sheet for MiSeq loading.	175
Table 7.8: Primer sequences for amplification of the <i>MSH6</i> c.3557-1G>C locus.	175
Table 7.9: Read-balancing the multiplex pool of molecular inversion probes.	183
Table 7.10: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v2 Micro Kit and 25 markers.	185
Table 7.11: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v3 Kit and 25 markers.	186
Table 7.12: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v2 Micro Kit and 10 markers.	187
Table 7.13: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v3 Kit and 10 markers.	188

List of Abbreviations

5-FU	5-fluorouracil
α FSP-Abs	anti-frameshift peptide antibody
<i>ACVR2A</i>	activin A receptor type 2A
<i>AIM2</i>	absent in melanoma 2
ALL	acute lymphoblastic leukaemia
<i>APC</i>	adenomatous polyposis coli
<i>APLNR</i>	apelin receptor
<i>ASTE1</i>	asteroid homolog 1
AUC	area under curve
β 2M	beta 2 microglobulin
<i>BANP</i>	BTG3 associated nuclear protein
<i>BAX</i>	BCL2 associated X
<i>BRAF</i>	V-Raf murine sarcoma viral oncogene homolog B
CAPP/CaPP	Cancer Prevention Program
CBS-K	Super ChemiBlock Heterophile Blocking Agent
<i>CDX2</i>	caudal type homeobox 2
CEA	carcinoembryonic antigen
cfDNA	cell free DNA
CIMP	CpG island methylator phenotype
CIN	chromosomal instability
CMMRD	constitutional mismatch repair deficiency
cMNR	coding mononucleotide repeat
CMS	consensus molecular subtype
CNA	copy number alteration
CNS	central nervous system
CRC	colorectal cancer
CT	computed tomography
ctDNA	circulating tumour DNA

CTL	cytotoxic lymphocyte
CTLA-4	cytotoxic T-lymphocyte-associated antigen 4
<i>CTNNB1</i>	catenin beta 1
<i>DCC</i>	deleted in colorectal cancer
DDR	DNA damage response
DG27	Diagnostic Guidance 27
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
DNR	dinucleotide repeat
dNTP	deoxy nucleoside triphosphate
<i>DYPD</i>	dihydropyrimidine dehydrogenase
EC	endometrial cancer
EDTA	ethylene-diamine-tetra-acetic acid
<i>EGFR</i>	epidermal growth factor receptor
ELISA	enzyme-linked immunosorbent assay
<i>EPCAM</i>	epithelial cell adhesion molecule
FAP	familial adenomatous polyposis
FCCTX	familial colorectal type X
FDA	Food and Drugs Administration
FFPE	formalin fixed paraffin embedded
FIT	faecal immunochemical testing
FLA	fragment length analysis
FSP	frameshift peptide
GI	gastrointestinal
HLA	human leukocyte antigen
HNPCC	hereditary non-polyposis colorectal cancer
HR	hazard ratio
HRM	high resolution melt curve analysis
IDL	insertion-deletion loop
IgG	immunoglobulin G
IHC	immunohistochemistry

indel	insertion-deletion mutation
InSiGHT	International Society for Gastrointestinal Hereditary Tumours
IRR	incident rate ratio
<i>JAK1</i>	Janus kinase 1
<i>KRAS</i>	Kirsten rat sarcoma viral oncogene homolog
LCL	lymphoblastoid cell lines
<i>LMAN1</i>	lectin, mannose binding 1
MANCOVA	Multivariate Analysis of Covariance
MFI	median fluorescence intensity
<i>MGMT</i>	methylguanine methyltransferase
MHC	major histocompatibility complex
MIP	molecular inversion probe
<i>MLH1</i>	MutL-Homolog 1
MMR	mismatch repair
MMRd	mismatch repair deficient
MMR-DCF	mismatch repair deficient crypt foci
MMRp	mismatch repair proficient
MNNG	N-Methyl-N'-Nitro-N-Nitrosoguanidine
MNR	mononucleotide repeat
MRI	magnetic resonance imaging
<i>MSH2</i>	MutS-Homolog 2
<i>MSH3</i>	MutS-Homolog 3
<i>MSH6</i>	MutS-Homolog 6
MSI	microsatellite instability
MSI-high	high levels of microsatellite instability
MSI-low	low levels of microsatellite instability
MSS	microsatellite stable
<i>MUM1</i>	melanoma associated antigen (mutated) 1
<i>MUTYH</i>	MutY DNA glycosylase
NCI	National Cancer Institute
<i>NF-1</i>	neurofibromatosis type 1

NGS	next generation sequencing
NHL	non-Hodgkin's lymphoma
NICE	National Institute of Health and Care Excellence
<i>NRAS</i>	neuroblastoma RAS viral oncogene homolog
OR	odds ratio
PBL	peripheral blood leukocytes
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PD-1	programmed cell death protein 1
PD-L1	programmed cell death protein ligand 1
PE	phycoerythrin
<i>PMS1</i>	post-meiotic segregation 1
<i>PMS2</i>	post-meiotic segregation 2
<i>POLE</i>	DNA polymerase epsilon
prWT	proportion of smSequences with WT microsatellite length
QALY	Quality Adjusted Life Year
QC	Quality Control
RNA	ribonucleic acid
ROC	receiver operating characteristic
<i>SLC22A9</i>	solute carrier family 22 member 9
<i>SMAP1</i>	small ArfGAP 1
smMIP	single molecule molecular inversion probe
smSequence	single molecule sequence
SNP	single nucleotide polymorphism
sPNET	supratentorial primitive neuroectodermal tumours
STARD	standards for reporting diagnostic accuracy
TAA	tumour-associated antigen
<i>TAF1B</i>	TATA-box binding protein associated factor, RNA polymerase I subunit B
TAT	turnaround time
<i>TGFBR2</i>	transforming growth factor- β receptor 2
TIL	tumour infiltrating lymphocyte

<i>TP53</i>	tumor protein P53
TSG	tumour suppressor gene
VAF	variant allele fraction
VUS	variant of unknown significance
WT	wild type

Chapter 1. Introduction

1.1. Maintaining Genomic Stability through Mismatch Repair

Genomic instability describes an abnormally high rate of change in the genome of an organism or cell, including large structural aberrations and alterations in the DNA base sequence (Negrini *et al*, 2010). Genomic instability can enable tumour growth through cellular mutation and acquisition of cancer hallmarks, including uncontrolled proliferation, evasion of cell death, angiogenesis and tissue invasion (Hanahan and Weinberg, 2011). Complex organisms have therefore evolved robust mechanisms to avoid and repair DNA damage to maintain genomic stability. One such mechanism, within the network of the DNA damage response (DDR), is the mismatch repair (MMR) system (Jackson and Bartek, 2009).

MMR is a multistep process that repairs base-base mismatches and insertion-deletion loops (IDLs), which are frequently generated by polymerase error during DNA replication. Lack of repair of these lesions produces substitution and insertion-deletion mutations (indels), respectively (Jiricny, 2006). MMR is conserved throughout evolution and was first characterised in *Escherichia coli* and *Saccharomyces cerevisiae*. In prokaryotes, the mismatch or IDL is recognised by MutS, a homodimer that binds to the damaged site and recruits a similar homodimeric “DNA-clamp” MutL. The complex of MutS and MutL coordinates repair through accessory enzymes, involving exonuclease excision of the nascent DNA strand, synthesis of an undamaged replacement by DNA polymerase III and nick-sealing by DNA ligase (Kunkel and Erie, 2005; Jiricny, 2006). The key MMR proteins in mammals are MutS-Homologs 2, 3 and 6 (MSH2, MSH3, MSH6), MutL-Homolog 1 (MLH1), and Post-Meiotic Segregation 2 (PMS2). These form heterodimers equivalent to bacterial MutS and MutL, specifically MutS α (MSH2-MSH6), MutS β (MSH2-MSH3) and MutL α (MLH1-PMS2) (Kunkel and Erie, 2005; Jiricny, 2006). MutS α and MutS β recognise and initiate repair of lesions of different sizes. MutS α is required for the repair of base-base mismatches and single nucleotide IDLs: both tumours and cell lines deficient in MSH6 have a high frequency of substitutions and single, but not multiple, nucleotide indels, repair of which can be restored by extracts containing MutS α (Drummond *et al*, 1995; Verma *et al*, 1999; Wu *et al*, 1999). In contrast, MutS β efficiently repairs a range of larger IDLs, but not the mismatches and single-bp IDLs repaired by MutS α (Figure 1.1) (Genschel *et al*, 1998). Whilst active throughout the cell cycle, MMR proteins accumulate and show highest activity during S-

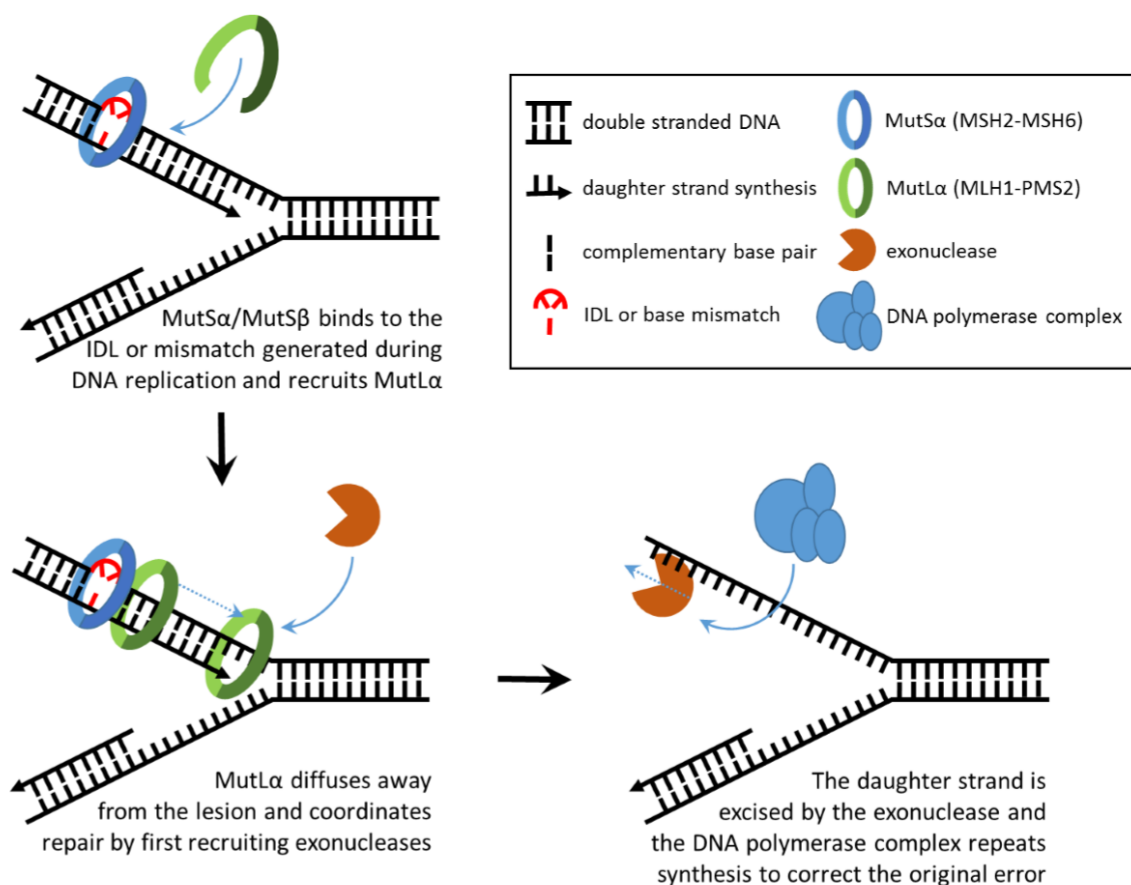


Figure 1.1. Simple schematic of the mismatch repair system. Insertion-deletion loops (IDLs) and mismatches are generated by polymerase error during DNA replication, distorting the DNA structure. These lesions are bound by the MutS heterodimer (MutS α or MutS β , depending on size of lesion), which clamps onto the DNA. The MutL α heterodimer is recruited by MutS and associates with the DNA to form a “sliding clamp”. The MutL α sliding clamp dissociates from MutS and diffuses along the replicating DNA duplex. It is theorised that association of MutL α with DNA replication machinery recruits additional repair proteins and coordinates exonuclease excision of the daughter strand. Subsequent to excision, DNA polymerase complex is recruited and synthesises a new daughter strand and in doing so corrects the IDL or mismatch (Jiricny, 2006).

phase to increase the fidelity of DNA replication (Edelbrock *et al*, 2009) and signal to the wider DDR to coordinate the cellular response to damage, for example through p53 to arrest the cell cycle at the G2/M checkpoint and promote apoptosis when damage persists (Aquilina *et al*, 1999; Hickman and Sansom, 1999). MutL α and other MutS and MutL homologs also play roles in recombination and mammalian meiosis (Lipkin *et al*, 2002).

Microsatellites are tandem repeats of short DNA sequences (1-6bp) that occur at hundreds of thousands of loci throughout the human genome. Microsatellites can be subdivided into mono-, di-, tri-, tetra-, penta- and hexanucleotide repeats depending on the length of the repeat unit (Subramanian *et al*, 2003; Ellergren, 2004). They are highly mutable

with three proposed mutational mechanisms: (1) polymerase slippage during DNA replication creating IDLs that are stabilised by the repetitive sequence (Kornberg *et al*, 1964), (2) homology-driven incorporation of retrotransposons (Nadir *et al*, 1996), and (3) unequal crossing over in meiosis (Huang *et al*, 2002). The degree of mutability depends on several factors including genomic locus and the structure of the microsatellite, such as the unit sequence and the number of repeats (Bacolla *et al*, 2008; Kelkar *et al*, 2008). Taking these factors into account, *in vivo* (Strand *et al*, 1993), *in vitro* (Schlötterer and Tautz, 1992) and *in silico* (Dieringer and Schlötterer, 2003) analyses have all shown that polymerase slippage is the best model for microsatellite mutation rate and, hence, it is accepted as the predominant mechanism (Fan and Chu, 2007). “Microsatellite instability” (MSI) is the term used to define this mutability, and is measurable by the rate at which indels are acquired in microsatellites. MMR reduces MSI by three orders of magnitude through repair of IDLs generated by polymerase slippage, ensuring accurate replication of microsatellites (Strand *et al*, 1993; Koi *et al*, 1994; Umar *et al*, 1997; Herman *et al*, 1998; Deng *et al*, 1999). Increased MSI is a well-established biomarker of MMR deficiency in human disease.

1.2. Mismatch Repair Deficiency and Microsatellite Instability in Colorectal Cancer and Implications for Prognosis and Response to Chemotherapy

Colorectal cancer (CRC) is the third most common cancer in Western society and is the second highest cause of cancer-related mortality (Siegel *et al*, 2017). In the early 1990s it was discovered that approximately 15% of CRCs had indels in an exceptionally large number of microsatellite loci, indicative of a particularly high rate of MSI – a phenotype denoted as MSI-high (Thibodeau *et al*, 1998). These cancers also had a diploid karyotype and therefore lacked the chromosomal instability (CIN) seen in the majority of CRCs. MSI-high CRCs were characterised by a better prognosis, an increase in tumour infiltrating lymphocytes (TILs), a higher frequency of proximal (right-sided) location and poorer cellular differentiation relative to microsatellite stable (MSS) CRCs, suggesting that they belonged to a distinct pathway of tumorigenesis (Ionov *et al*, 1993; Lothe *et al*, 1993; Thibodeau *et al*, 1993).

The MSI-high phenotype occurs in both hereditary non-polyposis CRC (HNPCC) and sporadic CRC (Aaltonen *et al*, 1993). To understand the origins of this phenotype, loss of function mutations were introduced into the MMR genes *MLH1*, *MSH2* and *PMS1* in yeast and were shown to increase the frequency of indels at microsatellite loci (Strand *et al*, 1993).

An association between MMR deficiency and increased MSI in CRC was also established when human *MSH2* was mapped to chr2p22.1, a locus known to segregate with MSI-high HNPCC (Fishel *et al*, 1993). Subsequently, pathogenic mutations of other MMR genes were also discovered in MSI-high HNPCC germline and tumour DNA, including *MLH1* (Bronner *et al*, 1994), *PMS2* (Nicolaidis *et al*, 1994) and *MSH6* (Miyaki *et al*, 1997). Sporadic MSI-high CRCs were shown to have methylation silencing of the *MLH1* promoter with resulting loss of *MLH1* expression (Herman *et al*, 1998; Deng *et al*, 1999). The causative link between MMR deficiency and MSI in cancer has been further demonstrated in several human CRC cell lines. In cell lines containing *MLH1* hypermethylation, inhibition of methyl transferases and demethylation of the *MLH1* promoter restores *MLH1* expression and microsatellite stability (Herman *et al*, 1998; Deng *et al*, 1999). MSI in the HCT116 CRC cell line, which has a hemizygous nonsense *MLH1* mutation (Papadopoulos *et al*, 1994), can be reduced by transfection with human chromosome 3 from normal fibroblasts, which contains the *MLH1* locus (Koi *et al*, 1994). The same result was achieved when transfecting HEC59 and HCT15 cell lines, which have biallelic mutation of *MSH2* and *MSH6* respectively, with normal chromosome 2, which contains the *MSH2* and *MSH6* genes (Umar *et al*, 1997).

MSI is not a feature specific to MMR deficient (MMRd) CRC as MMR proficient (MMRp) CRCs can have indel mutations in a minority of microsatellites (Thibodeau *et al*, 1998). Therefore, at the National Cancer Institute (NCI) workshop in December 1997, it was agreed to designate cancers as MSS if there was no evidence of MSI, MSI-low if <30% of markers analysed were mutated, or MSI-high if $\geq 30\%$ of markers analysed were mutated (Boland *et al*, 1998; Thibodeau *et al*, 1998). It was unknown if MSI-low tumours represented another pathway of tumorigenesis as there is evidence that they have slightly worse prognosis than MSS tumours; for example, the cancer-specific survival hazard ratio (HR) of MSI-low tumours relative to MSS tumours was 2.0 (95% CI: 1.1-3.6) in a study of 209 MMRp CRCs (Wright *et al*, 2005). However, analyses of large panels of microsatellites (up to 377 markers) showed that up to 79% of MMRp CRCs could be classed as MSI-low and that MSI-low samples had no unique clinical or molecular features relative to MSS samples irrespective of classification thresholds (Halford *et al*, 2002; Laiho *et al*, 2002). The conclusion of these studies is that MSS and MSI-low tumours originate from the same tumorigenesis pathways, with variation caused by the evolutionary history of the cancer and chance mutation; hence only MSI-high is a recognised biomarker of MMR deficiency.

The improved survival of MMRd CRC patients was evident in many early studies (Aaltonen *et al*, 1993; Ionov *et al*, 1993; Lothe *et al*, 1993; Thibodeau *et al*, 1993). In light of this, a meta-analysis by Popat *et al* defined an overall survival HR of 0.65 for a diagnosis of MSI-high (95% CI: 0.59-0.71) (Popat *et al*, 2005), and MMR deficiency has been confirmed as an independent prognostic indicator that enhances multivariate models of prognosis, which include established clinico-pathological features such as TNM staging (Dienstmann *et al*, 2017). A possible cause for the better prognosis is the increased immune cell infiltrate and reduced rate of metastasis of MMRd tumours, which are proposed consequences of a high mutational burden and generation of tumour associated antigens (Buckowitz *et al*, 2005). Furthermore, there is evidence that MMRd CRCs are resistant to the frontline, adjuvant chemotherapy 5-fluorouracil (5-FU), as shown by randomised trials (Ribic *et al*, 2003; Jover *et al*, 2006), and the poorer overall survival of patients treated with 5-FU observed in stage II disease (HR = 2.95; 95% CI = 1.02-8.54; p = .04) (Sargent *et al*, 2010). This is supported by a mechanism defined in cell lines, by which the MMR system is required to induce G2 cell cycle arrest and apoptosis through c-Abl/p73 α /GADD45 α following detection of 5-FU (incorporated into the DNA) mispairing with guanine (Li *et al*, 2009). However, no predictive value to MMR status for 5-FU response has been observed in other studies (Bertagnolli *et al*, 2009) and combination therapies containing 5-FU, such as irinotecan-5-FU-leucovorin, have contrarily been associated with improved disease free survival in MMRd CRCs (Bertagnolli *et al*, 2009). More recent trials of 5-FU combination therapies, covering up to 10 years follow up, also failed to identify MMR status as a significant indicator. However, it is not known if the better prognosis of MMRd cancers, their relatively low numbers in such studies, or the effect of the combined drugs, confounds the predictive value of MMR status for 5-FU therapy (André *et al*, 2015).

Promotion of apoptosis by the MMR system in response to DNA damage (Aquilina *et al*, 1999; Hickman and Sansom, 1999) also has therapeutic implications for the use of thiopurines and alkylating agents. Whilst these drugs are not used in the treatment of CRC, they are used in other cancers associated with MMRd cancer-predisposition syndromes (Wimmer *et al*, 2014; Section 1.6.2). Acute lymphoblastic leukaemia (ALL), for example, is typically treated with thiopurines (Vora *et al*, 2006), which both inhibit nucleoside metabolism to slow malignant cell growth, and are incorporated into DNA where the thiopurine lesion is recognised by MutS α to promote apoptosis (Karran and Attard, 2008). A

study of changes in gene expression associated with ALL relapse found that *MSH6* expression was inversely associated with sensitivity to mercaptopurine (a thiopurine) (Yang *et al*, 2008), suggesting that resistance to thiopurines can be acquired in ALL through evolution of MMR deficiency. Alkylating agents, such as N-Methyl-N'-Nitro-N-Nitrosoguanidine (MNNG) and temozolomide, covalently link alkyl groups to DNA bases. These alkyl-DNA adducts are mutagenic, through base-mispairing or blockage of replication, and are repaired by a variety of mechanisms, including direct reversion of alkylation, base excision repair, nucleotide excision repair, and MMR (Fu *et al*, 2012). Loss of direct repair mechanisms sensitises cells to alkylating agents; depletion of the methylguanine methyltransferase (MGMT) enzyme, for example, increases temozolomide toxicity (Zhang *et al*, 2012). MMR is necessary to induce apoptosis in response to some alkyl-DNA adducts. In brain tumours, for example, resistance to temozolomide in the absence of MGMT is acquired by additional loss of MMR. Normally, the MMR system detects mispairing between O⁶-methylguanine (the product of MNNG DNA methylation) and thymine, and initiates a futile repair cycle where thymine is repeatedly paired with O⁶-methylguanine, leading to promotion of apoptosis via the tumour suppressor p53 (Hickman and Sansom, 1999). Without MMR the base mispairing is not recognised and is tolerated (Thomas *et al*, 2017). MMR deficiency, therefore, is associated with resistance to numerous therapies for cancer, including 5-FU, thiopurines, and alkylating agents.

1.3. Microsatellite Instability generates Frameshift Mutations that drive Tumorigenesis

Genomic instability is a hallmark of cancer (Hanahan and Weinberg, 2011), but there has been contention regarding its functional relevance. There are two main arguments. First is that genomic instability is not a driver of tumorigenesis but is, instead, a passenger caused by oncogene-induced replicative stress, whereby an increased rate of cell division reduces the fidelity of DNA replication. In opposition is the idea that genomic instability is caused by an early event or mutation that leads to additional, functional mutations and, therefore, is a critical driver of tumorigenesis (Negrini *et al*, 2010). Mathematical models that test the likelihood that such mutator phenotypes contribute to tumorigenesis have shown that, assuming cancer progression to be a multi-step process requiring 4 or more events, mutator phenotypes facilitate carcinogenesis within biologically relevant timescales (Beckman and Loeb, 2006). Hence it would be expected that driver mutations in a tumour will reflect the mutational mechanism of its type of genomic instability.

The classical pathway of genetic changes in colorectal tumorigenesis was first described by Fearon and Vogelstein, including mutations in oncogenes such as *KRAS* and tumour suppressor genes (TSGs) such as *APC* and *p53* (Fearon and Vogelstein, 1990; Figure 1.2). These tumours are MMRp and their driver mutations are a combination of point mutations and frequent somatic copy number alterations (CNAs) caused by large deletions, duplications and other chromosomal rearrangements, which are characteristic of CIN.

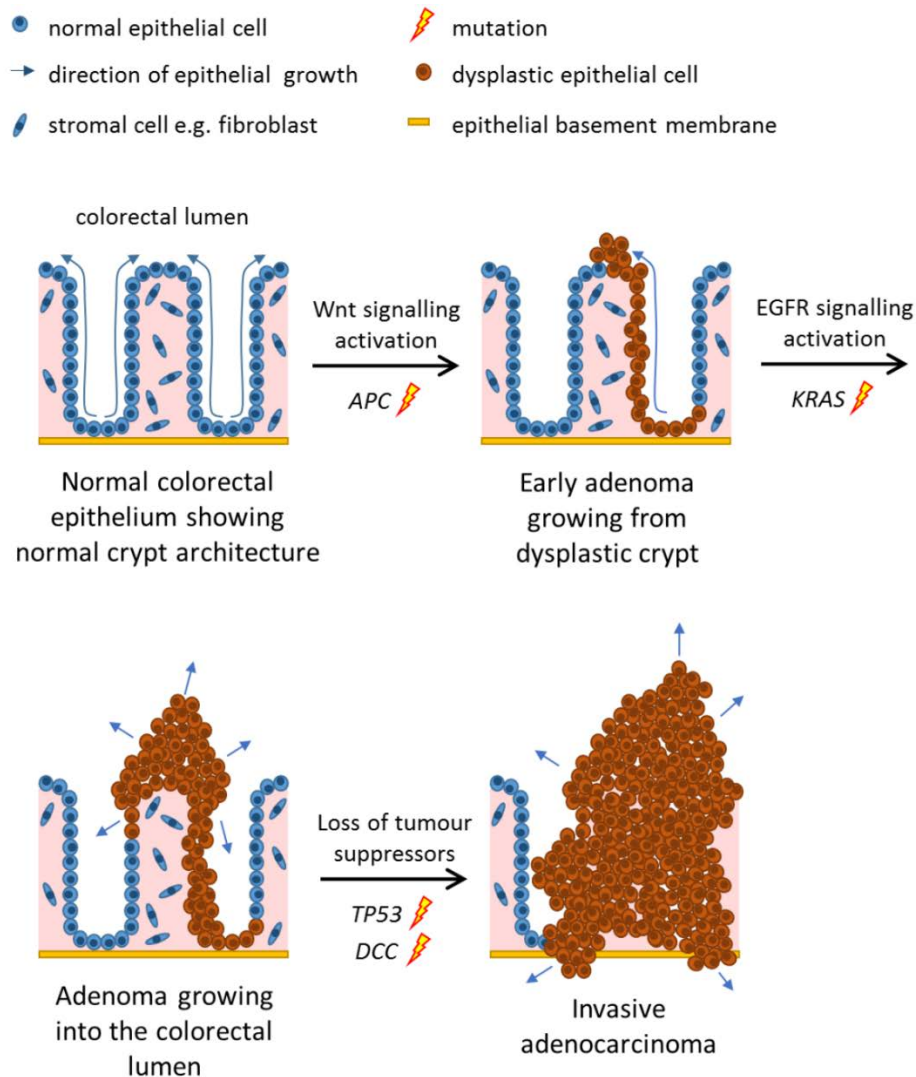


Figure 1.2. The classical pathway of genetic changes in colorectal tumorigenesis. Fearon and Vogelstein described the classical model of genetic changes that occur during colorectal tumorigenesis and their associations with each stage. In normal colorectal epithelium, stem cells in the base of the crypts reproduce to replenish the epithelium, with daughter cells moving up the walls of the crypt to the epithelium surface where they will eventually be shed and replaced. Activation of Wnt signalling by loss of function mutations in *APC* causes stem cells to over-proliferate, forming an early adenoma. Gain of function mutations in proto-oncogenes, such as *KRAS*, drive uncontrolled proliferation of epithelial cells, propagating adenoma growth. Loss of function mutations in tumour suppressor genes, such as *TP53*, and other genetic changes progress the adenoma into an invasive adenocarcinoma (Fearon and Vogelstein, 1990).

MMRd CRCs, however, lack CIN and have a diploid karyotype with a low number of CNAs. Instead, they have the MSI-high phenotype and many other small indels and point mutations, generally termed “hypermethylation” (Muzny *et al*, 2012). Furthermore, the genes mutated in MMRd CRCs are different to those mutated in CRCs with the CIN phenotype, evident in significantly lower rates of *APC* mutation (51% versus 81%, $p = 0.0023$) and *p53* mutation (20% versus 60%, $p < 0.0001$), a higher rate of *BRAF* V600E mutation, and mutations in genes such as *ACVR2A* and *TGFBR2* that are rarely seen in MMRp CRCs (Muzny *et al*, 2012). Furthermore, the type of mutations are different in MMRd versus MMRp CRCs, with a 50-fold increase in indels in coding mononucleotide repeats (cMNRs), which cause pathogenic frameshift mutations in the affected genes by introduction of early stop codons (Muzny *et al*, 2012). Also, where genes associated with the classical pathway of colorectal tumorigenesis are mutated in MMRd CRCs, there is a prevalence of cMNR frameshift mutations. For example, *APC* has a much higher incidence of cMNR frameshift mutations in MMRd versus MMRp CRCs ($p < 0.0002$), (Huang *et al*, 1996) and this observation extends to precancerous tumours with 14/26 MMRd versus 3/52 MMRp adenomas containing such frameshifts in *APC* (Sekine *et al*, 2017). cMNR frameshift mutations can also confer drug-resistance when MMRd but not MMRp CRC cell lines are exposed to selection by 6-thioguanine (Bhattacharyya *et al*, 1994).

The inherent mutability of microsatellites is proposed to explain their reduced density in coding relative to non-coding regions of the genome (Subramanian *et al*, 2003). Furthermore, there has been debate around the functional impact of cMNR frameshift mutations in the non-classical genes of colorectal tumorigenesis; are they drivers of tumorigenesis or passengers caused by this inherent mutability? Transforming growth factor- β receptor 2 (*TGF β 2*) has multiple regulatory roles in cellular homeostasis and growth, and has been linked with both progression and suppression in multiple types of cancer (Padua and Massagué, 2009). *TGF β 2* was shown to be absent in approximately 81% of MMRd CRCs but only 11% of MMRp CRCs, and the causative mutations in the MMRd samples were found to be frameshifts due to 1-2bp deletions in a 10bp poly-adenine (A10) tract of the *TGF β 2* gene, leading to an early stop codon and protein truncation (Markowitz *et al*, 1995) (Figure 1.3). To confirm functional impact, cell line HCT116, which lacks *TGF β 2* expression due to frameshift mutation in the A10 repeat, was transfected with a wild type (WT) copy of the gene; rescue of *TGF β 2* expression decreased clonogenicity in culture and

<i>TGFβR2</i> WT	A10	Codon N	125	126	127	128	129	130	131	132	...	568	
		Sequence	...	GAA	AAA	AAA	AAG	CCT	GGT	GAG	ACT	...	TGA
		Amino acid	...	E	K	K	K	P	G	E	T	...	stop, full length protein
<i>TGFβR2</i> -1 and -2 bp frameshift mutations	A9	Codon N	125	126	127	128	129	130	131	132	...	162	
		Sequence	...	GAA	AAA	AAA	AGC	CTG	GTG	AGA	CTT	...	TAG
		Amino acid	...	E	K	K	S	L	V	R	L	...	stop, truncated protein
	A8	Codon N	125	126	127	128	129	130					
		Sequence	...	GAA	AAA	AAA	GCC	TGG	TGA				
		Amino acid	...	E	K	K	A	W	stop, truncated protein				

Figure 1.3. Frameshift mutations identified in *TGFβR2*. -1 and -2bp deletions in an A10 repeat within the coding sequence of *TGFβR2* produce frameshift mutations that introduce early stop codons and cause protein truncation. Note that the frameshift also produces a novel sequence of amino acids in the C-terminus of the truncated protein.

decreased tumorigenicity in athymic mice (Wang *et al*, 1995). Similarly, the pro-apoptotic *BAX* gene had frameshift mutations in its 8bp poly-guanine (G8) tract in approximately 50% of MMRd CRCs, often in both alleles (Rampino *et al*, 1997), and presence of these *BAX* mutations affected survival of CRC clones inoculated into immune-deficient mice (Ionov *et al*, 2000). The data from *TGFβR2* and *BAX* support functional roles for cMNR frameshift mutations in colorectal tumorigenesis. Other cMNR frameshift mutations with similar functional evidence for being drivers of tumorigenesis have since been identified, including frameshift mutation of an A8 repeat in *ACVR2* (Deacu *et al*, 2004) and an A10 repeat in *AIM2* (Lee *et al*, 2012). Whole-genome and whole-exome sequencing, comparing the spectrum of mutations in colorectal versus endometrial cancers (ECs), has also shown that *TGFβR2*, *ACVR2A*, *AIM2*, *SLC22A9* and *SMAP1* all contain frameshift mutations in 50-70% of MMRd CRCs but <25% of MMRd ECs. MMRd ECs likewise have their own set of frequently frameshift-mutated genes which are not observed in CRCs, suggesting that cMNR frameshifts are subject to selection during tumorigenesis in the context of different tumour types (Kim *et al*, 2013).

Duval and Hamelin proposed that the biological relevance of cMNR frameshift mutations could be determined from their observed frequencies by assuming that all microsatellites, coding or non-coding, have a constant mutation rate dependent on length and, therefore, genes with cMNRs of a given length that have an over or under

representation of frameshift mutations in cancers are likely to be selected for or against respectively (Duval and Hamelin, 2002). Adapting this concept into a statistical model, Woerner *et al* have created SelTarbase, a database of genes that are likely targets of selection during MSI-driven tumorigenesis (www.seltarbase.org; Woerner *et al*, 2003; Woerner *et al*, 2005; Woerner *et al*, 2010). Of the 1793 MNRs analysed, 4.0% are predicted to be positively or negatively selected in CRC (Woerner *et al*, 2010). From these predictions and the previously described weight of evidence, it is clear that MMR deficiency and MSI constitute a mutator phenotype that drives colorectal tumorigenesis.

1.4. Frameshift Mutations stimulate an Anti-Tumour Immune Response

MMR deficiency was defined as a distinct molecular subtype of CRC by multiple clustering algorithms using transcriptomic data (Muzny *et al*, 2012; De Sousa E Melo *et al*, 2013; Sadanandam *et al*, 2013; Roepman *et al*, 2014). These methods were combined into one classification algorithm to define the four consensus molecular subtypes (CMS) of CRC. MSI and hypermutation were genetic hallmarks of CMS1. CMS1 CRCs also had an increased expression of gene signatures related to T helper 1 cells and cytotoxic T lymphocytes (CTLs), which are associated with an anti-tumour immune response, and, in opposition to this, activation of immune evasion mechanisms (Guinney *et al*, 2015). Characteristics of CMS1 agree with the strong association between hypermutation and MMR deficiency in genomic analyses of more than 100,000 cancers (Alexandrov *et al*, 2013; Chalmers *et al*, 2017) and the early observations of increased TILs in MSI-high CRCs indicating that these tumours are particularly immunogenic (Ionov *et al*, 1993; Lothe *et al*, 1993; Thibodeau *et al*, 1993).

Cells throughout the body present intracellular protein antigens to T cells by binding of peptide fragments to major histocompatibility complex (MHC) class I receptors on the cell surface, facilitating immune responses against cellular infection and dysfunction (Matsumura *et al*, 1992). Tumour associated antigens (TAAs) originate through several mechanisms including aberrant gene expression, infection by oncoviruses and abnormal post-transcriptional modification (Ilyas and Yang, 2015). Of particular interest is the generation of TAAs by mutation of proteins also expressed in normal cells, producing amino acid sequences that are recognised as “foreign” by the immune system. An early example is antigen LB33-B, presented on MHC class I receptors of melanoma cell line LB33-MEL.A. LB33-B was tracked back to a point mutation in an exon-intron junction of the *MUM1* gene,

producing a novel peptide by translation of the intronic sequence, expression of which stimulated cell lysis by CTLs (Coulie *et al*, 1995). Therefore, mutations can generate TAAs that stimulate an immune response, and it follows that there is a positive correlation between tumour mutational burden and the response rate of cancers to immunotherapy ($r = 0.74$, $p < 0.001$) (Yarchoan *et al*, 2017).

There is strong evidence of TAAs being generated as a consequence of the frameshift mutations at cMNRs that drive tumorigenesis of MMRd CRCs. It was initially theorised that the affected genes could be translated to produce truncated proteins containing immunogenic frameshift peptides (FSPs) at their C termini due to the change in reading frame downstream of the mutation (Figure 1.3). To test this, Linnebacher *et al* selected FSP antigens from common frameshift mutations in MMRd CRCs that were also predicted *in silico* to bind to MHC class I receptors for antigen presentation to immune cells. CTLs specific to these antigens were generated using synthetic FSPs presented on CD40-activated B cells. They found that three of the anti-FSP CTL lines were able to lyse cells loaded with the respective peptide. Most importantly, CTLs targeting TGF β 2-derived FSPs could lyse HCT116, an MMRd CRC cell line which contains the associated frameshift mutation in *TGF β 2* (Linnebacher *et al*, 2001). CTLs expanded from the TIL population of MSI-high CRCs were similarly able to lyse MSI-high, but not MSS, CRC cell lines, and T cells isolated from the peripheral circulation from MSI-high, but not MSS, CRC patients were activated by FSP-loaded autologous B cells, suggesting that activation of T cells by FSPs also occurs *in vivo* (Saeterdal *et al*, 2001; Schwitalle *et al*, 2008). The FSPs required to stimulate peripheral CTLs were shown to match the cMNR frameshift mutations present in the tumour, and the number of frameshift mutations correlated with the density of TILs and their CTL component (Tougeron *et al*, 2009; Maby *et al*, 2015). Finally, enzyme-linked immunosorbent assay (ELISA) has been used to identify anti-FSP antibodies in the serum of MSI-high CRC patients (Reuschenbach *et al*, 2010), which suggests that both cytotoxic and humoral immune responses can be generated against the intrinsic FSP antigens of MMRd CRCs.

1.5. Mismatch Repair Deficient Cancers respond to Immunotherapy

The immunoediting model of tumour evolution proposes that interaction between tumours and the immune system shapes tumour evolution by Darwinian selection (Greaves and Maley, 2012). There are three stages to immunoediting. Foremost is “elimination” by which

immunosurveillance destroys early tumour cells. “Equilibrium” is reached when clones surviving initial elimination continue to propagate but remain asymptomatic under the continuing selection pressure of immune destruction, driving tumour evolution toward the final stage of “escape”. During escape, mechanisms that allow the tumour to evade immune destruction are evolved and disease progresses (Dunn *et al*, 2002; Mittal *et al*, 2014). Tumour escape of the immune system can be achieved by several mechanisms. These include loss of antigen presentation through *beta2-microglobulin* ($\beta 2M$) mutation or human leukocyte antigen (HLA) copy number variation (Paschen *et al*, 2003; McGranahan *et al*, 2017), disruption of antigen processing through loss of tapasin (Sokol *et al*, 2015), loss of *APLNR* function, which regulates interferon- γ stimulation of immune cells via JAK1 signalling (Patel *et al*, 2017), inhibition of natural killer cells by intra-tumoral *Fusobacterium nucleatum* (Gur *et al*, 2015), and tumour or stromal expression of immune checkpoint proteins, such as programmed cell death protein ligand 1 (PD-L1) and cytotoxic T-lymphocyte-associated antigen 4 (CTLA-4), which inhibit T cell anti-tumour activity (Pardoll *et al*, 2012). Both loss of MHC class I and class II antigen presentation has been observed in MMRd CRC (Kloor *et al*, 2007; Surmann *et al*, 2015), and MMR deficiency is associated with increased PD-L1 expression in numerous cancer types ($p = 0.01$) (Kim *et al*, 2017). Two years ago, the immunology of MMRd cancers was thoroughly reviewed (Kloor and von Knebel Doeberitz, 2016).

Immune checkpoint proteins can be found on the surface of both immune and non-immune cells. CTLA-4 is expressed by T cells and auto-regulates activity by increasing its concentration on the T cell surface in proportion to the strength of T cell receptor activation, where it antagonises further stimulation by competitive binding with the stimulatory receptors of antigen presenting cells. Programmed cell death protein 1 (PD-1) is expressed on the surface of activated T cells and binding to PD-L1 expressed on tumour or stromal cells promotes T cell exhaustion (Pardoll *et al*, 2012). In cancers where TAA-reactive T cells have been exhausted by immune checkpoints, blocking of the checkpoint signal using small molecule or antibody inhibitors has proven to be an effective and durable therapy by releasing exhaustion and allowing proliferation of the suppressed T cells (Gubin *et al*, 2014). Immune checkpoint blockade by pembrolizumab, an antibody that binds PD-1, had a disease control rate (which includes stable disease, partial response or complete response) of 90% in MMRd CRCs and 71% in MMRd non-colorectal cancers, whereas MMRp CRCs only had an

11% disease control rate (Le *et al*, 2015). Expansion of pembrolizumab treatment to 12 different cancer types found complete response in 21% of MMRd cancers and a disease control rate of 77%. Overall survival at 2 years was 64% in MMRd cancers despite the advanced stage of disease (Le *et al*, 2017). Analysis of pembrolizumab's mechanism of action revealed expansion of T cell clones that were specifically reactive to FSP antigens related to cMNR mutations found in the respective tumours (Figure 1.4), confirming the association of response with MMR deficiency (Le *et al*, 2017). Whilst these initial results are very promising, the efficacy of pembrolizumab is still to be confirmed by randomised clinical trials (Cummings and Garon, 2017). Hence, the US Food and Drugs Administration (FDA) have approved pembrolizumab as a second line treatment in all MSI-high cancers refractory to primary treatment (MERCK & Co. Inc, 2017).

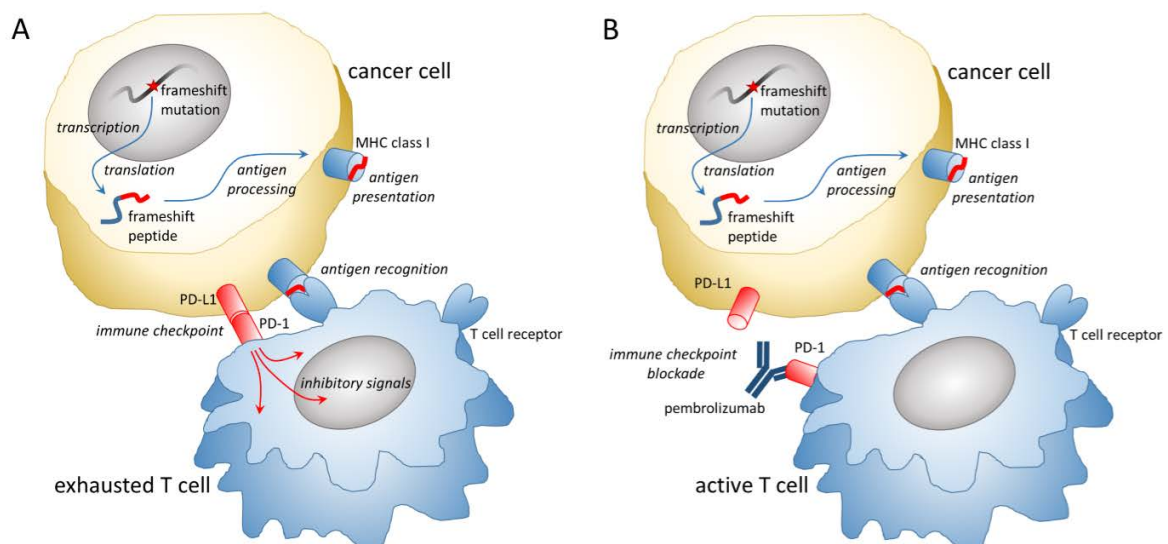


Figure 1.4. The mechanism of action of immune checkpoint blockade by pembrolizumab in mismatch repair deficient cancers. Mismatch repair deficient cancers frequently contain frameshift mutations in coding mononucleotide repeats. Expression of the mutated gene produces frameshift peptide antigens that are recognised as “foreign” by the patient’s immune system, stimulating an anti-tumour immune response. **(A)** To escape immune-mediated cytotoxicity, cancer cells express programmed cell death ligand 1 (PD-L1), which binds to programmed cell death protein 1 (PD-1) expressed on active T cells. The binding of PD-L1 with PD-1 sends inhibitory, or immune checkpoint, signals to the T cell to prevent clonal expansion and cytotoxic activity. This inactive state is referred to as T cell exhaustion. **(B)** Blockade of the PD-L1 to PD-1 immune checkpoint by pembrolizumab binding to PD-1 releases T cell suppression and exhausted T cells become active, expanding and renewing cytotoxicity.

1.6. Cancer-predisposition Syndromes associated with Mismatch Repair Deficiency

Familial adenomatous polyposis (FAP) was recognised in the early 20th century as a high penetrance, but surgically curable, cause of CRC, characterised by tens to thousands of macroscopically visible polyps of the colorectal epithelium (Gardner, 1951; Dukes, 1952). It was later recognised that the majority of hereditary CRC lacked this polyposis phenotype, but it was not until Henry Lynch and colleagues retraced Warthin's family G, which was originally reported in 1913, and identified others that the clinical condition of HNPCC became accepted (Lynch and Krush, 1971; Lynch *et al*, 1998; Douglas *et al*, 2005). While the FAP gene was being sought (Groden *et al*, 1991), an international consortium similarly began collecting families with HNPCC with a view to identification of the underlying genes (Section 1.2). For example, Dunstone and Knaggs (1972) described a family in North East England, similar to Warthin's family G, with 45 cancers in 104 individuals, later shown to have an *MLH1* mutation (John Burn, personal communication). Initially, Lynch described two types of HNPCC depending on the presence of other cancers, typically of the endometrium, but the discovery of the MMR genes as the underlying cause made it clear that the clinical phenotypes, and the condition known as Muir Torre syndrome which includes types of skin cancer, were all variants of the same condition (Lynch *et al*, 1985). Given the wider range of cancers, it was decided to use the diagnostic label of Lynch syndrome instead of HNPCC (Lynch *et al*, 2009). Currently, HNPCC is considered an umbrella term shared by multiple distinct cancer syndromes that can be separated by testing the MSI status of tumours and by their distinct genetic aetiologies (Figure 1.5).

The majority of MSI-high HNPCC is attributable to Lynch syndrome, which accounts for 2.4-3.7% of all CRC cases (Hampel *et al*, 2005a; Hampel *et al*, 2008; Canard *et al*, 2012; Moreira *et al*, 2012; Pérez-Carbonell *et al*, 2012; van Lier *et al*, 2012) and, by extension, 16-25% of all MMRd CRC. Lynch syndrome is caused by pathogenic germline mutation affecting one of four MMR genes, *MLH1*, *MSH2*, *MSH6* and *PMS2*, and is a significant burden to healthcare services (Lynch *et al*, 2009). Where a genetic diagnosis of Lynch syndrome is not made, MSI-high HNPCC is termed Lynch-like syndrome. Lynch-like syndrome has a CRC risk intermediate between Lynch syndrome and the general population (Rodríguez-Soler *et al*, 2012) and may be linked to non-MMR germline variants, such as in *MUTYH* or *POLE*, that result in somatic MMR mutation (Castillejo *et al*, 2014; Morak *et al*, 2014), or may be attributable to a heterogeneous population of Lynch syndrome cases with un-characterised

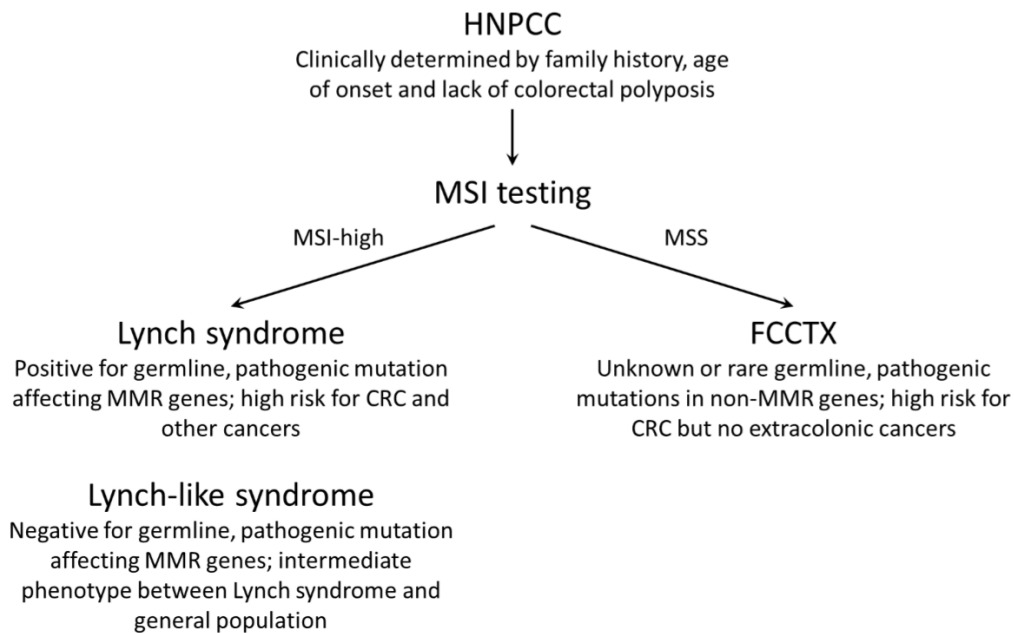


Figure 1.5. The colorectal cancer syndromes within hereditary non-polyposis colorectal cancer (HNPCC). MSI testing can separate HNPCC into distinct cancer syndromes, with MSI-high HNPCC being associated with Lynch and Lynch-like syndromes, and MSI-low or MSS HNPCC being associated with familial colorectal cancer type X (FCCTX) and polymerase proof reading polyposis. Adapted from Carethers and Stoffel, 2015.

MMR mutations (Clendenning *et al*, 2011; Borràs *et al*, 2013; Rhees *et al*, 2014; Liu *et al*, 2016) mixed with double somatic MMR mutations that appear more like Lynch syndrome CRCs than the majority of sporadic MMRd CRCs (Geurts-Giele *et al*, 2014; Haraldsdottir *et al*, 2014; Mensenkamp *et al*, 2014) (Figure 1.5). MSS HNPCC is a poorly characterised phenotype of a heterogeneous population, and is given the name familial CRC type X (FCCTX) (Lindor *et al*, 2005) (Figure 1.5). Efforts to identify causative germline variants have found that known cancer predisposition genes are rarely implicated in FCCTX, and the majority of candidate genes identified have not been validated (Lorans *et al*, 2018). It has also been suggested that polygenic, rather than monogenic, inheritance may account for a large proportion of FCCTX (Ku *et al*, 2012). Constitutional MMR deficiency (CMMRD), also known as biallelic MMR deficiency (Durno *et al*, 2012), is a very rare childhood cancer syndrome caused by germline, biallelic mutation in the same MMR gene (Wimmer *et al*, 2014).

1.6.1. Lynch syndrome biology and clinical management

An analysis of 1112 CRC patients genetically tested on suspicion of Lynch syndrome in 2012-2013 identified 114 (10.3%) patients were germline heterozygous for pathogenic *MMR* gene

(*path_MMR*) mutations. Of these mutations, 27% were *path_MLH1*, 35% were *path_MSH2*, 3% were EPCAM 3' deletions, which leads to silencing of downstream *MSH2* (Ligtenberg *et al*, 2009), 23% were *path_MSH6*, and 12% were *path_PMS2* (Yurgelun *et al*, 2015). Lynch syndrome gene carriers (meaning individuals carrying an MMR mutation but not necessarily presenting with disease) have an increased risk of multiple cancers, in particular CRC and EC, with disease penetrance depending on which MMR gene is affected, as estimated by prospective data (Figure 1.6). Whilst *path_PMS2* mutations are known to cause Lynch syndrome (Nicolaidis *et al*, 1994), there is currently insufficient prospective data to accurately estimate cancer risks in these patients, although these are significantly lower than other *path_MMR* mutations (Møller *et al*, 2017b). Therefore, whilst *PMS2* mutations account for a minority of Lynch syndrome cases (Yurgelun *et al*, 2015), *path_PMS2* variants may occur at a higher frequency in the population than variants in any other MMR gene.

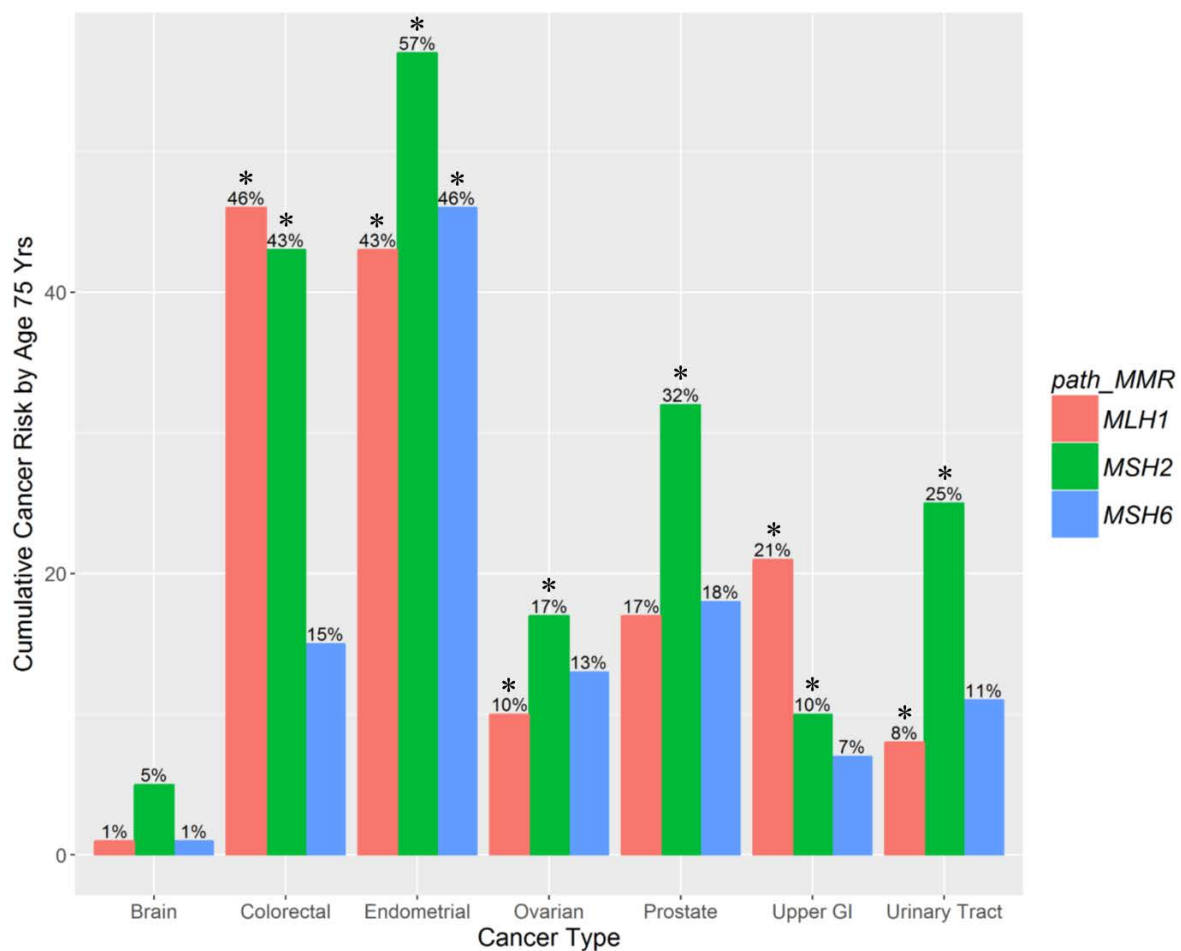


Figure 1.6. The cumulative cancer risks in Lynch syndrome gene carriers by age 75 years. Significant increases in risk relative to the general population are marked with *, using results from the Prospective Lynch Syndrome Database (Møller *et al*, 2017b). GI: gastrointestinal.

The increased cancer risk in path_ *MMR* carriers follows the two-hit hypothesis for loss of tumour suppressor gene function (Knudson, 2001) as 90-100% of Lynch syndrome CRCs contain a “second hit” in the germline-affected *MMR* gene causing *MMR* deficiency in the tumour (Leach *et al*, 1996; Liu *et al*, 1996; Thibodeau *et al*, 1996; Hampel *et al*, 2008). Initially, it was believed that *MMR* deficiency was a secondary event in pre-established adenomas due to the similar adenoma incidence and histology in Lynch syndrome families and age-matched autopsy populations, which suggested a similar initiation and progression of colorectal tumours. Combined with the observed increase in high grade dysplasia, size, and only slightly earlier onset of adenomas in Lynch syndrome gene carriers, and a lack of a polyposis phenotype, it was therefore suggested that loss of *MMR* is an accelerator rather than initiator of tumour progression (Jass and Stewart, 1992; Jass *et al*, 1994). However, subsequent studies on larger populations have shown that there is a 2-3-fold increase in the adenoma burden in Lynch syndrome gene carriers relative to age-matched controls, with an increased incidence ($p = 0.0001$) of villous or tubulovillous histology (de Jong *et al*, 2004a). Furthermore, the discovery of *MMRd* crypt foci (*MMR*-DCF) in the normal colorectal epithelium of Lynch syndrome gene carriers revealed that *MMR* deficiency can occur in phenotypically normal cells (Kloor *et al*, 2012). *MMR*-DCF can cover multiple crypts, can have aberrant histology and can have cMNR frameshift mutations in genes such as *AIM2* and *BAX* (Staffa *et al*, 2015). This raises the possibility that *MMR* deficiency could be the initiating event of colorectal tumorigenesis in Lynch syndrome. This has recently been supported by the discovery of *MMRd* adenomas outgrowing from *MMR*-DCF (Figure 1.7), and by the high frequency of cMNR frameshift and *MMR* deficiency related substitution mutations in *APC*, loss of which initiates adenoma formation (Sekine *et al*, 2017; Ahadova *et al*, 2018).

Most significantly, the Lynch syndrome pathway of colorectal tumorigenesis is distinct from the pathway associated with sporadic *MMRd* CRC. *MMR* deficiency in sporadic CRC is associated with the CpG island methylator phenotype (CIMP), whereby widespread hypermethylation of CG dinucleotides in the tumour DNA leads to promoter-methylation and aberrant silencing of many genes, often including *MLH1* (Young *et al*, 2001). Weisenberger *et al* (2006), in a study of 195 CpG methylation sites in 295 CRCs, showed that CIMP characterises a distinct tumour subtype and accounts for nearly all CRCs with *BRAF* V600E mutations (odds ratio (OR) = 203). CIMP has also been associated with the serrated pathway of tumorigenesis in which cancers arise from adenomas with a serrated histology

(Yamane *et al*, 2014) rather than the traditional adenomas characteristic of Lynch syndrome and most MMRp tumours (Figure 1.8). It follows, therefore, that Lynch syndrome CRCs can be distinguished from sporadic MMRd CRCs by key molecular changes as they make up a large proportion of those that do not have CIMP or *BRAF* V600E mutation (Kambara *et al*, 2004).

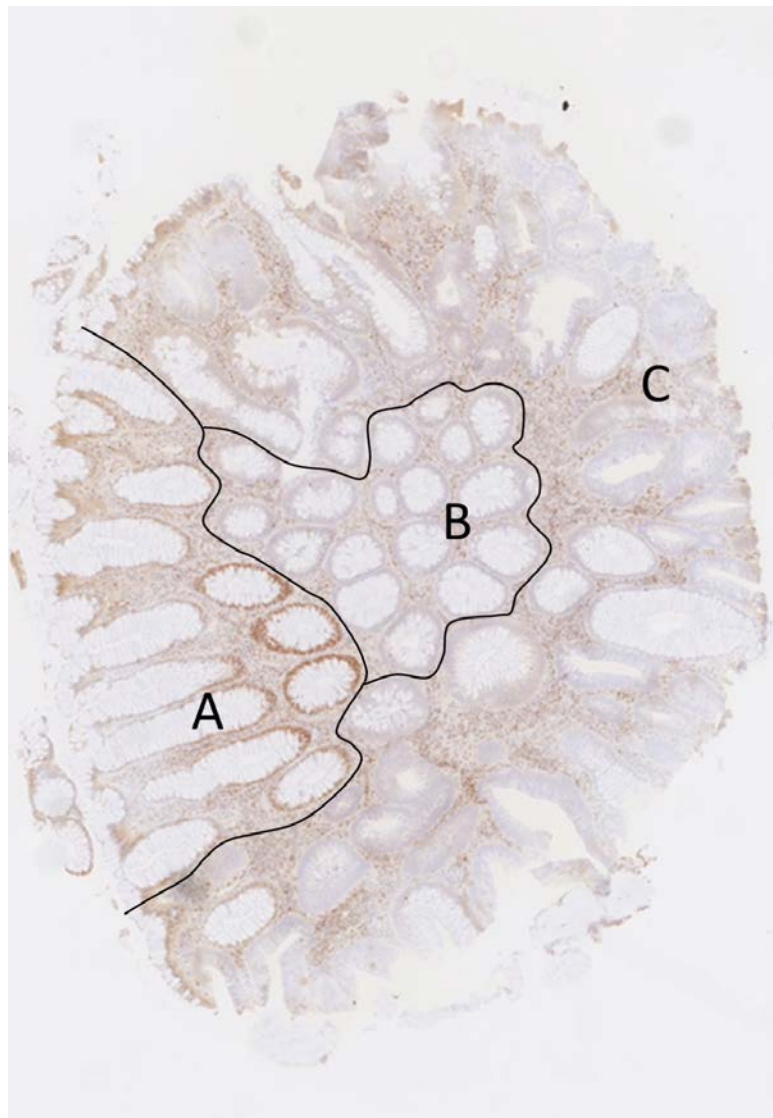


Figure 1.7. Mismatch repair deficient adenoma associated with mismatch repair deficient (MMRd) colorectal crypt focus. A 3 μ m section of FFPE adenoma tissue resected from a Lynch syndrome patient with a germline path_ *MSH2* mutation was stained by immunohistochemistry (IHC) for MSH2 expression. **(A)** Histologically normal and MMR proficient colorectal crypts, showing MSH2 positive staining in the nuclei of dividing cells at the base of the crypts. **(B)** MMRd colorectal crypts, showing loss of MSH2 expression but otherwise normal histology. **(C)** Dysplastic and MMRd tissue of the adenoma directly adjacent to the MMRd colorectal crypts. Adapted from Ahadova *et al*, 2018*.
* I am joint first author on this publication, however this work is not described in this thesis. Please see Appendix L for more details.

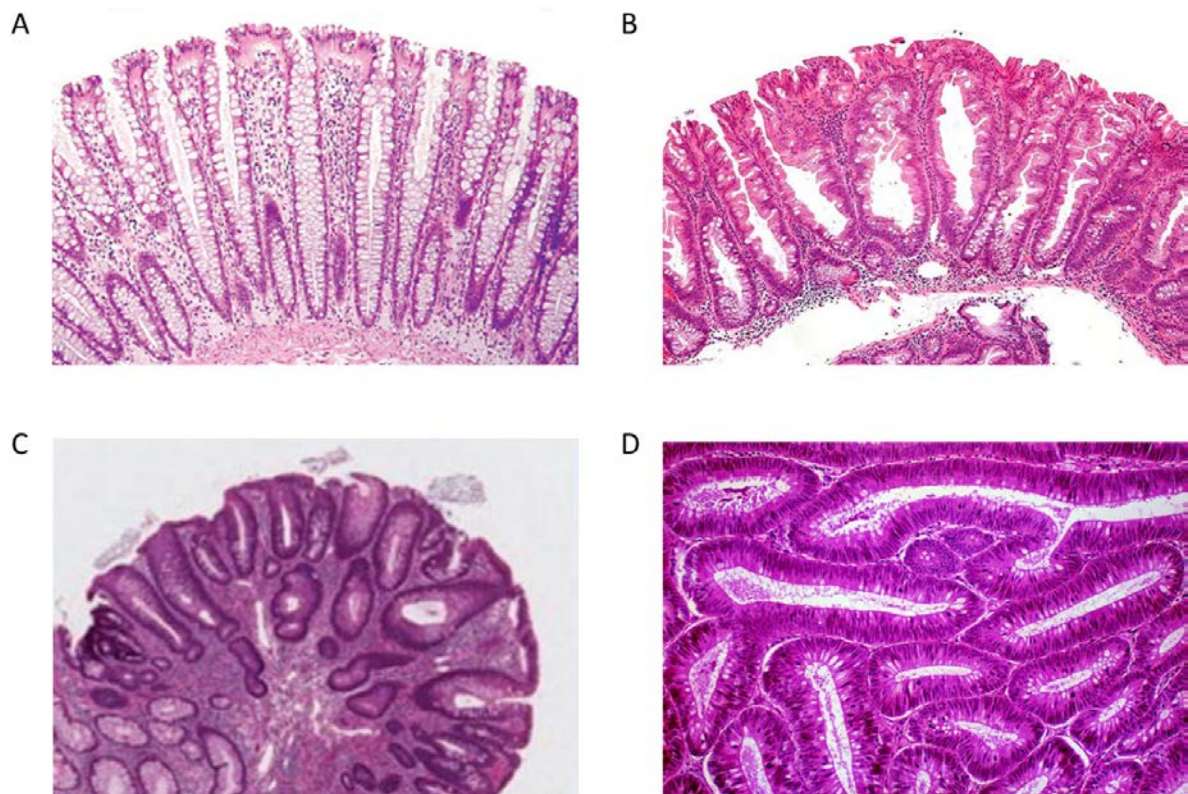


Figure 1.8. Comparative histology of colorectal adenomas. Haematoxylin and eosin staining of colorectal tissues to show differences in histology. **(A)** Normal colorectal mucosa showing mucus-producing epithelial cells lining uniformly-structured crypts. **(B)** Serrated adenoma with the eponymous serrated or saw-tooth edge to the crypts. These grow as sessile and flat lesions that do not obviously protrude from the colorectal wall (Bartley *et al*, 2010). **(C)** Traditional adenomas, with villous, tubular or tubulovillous histology, are typically polypous and extend from the colorectal wall into the lumen (Buchanan *et al*, 2011). **(D)** Close up of tubular histology in a colorectal adenoma, showing degradation of crypt architecture and lack of serration.

A diagnosis of Lynch syndrome has implications for the clinical management of the patient and their family to mitigate their increased risk for colorectal, endometrial and other cancers. Management guidelines include surveillance, prophylaxis and genetic counselling (Vasen *et al*, 2013). Numerous studies have assessed the efficacy of surveillance for the two most common Lynch syndrome cancers, CRC and EC. Colonoscopy with polypectomy every 1-2 years is highly effective at reducing CRC risk and mortality in Lynch syndrome families. For example, Järvinen *et al* (2000) observed a 62% reduction in CRC incidence in patients under endoscopic surveillance, and 0 versus 9 CRC-related deaths in surveillance and control groups respectively. In agreement, there was a statistically significant decrease in the standardised mortality ratio associated with colonoscopic surveillance (de Jong *et al*, 2006). However, no benefit has been observed for endometrial surveillance (de Jong *et al*, 2006).

This is perhaps due to the low mortality rates of EC due to symptomatic stage 1 disease that is curable by surgical resection (Boks *et al*, 2002). From these observations, colonoscopic surveillance is recommended in 1-2 yearly intervals starting at age 20-25 years (Vasen *et al*, 2013), although the optimal time interval is still being debated. Recent comparisons of CRC incidence in path_*MLH1* carriers shows no difference in the rate of interval cancers between 1-3 year intervals, reasoning for less frequent surveillance (Seppälä *et al*, 2017). Surveillance may also be tailored by MMR gene in the future, with proposals that path_*PMS2* carriers should start colonoscopic surveillance at 35-40 years of age, due to lower disease penetrance in these patients (Ten Broeke *et al*, 2018). Prophylactic surgery to remove at risk organs is another option for disease management in Lynch syndrome gene carriers, in particular partial or sub-total colectomy, and hysterectomy after completion of child-bearing (Vasen *et al*, 2013). Due to the high risk of metachronous CRC (Aarnio *et al*, 1995), known Lynch syndrome patients can choose partial or total colectomy at surgical resection of the first tumour, based on evidence that more extensive surgery reduces the risk of metachronous CRC by 31% (95% CI: 12-46%, $p = 0.002$) for every 10cm of colorectum removed (Parry *et al*, 2011). Parry *et al* (2011) also found no metachronous CRCs were diagnosed in study patients opting for a full colectomy (incident rate ratio (IRR): 0.0, 95% CI: 0.0-7.2 per 1000 person years). Chemoprevention of CRC and other Lynch-spectrum cancers is also effective. For example, Lynch syndrome gene carriers with a daily intake of 600mg of aspirin were shown to have an IRR for CRC of 0.37 (95% CI: 0.18-0.78, $p = 0.008$) relative to those randomised to placebo, after a median 55.7 months of follow up in the CAPP2 clinical trial (Burn *et al*, 2011). Lifestyle has also been associated with CRC risk in Lynch syndrome, including increased risk (HR = 2.34; 95% CIs = 1.17-4.67; $p = 0.02$) in obese patients (Movahedi *et al*, 2015), and decreased risk (HR: 0.71, 95% CI: 0.53-0.96, $p = 0.02$) in those exercising regularly (Dashti *et al*, 2018). Therefore, a diagnosis of Lynch syndrome allows optimised clinical management and patients to modify their lifestyle to reduce their cancer burden. Finally, immunotherapies are applicable to Lynch syndrome cancers due to the high rate of MMR deficiency and associated immune response in these tumours (Westdorp *et al*, 2016). For example, immune checkpoint blockade by pembrolizumab showed high response rates and overall survival at 2 years in MMRd cancers from both Lynch syndrome and sporadic patients (Le *et al*, 2017).

1.6.2. Constitutional mismatch repair deficiency biology and clinical management

Turcot's syndrome is a familial condition that includes cancer and polyposis of the colorectum and tumours of the central nervous system (CNS), but its genetic aetiology was unknown for several decades (Turcot *et al*, 1959). Patients from 14 families fulfilling clinical criteria for Turcot's syndrome were shown to harbour germline mutations in either *APC* or MMR genes *MLH1* and *PMS2*. Families lacking a genetic diagnosis had MSI-high tumours characteristic of Lynch syndrome, suggesting that they too had a causative MMR defect (Hamilton *et al*, 1995). Furthermore, these families could be segregated based on the type of CNS tumour, with medulloblastomas versus glioblastomas predominating in *APC*-associated and MMR-associated Turcot's syndrome respectively (Hamilton *et al*, 1995). Subsequently, multiple case studies of Turcot's syndrome, or of offspring of consanguineous marriages in HNPCC kindreds, found biallelic MMR gene mutation in the affected patients, including *MLH1* (Wang *et al*, 1999; Ricciardone *et al*, 1999; Gallinger *et al*, 2004), *MSH2* (Whiteside *et al*, 2002; Toledano *et al*, 2009), *PMS2* (De Rosa *et al*, 2000; De Vos *et al*, 2006; Krüger *et al*, 2008) and *MSH6* (Menko *et al*, 2004; Ripperger *et al*, 2010), covering all MMR genes involved in Lynch syndrome. The spectrum of cancers in these patients was diverse, including brain, haematological and gastrointestinal (GI), and are typically diagnosed in childhood to adolescence. Due to their cancer burden, mortality in these case studies was high and often at a young age. Pre-malignant or benign phenotypes prevalent in these patients included colorectal polyps despite their young age, and features of neurofibromatosis type 1 (NF-1), including benign neurological tumours (such as neurofibromas), and skin markings (such as café-au-lait maculae, freckling, and hypopigmentation). It was proposed that this rare childhood cancer syndrome be called CMMRD in reference to its underlying aetiology (Wimmer *et al*, 2008).

Lynch syndrome and CMMRD have a common cause in pathogenic variants in MMR genes, but the representation of gene variants in the two syndromes differs. A collation of 146 genetically confirmed CMMRD cases showed that 58% are caused by *PMS2* mutations, 20% by *MSH6* mutations and only 22% by *MLH1* or *MSH2* mutations (Wimmer *et al*, 2014). This is in contrast to the 6-12% of Lynch syndrome CRCs associated with *PMS2* mutation (Borràs *et al*, 2013; Yurgelun *et al*, 2015). Diagnosing CMMRD caused by *PMS2* mutation can be complicated by *PMS2* pseudogenes on chromosome 7 (Nicolaidis *et al*, 1995). These contain paralogous copies of all *PMS2* exons and make accurate sequencing and variant

calling difficult (Nakagawa *et al*, 2004). Cases of CMMRD may be misdiagnosed as Lynch syndrome or otherwise, due to failure to detect a pathogenic variant. For example, when genotyping young CRC patients that lacked PMS2 expression in the tumour, path_*PMS2* nonsense mutations affecting the second allele were initially missed and only recognised after repeat sequencing using alternative methods (De Vos *et al*, 2004). Furthermore, 8% of supposed monoallelic path_*PMS2* carriers had CRC diagnosed below 30 years of age and all on the left-side of the colorectum, a feature more common to CMMRD than Lynch syndrome (Goodenberger *et al*, 2016). Given that only 9% of MMR variants listed in the InSiGHT database of CRC-related gene variants affect *PMS2*, additional knowledge of *PMS2* variants is needed to reduce the uncertainty in its diagnosis (Blount and Prakash, 2017).

Clinical details of 197 CMMRD patients were used to define the neoplastic and benign features of CMMRD, one of the largest collections of CMMRD data due to this syndrome's rarity (Wimmer *et al*, 2017). Based on 321 tumours across 34 tumour types, haematological malignancies were diagnosed in 38.6%, brain and CNS tumours in 54.8%, and Lynch-spectrum cancers in 51.8% of patients. Within the haematological malignancies, non-Hodgkin's lymphoma and lymphoid leukaemia were the most common, being diagnosed in 19.8% and 7.6% of all patients, respectively. The vast majority of CNS tumours are high grade glioblastomas, found in 40.6% of all patients, and CRC is the predominant Lynch-spectrum cancer being diagnosed in 38.1%. The distribution in age of diagnosis depends on tumour type, but haematological malignancies have been diagnosed in patients younger than 1 year of age and CNS tumours in patients as young as 2 years (Figure 1.9).

The Lynch-spectrum cancers of CMMRD patients, the adenomatous histology of their colorectal adenomas, and the presence of the MSI-high phenotype in their CRCs, suggests that CMMRD cancers develop by tumorigenesis pathways similar to Lynch syndrome, albeit accelerated due to the younger age of onset (Aronson *et al*, 2016). However, the presence of a polyposis-like phenotype with histology reminiscent of juvenile polyposis suggests alternative progression also occurs in CMMRD (Levi *et al*, 2015; Aronson *et al*, 2016). CMMRD haematological malignancies are frequently MSI-high but, interestingly, CNS tumours rarely are (Bakry *et al*, 2014). The lack of MSI despite MMR deficiency is particularly associated with glioblastomas (Bougeard *et al*, 2003; Leenen *et al*, 2011) even when other tumours in the same patient are MSI-high (Merlo *et al*, 1996), suggesting glioblastomas progress by a pathway that is not driven by MSI. Indeed, Shlien *et al* analysed the mutation

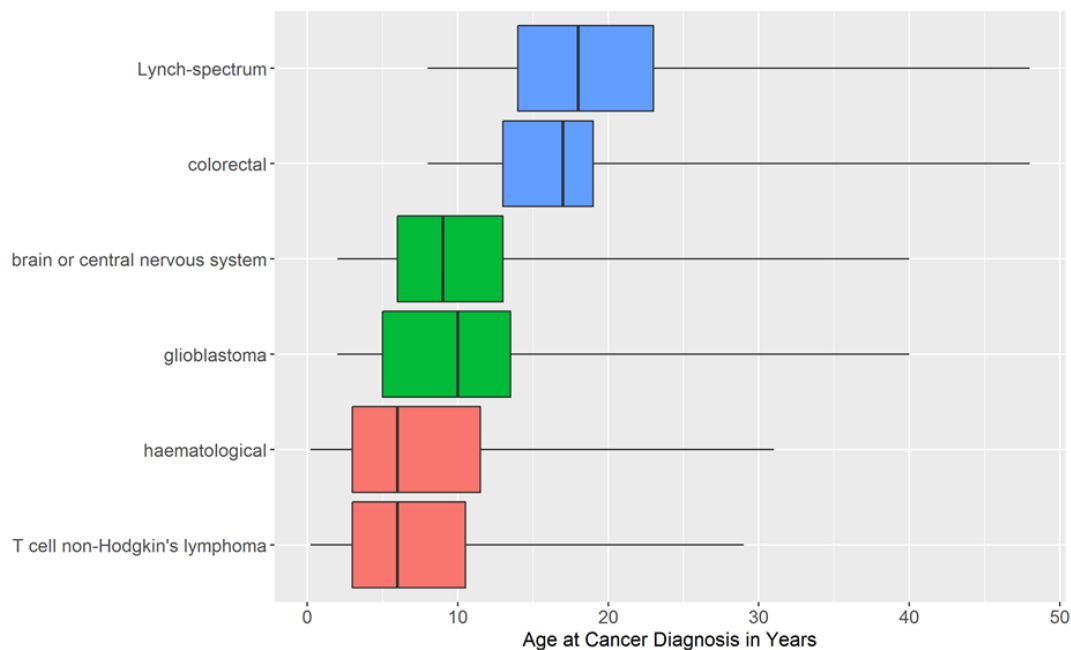


Figure 1.9. The ages of 197 constitutional mismatch repair deficiency patients at cancer diagnosis. Including Lynch-spectrum cancers, brain or central nervous system tumours, and haematological malignancies, and the most common cancer types within these groups (Wimmer *et al*, 2017).

spectrum of CMMRD-associated glioblastomas and found them to be ultra-hypermuted with an anticipated 600 mutations acquired per cell division, in particular substitution mutations, leading to rapid progression of a polyclonal tumour, and insufficient time or clonal homogeneity to develop a detectable MSI signal. Concurrent mutation in polymerase ϵ and polymerase δ , and complete loss of replication-associated repair, was proposed to cause this aggressive phenotype (Shlien *et al*, 2015).

CMMRD can also be recognised by its non-neoplastic features, such as the high frequency of skin cafe-au-lait maculae or hyperpigmentation that occurs in over 60% of patients (Wimmer *et al*, 2014). These overlap with the NF1 phenotype. Functional association between MMR deficiency and *NF1* gene mutation has been found, with 40% of MMRd cell lines and primary MMRd tumours shown to contain *NF1* mutations (Wang *et al*, 2003). However, only one patient has been identified with mutated *NF1* in blood (Alotaibi *et al*, 2008) despite this being explored in several studies (e.g. Menko *et al*, 2004; Østergaard *et al*, 2005). Therefore, the cause of the NF1-like phenotype of CMMRD is still to be determined. Additional non-neoplastic features are highly varied and include agenesis of the corpus callosum and grey matter hypertopia (Baas *et al*, 2013) and impaired immunoglobulin class switch recombination (Peron *et al*, 2008). A distinct molecular feature of CMMRD patients is the lack of MMR in all tissues. This can be observed by immunohistochemistry

(IHC) where mutation leads to loss of protein expression (Bakry *et al*, 2014). A review of MSI testing of tumour and normal tissues by Wimmer *et al* in 2008 found that somatic indels in microsatellites were detectable in the normal tissue of CMMRD patients when using highly sensitive small pool PCR (Parsons *et al*, 1995), but not MSI detection techniques normally applied to cancer diagnostics such as PCR fragment length analysis (Bacher *et al*, 2004).

The case study of a 43 years old female with biallelic *PMS2* mutation describes a clinical history in which the severe cancer risk of CMMRD patients can be combated by intense surveillance and extensive surgery. In her lifetime, the patient had been diagnosed with 9 different cancers from age 10 years and had multiple GI and gynaecological surgeries. Surveillance included upper and lower endoscopy, computed tomography (CT) and magnetic resonance imaging (MRI) (Sjursen *et al*, 2009). Prophylactic surgery and endoscopic GI surveillance are proven to be effective in Lynch syndrome (Järvinien *et al*, 2000; Parry *et al*, 2011) and other CMMRD case studies have reported early detection of CRC and a lack of mortality due to GI malignancy in patients under endoscopic surveillance (Durno *et al*, 2012; Aronson *et al*, 2016). Brain CT or MRI is routinely used in the diagnosis of brain tumours in children (Perkins *et al*, 2011) and can be used in surveillance for early and asymptomatic CNS tumours in CMMRD (Durno *et al*, 2012). Haematological malignancies can be detected by ultrasound of the abdomen, to assess the liver and spleen (Siniluoto *et al*, 1991), and blood counts (Juliusson and Liliemark, 1993). These observations were used by the European Care for CMMRD (C4CMMRD) consortium to compile surveillance and management guidelines for CMMRD (Table 1.1) (Vasen *et al*, 2014).

The use of aspirin as a chemopreventive has also been debated given its efficacy in Lynch syndrome (Burn *et al*, 2011), with current recommendations from European experts that prescription to daily aspirin be considered from first diagnosis of CMMRD, whilst clinicians should be cognizant of its risks, in particular cranial bleeds given the frequency of brain tumours (Leenders *et al*, 2018). Immunotherapy has also been considered in CMMRD patients but studies are needed to confirm safety and efficacy (Westdorp *et al*, 2017). Indeed, immune checkpoint blockade therapy had durable response in two siblings with CMMRD and ultra-hypermutated glioblastomas (Bouffet *et al*, 2016), so this is a promising avenue of research in CMMRD cancer therapy. Therefore, identification of patients with germline MMR gene defects, whether causative of Lynch syndrome or CMMRD, is critical to provide these patients with personalised clinical management.

Cancer	Start age	Procedure, interval
NHL/other lymphoma	1 year	Clinical examination, 1 per 6months Abdominal ultrasound (optional), 1 per 6months
Leukaemia	1 year	Blood count, 1 per 6months
Brain tumours	2 years	Brain MRI, 1 per 6-12months
CRC	8 years	Ileocolonoscopy, 1 per year
Small bowel cancer	10 years	Video capsule/upper GI endoscopy, 1 per year
Other Lynch-spectrum* cancer	20 years	Gynaecological examination/transvaginal ultrasound/pipelle curettage, 1 per year Urine cytology, 1 per year

*Colorectal, endometrial, small bowel, ureter, renal pelvis, biliary tract, stomach, bladder carcinoma.

Table 1.1. Surveillance recommendations for constitutional mismatch repair deficiency.

Recommendations agreed by the C4CMMRD consortium. Adapted from Vasen *et al*, 2014.

1.7. The Utility of Biomarkers in Colorectal Cancer and Cancer-predisposition Syndromes

A biomarker is defined as “A characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” (Biomarkers Definitions Working Group, 2001). Testing for biomarkers can therefore inform healthcare practice. There are two considerations when adopting a biomarker test into clinical practice: its analytical validity and its clinical utility (Hayes, 2018). Analytical validity defines the ability of the test to detect the biomarker and can be summarised by several quantifiable parameters (Ray *et al*, 2010; Henry and Hayes, 2012). These include, but not exclusively:

- sensitivity: the proportion of biomarker-positive cases correctly identified
- specificity: the proportion of biomarker-negative cases correctly identified
- accuracy: the proportion of all cases correctly identified
- concordance: the proportion of results in agreement from repeat testing of the same samples
- robustness: the ability of the test to cope with relevant sample variables, which can be measured by multiple parameters
- validation: where an assay is developed using one cohort of samples, analytical validity (e.g. by the above criteria) must be shown in an independent set of samples

Furthermore, the standards for reporting diagnostic accuracy (STARD) compiled a list of 25 items to guide assessment of the analytical validity of a new biomarker test, which includes clear statement of the study aim, description of participant selection, number and demographic, sample processing, defining the comparator “gold standard” assay used, technical aspects of the biomarker test and the method of statistical analysis (Bossuyt *et al*, 2003). Clinical utility of a biomarker test is determined by the association between the test result and disease outcomes, and the feasibility of its deployment (Ray *et al*, 2010; Henry and Hayes, 2012). Key points to consider include:

- whether or not the results of the biomarker test influence clinical decisions
- the incidence or rate of side effects or adverse events in patients subject to the biomarker test
- the cost effectiveness of the biomarker test and interventions informed by the test result, which is frequently measured as the economic cost per patient life-year gained
- improvement in measurable clinical outcomes resulting from biomarker testing

By quantifying the analytical validity and clinical utility of a biomarker test it can be compared to alternative methods and strategies to select the most appropriate diagnostic tool.

Biomarker tests are available to assess various aspects of CRC management, from risk estimation to disease monitoring (Table 1.2). Early detection of cancer is viewed as the “holy grail” of cancer diagnostics due to superior prognosis and more favourable treatment options available for early stage disease (Etzioni *et al*, 2003). For example, 1 year survival rates in CRC are 98% in stage 1, 93% in stage 2, 89% in stage 3, but only 44% in stage 4, or metastatic, disease, based on the UK population in 2014-2015 (Broggio and Bannister, 2016). Biomarkers capable of early detection of CRC in cancer-predisposition syndromes are of particular interest. In Lynch syndrome and CMMRD, clinical guidelines state that gene carriers should have 1-2 yearly colonoscopies to screen for CRC and to prevent disease by polypectomy of precancerous adenomas (Vasen *et al*, 2013; Vasen *et al*, 2014). However, colonoscopy has its risks, such as perforation of the bowel in 0.5% patients, post-endoscopy bleeding in 0.26% patients and mortality in 0.003% patients (Reumkens *et al*, 2016), and is considered to be a highly invasive procedure (Fisher *et al*, 2011). In a prospective study of CRC incidence and mortality in Lynch syndrome patients, only 42.2% of patients were

Example Biomarker	Clinical Use	Reference
Germline <i>APC</i> mutation	Cancer-risk estimation	Groden <i>et al</i> , 1991
Faecal immunohistochemical testing	Early detection/screening	Lee <i>et al</i> , 2014
Tumour Dukes' or TNM staging and grading	Determine prognosis	Labianca <i>et al</i> , 2010
Tumour <i>KRAS</i> mutation (anti-EGFR antibody)	Predict therapeutic response	Allegra <i>et al</i> , 2009
Serum CEA	Monitor disease progression	Locker <i>et al</i> , 2006

CEA, carcinoembryonic antigen; EGFR, epidermal growth factor receptor; *KRAS*, Kirsten rat sarcoma viral oncogene homolog; MMR, mismatch repair.

Table 1.2. The utility of biomarkers in colorectal cancer (CRC)-related healthcare. CRC biomarkers are useful in the assessment of multiple aspects of disease. This is not an exhaustive list.

compliant with colonoscopic screening at the recommended 1-2 year intervals (Stuckless *et al*, 2012). Furthermore, although colonoscopic surveillance increases the median age at first CRC diagnosis in Lynch syndrome patients, it does not prevent all CRCs: it has been shown that 20% of males and 7% of females develop an interval CRC within 2 years of the previous colonoscopy (Stuckless *et al*, 2012). Comparison of surveillance protocols found no difference in CRC incidence or time to diagnosis since last colonoscopy between Lynch syndrome patients following longer (3 yearly) and shorter (1-2 yearly) intervals between colonoscopies (Møller *et al*, 2017a; Seppälä *et al*, 2017). This suggests that the rate of tumorigenesis is independent of the surveillance interval and therefore unaffected by an increase frequency of prophylactic polypectomy. The quality of colonoscopy is another factor to consider, as pre-cancerous lesions may be missed if the full extent of the colorectum is not visualised especially when there is a high rate of proximal (right-sided) CRCs in Lynch syndrome (Lynch *et al*, 2009). However, in a study of colonoscopy quality in Lynch syndrome, only 9% of interval cancers were detected in sections of the colorectum un-visualised in the previous colonoscopy, and only 21% of cancers were in the same location as an adenoma removed in the previous colonoscopy (suggesting incomplete polypectomy), leaving the origin of approximately 70% of interval cancers unexplained (Haanstra *et al*, 2013). Ahadova *et al* showed that 17.4% (95% CI: 7.8-31.4%) of Lynch syndrome CRCs have

mutations in *CTNNB1* that encodes β -catenin, a component of the Wnt signalling pathway. The majority of these CRCs (62.5%) lacked any evidence of polypous growth. Instead, they appeared to invade directly into the colorectal wall and so may be undetectable by colonoscopy and, therefore, are a plausible explanation for interval cancers in Lynch syndrome patients under colonoscopic surveillance (Ahadova *et al*, 2016). In addition, the thousands of MMR-DCF in the colorectum of healthy Lynch syndrome gene carriers (Kloor *et al*, 2012) have been associated with cMNR frameshift mutations (Staffa *et al*, 2015) and evolution into MMRd tumours (Ahadova *et al*, 2018), suggesting that these may also contribute to the undetectable and pre-cancerous lesions during colonoscopy. Therefore, less invasive methods, such as biomarker tests, for the early detection of MMRd CRC would greatly benefit both Lynch syndrome gene carriers and, most likely, CMMRD patients.

To screen patients with germline MMR gene defects for biomarkers of MMRd CRC, the relevant patients must first be identified. Currently Lynch syndrome is severely underdiagnosed, with an estimated 1.2% of Lynch syndrome gene carriers known to medical services in 2011 (Hampel and de la Chapelle, 2011). As will be discussed in more detail later, testing any and all CRCs for biomarkers of MMR deficiency can be used to detect potential cases of Lynch syndrome (Newland *et al*, 2017), given that Lynch syndrome accounts for approximately 23% of all MMRd CRCs (Hampel *et al*, 2008). MMR deficiency testing of cancers also informs the use of immunotherapy, with MMRd cancers of any type responding to immune checkpoint blockade (Le *et al*, 2017). Finally, the lack of MMR in all tissues of CMMRD patients can also be detected to complement genetic diagnosis (Bodo *et al*, 2015). Thus, there are several distinct clinical needs for biomarker tests of MMR deficiency.

1.8. Biomarkers for the Early Detection of Colorectal Cancer

Current methods (using biomarkers or otherwise) for early detection of CRC include analysis of tumour-derived nucleic acids in blood, detection of specific proteins (or proteomic signatures) in liquid biopsies such as blood or urine, and clinical examination to name just a few. However, each of these biomarkers have their limitations for screening and early detection, due to either a lack of analytical validity or clinical utility. Here I will discuss a few examples.

Circulating tumour DNA (ctDNA) constitutes a fraction of cell free DNA (cfDNA), which consist of 180bp fragments of genomic DNA released into circulation by apoptotic

cells, with fragments defined by the DNA structure around histone bodies (Jahr *et al*, 2001). ctDNA has conventionally been used for longitudinal monitoring of cancer progression and relapse (Taly *et al*, 2013 ; Schøler *et al*, 2017), and evolution of therapeutic resistance, for example mutation of *KRAS* to confer resistance to anti-EGFR therapy in CRC (Diaz *et al*, 2012; Siravegna *et al*, 2015). It was suggested ctDNA could be used for early detection in CRC when it was shown that 47% of stage 1 cancer patients had detectable ctDNA across multiple cancer types, including CRC (Bettegowda *et al*, 2014). However, due to its low abundance in a background of cfDNA from non-neoplastic cells, detection of ctDNA requires very sensitive techniques. Picodroplet digital PCR is one such technique that has been effective in CRC diagnostics (Taly *et al*, 2013; Bettegowda *et al*, 2014), and more recently next generation sequencing (NGS)-based methods using very high read depths have been employed (Shu *et al*, 2017). Unfortunately, these techniques are costly and time-consuming. Furthermore, ctDNA is usually quantified by the fraction of cfDNA which contains mutations present in the cancer (Bettegowda *et al*, 2014), and prior knowledge of the tumour is incompatible with screening for early diagnosis. Instead, extensive gene panels of frequently mutated genes can be assessed, such as the 382 gene panel used by Shu *et al* (2017), but this will further increase cost. Additional challenges include distinguishing between mutations that occur during natural aging from those associated with malignancy, and defining cancer location (Aravanis *et al*, 2017). Circulating micro RNAs have also been shown to detect advanced colorectal adenomas with 73.0% sensitivity and 79.7% specificity (Huang *et al*, 2010) and aberrant methylation of *APC*, *MGMT* and other genes in ctDNA has 86.5% sensitivity and 92.1% specificity for stage 1 and 2 CRC (Lee *et al*, 2009), highlighting that there are several avenues of research into use of circulating nucleic acids for the early detection of CRC.

Single protein biomarkers from liquid biopsy often have poor sensitivity, particularly for early stage disease (Borrebaeck *et al*, 2017). For example, carcinoembryonic antigen (CEA) is a clinically accepted serum biomarker for multiple cancers, but its sensitivity is only 21% for stage 1 CRC and 37% overall, irrespective of CRC stage (Su *et al*, 2012). Combinations of protein biomarker can improve detection; Zhang *et al* reported sensitivity of 94% and specificity of 98% for CRC using a panel of 4 serum peptides CA199, CA242, CA125, and CA153 (Zhang *et al*, 2016). However, meta-analyses have highlighted that studies conducted in a clinical setting such as that of Zhang *et al* where patients have consulted a clinician due to related symptoms or disease, rather than screening in average-risk individuals, can bias

results, with receiver operating characteristic (ROC) area under curve (AUC) values being higher in clinic- versus screening-based studies (AUC range 0.68-0.996 versus 0.62–0.78 respectively) (Bhardwaj *et al*, 2017). New approaches harnessing proteomics are in development in numerous cancer types, with a view to analysing protein biomarker signatures, using methods such as multiplexed ELISA and bead-based arrays (Borrebaeck *et al*, 2017). Non-blood based approaches for CRC screening include faecal immunochemical testing (FIT), which quantifies the micrograms of haemoglobin per gram of stool as a biomarker of colorectal bleeding. As a screening tool for early detection of cancer in average risk populations, FIT has a sensitivity of 79% (95% CI: 69-86%), and specificity of 94% (95% CI: 92-95%) using thresholds of >20µg/g, based on a meta-analysis of nineteen studies (Lee *et al*, 2014). However, FIT sensitivity for right-sided (proximal) CRC is as low as 20% (95% CI: 11-31%) (Haug *et al*, 2011). Therefore, like circulating nucleic acids, protein biomarkers of CRC can have very low sensitivity for early stage or pre-malignant tumours, and research is ongoing to identify novel markers and develop multi-marker panels to increase sensitivity (Borrebaeck *et al*, 2017).

1.9. Early Detection of Mismatch Repair Deficient Colorectal Cancer using Autoantibodies

The immune response against MMRd CRC may provide a novel source of biomarkers for early detection of CRC (Reuschenbach *et al*, 2010). Immune cells and signals, such as inflammatory cytokines, circulate throughout the body. For example, T cells reactive to FSP TAAs have been isolated from the peripheral circulation as well as the TIL population (Saeterdal *et al*, 2001; Schwitalle *et al*, 2008). It may be feasible to detect immunological biomarkers from liquid biopsy for early detection of MMRd CRC, which would be a less invasive surveillance method than the currently recommended colonoscopy.

The humoral immune response and the generation of autoantibodies (antibodies targeting antigens derived from self-molecules) against TAAs has been explored in numerous cancer types. For example, *TP53* mutations frequently lead to over expression of p53 protein in tumour cells and so it was hypothesised that p53 autoantibodies would be detectable in CRC patients. Angelopoulou *et al* (1997) found that 53/229 (23%) of CRC patients, but no controls, had p53 autoantibodies in their serum. Hammel *et al* (1997) found similar results with p53 autoantibodies in sera from 14/54 (26%) CRC patients, but not from controls. In addition, they showed over-expression of p53 in tumours from 22 patients, 10 of whom had

p53 autoantibodies, suggesting that p53 over-expression can, but is not necessary to, stimulate autoantibody production (Hammel *et al*, 1997). In three autoantibody-positive patients, antibody titre correlated with CEA concentration over 10-16 months of clinical follow up (Angelopoulou *et al*, 1997), and in 11/13 patients p53 autoantibodies decreased following surgical resection of the tumour (Hammel *et al*, 1997). Agreement with these results has been found in other cancer types. For example, in a study of breast cancer, 9% of patients had p53 autoantibodies, but analysis of p53 accumulation in tumour tissues showed only a weak association between over-expression of p53 and detectable autoantibodies ($p = 0.05$) (Angelopoulou *et al*, 2000). A meta-analysis of >130 publications has concluded that p53 autoantibodies have 30% sensitivity and 96% specificity in cancers tested, and that the signal is associated with p53 accumulation in the tumour and missense mutations (Soussi *et al*, 2000).

The low sensitivity of a single antibody assay can be improved by panel testing, for example using microarrays (Robinson *et al*, 2002). A 22-TAA panel for autoantibody detection in sera was 81.6% sensitive and 88.2% specific for prostate cancer, with a higher ROC AUC than prostate serum antigen (0.93 versus 0.80) (Wang *et al*, 2005) and analysis of autoantibodies against a panel of 3 TAAs had 55% sensitivity and 95% specificity for preclinical lung cancer (Pereira-Faca *et al*, 2007). These and other studies suggest that autoantibodies can be used to monitor disease and may be applicable to early detection (Desmetz *et al*, 2011). As proof of principle, antibody titres against TAAs have previously been used to predict cancer incidence: increased serum titres of p53 autoantibodies were significantly associated with lung cancer incidence in a high risk group, with an average lead time to diagnosis of 3.5 years (Li *et al*, 2005).

Currently, tests for autoantibodies, which include ELISA and microarrays, lack the analytical validity required of biomarkers for early detection due to generally low sensitivity and inadequate specificity – such screening tests need to have exceptionally high specificity to avoid over-diagnosis, with a recommended ROC AUC >0.95 (Hartwell *et al*, 2006). However, autoantibodies are still attractive candidate biomarkers for several reasons. For instance, the immune response to TAAs amplifies the signal from transient antigens that would be near impossible to detect otherwise. Also, handling and storage of samples is simplified by the stability of antibodies in serum, due to resistance to proteolysis that affects other peptides, and antibodies are particularly stable in serum *in vivo* with half-lives >7 days,

meaning that the timing of sample collection is not critical (Anderson and LaBaer, 2005). Finally, autoantibodies may be more sensitive and specific in immunogenic cancer types.

Autoantibodies against FSPs (α FSP-Abs) associated with MMRd cancer have been detected. Ishikawa *et al* (2003) generated a λ phage-display library from cDNA of three MMRd CRC cell lines and screened for TAAs using sera from an MMRd CRC patient. Serum antibodies were present against 64 antigens, 49 of which were shown to be specific to MMRd CRC by exposure of these antigens to sera from controls and other cancer patients. Significantly, one antigen was associated with a frameshift mutation in a G7 tract of *CDX2*, which would lead to a novel 30 amino acid sequence at the C terminus of the CDX2 protein. This frameshift mutation was also found in the tumour of the patient with serum antibodies against this FSP. Using ELISA, Reuschenbach *et al* (2010) exposed 6 FSPs associated with common cMNR frameshift mutations in MMRd CRCs to sera from 69 MMRd CRC patients, and autoantibodies against FSPs derived from TAF1B and TGF β R2 were observed in 8 (11.6%) and 7 (10.1%) patients, but only 3 (5.8%) and 1 (1.9%) controls respectively. Whilst these anti-FSP antibodies (α FSP-Abs) individually had very low sensitivity, a multiplexed, bead-based method, which assesses 32 α FSP-Abs simultaneously, has been developed that may be able to address the low sensitivity of single autoantibody biomarkers. As an initial test of this method, it was shown to be able to detect ASTE1-FSP and TAF1B-FSP autoantibodies in patients vaccinated with the respective synthetic peptide, but the number of sera tested was limited (Reuschenbach *et al*, 2014).

Autoantibody tests for early detection of MMRd CRC is an intriguing prospect. Due to the frequent frameshift mutations in multiple cMNRs that are intrinsic to tumour progression (Woerner *et al*, 2010), it is possible to build a panel of synthetic FSPs with confidence that some of the corresponding mutations will exist in preclinical tumours, such as in the 32 FSP panel designed by Reuschenbach *et al* (2014). Once detected, α FSP-Abs may also be able to monitor disease, as has been observed with p53 autoantibodies (Angelopoulou *et al*, 1997; Hammel *et al*, 1997) and observed by loss of the anti-CDX2 FSP antibody from one patient's serum 7 years after CRC-resection (Ishikawa *et al*, 2003). Another potential advantage of analysing α FSP-Abs is that they are proof of an immune response against the cancer, and may provide information on prognosis and therapeutic response. For example, the immunoscore method of characterising TILs is a better predictor of patient prognosis in CRC than MMR deficiency (Mlecnik *et al*, 2016). This is particularly

significant as not all MMRd cancers respond to immune checkpoint blockade and, whilst some may be explained by evolution of alternative immune evasion mechanisms (Sade-Feldman *et al*, 2017), this may be due to an inadequate anti-tumour immune response. Quantification of the immune response by α FSP-Abs as a mechanism-driven biomarker of therapeutic response would, therefore, be superior to qualifying mutational load by MMR deficiency testing alone and assuming immunogenicity (Topalian *et al*, 2016). This is particularly significant as immune checkpoint blockade produces numerous side effects similar to auto-immune disease, some of which can be severe (Postow *et al*, 2018), meaning it is critical to target suitable patients. However, the analytical validity and clinical utility of α FSP-Abs needs to be explored.

1.10. Mismatch Repair Deficiency testing to identify Cancer-predisposition Syndromes

Biomarkers of MMR deficiency can be used in the diagnosis of Lynch syndrome and CMMRD.

1.10.1. Diagnosing Lynch syndrome

Lynch syndrome accounts for a large proportion of hereditary CRC and the availability of disease-preventing and, ultimately, life-saving options make the identification of Lynch syndrome gene carriers an important task for healthcare providers (Vasen *et al*, 2013). However, in 2011 it was estimated that only 1.2% of all Lynch syndrome gene carriers were known (Hampel and de la Chapelle, 2011) despite frequency-estimates of one carrier per 370-1000 of the population, based on Finnish and American statistics (Aaltonen *et al*, 1998; Hampel and de la Chapelle, 2011). This is equivalent to one million carriers in Europe (Vasen *et al*, 2010). In the Icelandic population, which has experienced a genetic bottleneck, the frequency is as high as one carrier per 226 (Haraldsdottir *et al*, 2017).

Historically, Lynch syndrome was diagnosed following clinical indicators defined in the Amsterdam criteria, including a family history of Lynch-spectrum cancers and early onset of disease (Vasen *et al*, 1991; Vasen *et al*, 1999). The Bethesda guidelines, defined at the NCI meeting in 1996, used less stringent familial criteria but included assessment of other disease features, such as adenoma incidence (Table 1.3). Amsterdam and Bethesda criteria performance has been tested in registered Lynch syndrome families, revealing sensitivities of 23% and 70% respectively (Terdiman *et al*, 2001). However, testing in families previously identified by family history and age of disease onset creates an ascertainment bias and

confounds results, an issue recognised during revision of the Bethesda guidelines in 2002 (Umar *et al*, 2004). Alternative screening strategies were encouraged (Umar *et al*, 2004), and it was concurrently suggested that tumours should be tested for MMR deficiency to select patients for MMR gene testing to identify pathogenic germline variants (Rodriguez-Bigas *et al*, 1997). Aaltonen *et al* (1998) showed that fragment length analysis of PCR-amplified microsatellites (MSI FLA) of an unselected cohort of 509 CRCs detected 63 (12%) MMRd tumours and, by germline genetic testing of this selected population, identified 10 Lynch syndrome cases with mutations in *MLH1* or *MSH2*. A direct comparison of molecular and clinical screening strategies for the detection of path_*MLH1* or path_*MSH2*, carriers in a cohort of 1222 CRC patients from the EPICOLON I study, found sensitivities and specificities of 90.9% and 93.9% for MSI FLA, 81.8% and 94.2% for immunohistochemistry to detect loss of MMR protein expression (MMR IHC), and 90.9% and 77.1% for Bethesda criteria (Piñol *et al*, 2005). This showed that testing for biomarkers of MMR deficiency in CRC is

Guidelines	Criteria
Amsterdam II	<p>At least 3 relatives with a Lynch-associated cancer (CRC, cancer of the endometrium, small bowel, ureter, or renal pelvis). All of the following criteria should be met:</p> <ul style="list-style-type: none"> • One should be the first-degree relative of the other 2 • At least 2 successive generations should be affected • At least 1 CRC should be diagnosed before age 50 yr Familial adenomatous polyposis should be excluded
Bethesda	<p>Individuals with cancer in families that fulfil the Amsterdam criteria</p> <p>Individuals with 2 Lynch-related cancers, including synchronous or metachronous CRCs or associated extracolonic cancers</p> <p>Individuals with CRC and a first-degree relative with CRC and/or Lynch-related extracolonic cancer and/or colorectal adenoma; 1 of the cancers diagnosed at age <45 yr and the adenoma diagnosed at <40 yr</p> <p>Individuals with CRC or endometrial cancer diagnosed at <45 yr</p> <p>Individuals with right-sided CRC with an undifferentiated pattern (solid/cirbriform) on histopathology diagnosed at <45 yr</p> <p>Individuals with signet ring cell–type CRC diagnosed at <45 yr</p> <p>Individuals with adenomas diagnosed at <40 yr</p>

Table 1.3. Clinical criteria for Lynch syndrome screening. The Amsterdam II and Bethesda criteria for the identification of Lynch syndrome families, as summarised by Terdiman *et al*, 2001.

superior to clinical criteria for identification of Lynch syndrome. A more recent comparison of screening strategies has confirmed the superiority of MSI FLA or MMR IHC over Bethesda criteria using a cohort of 2093 CRC patients, with molecular screening of CRCs having 100% sensitivity and 92% specificity, whilst Bethesda criteria only had 86% sensitivity and 78% specificity for Lynch syndrome (Pérez-Carbonell *et al*, 2012).

The lower sensitivity and specificity of clinical screening using family history and age of onset of disease is likely due to the criteria being based on patient characteristics influenced by ascertainment bias. Early data suggested the median age of CRC diagnosis in Lynch syndrome gene carriers was <50 years, but this was largely based on probands. In contrast, the median age of diagnosis is 61.2 years for CRC and 62.0 years for EC in mutation positive members of Lynch families when probands are excluded (Hampel *et al*, 2005b). This older age of disease onset was confirmed when molecular screening of 1117 CRCs diagnosed <70 years of age showed that 70% of Lynch syndrome CRCs were diagnosed in patients over 50 years (van Lier *et al*, 2012). Also, the lower penetrance of path_*MSH6* and path_*PMS2* mutations mean that family histories are less obvious than in path_*MLH1* and path_*MSH2* families (Kariola, 2004; Sjursen *et al*, 2010), and familial criteria do not exclude the broader HNPCC phenotype, which includes patients associated with FCCTX (Figure 1.5; Lindor *et al*, 2005). Furthermore, the presence of *BRAF* mutation and CIMP in sporadic MMRd, but not Lynch CRCs, can be used to improve the specificity of molecular screening for Lynch syndrome (Parsons *et al*, 2012). *BRAF* V600E testing, for example, allows identification and removal of approximately 40% of sporadic MMRd CRCs from Lynch syndrome screening pipelines (Domingo *et al*, 2004). Similarly, *MLH1* methylation testing of MMRd CRCs removes up to 78% of sporadic cases (Pérez-Carbonell *et al*, 2010). However, despite its greater specificity for Lynch syndrome over *BRAF* V600E testing, *MLH1* methylation testing reduces sensitivity by exclusion of Lynch syndrome tumours with *MLH1* methylation as the second hit, which occurs in 53% of CRCs arising in path_*MLH1* gene carriers (Kaz *et al*, 2007), and Lynch syndrome tumours associated with germline *MLH1* epimutation (Suter *et al*, 2004).

The cost-effectiveness of molecular screening is agreed across multiple studies, accounting for the cost of screening and cascade testing of family members of genetically confirmed probands against the benefits of surveillance and prophylaxis (Mvundura *et al*, 2010; Ladabaum *et al*, 2011). A comprehensive economic evaluation of multiple screening strategies (Snowsill *et al*, 2014) concluded that MSI FLA or MMR IHC, followed by *BRAF*

V600E or *MLH1* methylation testing, in CRCs diagnosed <70 years of age is a cost effective medical intervention, with incremental cost-effectiveness ratios as low as £5,491 per Quality Adjusted Life Year (QALY)-gained, which is below the £20,000 threshold set by UK National Institute of Health and Care Excellence (NICE). The UK Royal College of Pathologists included MMR deficiency testing of all CRCs diagnosed <50 years of age in their 2014 Dataset for colorectal cancer histopathology reports (Loughrey *et al*, 2014) and, subsequently, NICE published its Diagnostic Guidance 27 (DG27) stating that all CRCs should be tested for MMR deficiency, by MSI FLA or MMR IHC, followed by *BRAF* V600E or *MLH1* methylation, to select patients for germline MMR gene testing (Newland *et al*, 2017). Similar guidelines can be found from the American Society for Clinical Pathology, and the European Society for Medical Oncology (Balmana *et al*, 2013; Stoffel *et al*, 2015). Despite these guidelines and the severe under-diagnosis of Lynch syndrome gene carriers (Hampel and de la Chapelle, 2011), only 28.2% of 152,993 CRCs diagnosed in the US were tested for MMR deficiency during 2010-2012 (Shaikh *et al*, 2018). Whilst rates were increasing, from 22.3% in 2010 to 33.1% in 2012, these statistics are of concern given that there is over a decade of literature and guidelines supporting MMR deficiency testing to screen for Lynch syndrome (Hamilton, 2018). These rates are also low in comparison to germline genetic testing for *BRCA1/2* in young (aged ≤45 years) breast cancer patients, for whom testing is recommended by US guidelines, with 65.3% of patients aged 41-45 years being tested in 2012, and 72.9% of those aged ≤40 years (Kehl *et al*, 2016).

1.10.2. Diagnosing constitutional mismatch repair deficiency

CMMRD is a less significant healthcare burden compared to Lynch syndrome due to its rarity. However, the near certainty of cancer diagnosis in these patients, and the availability of surveillance guidelines makes identification critical for their clinical management (Vasen *et al*, 2014). Currently, clinical criteria can be used to select patients for CMMRD genetic testing based on their malignant, pre-malignant, and non-neoplastic features, as defined by the C4CMMRD consortium in 2014 (Table 1.4). Unfortunately, the pleiotropic phenotype of CMMRD requires many criteria to be considered, and this is further complicated by overlap with other syndromes, in particular NF1. Also, the families of biallelic *PMS2* patients only have a low incidence, or even lack, of Lynch-spectrum cancers (De Vos *et al*, 2006; Urganci *et al*, 2015), which can be explained by the much lower penetrance of Lynch syndrome in

path_ *PMS2* heterozygotes compared to other MMR genes (Møller *et al*, 2017b). Family history and patient phenotype, as in Lynch syndrome screening, is not always a useful criterion for CMMRD diagnosis. Therefore, germline genetic testing by diagnostic sequencing of all MMR genes is required to confirm diagnosis, and is a viable frontline test, given the rarity of the disease. However, *PMS2* mutations account for nearly 60% of CMMRD (Wimmer

Criteria	Points
Indication for CMMRD testing in a cancer patient, add points from malignancies/pre-malignancies and additional features listed below	≥3
Carcinoma from the Lynch-spectrum* at age <25 years	3
Multiple bowel adenomas at age <25 years and absence of <i>APC/MUTYH</i> mutation(s) or a single high-grade dysplasia adenoma at age <25 years	3
WHO grade III or IV glioma at age <25 years	2
NHL of T-cell lineage or sPNET at age <18 years	2
Any malignancy at age <18 years	1
Clinical sign of NF1 and/or ≥2 hyperpigmented and/or hypopigmented skin alterations Ø>1 cm in the patient	2
Diagnosis of Lynch syndrome in a first- or second-degree relative	2
Carcinoma from Lynch-spectrum* before the age of 60 in first-degree, second-degree, and third-degree relative	1
A sibling with carcinoma from the Lynch-spectrum*, high-grade glioma, sPNET or NHL	2
A sibling with any type of childhood malignancy	1
Multiple pilomatricomas in the patient	2
One pilomatricoma in the patient	1
Agenesis of the corpus callosum or non-therapy-induced cavernoma in the patient	1
Consanguineous parents	1
Deficiency/reduced levels of IgG2/4 and/or IgA	1
*Colorectal, endometrial, small bowel, ureter, renal pelvis, biliary tract, stomach, bladder carcinoma. sPNET, supratentorial primitive neuroectodermal tumours.	

Table 1.4. Clinical criteria for constitutional mismatch repair deficiency (CMMRD) screening.

Germline testing of MMR genes for biallelic mutation is considered when patients present with one of the listed malignancies or pre-malignancies and have points ≥3, by addition of points from multiple features. These consensus criteria were agreed by the C4CMMRD consortium. Adapted from Wimmer *et al*, 2014.

et al, 2014) and is a known dead zone of diagnostic sequencing (Mandelker *et al*, 2016). The difficulty in sequencing and interpreting *PMS2* variants can cause uncertain genetic diagnosis (Nakagawa *et al*, 2004), as can any other MMR variant of unknown significance. Assays are needed to segregate CMMRD from phenotypically-similar syndromes and confirm genetic diagnoses by proof of functional impact of any variants detected.

IHC to show a loss of MMR in non-neoplastic tissues has frequently been used to clarify CMMRD diagnosis. Alternatively, the detection of MSI in non-neoplastic tissues is another functional assessment of CMMRD, but MSI cannot be detected by standard FLA techniques due to the very weak signal. Small pool PCR of microsatellite markers allows FLA of products from single or low copy number templates to achieve sufficient sensitivity, but the protocol is laborious as it requires analysis of hundreds of PCR products per sample (Parsons *et al*, 1995). To improve biomarker tests of MMR deficiency in non-neoplastic tissues of CMMRD patients, Ingham *et al* developed the germline MSI (gMSI) assay to detect MMR deficiency in peripheral leukocytes from blood. The assay uses multiplexed PCR to amplify and fluorescently label three dinucleotide repeats (DNRs), and capillary electrophoresis traces of amplicons are analysed by comparing heights of the major peak (which varies from patient to patient due to polymorphisms of marker length) with the +1repeat/+2bp peak (defined relative to the major peak). Individual markers achieved up to 100% sensitivity and 98.9% specificity in 8 CMMRD patients with biallelic *PMS2* or *MSH2* mutations and 90 controls (Ingham *et al*, 2013). *Ex vivo* MSI (*evMSI*) is an alternative assay that detects MSI in primary lymphoblastoid cell lines (LCLs) derived from patient leukocytes, using a similar fluorescently-labelled FLA method as gMSI. However, three MNR rather than DNR markers are used, including two markers found in the Promega MSI Analysis System. In a cohort of 14 genetically confirmed CMMRD patients, including biallelic *MLH1*, *MSH6* and *PMS2* mutations, together with 23 controls (including 12 Lynch syndrome gene carriers), *evMSI* was 100% sensitive and specific (Bodo *et al*, 2015). A technique explored in parallel by Bodo *et al* was methylation tolerance of LCLs, based on the logic that MMR is required to initiate cell cycle arrest and apoptosis in response to methylation damage to the DNA (Karran and Stephenson, 1990). Quantifying LCL tolerance of the methylating agent MNNG by cell survival assays had equal sensitivity and specificity to *evMSI* (Bodo *et al*, 2015).

1.11. The Inadequacies of Current Biomarker Tests for Mismatch Repair Deficiency

Immune checkpoint blockade has proven to be a highly effective therapy in MMRd cancers (Le *et al*, 2015; Le *et al*, 2017), and MMR deficiency is associated with cancer-predisposition syndromes Lynch syndrome and CMMRD, both of which have extensive guidelines for disease management (Vasen *et al*, 2013; Vasen *et al*, 2014). MMR deficiency testing is therefore applicable to three scenarios; **1**, selection of patients eligible for immune checkpoint blockade therapies such as pembrolizumab, which has been approved by the FDA as a second line therapy in any MSI-high cancer (MERCK & Co. Inc, 2017), **2**, screening of all CRCs to identify Lynch syndrome, which is currently severely underdiagnosed, as per NICE DG27 (Newland *et al*, 2017), and **3**, assessment of non-neoplastic tissue to clarify CMMRD diagnosis, which is otherwise complicated by a pleiotropic phenotype and sometimes difficult genetic diagnosis (Wimmer *et al*, 2014).

MMR deficiency tests used in routine clinical practice include MSI FLA and MMR IHC. MSI FLA initially used PCR amplification of a variety of microsatellite markers, including MNRs, DNRs and others, and PCR amplicons were analysed by Southern blotting. Instability at a marker was observed by the presence of amplicons of a novel length relative to the lengths observed in tissue from matched normal (i.e. non-neoplastic) DNA. Matched normal DNA was important to account for length polymorphisms in the germline, which are common to microsatellites, and to help resolve the “stutter bands” produced by PCR error (Figure 1.10) (Aaltonen *et al*, 1993). However, different laboratories favoured different markers and so an optimal panel of microsatellite markers to unify MSI diagnostics was defined by the NCI, including two MNRs (BAT25, BAT26) and three DNRs (D5S346, D2S123, D17S250) (Boland *et al*, 1998). These markers were chosen from panels of approximately 30 markers based on sensitivity for MMR deficiency and ease of interpretation (Bocker *et al*, 1997; Dietmaier *et al*, 1997). Diagnostic thresholds were specified as follows: tumours with $\geq 30\%$ of markers showing instability were to be classified as MSI-high, tumours showing instability in $< 30\%$ of markers as MSI-low, and tumours showing no markers with instability as MSS (Thibodeau *et al*, 1998). However, using the NCI panel of two MNRs and three DNRs, MSH6 deficient tumours were frequently misclassified as MSI-low or MSS (Wu *et al*, 1999) due to MSH6 being critical to maintaining MNR stability but not microsatellites with longer repeat units due to redundant repair by MutS β (Verma *et al*, 1999). Adoption of panels exclusively composed of MNRs correctly classifies 97.7-100% of MSH6 deficient tumours as

MSI-high (You *et al*, 2010). Furthermore, the quasi-monomorphism of the MNRs used reduces the need for matched normal DNA (Suraweera *et al*, 2002), and MNRs produce less PCR stutter relative to DNRs (Buhard *et al* 2004); hence MNRs can be used to re-classify MSI-low tumours as MSS due to removal of ambiguous results from DNRs (Murphy *et al*, 2006). A multiplexed, fluorescently-labelled FLA of 5 MNRs by capillary electrophoresis achieved near 100% accuracy for MMR deficiency diagnosis (Bacher *et al*, 2004), and has been developed into the widely used MSI Analysis System (Promega) (Figure 1.10). A recent meta-analysis using nine high quality studies, for example by excluding those that measure diagnostic accuracy in case-control populations as this can inflate results (Rutjes *et al*, 2005), found that the sensitivity of MSI FLA testing CRCs for the detection of Lynch syndrome ranged from 67% to 100% (Coelho *et al*, 2017). This wide variability could be due to rates of MSI-high in Lynch syndrome CRCs, however, all of the studies in this meta-analysis used out-dated MSI marker panels, such as mixtures of mono-, di- and tetra-nucleotide repeats, or the NCI marker panel of two MNRs and three DNRs (Boland *et al*, 1998), which is the likely explanation for lower estimates of sensitivity, highlighting the importance of using appropriate markers.

MMR IHC is used to detect loss of MMR protein expression resulting from loss of function mutations (Leach *et al*, 1996; Thibodeau *et al*, 1996), and has been accepted as an alternative to MSI FLA (Boland *et al*, 1998). Comparisons of IHC and FLA found 91-98% concordance of results, with lower sensitivity using IHC (90.6-95.2%) (Thibodeau *et al*, 1998; Chapusot *et al*, 2002; Chapusot *et al*, 2004). However, in these early studies only MLH1, MSH2 and, variably, MSH6 proteins were analysed, reducing sensitivity. Also, a meta-analysis of IHC performance observed only 74% sensitivity for loss of MLH1 function (Shia, 2008). This low sensitivity of IHC for MLH1 deficiency was often caused by loss of function missense mutations that retained protein antigenicity (Salahshor *et al*, 2001; Wahlberg *et al*, 2002). Staining for the full complement of MMR proteins increases the sensitivity and specificity of IHC for MMR deficiency to near 100%, equivalent to MSI FLA (Hampel *et al*, 2005a; Southey *et al*, 2005). This is partly due to negative PMS2 staining identifying those cases where *MLH1* mutations produce non-functional but antigenic MLH1 protein, due to degradation of PMS2 that fails to complex with the non-functional MLH1 to form the MutL α heterodimer (de Jong *et al*, 2004b; Mangold *et al*, 2005).

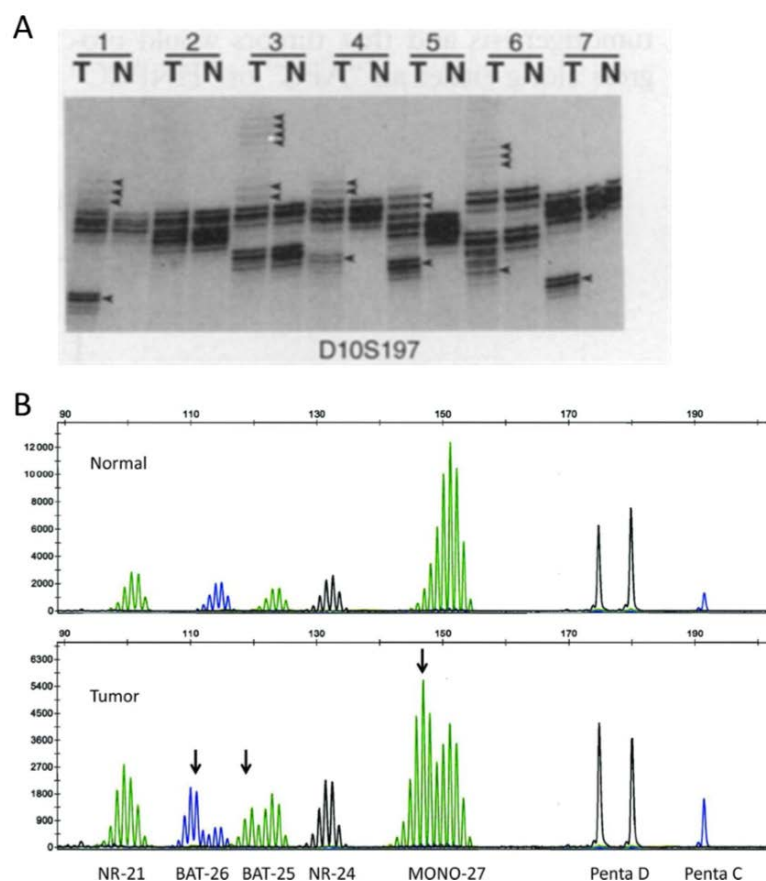


Figure 1.10. Microsatellite instability (MSI) detection by fragment length analysis (FLA). (A) PCR and FLA by Southern blot of a dinucleotide repeat shows that some tumours (T) have amplicons of lengths not detected in matched normal DNA from the same patient (N), as indicated by arrows. These length changes are a signal of indel mutations in the microsatellite, and a high burden of these (in $\geq 30\%$ of markers) is referred to as MSI-high, a biomarker of mismatch repair (MMR) deficiency. Note the stutter bands from both matched normal and tumour DNAs due to PCR error. Taken from Aaltonen *et al*, 1993. **(B)** PCR of the MSI Analysis System (Promega) panel of five mononucleotide repeats and FLA of fluorescently-labelled amplicons using capillary electrophoresis, of DNA from a MMR deficient endometrial cancer and matched normal DNA. Again, amplicons from the tumour DNA of lengths not detected from the matched normal DNA are indicated by arrows and are a signal of indel mutations in the microsatellites. Penta-C and Penta-D are pentanucleotide repeats used in genetic fingerprinting and confirm identity of tumour and matched normal DNA. Stutter peaks are caused by PCR error. Taken from Tafe *et al*, 2014.

MMR deficiency testing, by MSI FLA or MMR IHC, has established analytical validity over many years of use in clinical practice. Both tests have high diagnostic accuracy for MMR deficiency, irrespective of which MMR gene is affected, when appropriate methods are used, as described above. With respect to reproducibility, MMR IHC can give variable results due to subjective interpretation of staining patterns, which can be difficult with poor tissue quality, and to differences in protocols and antibodies used between laboratories (Shia,

2008). In contrast, MSI FLA is considered highly reproducible, with >95% concordance of results generated by independent pathology laboratories (Boyle *et al*, 2014b). Both techniques are considered robust, being applicable to fresh, high quality material, and to formalin fixed paraffin embedded (FFPE), low quality material (Chapusot *et al*, 2002; Bacher *et al*, 2004). Another sample variable that can affect diagnostic test results is sample purity and whether or not the analyte of interest is at a high enough representation in the sample to be detected by the assay being used, often referred to as the lower limit of detection (LLoD) of the assay (Armbruster and Pry, 2008). In the case of MMR deficiency testing of CRCs, this is equivalent to the MMRd tumour cell content of the tissue being analysed or from which DNA was extracted. MSI FLA can detect 10% MMRd tumour cell content (Berg *et al*, 2000) and focal absence of MMR staining is detectable by IHC (Chapusot *et al*, 2002), therefore both tests have a LLoD that is lower than the typical >20% tumour cell content used for diagnostic sequencing (Smits *et al*, 2014).

The clinical utility of MSI FLA and MMR IHC is evident in the impact a diagnosis of MMR deficiency in CRC has on clinical decisions, the benefit to patients, and the cost-effectiveness of MMR deficiency testing to screen for Lynch syndrome, as summarised in the economic evaluation of Snowsill *et al* (2014). However, the demands of cancer diagnostics have changed with the release of NICE diagnostic guidance, which recommend MMR deficiency testing of all 41,000 CRCs detected per annum in the UK (Cancer Research UK Statistics, 2015; Newland *et al*, 2017), and with the FDA approval of MMR deficiency testing as a companion diagnostic for immune checkpoint blockade in all cancers (MERCK & Co. Inc, 2017). As discussed earlier, clinical uptake of testing has been poor (Shaikh *et al*, 2018), which suggests current methods of MMR deficiency testing are inadequate. Methods that are not only accurate, reproducible, robust and cost-effective, but also applicable to high throughput testing, are now needed. In this context the key disadvantage of both MSI FLA and MMR IHC is the reliance on expert interpretation of results on a case-by-case basis, which is feasible for low numbers of samples, but time consuming and costly otherwise (Shia, 2008; Zhang, 2008). In particular, the stutter bands/peaks and quasi-monomorphism of the markers used in MSI FLA ideally require matched normal DNA and multiple interpreters to ensure correct classification (Lindor *et al*, 2006; Zhang, 2008), while IHC requires trained and skilled pathologists to both process samples and analyse tumour

histology and variable staining patterns (Shia *et al*, 2008). Novel biomarker tests of MMR deficiency are therefore needed to meet new clinical demand.

With respect to CMMRD diagnostics, IHC is insensitive to variants that produce antigenic but non-functional proteins (Sjursen *et al*, 2009; Mork *et al*, 2016), and ideally requires internal positive control staining of TILs or other stromal cells (Shia *et al*, 2008), which is clearly not possible in CMMRD. Also, IHC relies on solid tissue biopsies, which are difficult to acquire for normal tissue. MSI FLA fails to detect CMMRD in non-neoplastic tissues unless small pool PCR is used (Wimmer *et al*, 2008), which is a laborious technique that requires careful dilution of template DNA and hundreds of PCRs per sample (Parsons *et al*, 1995). A significant advantage of the gMSI assay presented by Ingham *et al* is the simple laboratory workflow and automatable analysis. However, due to the use of DNRs, gMSI cannot detect patients with biallelic *MSH6* mutations (Ingham *et al*, 2013). *evMSI* uses MNRs and can detect biallelic *MSH6* mutations, but is an expensive and time consuming technique, requiring approximately 120 days of cell culture post immortalisation of LCLs to develop the MSI signal (Bodo *et al*, 2015). Similarly, time and cost of LCL culture hinders clinical utility of MNNG tolerance assays (Bodo *et al*, 2015). Ideally, an assay of MSI in normal tissue would combine the simple workflow and analysis of gMSI with the sensitivity and specificity of LCL *evMSI* or MNNG tolerance, which is capable of detecting deficiency of all MMR genes.

1.12. Detection of Microsatellite Instability using Next Generation Sequencing

Informative biomarkers can be found in the cancer genome. As an example from CRC, gain of function mutations in the *KRAS* oncogene predict resistance to anti-epidermal growth factor therapies such as cetuximab (Allegra *et al*, 2009). Frampton *et al* (2013), for example, showed the power of NGS to test for these genetic biomarkers by detection of base substitutions, indels, CNAs and gene fusions across 287 cancer-associated genes. 83 cell lines, with thoroughly characterised mutations, were used for analytical validation of the pipeline, showing 95-99% sensitivity and >99% specificity across mutation types. When the same NGS gene panel and analysis pipeline was applied to 2,221 cancers from routine clinical services, actionable mutations were identified in 76% of cancers, which was approximately a 3-fold higher yield than established, non-NGS-based biomarker tests. Significantly, NGS can be deployed for high throughput cancer diagnostics due to simple

laboratory protocols for library preparation, and automated analysis of potentially thousands of samples in parallel (Frampton *et al*, 2013).

Several software packages have been developed to classify MSI status based on NGS data. Lu *et al* (2013) identified 505,657 microsatellites from Ref-seq transcriptomic data (O'Leary *et al*, 2016) that could be assessed in RNA-seq of tumours. The ratio of indels at microsatellites versus indels detected at non-microsatellite loci was used to determine MMR status with 100% accuracy relative to MSI FLA in a cohort of 14 MMRd and 14 MMRp CRCs (Lu *et al*, 2013). Another package, MSIsensor (Niu *et al*, 2014), applicable to whole genome, whole exome or gene panel sequencing data, computes the sequencing read count distribution across different microsatellite lengths detected at each locus in the tumour, and compares these to read count distributions from matched normal DNA. Chi-squared tests for significant differences in these distributions determines if a locus is unstable or stable, and the proportion of unstable loci is used to classify MMR status. When implemented in exome sequencing data from 242 ECs, MSIsensor had 98.6% sensitivity and 97.6% specificity relative to MSI FLA with a threshold of 3.5% unstable loci. mSINGS (Salipante *et al*, 2014) similarly analyses the proportion of reads of different microsatellite lengths at multiple loci and, using exome data from three different panels covering from 15 to 2957 microsatellite loci, achieved 96.4-100% sensitivity and 97.2-100% specificity relative to MSI FLA. A comparison of MSI-classification software, including MSIsensor, mSINGS and MSI-ColonCore (which uses a similar analysis), using sequencing data from the ColonCare gene panel in 54 MMRd and 37 MMRp CRCs, showed equivalent performance, with accuracy ranging from 96.7-98.9% (Zhu *et al*, 2018). This testifies that, despite subtle differences in analysis pipelines and input data, analysing the distribution of sequencing reads associated with microsatellite length is an accurate method of detecting MSI-high tumours.

Software to classify MSI status have been used on NGS gene panels that (non-exclusively) sequence MMR genes including *MLH1*, *MSH2*, *MSH6* and *PMS2*, *EPCAM* (to detect 3' deletions that silence *MSH2*), and *BRAF*, as a high throughput method for Lynch syndrome screening (Gray *et al*, 2018; Hampel *et al*, 2018). This approach reduces the Lynch syndrome screening pipeline to just two steps; tumour-sequencing followed by genetic testing to confirm any MMR mutations in the germline. Additional, clinically actionable genetic markers can also be included within the gene panel, such as *RAS* gene mutations that confer resistance to anti-EGFR therapy, and mutation of *DYPD* that is associated with 5-FU

toxicity (Hampel *et al*, 2018). Parallel sequencing of tumour and germline DNA, whilst doubling the initial cost, further reduces Lynch syndrome screening to one step, and determination of somatic origin of MMR mutations is feasible in one analysis pipeline (Gray *et al*, 2018); identifying Lynch-like tumours with double somatic MMR mutations avoids the unnecessary management of these patients as Lynch syndrome cases (Mensenkamp *et al*, 2014). However the cost of tumour-sequencing is a barrier to its deployment, with an estimated cost in clinical practice of 607±207€ per sample in a recent French, nationwide, study (Marino *et al*, 2018).

To reduce costs, NGS-based assays that target a small number of clinically actionable hotspot loci have also been developed. MSIplus (Hempelmann *et al*, 2015) is a PCR-based assay that amplifies 11 MNR loci, *RAS* gene mutation hotspots and the *BRAF* V600E locus for amplicon sequencing using Illumina platforms. mSINGS is used to analyse MSI status from the 11 MNRs. In 78 tumours, the assay achieved 97% sensitivity and 100% specificity relative to MSI FLA (Hempelmann *et al*, 2015). However, MSIplus is only partially multiplexed, and uses long (up to 28bp) MNRs that are known to be both prone to PCR and sequencing error (Fazekas *et al*, 2010), and are more likely to be polymorphic (Ananda *et al*, 2013). Furthermore, the robustness of MSIplus to sample variables has not been tested, and 5% of samples failed to give interpretable scores. These uninterpretable samples were excluded from accuracy calculations and remained unresolved (Hempelmann *et al*, 2015). Recently, our research group has developed a PCR and NGS-based MSI test using a novel marker panel of short (7-12bp) MNRs, and a novel analysis pipeline that uses both the proportion of reads containing deletion mutations in the microsatellite markers, and the allelic bias of these deletions, to classify MSI status. In an analysis of 209 CRCs, the assay gave 98% sensitivity and 98% specificity relative to the MSI Analysis System (Redford *et al*, 2018). As is the case for many NGS-based MSI tests, the analytical validity of this assay needs to be proven by assessment of assay reproducibility and robustness (Jennings *et al*, 2017). Also, the current assay requires two rounds of PCR and isn't fully multiplexed. Continued development of the protocol toward high throughput cancer diagnostics is needed to optimise clinical utility.

Another advantage of NGS is its ability to detect somatic variants at very low frequencies within template DNA. This is of particular interest for diagnosis of CMMRD through detection of low level MSI in non-neoplastic tissues, which normally requires highly sensitive but laborious techniques like small pool PCR (Parsons *et al*, 1995), or time-

consuming primary cell culture to develop an MSI signal (Bodo *et al*, 2015). For instance, using a read depth >1000x, Spencer *et al* (2014) showed that they could detect variant allele fractions (VAFs) ranging from 25% down to 2.5% using mixtures of HapMap DNAs. However, reliable calling of VAFs lower than this, likely to be required for NGS-based CMMRD detection, is difficult due to the 1-1.5% error rate of NGS platforms (Shendure and Ji, 2008).

Several techniques have been developed that allow accurate detection of VAFs of <1% by molecular barcoding of the DNA molecules to be sequenced (Marx, 2016). To use molecular barcoding, the sample DNA must be sequenced to a redundant read depth such that the total number of reads is greater than the number of original DNA molecules captured for sequencing. Sequencing errors can then be discriminated from true mutations by assessing the concordance of variants from multiple reads that are all derived from one original DNA molecule – those variants found in the majority of, or all, redundant reads are likely to be true variants present in the original DNA molecule, whereas variants found only in single, or minority of, redundant read(s) are likely to be errors. The molecular barcode facilitates the grouping of redundant reads (Figure 1.11). For example, Safe-SeqS shears template DNA and ligates short oligonucleotides containing a section of 12-14 random nucleotides onto each end of the DNA fragments before amplification and sequencing. The 12-14 random nucleotides constitute the molecular barcode, with approximately 17-268 million possible sequences per barcode. Amplification of these fragments incorporates the molecular barcode into each and every amplicon. Amplicon sequencing covers both region of interest and molecular barcode such that reads with the same barcode from the same locus can be grouped in downstream analyses for variant versus error discrimination (Kinde *et al*, 2011). More recent methods of molecular barcoding of sample DNA fragments use double stranded molecular barcodes so that reads from complementary strands from the same DNA duplex can be grouped to discriminate against strand specific lesions, allowing detection of VAFs as low as 2.4×10^{-7} . These include Duplex Sequencing (Schmitt *et al*, 2012; Kennedy *et al*, 2014) and CypherSeq (Gregory *et al*, 2016). In the development of NGS-based MSI tests it is, therefore, feasible to incorporate molecular barcodes into reads to increase sensitivity for low-level MSI, which may be able to detect MSI in the non-neoplastic tissues of CMMRD patients.

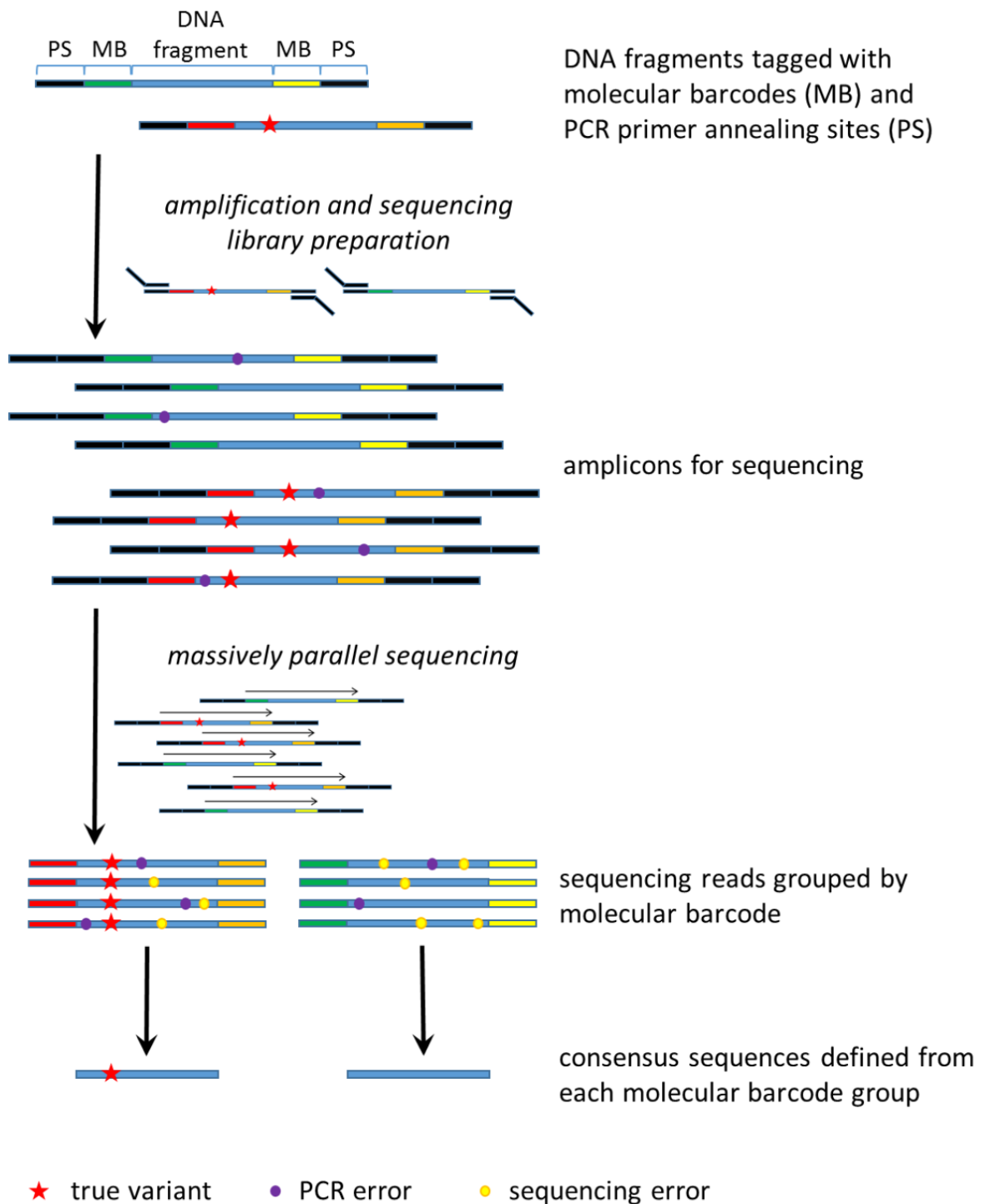


Figure 1.11. Utilising molecular barcodes (MBs) in next generation sequencing (NGS) to discriminate true variants from PCR and sequencing errors. DNA fragments containing the sequence of interest are tagged with MBs (different colours represent unique barcode sequences) and primer annealing sites. PCR amplifies the DNA fragments with their associated MBs and the amplicons are sequenced. Each read has a MB that traces it back to the original DNA fragment and reads of the same locus and same MB can be grouped. Therefore PCR and sequencing errors that occur randomly within the minority of reads in a group can be discriminated from true variants that will be present in the majority to all of the reads in a group.

1.13. Summary and Aims

MMR deficiency defines a distinct molecular subtype of CRC that can be identified by its MSI-high phenotype (Guinney *et al*, 2015). These CRCs respond well to immunotherapy (Le *et al*, 2015) and are associated with cancer-predisposition syndromes: Lynch syndrome and CMMRD (Lynch *et al*, 2009; Wimmer *et al*, 2014).

Accurate biomarker tests for the early detection of MMRd CRCs are yet to be realised and CRC screening in Lynch syndrome and CMMRD currently relies on colonoscopy (Vasen *et al*, 2013; Vasen *et al*, 2014), which is invasive, leading to lack of compliance to surveillance protocols (Stuckless *et al*, 2012), and is insensitive to a significant proportion of CRCs in Lynch syndrome (Seppälä *et al*, 2017). α FSP-Abs have been detected in the peripheral circulation of MMRd CRC patients (Reuschenbach *et al*, 2010) and a multiplexed, but unproven, method to quantify α FSP-Abs in serum has recently been developed (Reuschenbach *et al*, 2014), suggesting it may be used as a novel, non-invasive biomarker test for the early detection of MMRd CRC.

For such surveillance and other preventative measures to be applied, Lynch syndrome gene carriers and CMMRD patients must first be identified. NICE DG 27 states that all CRCs should be tested for MMR deficiency, followed by *BRAF* V600E or *MLH1* methylation testing (Newland *et al*, 2017), to select patients for germline genetic testing for Lynch syndrome. However, current MMR deficiency tests are not scalable for high throughput cancer diagnostics, evident from the poor uptake of testing (Shaikh *et al*, 2018), and novel approaches are required. NGS is amenable to automated and scalable MSI testing (Zhu *et al*, 2018) and our research group has developed a cheap, PCR and NGS-based assay with high sensitivity and specificity for MSI-high CRCs that requires optimisation and further analytical validation (Redford *et al*, 2018).

CMMRD diagnosis is complicated by a pleiotropic phenotype that overlaps with otherwise unrelated syndromes, variable family history and sometimes difficult genetic diagnosis due to the presence of pseudogenes that interfere with *PMS2* sequencing (Wimmer *et al*, 2014). Assaying MMR deficiency in the non-neoplastic tissues of suspected CMMRD patients would clarify diagnosis, but current molecular techniques are either insensitive for specific mutations or laborious and expensive (Ingham *et al*, 2013; Bodo *et al*, 2015). NGS-based MSI tests can use high read depths and molecular barcoding to detect rare

somatic variants (Marx *et al*, 2016), which may be applicable to the low-level MSI in the non-neoplastic tissues of CMMRD patients.

The aims of the work outlined in this thesis were to:

1. Quantify the α FSP-Abs titres in the serum of Lynch syndrome gene carriers and test the association of α FSP-Abs with MMRd CRC, using the multiplexed, bead-based methodology of Reuschenbach *et al* (2014).
2. Develop a high throughput and accurate NGS-based MSI test for CRC diagnostics that is reproducible and robust to sample variables, with a view to improving the uptake of MMR deficiency testing in Lynch syndrome screening and providing a companion diagnostic for immune checkpoint blockade therapy.
3. Adapt this NGS-based MSI test for the detection of low-level MSI in the non-neoplastic tissues of CMMRD patients as a functional, molecular assay to clarify uncertain diagnoses.

Chapter 2. Materials and Methods

2.1. Ethical Approval for Research Conducted

Human sera were obtained from the CaPP3 clinical trial biobank with patient consent for the use of collected material in research at Newcastle University. The CaPP3 trial is an ongoing study analysing the optimal dose of aspirin for chemoprevention of cancer in a cohort of Lynch syndrome gene carriers (ISRCTN16261285).

Human CRC tissue samples were obtained, either as formalin fixed paraffin embedded (FFPE) tissues, or as DNA extracted from FFPE tissues, following ethical review (REC reference 13/LO/1514), which was extended until January 2019.

Samples received from Division of Human Genetics, Medical University of Innsbruck, Austria, were collected and used with consent of the patient and following local ethical review.

2.2. Human Tissue and DNA Samples

2.2.1. *Samples for the assessment of immunological biomarkers*

494 serum samples collected from the first 500 recruits to the CaPP3 clinical trial were assessed for reactivity against FSPs as a measure of α FSP-antibody titres. The serum samples were all collected during patient consultation at trial entry and randomisation (year 0), according to the CaPP3 Study Protocol (ISRCTN16261285).

2.2.2. *Samples for development of a smMIP-based MSI assay for cancer diagnostics*

19 anonymised CRC DNAs, originally extracted from FFPE tissue, were provided by the Department of Molecular Pathology, University of Edinburgh, UK. 73 anonymised CRC DNAs, originally extracted from FFPE tissue, were provided by the Genetics Service of the Complejo Hospitalario de Navarra and Hereditary Cancer Group, Biomedical Research Institute of Navarra, Spain. These 92 samples were residual stocks from the work described by Redford *et al* (2018) and were used in the MSI classifier training cohort.

105 anonymised CRC samples, either as DNA extracted from FFPE tissue or 10 μ m-thick FFPE tissue sections, were provided by the Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust, UK. 6 of these CRC samples were used with samples from Edinburgh and Spain in the MSI classifier training cohort, and the remaining 99 were used to validate the MSI assay and classifier. All samples are listed in Appendix A.

All samples were independently tested for MSI status using MSI Analysis System v1.2 (Promega) by the contributing pathology laboratory. *BRAF* V600E status was also tested in 46 of the MSI-high samples from the Northern Genetics Service, Newcastle, by high resolution melt curve analysis (HRM) on a LightCycler 480 (Roche) according to Nikiforov *et al* (2009).

DNA extracted from fresh tissues of an MMRd CRC and a biopsy of normal colorectal mucosa, taken 10cm from the tumour margin in the same patient, were used as MSI-high and MSS controls, and were originally provided by the Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust, UK. These were taken from residual stocks of work previously conducted by our research group (Alhilal PhD Thesis, 2016).

2.2.3. Samples for development of a smMIP-based MSI assay for constitutional mismatch repair deficiency

94 germline DNA samples, extracted from peripheral blood leukocytes (PBLs), of 94 anonymised control patients were provided by Dr Katharina Wimmer, Division of Human Genetics, Medical University of Innsbruck, Austria. These control samples were selected from patients consulted for non-cancer related reasons, and consenting to use of residual DNA samples in assay development.

36 germline DNA samples, extracted from peripheral blood leukocytes (PBLs), of 32 genetically confirmed CMMRD patients were provided by Dr Katharina Wimmer, Division of Human Genetics, Medical University of Innsbruck, Austria. Samples were collected by Dr Wimmer from several clinicians; sample data, including source, can be found in Appendix B. All patients were consented for use of samples in assay development.

40 germline DNA samples, extracted from PBLs, of 40 genetically confirmed Lynch syndrome gene carriers, with pathogenic germline variants in *MLH1* (n =9), *MSH2* (n = 21), *MSH6* (n = 8), and *PMS2* (n =1) (1 patient had not disclosed their MMR variant), were provided by the CaPP3 clinical trial biobank (ISRCTN16261285).

2.3. Cell Line Samples and Cell Culture Protocols

Genomic DNA from embryonic stem cell H9 (Thomson *et al*, 1998) was a gift from L. Lako (Institute of Genetic Medicine, Newcastle University, UK) and was used as MSS control DNA during smMIP-based MSI assay development.

Both HCT116 and K562 cell lines were gifted by J. Irving (Northern Institute for Cancer Research, Newcastle University, UK). HCT116 is an MMRd CRC cell line, containing a hemizygous *MLH1* truncation S252X (Papadopoulos *et al*, 1994; Boyer *et al*, 1995), and provided MSI-high control genomic DNA during smMIP-based MSI assay development. K562 is an MMRp chronic myeloid leukaemia cell line (Klein *et al*, 1976) and provided MSS control genomic DNA during smMIP-based MSI assay development. HCT116 and K562 cells were both grown in RPMI growth medium containing 2mM L-glutamine (Gibco), 10% fetal bovine serum (Gibco) and 60µg/ml penicillin and 100µg/ml streptomycin (Gibco), at 37°C and 5% CO₂. HCT116 cells grow as a monolayer and were passaged and/or harvested at 80-90% confluence by decanting expired growth medium, washing the monolayer in 5ml PBS (Gibco), detaching the cells using 0.05% Trypsin-EDTA in PBS (Gibco) for 2-5min, before re-suspending the cell pellet in fresh growth medium for passaging. K562 cells grow in suspension and were passaged and/or harvested at a density of 1x10⁶cells/ml. To harvest HCT116 or K562 cells, cell suspension was centrifuged at 1500g for 5min and the supernatant discarded, the pellet washed in 5ml PBS, and again centrifuged at 1500g for 5min and the supernatant discarded. The cell pellet was used for DNA extraction immediately, or stored at -80°C until ready for DNA extraction.

2.4. DNA Extraction

Genomic DNA extraction from FFPE CRC tissue used GeneRead DNA FFPE Kit (QIAGEN), following the manufacturer's protocol.

Genomic DNA extraction from cell lines used Wizard Genomic DNA Purification Kit (Promega), following the manufacturer's protocol.

2.5. DNA Quantification and Dilution

Sample template DNAs and amplicons were quantified using QuBit 2.0 Fluorometer and QuBit dsDNA BR or QuBit dsDNA HS Kits (Invitrogen), following the manufacturer's protocol. DNA ng/µl was converted to nanomolar concentration (nM) using the following equation, where 660g/mol/bp is the average molar mass of one base pair of DNA and N bp is the number of base pairs in the DNA molecule of interest:

$$\text{concentration} = \frac{\text{density} \times 10^6}{660 \text{ g/mol/bp} \times N \text{ bp}}$$

Dilutions of template DNAs and amplicons used 10mM Tris-Cl at pH8.5.

2.6. Generation of Samples containing Known Proportions of MSI-high DNA

To assess the lower limit of detection of the smMIP-based MSI assay for cancer diagnostics, DNA samples containing different proportions of MSI-high DNA were generated. Pure MSI-high DNA extracted from HCT116 (see section 2.3) was diluted to 25ng/ μ l. 19 MSS DNAs, originally extracted from control patient PBLs (see Section 2.2.3) and confirmed to be free of length polymorphisms in any of the microsatellites analysed during the course of this work (see Section 5.4), were mixed in equal quantity to provide sufficient stock of pure MSS DNA, which was also diluted to 25ng/ μ l. In triplicate, the series of samples containing varying MSI-high content were generated by serial dilution as described in Table 2.1. These sample series were used to assess the lower limit of detection of the smMIP-based MSI assay.

MSI-high Content	Mixture
50.00%	10 μ l of MSI-high DNA + 10 μ l of MSS DNA
25.00%	10 μ l of 50.00% mixture + 10 μ l of MSS DNA
12.50%	10 μ l of 25.00% mixture + 10 μ l of MSS DNA
6.25%	10 μ l of 12.50% mixture + 10 μ l of MSS DNA
3.13%	10 μ l of 6.25% mixture + 10 μ l of MSS DNA
1.56%	10 μ l of 3.13% mixture + 10 μ l of MSS DNA
0.78%	10 μ l of 1.56% mixture + 10 μ l of MSS DNA

Table 2.1: Generation of samples with varying MSI-high DNA content.

2.7. Detection of Frameshift Peptide Serum Reactivity

2.7.1. Generation of median fluorescence intensity data for frameshift peptides

28 FSPs and a control FLAG peptide were analysed per serum sample. FSPs are listed in Appendix C. All laboratory work was conducted by Jonathan Dörre and Dr Miriam Reuschenbach at the Department of Applied Tumour Biology, Heidelberg University Hospital, Germany, and the raw data was provided in *.xlsx format by Dr Reuschenbach and Dr Kloor (Heidelberg University) for processing and analysis by me in Newcastle. The following laboratory protocol is adapted directly from Reuschenbach *et al* (2014), with additions from personal communication and my own experience with the protocol.

The amino acid sequences of FSPs were predicted *in silico* from cMNR frameshift mutations identified in >60% of MMRd CRCs (Woerner *et al*, 2010). Each FSP was synthesised with an N-terminal biotin tag (connected by a 6-aminohexanoic acid linker) and a C-terminal FLAG octapeptide. A FLAG-only control peptide was also synthesised, containing the N-terminal

biotin tag (and linker) covalently bound directly to the FLAG octapeptide, with no intermediate FSP sequence. All peptides were HPLC-purified to >95% purity (Genaxxon Bioscience), and were dissolved in DMSO and stored at 5mg/ml at -80°C.

Polystyrene beads containing fluorescent dyes and coated with avidin were purchased at 2.5million beads/ml in phosphate buffer saline (PBS) containing BSA, Tween-20, and sodium azide (LumAvidin, Luminex Corp). Prior to use, beads were spun down at 13,000rpm for 2min, and washed in PBS containing 0.1 % casein. Peptides were diluted in PBS and 0.1% casein from DMSO stocks to 400nM. The FSPs and FLAG-only control peptide were bound to the surface of the beads through biotin-avidin conjugation: the washed bead pellet was sonicated and resuspended in the peptide dilution, and incubated for 30min on a shaker protected from light. Importantly, the beads contained different fluorescent dyes, and each FSP was bound to beads of a specific fluorescence. Beads were spun down at 13,000rpm, and washed three times in PBS plus 0.1 % casein, and incubated for 30 min in PBS, 0.1 % casein and 1µM biotin (Sigma-Aldrich) to block free avidin. After an additional washing step, peptide-coated beads were resuspended in PBS containing 0.1 % casein and 0.05 % sodium azide, and stored at 4°C in the dark to protect the fluorophores.

Patient sera were preincubated at a dilution of 1:50 in 0.5 % casein-PBS blocking buffer containing 0.5 % polyvinylalcohol, 0.8 % polyvinylpyrrolidone, and 2.5 % CBS-K (MERCK) to suppress non-specific binding of sera to the beads. Each serum was then diluted two fold, to a final concentration of 1:100, in a filter plate containing a multiplex of FSP- and FLAG control-bead conjugates, with each peptide represented by 3,000 beads. The mix of beads and sera were incubated for 30min on a shaker protected from light. The filter plate was washed three times in PBS containing 1% Tween-20. Beads were incubated in PBS, 0.1 % casein and 1µM biotin (Sigma-Aldrich) to block free avidin, for 30min on a shaker protected from light. The filter plate was washed once in PBS containing 1% Tween-20. Beads were incubated in PBS, 0.1 % casein and 1:2000 dilution of anti-human IgG conjugated to a phycoerythrin (PE) fluorophore to label each bead with a fluorophore for quantification of primary antibody binding, for 30min on a shaker protected from light. The filter plate was washed three times in PBS containing 1% Tween-20. Beads were resuspended in PBS and 0.1% casein and the plate was loaded into the Luminex-100.

Using the Luminex-100, each reaction was analysed by passing the beads through a channel, one bead at a time, with two measurements taken per bead: the fluorescence of the bead

dye was used to determine the FSP being measured, and the fluorescence of the PE fluorophore was used to quantify the bound antibodies; the raw data output for a serum sample is the median fluorescence intensity (MFI) of PE detected for each FSP.

2.7.2. Analysis of serum reactivity against frameshift peptides

Anonymised patient data was available from the CaPP3 database extract taken on 17.02.2016. Normalisation of MFI data (see Section 3.5) gave a measure of “serum reactivity” for each FSP. Analyses of serum reactivity and patient variables used custom R scripts and R packages *quantreg* for quantile regression and *pvc* for FSP clustering.

2.8. Design of Single Molecule Molecular Inversion Probes

2.8.1. Selection of marker loci for the smMIP-based MSI assay

A total of 27 short mononucleotide repeats (7-12bp in length) with neighbouring SNP, previously identified by our lab (Redford *et al*, 2018; Section 4.1), were selected to assess MSI status as a biomarker of MMR deficiency (Appendix D). The SNPs allow analysis of allelic bias of deletions in the microsatellite markers in heterozygous patients. 17 of these markers were previously confirmed as highly sensitive and specific for MSI (Redford *et al*, 2018) and 10 additional markers were selected to provide more options for panel optimisation. *BRAF* V600E locus (Appendix D) was included so that the assay can diagnose MMR deficiency and screen for Lynch syndrome in one test. Sporadic MMRd CRCs frequently have *BRAF* V600E whilst Lynch syndrome CRCs do not, and so *BRAF* V600E positive CRCs can be excluded from germline genetic testing (Domingo *et al*, 2004).

The *KRAS* codons 12 and 13 mutation hotspot locus (Appendix D) was included as a proof of principle of the modularity of a smMIP-based assay and the ease with which it could be expanded to include other clinically actionable genetic markers. *RAS* gene mutations are of clinical relevance as they confer resistance to anti-EGFR therapeutics, such as cetuximab (De Roock *et al*, 2010a), and *RAS* gene testing is required before application of anti-EGFR therapies (Cooper *et al*, 2017).

2.8.2. Design of smMIPs using MIPgen

MIPgen (Boyle *et al*, 2014a) was used to generate smMIPs for each marker using the inputs: hg19 as a .fasta file and indexed by SAMtools v1.3 and BWA v0.7.12, and a .bed file of

marker loci, and parameters: tag size 6,0, minimum capture size 120 and maximum capture size 150. These parameters, respectively, determine the lengths (N nucleotides) of the molecular barcode tags neighbouring the extension arm and ligation arm of the smMIP, and the size of the region of interest to captured (which includes sequence within the extension and ligation arms of the smMIP). Molecular barcodes of length N_6 provided 4096 possible barcode sequences, which would be sufficient to represent unique template DNA molecules, given target read depths ~ 5000 reads per marker and the redundant coverage of template molecules in amplicon sequencing (i.e. generation of multiple reads per molecular barcode; Casbon *et al*, 2011).

Final smMIP sequences (Appendix E) were selected by the following *in silico* criteria: successful capture of marker and associated SNP (for microsatellite loci), no SNPs in the smMIP extension or ligation arms and logistic score >0.8 .

All smMIPs and smMIP protocol-associated primers (Appendix E) were synthesised by and purchased from Metabion GmbH (Planegg, Germany).

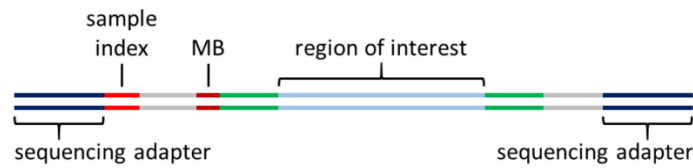
2.8.3. Validation of smMIP designs and amplicons

smMIPs designs were validated in the laboratory by successful generation of smMIP amplicons of the expected size (240-270bp) and confirmation that these amplicons contained the correct sequence by PCR amplification of a region internal to the amplicon. Primers were designed and paired such that a forward primer targeting the “backbone” sequence of the smMIP amplicon (i.e. the sequence common to all smMIP amplicons) could be paired with a reverse primer targeting the “internal” sequence specific to the marker locus, and vice versa, to ensure the primers were specific to the amplicon and would not amplify any contaminating genomic DNA (Figure 2.1).

Primers for validation of smMIP amplicons (Appendix E) were designed using Primer3 (Untergasser *et al*, 2012; <http://primer3.ut.ee/>).

PCR amplification to validate smMIP amplicons used 1x Herculase II Reaction Buffer (Agilent), 1.25U Herculase II Fusion DNA Polymerase (Agilent), 6.25nmol dNTPs (Agilent), 6.25pmol forward primer, 6.25pmol reverse primer, and 1 μ l of a 1 in 1000 dilution of purified smMIP amplicon in a 25 μ l reaction volume. Reactions were incubated at 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 57°C for 30 seconds, and 72°C for 30 seconds, followed by 72°C for 2 minutes.

smMIP amplicon structure:



Primers for amplicon verification:

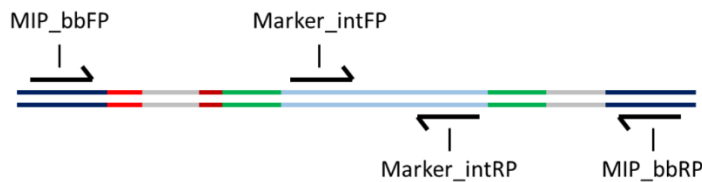


Figure 2.1: Design of PCR primers to verify sequence content of smMIP amplicons. MIP_bbFP: MIP backbone forward primer. MIP_bbRP: MIP backbone reverse primer. Marker_intFP: marker-specific forward primer. Marker_intRP: marker-specific reverse primer. For full details of smMIP amplicon structure see Figure 4.1

2.9. Single Molecule Molecular Inversion Probe Amplification Protocol

2.9.1. Probe phosphorylation

smMIPs were individually phosphorylated using 10U of T4 Polynucleotide Kinase (NEB), 1X T4 DNA Ligase buffer (NEB) and 1 μ M of unphosphorylated smMIP in a 100 μ l reaction volume, and incubated at 37°C for 45 minutes and 80°C for 20 minutes. Phosphorylated smMIPs were diluted 1:10,000 using TE buffer (Sigma) in a multiplex pool, such that each smMIP was at 0.1nM (0.1fmol/ μ l).

2.9.2. Target capture and amplification

Markers were amplified in singleplex or multiplex following the smMIP-based protocol of Hiatt *et al* (2013) using a SensoQuest thermocycler (SensoQuest GmbH). smMIPs were annealed to template DNA using 1x AmpLigase Reaction Buffer (Lucigen), 0.125fmol of each smMIP and 50-200ng of template DNA (unless stated otherwise) in a 10 μ l reaction volume, incubated at 98°C for 3 minutes, 85°C for 30 minutes, 60°C for 60 minutes, and 56°C for 120 minutes. Gap-fill and ligation captured target sequence in circularised smMIPs using 1x AmpLigase Reaction Buffer (Lucigen), 5U AmpLigase DNA Ligase (Lucigen), 3.2U Herculase II Fusion DNA Polymerase (Agilent), 300pmol dNTPs (Agilent) in 10 μ l added to each reaction for a total 20 μ l reaction volume, incubated at 56°C for 60 minutes and 72°C for 20 minutes.

Non-circularised smMIPs and template DNA were digested using 1x Ampligase Reaction Buffer (Lucigen), 20U Exonuclease I (NEB) and 100U Exonuclease III (NEB) in 3µl added to each reaction for a total 23µl reaction volume, incubated at 37°C for 60 minutes and 95°C for 2 minutes. Following digestion of linear DNAs, 10µl of this target capture reaction was mixed with 1x Herculase II Reaction Buffer (Agilent), 1.25U Herculase II Fusion DNA Polymerase (Agilent), 6.25nmol dNTPs (Agilent), 6.25pmol MIP amplification forward primer and 6.25pmol MIP amplification reverse primer in a 25µl reaction volume, incubated at 98°C for 2 minutes, followed by 30 cycles of 98°C for 15 seconds, 60°C for 30 seconds and 72°C for 30 seconds, followed by 72°C for 2 minutes. smMIP amplification reverse primers contain unique sample index sequences (Appendix E); different samples in the same sequencing run used different smMIP amplification reverse primers for sequencing read de-multiplexing. Remaining target capture reaction was stored at -20°C.

smMIP reaction products (smMIP amplicons at 240-270bp) were analysed using 3% Agarose gel electrophoresis at 80mV for 60 minutes or QIAxcel (QIAGEN) using method AL420.

2.10. Library Preparation for Amplicon Sequencing

Each sequencing run was planned according to the desired mean number of reads per marker per amplicon, the number of markers amplified per sample by smMIP protocol and the number of samples, using the following equation:

$$\text{reads/marker/sample} = 0.75 \times \frac{\text{sequencing kit read capacity}}{N \text{ markers} \times N \text{ samples}}$$

The 0.75 factor accounts for generation of non-specific reads, based on the findings of Niedzicka *et al* (2016) and confirmed by our own data (see Section 4.4). Sequencing kit read capacity is the expected number of reads generated from a MiSeq v3 or v2 kit.

4nM sequencing libraries were prepared by purification of smMIP amplicons using Agencourt AMPure XP Beads (Beckman Coulter) following manufacturer's protocols, diluting purified amplicons to 4nM in 10mM Tris pH 8.5 and pooling 4nM amplicons in equal volumes to create the final 4nM DNA library.

2.11. smMIP Amplicon Sequencing on the Illumina MiSeq

4nM libraries of smMIP amplicons were sequenced using the MiSeq platform (Illumina) following the manufacturer's protocol and using the GenerateFastq workflow, paired end sequencing and smMIP custom sequencing primers (Appendix E), as specified in the MiSeq

Sample Sheet (Appendix F), according to the protocol of Hiatt *et al* (2013). Sequencing run metrics, such as the % of base calls with quality >Q30, were acquired from basepace.illumina.com.

2.12. Sequencing Read Analysis

2.12.1. Generation of Marker Result tables

Unprocessed reads contained in fastq files generated by the MiSeq were aligned to reference genome hg19 using BWA v0.6.2 (Li and Durbin, 2010). Marker loci were analysed from .sam files and, for each marker, microsatellite lengths and SNPs observed in both orientations, i.e. concordant in both forward and reverse reads, were counted and summarised in Marker Result tables using custom R scripts written by Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University; Figure 2.2).

For *BRAF* and *KRAS*, the same Marker Result tables were generated with columns representing the base detected at the mutation hotspot. As there was no microsatellite associated with these mutation hotspots, microsatellite length was determined from an arbitrary locus and Marker Result table rows were not used in analyses.

Optionally, molecular barcodes could be analysed from the sequencing reads and Marker Result tables could contain marker data generated from single molecule sequences (smSequences) rather than all reads (see Section 5.4). Analysis of molecular barcodes and generation of smSequence Marker Result tables used custom R scripts written by Dr Santibanez-Koref.

The microsatellite lengths and SNP alleles summarised in these Marker Result tables were used for multiple downstream analyses, including sample classification.

	A	G
-2	2	1
-1	10	11
0	890	905
+1	5	6

Figure 2.2: Example Marker Result table. The Marker result table contains a count of reads at a specific marker, with reads distributed to table cells according to the SNP identified (columns) and the length of the microsatellite relative to the reference genome hg19 (rows).

2.12.2. MSI classification using a naïve Bayes approach

Sample classification was performed as described by Redford *et al* (2018). In summary, sequencing reads from 24 short MNRs were analysed for both microsatellite deletions, and the allelic bias of these deletions, to estimate ratio of the posterior probabilities that the data were generated by an MSI-high or MSS phenotype (a Bayes factor). This could be presented as a sample score by the following equation:

$$score = \log_{10} \frac{P(MMRd|O)}{P(MMRp|O)}$$

Samples with scores > 0 are considered MSI-high and with scores < 0 as MSS. The classifier parameters were determined from a training cohort of 98 CRCs of known MSI status (see Sections 2.2.2 and 4.5). The MSI classifier was then validated in a second, independent cohort of 99 CRCs. Training and execution of the MSI classifier used custom R scripts written by Dr Santibanez-Koref and the Marker Result tables as input. For classifier training, MSI-low samples were considered equivalent to MSS samples as described in the literature (Halford *et al*, 2002; Laiho *et al*, 2002).

2.12.3. Read and variant counting

During development of the smMIP-based assay for cancer diagnostics and CMMRD detection, counts of total reads for each marker and the number of reads containing variant or wild type (WT) microsatellite length, or number of reads containing different SNPs, were used. Read and variant counting used custom R scripts and the Marker Result tables as input.

2.12.4. Hotspot mutation calling

The relative frequency of *BRAF* and *KRAS* variants could be analysed to determine mutation status. Hotspot mutation calling used custom R scripts and Marker Result tables as input.

2.12.5. CMMRD classification

A method to detect CMMRD by microsatellite length variants in germline DNA extracted from PBLs was developed. In summary, 40 germline DNAs from anonymised controls were used to define the distribution of microsatellite lengths detected in a non-CMMRD population. Samples are then scored by the probability that the observed lengths of

microsatellites belong to these control distributions, with high scores indicating a high probability the sample is not from a control population (see Section 5.5).

CMMRD classification used custom R scripts, ExtDist and metap packages, and the Marker Result tables as input.

2.13. Germline Confirmation of *MSH6* c.3557-1G>C Mutation

MMR gene mutation was confirmed using Sanger sequencing. Primer sequences (Appendix G) were provided by Dr Katharina Wimmer (Division of Human Genetics, Medical University of Innsbruck, Austria) and primers were synthesised by and purchased from Metabion. MMR gene mutation loci were amplified using a SensoQuest thermocycler (SensoQuest GmbH). Each reaction contained 1x Herculase II Reaction Buffer (Agilent), 1.25U Herculase II Fusion DNA Polymerase (Agilent), 6.25nmol dNTPs (Agilent), 6.25pmol forward primer and 6.25pmol reverse primer in a 25µl reaction volume, and was incubated at 98°C for 2 minutes, followed by 30 cycles of 98°C for 15 seconds, 54°C for 20 seconds and 72°C for 30 seconds, followed by 72°C for 3 minutes. To digest dNTPs and single stranded DNAs, 10U Exonuclease I (NEB) and 1U Shrimp Alkaline Phosphatase (NEB) were added to 5µl of PCR product and incubated at 37°C for 15 minutes and 80°C for 15 minutes. Fluorescence-labelled termination fragments were generated from the exonuclease and phosphatase treated PCR product using BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems), following manufacturer's protocols, and were analysed using a SeqStudio Genetic Analyser (Applied Biosystems), following manufacturer's protocols.

2.14. Statistical Analyses and Graphics

Unless stated otherwise, all statistical analyses used base R, version 3.3.1.

Proportions and frequencies were modelled using the binomial distribution to determine confidence intervals, calculated in R, using the Hmisc package.

Graphs were plotted in R using the ggplot2, grid and gridExtra packages.

Chapter 3. The Utility of Anti-Frameshift Peptide Antibodies in the Serum to detect Mismatch Repair Deficient Colorectal Cancer

3.1. Introduction

Circulating α FSP-Abs are generated in patients with MMRd CRC in response to novel peptides expressed from genes containing cMNR frameshift mutations (Ishikawa *et al*, 2003; Reuschenbach *et al*, 2010). The possibility of using α FSP-Abs as a liquid-biopsy biomarker for the early detection of MMRd CRC is appealing for Lynch syndrome surveillance, which currently relies on the invasive and frequently insensitive technique of colonoscopy (Stuckless *et al*, 2012; Møller *et al*, 2017a; Seppälä *et al*, 2017). Multiplexed methods of detecting α FSP-Abs in the serum is particularly appealing as MMRd cancers are susceptible to multiple cMNR frameshift mutations, meaning that multiple α FSP-Abs may be generated against any one cancer, and singleplex detection of cancer-associated autoantibodies generally has low sensitivity for disease (Robinson *et al*, 2002). Previously, Reuschenbach *et al* (2014) used a novel multiplex method to analyse antibody titres against a panel of 32 synthetic FSPs (selected by frequency of cMNR frameshift mutation), using sera from 20 MSI-high CRC patients prior to surgical resection of their tumours, and serum from one MSI-high CRC patient post-surgical resection of their tumour. This latter patient had also been enrolled on the Micoryx clinical trial (NCT01461148), designed to test the hypothesis that vaccination by FSPs may prevent cancer relapse, meaning they had been vaccinated with 2 of the 32 FSPs analysed. They found that an antibody signal was evident in many of the MSI-high CRC patients, and relatively strong signals were generated in the Micoryx trial patient for the two FSPs with which they had been vaccinated (Figure 3.1). To assess reproducibility of the method, Reuschenbach *et al*, repeat tested the 20 sera and showed that the regression R^2 of original versus repeat results was >0.98 in all FSPs and all sera.

Whilst these results are promising, the study was a presentation of method and not designed to answer a biological question. With the small number of patients analysed and the lack of cancer-free controls, the observed antibody signal may be background noise rather than a quantification of α FSP-Ab titre in the serum. Also, the antibody signals in the MSI-high cancer patients were approximately 10-fold weaker than antibody signals from the two vaccine FSPs in the Micoryx trial patient, which remained unexplained (Figure 3.1). To better understand the association between MSI-high CRC and α FSP-Ab titre a larger cohort

of patients would be needed. Also, α FSP-Abs have previously been detected in the serum of cancer-free Lynch syndrome gene carriers using the alternative technique of ELISA (Reuschenbach *et al*, 2010), suggesting that background antibody signal may differ in Lynch syndrome patients compared to the general population. Therefore, for the early detection of MMRd CRC in the context of Lynch syndrome surveillance, this method would need to be tested in a cohort of Lynch syndrome gene carriers.

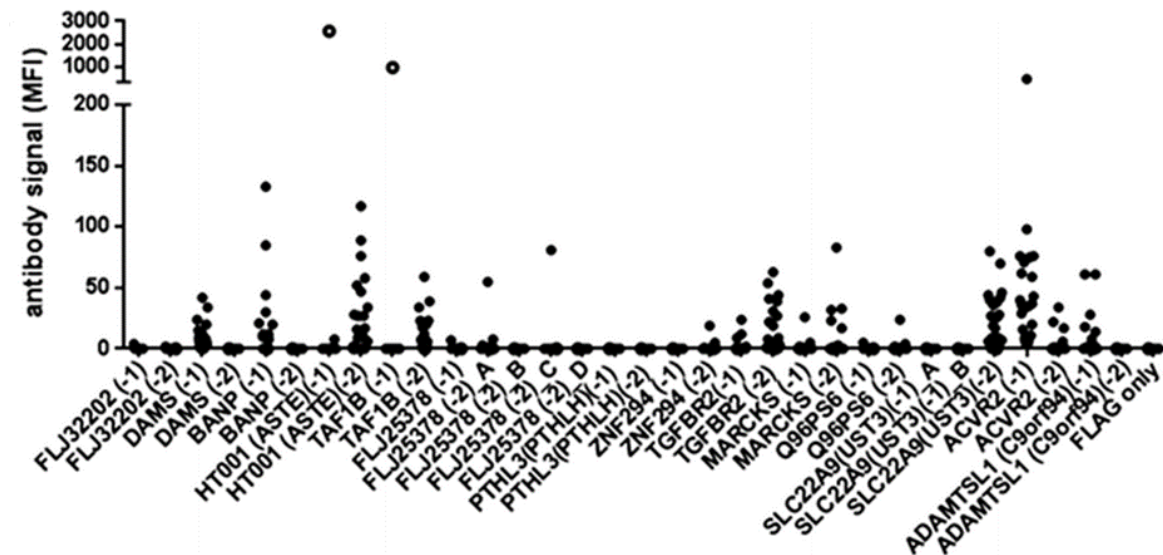


Figure 3.1: Detection of antibody signal against a panel of 32 frameshift peptides (FSPs). FSPs are named by “Gene(frameshift deletion size)”. 20 MSI-high colorectal cancer patients (filled dots) and one patient from the Micoryx trial (open circles), vaccinated with ASTE(-1) and TAF1B(-1) FSPs, were tested by the multiplex method of Reuschenbach *et al* and the antibody signal against each FSP quantified. The Micoryx trial patient has minimal antibody signal against all FSPs except for ASTE(-1) and TAF1B(-1). The 20 MSI-high CRC patients have comparatively weaker antibody signals against multiple other FSPs. Figure taken from Reuschenbach *et al*, 2014.

CaPP3 is an ongoing clinical trial that is recruiting Lynch syndrome gene carriers to analyse the optimal dose of aspirin for chemoprevention of cancer, by double-blind randomisation of patients to 100, 300 or 600mg of daily aspirin (ISRCTN16261285). It is run from Newcastle University under the guidance of its Chief Investigator, Prof Sir John Burn, and Programme Manager, Dr Gill Borthwick, and is building a biobank of material from its patients, including FFPE cancer tissues, germline DNAs, and, most significantly, sera. Serum samples are taken from patients at trial entry (year 0), at year 2, and at year 5, with a view to analysing the longitudinal effect of aspirin on circulating cytokines and other biomarkers, including α FSP-Abs. Therefore the hundreds of patients recruited and the detailed collection

of patient data and patient follow up, make the CaPP3 trial a useful resource for the exploration of α FSP-Ab titres for the early detection of MMRd CRC.

3.2. Aims

To analyse the association between MMRd CRC diagnosis and α FSP-Ab titres, with a view to α FSP-Ab titres being used in Lynch syndrome surveillance, I aimed to:

1. Quantify α FSP-Ab titres in the serum of the first 500 Lynch syndrome gene carriers recruited to the CaPP3 clinical trial using the multiplex method Reuschenbach *et al.*
2. Correlate α FSP-Ab titres with patient variables, in particular incidence of CRC.

3.3. Cohort Description and Justification of Method

For this work, I was given access to the CaPP3 clinical trial biobank and anonymised patient data. From the first 500 patients recruited, sera from 494 patients were available for analysis. Of these 494 samples, 464 were collected within trial protocol (0-4 days from blood draw to long term storage at -80°C), ensuring serum quality based on preliminary analyses by Dr Miriam Reuschenbach and Dr Matthias Kloor (personal communication). Patient details and variables of interest are shown in Table 3.1. A history (i.e. before blood draw) or on-trial (i.e. after blood draw) diagnosis of CRC or Lynch spectrum cancers (see Table 3.1 legend), were of interest due to the high frequency of MMR deficiency and associated cMNR frameshift mutations that lead to expression of antigenic FSPs. Measures of cancer prophylaxis taken by the patient, either by surgery or daily aspirin intake, were also of interest as these may indirectly impact α FSP-Ab titres by suppression of latent MMRd lesions, whether benign or malignant.

The 28 FSPs analysed in this study contained some of the 32 FSPs used by Reuschenbach *et al* in their 2014 publication, as well as additional FSPs. These FSPs were selected by two criteria: 1. The cMNR frameshift mutation was present in $>60\%$ of MMRd CRCs using frequencies reported in the SelTarBase database (Woerner *et al*, 2010), and 2. The FSP produced an antibody signal from sera of MSI-high CRC patients in preliminary analyses (Reuschenbach and Kloor, personal communication). Due to the different spectra of cMNR frameshift mutations in different cancer types (Kim *et al*, 2013), the low incidence of on trial CRCs at the time of analysis (Table 3.1), and the long half-life of antibody production due to immunological memory (Dörner and Radbruch, 2007), an analysis of the association of α FSP-Ab titres with a history of CRC diagnosis was the primary focus of this study.

Variable	Patients (n = 464)	
MMR gene mutation (n)	<i>MLH1</i> :	136
	<i>MSH2</i> :	181
	<i>MSH6</i> :	83
	<i>PMS2</i> :	29
	<i>EPCAM</i> 3' del:	5
	Not Disclosed:	30
Age at blood draw (years)	Median:	47
	Range:	19-77
Sex (n)	Male:	193
	Female:	262
	Not Disclosed:	9
History of cancer (n)	CRC:	112
	<i>Median</i> :	78
	<i>Range</i> :	5-440
	Lynch spectrum:	147
	<i>Median</i> :	71
	<i>Range</i> :	5-440
On trial cancer (n)	CRC:	3
	<i>Median</i> :	12
	<i>Range</i> :	2-19
	Lynch spectrum*:	10
	<i>Median</i> :	4.5
	<i>Range</i> :	1-19
Surgical removal of at-risk tissues (n)	Colorectal:	99
	Colorectal or gynaecological:	187
	None:	277
Pre-trial chemoprophylaxis (n)	Daily aspirin (ever)	114
	Daily aspirin (≥ 2 years)	31
	Never daily aspirin	350

Table 3.1: Patient details of the Lynch syndrome gene carrier cohort. The patient variables presented include age, sex, and cancer incidence, as well as other variables that may affect anti-frameshift peptide antibody titres in the serum.

**Note: Lynch spectrum cancers include: colorectal cancer (CRC), endometrial cancer, ovarian cancer, gastric cancer, small intestinal cancer, uroepithelial cancer, glioblastoma, sebaceous gland carcinoma, keratocanthomas (Lynch et al, 2009).*

For each patient sample, the 28 FSPs and a FLAG-only control were analysed by the multiplex method of Resuchenbach *et al* (2014; Figure 3.2; Section 2.7.1). Laboratory work was conducted by Jonathan Dörre and Dr Miriam Reuschenbach at the Department of Applied Tumour Biology, Heidelberg University Hospital. The multiplex protocol mixes

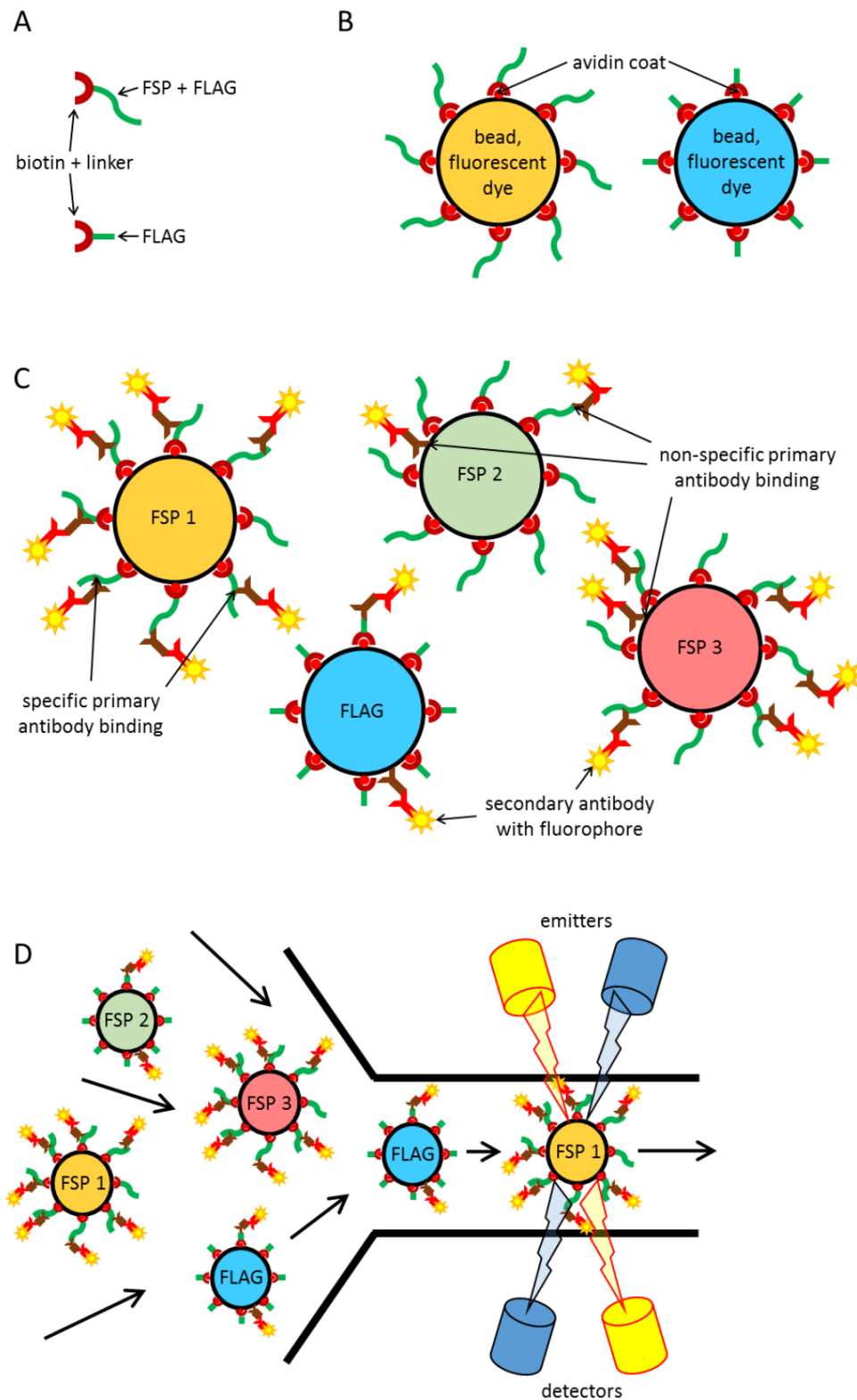


Figure 3.2: A multiplex method of detecting anti-frameshift peptide antibodies (α FSP-Abs) in serum (Reuschenbach *et al*, 2014). (A) Synthetic FSP and C-terminal FLAG with biotin and linker, and FLAG-only control with biotin and linker, are (B) bound to fluorescent beads coated with avidin. Bead preparations are pooled and incubated with patient serum to bind α FSP-Abs. (C) Subsequently, anti-human IgG secondary antibodies with phycoerythrin (PE) fluorophores are bound to the α FSP-Abs. The FLAG-only control accounts for non-specific binding. (D) Beads are analysed by a Luminex-100. Beads are passed, one by one, through a channel where fluorescence is measured to identify the FSP by the bead dye and quantify the bound antibodies.

patient sera with synthetic FSPs that are conjugated to beads containing fluorescent dyes, with each FSP represented by a specific fluorescence. α FSP-Abs from the patient sera bind to the FSP-bead conjugates and are detectable by secondary binding of anti-human IgG antibodies with covalently bound phycoerythrin (PE) fluorophores. The raw data output for a sample is the median fluorescence intensity (MFI) of PE for each FSP. MFI data was provided by Dr Reuschenbach and Dr Kloor for analysis.

In addition to the data from the 464 samples from the Lynch syndrome gene carriers, Dr Reuschenbach and Dr Kloor provided MFI data for 11 positive control samples, all taken at different time points from two MMRd CRC patients that had been vaccinated with FSPs ASTE1(-1) and TAF1B(-1) as part of the Micoryx clinical trial. However, the study cohort had its limitations, most notably was a lack of any other controls. Given that Lynch syndrome gene carriers may have α FSP-Abs irrespective of MMRd cancer incidence (Reuschenbach *et al*, 2014) and may have a different background compared to the general population, control samples from cancer-free and MMR mutation-negative patients would have been valuable. Furthermore, the positive controls were only vaccinated with two of the 28 FSPs analysed, hence positive controls were lacking for the other FSPs, which would be needed for an accurate definition of α FSP-Ab concentration. Despite this lack of controls, analysis of sera from the 464 Lynch syndrome gene carriers could still be used to answer the question of whether or not α FSP-Ab titres are associated with a history of MMRd CRC.

3.4. Subtraction-based Normalisation does not equalise Baseline FSP Serum Reactivity

As is common for assays of substrate binding, MFI data needed to be normalised relative to a control to account for non-specific binding and background fluorescence. Therefore, before analysing the data with respect to patient variables, I wanted to assess its structure and the validity of the normalisation method. Data normalisation by Reuschenbach *et al* (2014) subtracted the MFI of the FLAG-only control from the MFI of each FSP, and each resulting MFI was defined as an “antibody signal”. However, during my analysis I opted to use the term “serum reactivity” for normalised data as the detected signal may not solely be due to α FSP-Ab binding to the FSP-bead conjugates. Furthermore, the data normalisation method of Reuschenbach *et al* assumes the FLAG-only control MFI is equally representative of the non-specific binding and fluorescence for each FSP, and that signal from α FSP-Ab binding is additive to this background fluorescence. However, there are several reasons why this assumption might not be valid. Each FSP is conjugated to a bead containing a different

fluorescent dye, which may interfere with absorbance and fluorescence of the PE fluorophore used to quantify binding. Also, the structures of the FSPs in the multiplex are very different, and each FSP may have a unique level of non-specific binding relative to the FLAG-only control. Finally, the assumption that fluorescence from any α FSP-Ab is additive to the background may not be true, and a multiplicative model may be more appropriate. To test these assumptions, FSP MFI was plotted against FLAG-only control MFI. There was clear evidence of heteroscedasticity in the relationship between FSP and control MFI as the variance in FSP MFI increased with increasing control MFI (Figure 3.3A; Appendix H). This heteroscedasticity shows that an additive model is not an appropriate basis for the normalisation method as subtraction is not equivalent for higher control MFI due to increased variance. Positive skew was also observed in the MFI data for all peptides (Figure 3.3B); to better fit the MFI data to a normal distribution, a transformation by the natural logarithm was used (Figure 3.3C). A comparison of $\ln(\text{MFI})$ for each FSP and the control $\ln(\text{MFI})$ showed a reduction in heteroscedasticity (Figure 3.3D; Appendix H). Normalisation could then be achieved by subtracting the control $\ln(\text{MFI})$ from each FSP $\ln(\text{MFI})$ for a patient – due to the log transformation this is equivalent to normalisation by division, suggesting that a multiplicative model of control versus FSP MFI is suitable.

Before analysing the data I wanted to validate the method of normalisation. Criteria for this validation were required, based on clear assumptions and clear expectations of the range of serum reactivity that should be generated. Normalisation of the data should allow fair comparison within patients between FSPs, and within FSPs between patients. Therefore, it would be expected that normalisation produces equivalent base-line serum reactivity for each patient or FSP. A critical assumption of the data is that, for any FSP, the majority of patients will not have serum reactivity due to specific binding of α FSP-Abs. This assumption is based upon the observation that only a minority of Lynch syndrome gene carriers and MSI-high CRC patients generated α FSP-Ab signals using an independent technique, ELISA (Reuschenbach *et al*, 2010). Translated to a criterion for data normalisation, I expected that the median serum reactivity for any FSP or patient should be approximately 0, with the majority of data falling within a short range either side of this. To test the validity of subtraction of the FLAG-only control $\ln(\text{MFI})$ from FSP $\ln(\text{MFI})$ as a method of data normalisation, I analysed the distribution of serum reactivity across FSPs and across patients. It was clear that the median serum reactivity varied widely between FSPs as for 13/28 FSPs >75% of the data fell entirely above or below 0 (Figure 3.4A). The median serum

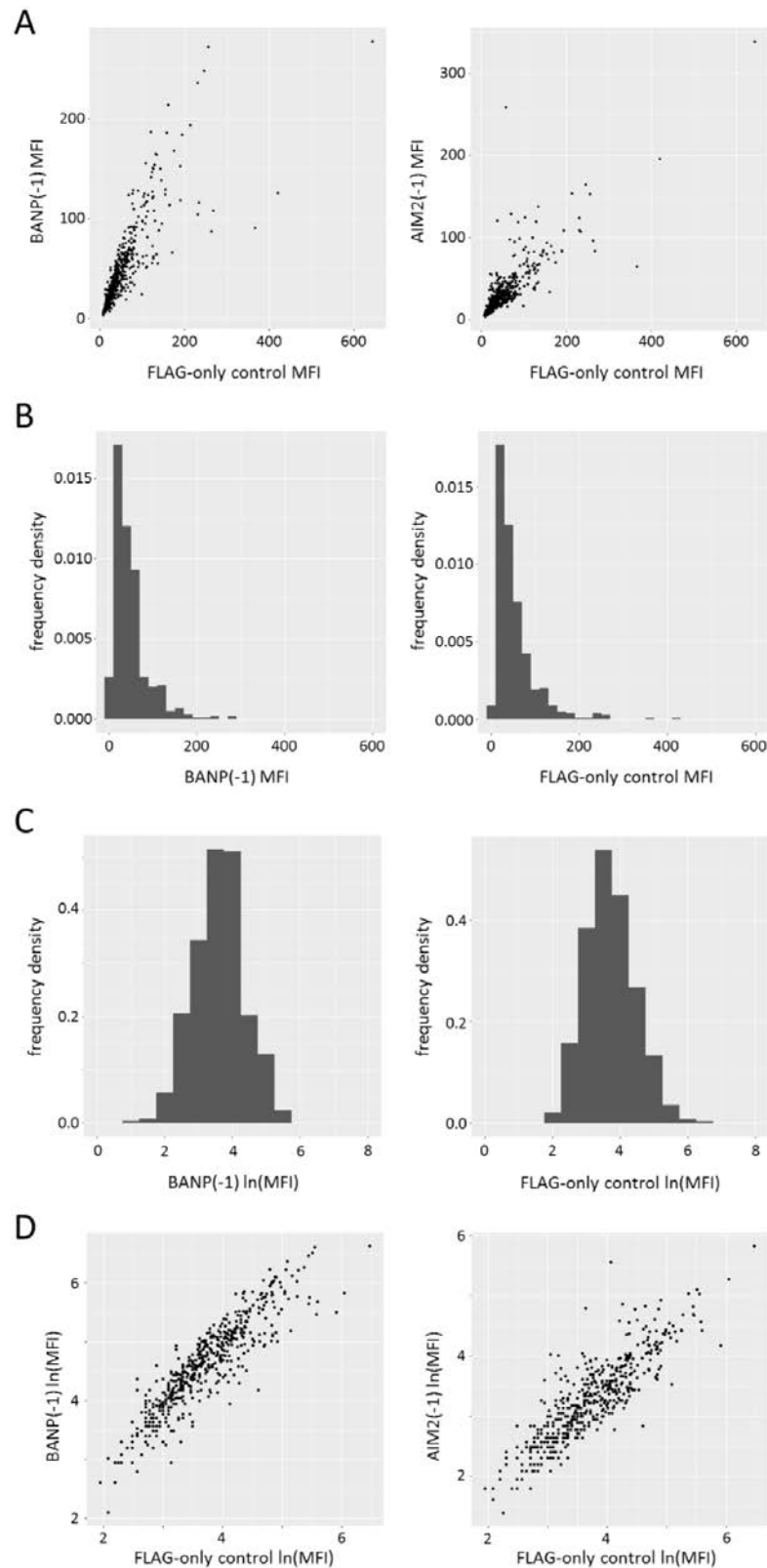


Figure 3.3: Data structure for median fluorescence intensity (MFI). (A) Comparison of frameshift peptide (FSP) MFI and the FLAG-only control MFI in example frameshift peptides (FSPs) BANP(-1) and AIM2(-1), with each data point representing one patient. (B) The distribution of MFI for BANP(-1) and for the FLAG-only control. (C) The distribution of log transformed MFI ($\ln(\text{MFI})$) for BANP(-1) and for the FLAG-only control. (D) Comparison of FSP $\ln(\text{MFI})$ and the FLAG-only control $\ln(\text{MFI})$ in example FSPs BANP(-1) and AIM2(-1), with each data point representing one patient.

reactivity for patients was also highly variable, and when compared with the control $\ln(\text{MFI})$, there was a significant correlation between the two ($\beta = -0.179$, $p < 10^{-16}$, $R^2 = 0.15$); which can be interpreted as a decrease in median serum reactivity as the non-specific fluorescence increases for a sample (Figure 3.4B). The differences in median serum reactivity between FSPs and between patients, and the correlation of serum reactivity with the control MFI do not fulfil the criteria specified for a valid method of data normalisation.

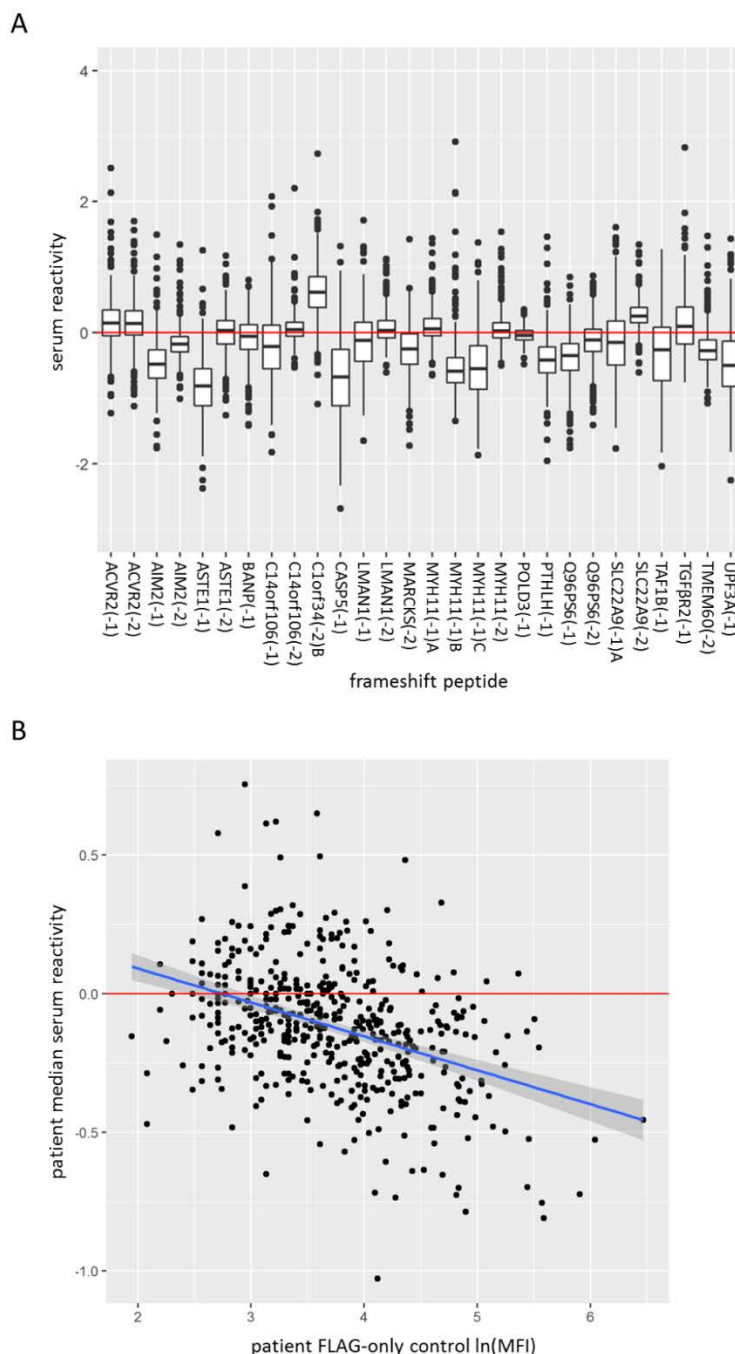


Figure 3.4: Distribution of serum reactivity between frameshift peptides (FSP) and between patients. FSP $\ln(\text{MFI})$ was normalised by subtraction of FLAG-only control $\ln(\text{MFI})$. **(A)** Distribution of serum reactivity is shown for each FSP. **(B)** Median serum reactivity for each patient relative to their FLAG-only control $\ln(\text{MFI})$; blue line = linear regression line, grey area = 95% CI for linear regression.

3.5. Regression-based Normalisation equalises Baseline FSP Serum Reactivity

For normalisation by division to be valid, it would be expected that for each FSP the $\ln(\text{MFI})$ should increase one-to-one in relation to the FLAG-only control $\ln(\text{MFI})$ when no $\alpha\text{FSP-Abs}$ are present. Based on the assumption that the majority of patients do not have $\alpha\text{FSP-Abs}$, median regression can be used to test the relationship between FSP and FLAG-only $\ln(\text{MFI})$ across all patients. Median regression was used as the median will not be affected by any extreme values from (the assumed minority of) individuals with $\alpha\text{FSP-Abs}$ against the FSP. Knowing that data normalisation by division was not appropriate, I expected to see the coefficient of the quantile regression deviating from 1. Indeed, median quantile regression of $\ln(\text{MFI})$ for each FSP against $\ln(\text{MFI})$ of the FLAG-only control showed a significant deviation of the regression coefficient from 1 in 22/28 of the FSPs (Table 3.2), confirming that the control MFI does not represent background fluorescence equally for each of the FSPs.

The median quantile regression statistics for each FSP also provided an alternative method of normalisation. For a patient (P) and FSP (F), the serum reactivity can be calculated using the following equation, where m_F is the coefficient derived from median quantile regression of the FSP $\ln(\text{MFI})$ versus the control $\ln(\text{MFI})$:

$$\text{serum reactivity}_{F,P} = \ln(\text{MFI})_{F,P} - m_F \cdot (\ln(\text{MFI}))_{\text{Control},P}$$

FSP	β	95% CIs	FSP	β	95% CIs
ACVR2(-1)	0.94	0.90-0.98	MYH11(-1)A	0.92	0.90-0.96
ACVR2(-2)	0.91	0.88-0.94	MYH11(-1)B	0.92	0.90-0.95
AIM2(-1)	0.84	0.82-0.88	MYH11(-1)C	0.71	0.63-0.74
AIM2(-2)	1.11	1.09-1.14	MYH11(-2)	0.92	0.90-0.94
ASTE1(-1)	0.73	0.69-0.78	POLD3(-1)	0.96	0.95-0.98
ASTE1(-2)	1.03	0.98-1.07	PTHLH(-1)	0.90	0.87-0.93
BANP(-1)	1.01	0.97-1.05	Q96PS6(-1)	1.08	1.03-1.11
C14orf106(-1)	0.63	0.53-0.69	Q96PS6(-2)	1.06	1.03-1.10
C14orf106(-2)	1.00	0.99-1.03	SLC22A9(-1)A	0.64	0.57-0.68
C1orf34(-2)B	0.77	0.72-0.81	SLC22A9(-2)	1.01	0.98-1.03
CASP5(-1)	0.47	0.42-0.53	TAF1B(-1)	0.41	0.35-0.46
LMAN1(-1)	0.64	0.60-0.69	TGFBR2(-1)	0.62	0.60-0.65
LMAN1(-2)	0.90	0.87-0.92	TMEM60(-2)	0.99	0.96-1.01
MARCKS(-2)	1.03	0.98-1.08	UPF3A(-1)	0.67	0.61-0.70

Table 3.2: Median regression of each frameshift peptide (FSP) compared to the FLAG-only control. For each FSP the $\ln(\text{MFI})$ was compared to $\ln(\text{MFI})$ of the FLAG-only control by median regression and the regression coefficient (β) determined with 95% confidence intervals (95% CIs).

Using normalisation by regression, each FSP had a median serum reactivity of approximately 0 and no correlation was observed between the median serum reactivity for a patient and the $\ln(\text{MFI})$ of the FLAG-only control (Figure 3.5). However, the variation in median serum reactivity for each patient remained high. For instance, the range of median serum reactivity

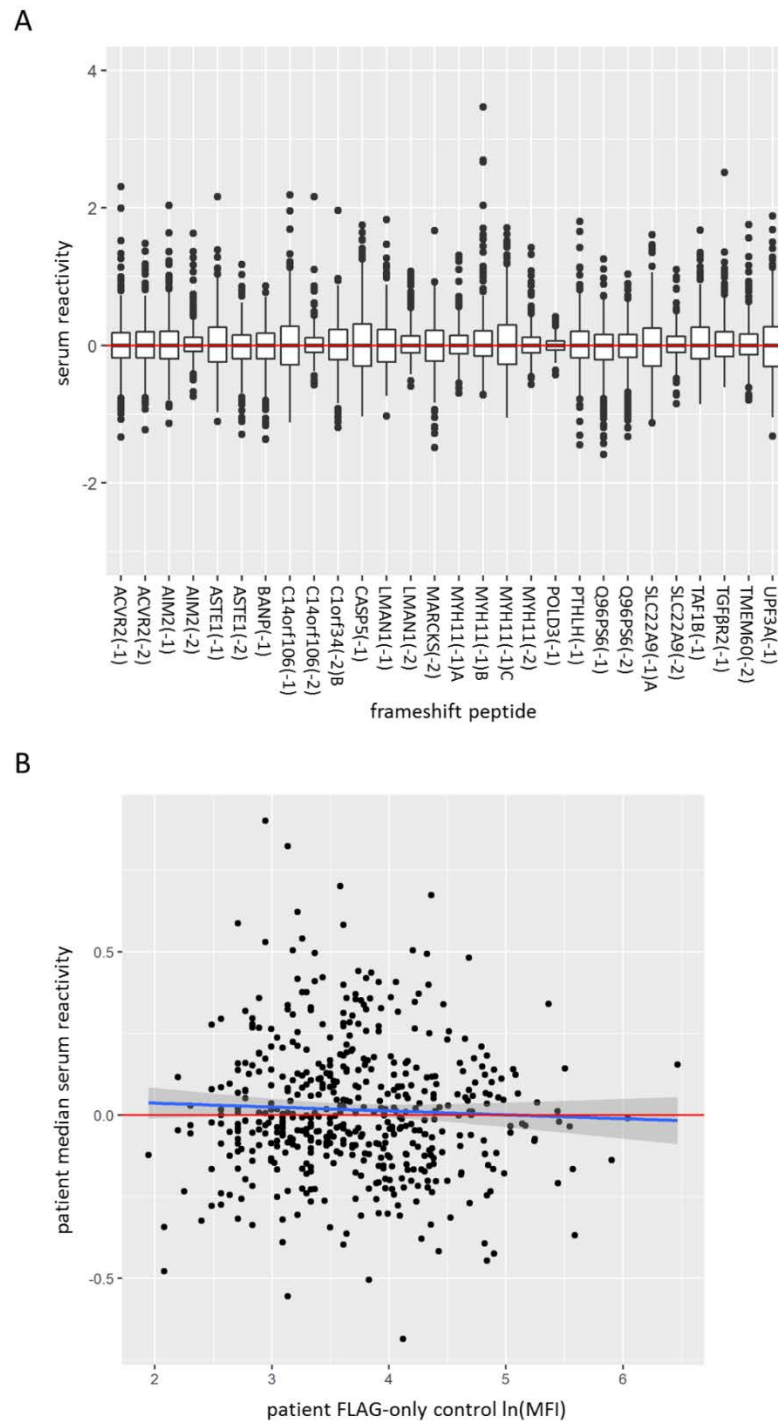


Figure 3.5: Distribution of serum reactivity between frameshift peptides (FSP) and between patients. FSP $\ln(\text{MFI})$ was normalised by the described method of quantile regression. **(A)** Distribution of serum reactivity is shown for each FSP, and **(B)** the median serum reactivity for each patient is shown in comparison to their FLAG-only control $\ln(\text{MFI})$.

values is from -0.7 to approximately 0.9, yet the interquartile range for serum reactivity values for each FSP were between -0.5 and 0.5. Furthermore, patients with median serum reactivity >0 had increased serum reactivity values across all FSPs compared to patients with median serum reactivity <0 (Mann-Whitney U test, $p < 10^{-15}$). This is unexpected based, again, on the assumption that only a minority of Lynch syndrome gene carriers should have serum α FSP-Abs against a minority of FSPs (Reuschenbach *et al*, 2010). Due to the limitations of the cohort analysed, specifically a lack of suitable controls as discussed in Section 3.3, the source of this could not be explored with confidence, and it was assumed to be a technical artefact caused by variation between sample reactions. Therefore, a “per patient” correction factor was needed to resolve the broad range of median serum reactivity values observed between patients. Following the regression-based normalisation, the median serum reactivity for each patient was subtracting from the serum reactivity of each FSP for that patient, therefore normalising serum reactivity between patients such that their median serum reactivity was equal to 0. I checked that this additional step in normalisation did not negatively affect the distribution of serum reactivity values for each FSP, and found that the median serum reactivity for each FSP remained close to 0 (Figure 3.6).

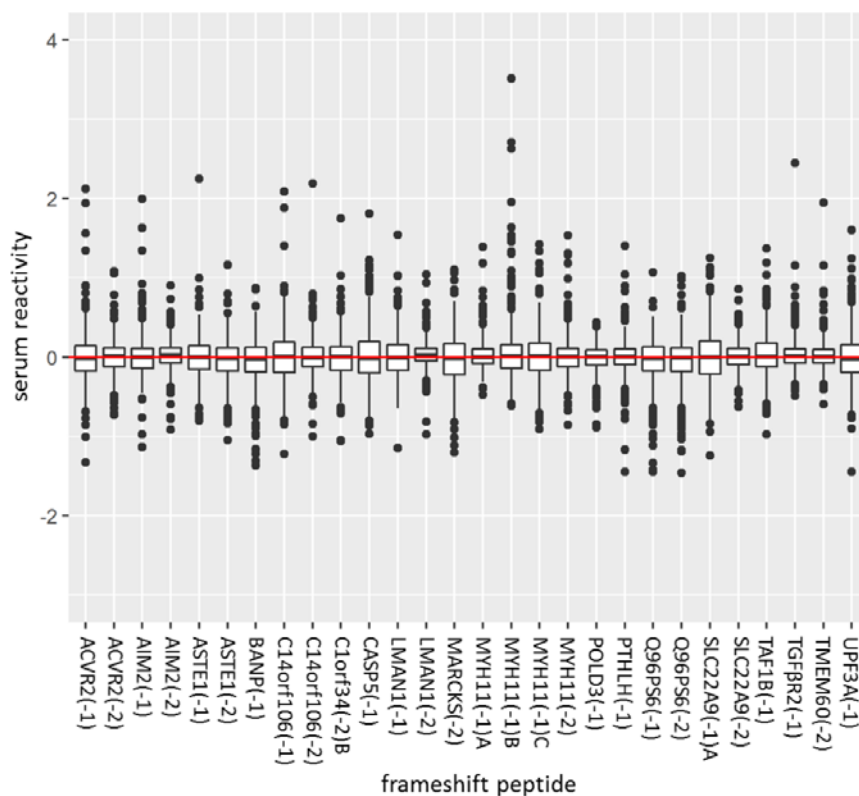


Figure 3.6: Distribution of serum reactivity between frameshift peptides (FSPs). FSP In(MFI) was normalised by the described method of quantile regression, followed by per patient correction. The distribution of serum reactivity is shown for each FSP.

As a final check of the method, I normalised the $\ln(\text{MFI})$ data from 11 positive control samples from two patients vaccinated with ASTE1(-1) and TAF1B(-1) either by subtraction or by regression and per patient correction. A comparison of serum reactivity from the two methods shows that FSPs for which the patients were not vaccinated cluster closer to 0 using the novel method, and the serum reactivity of the two FSPs the patients were vaccinated with is increased using the novel method (Figure 3.7). Ideally the method would also be validated in cancer-free and MMR mutation-negative controls and across repeats, but given the limitations of the cohort and the resources available it was not feasible to run this validation. However, as there was an improvement of normalisation by regression and per patient correction compared to normalisation by subtraction of control $\ln(\text{MFI})$, I decided to use this novel method of normalisation in all subsequent analyses.

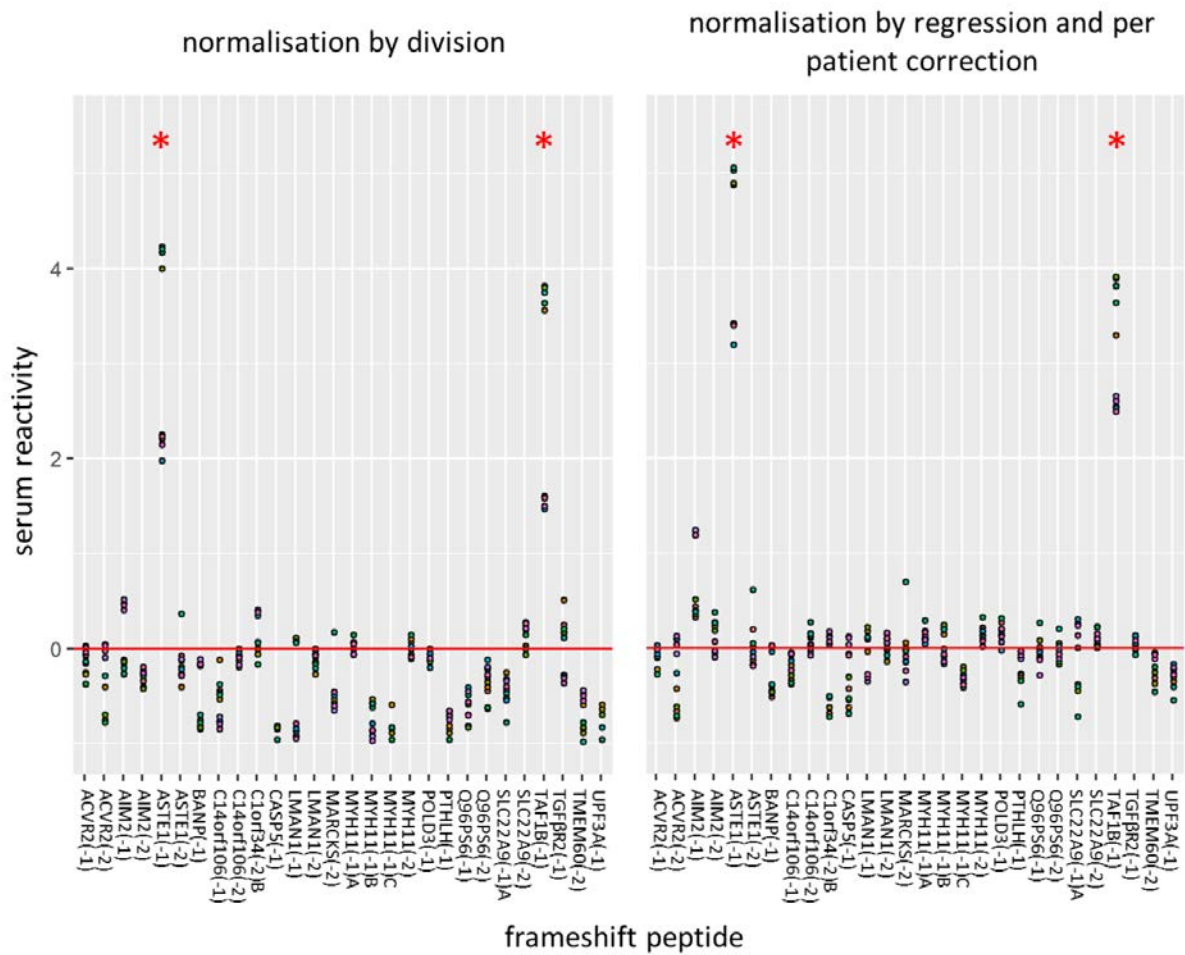


Figure 3.7: Distribution of serum reactivity between frameshift peptides (FSPs) in positive control samples. FSP $\ln(\text{MFI})$ was normalised by two methods, subtraction (equivalent to division; left-hand panel) and the novel method of quantile regression, followed by per patient correction (right-hand panel). The distribution of serum reactivity is shown for each FSP, and the two FSPs with which the patients were vaccinated, ASTE1(-1) and TAF1B(-1), are highlighted by an asterisk (*).

3.6. Frameshift Peptides cluster by Serum Reactivity

The generation of α FSP-Abs would require a complex interaction between mutations in the DNA, expression of the FSP, presentation of the FSP antigen to the immune system, and stimulation of a humoral immune response. At each stage multiple factors could influence the ultimate generation of α FSP-Abs. Given these potential influences on α FSP-Ab titres, it was of interest to see if any associations existed between the serum reactivity of the FSPs. This was tested by un-supervised clustering of FSPs by the serum reactivity detected for each, and four statistically significant cluster groups were identified (Figure 3.8).

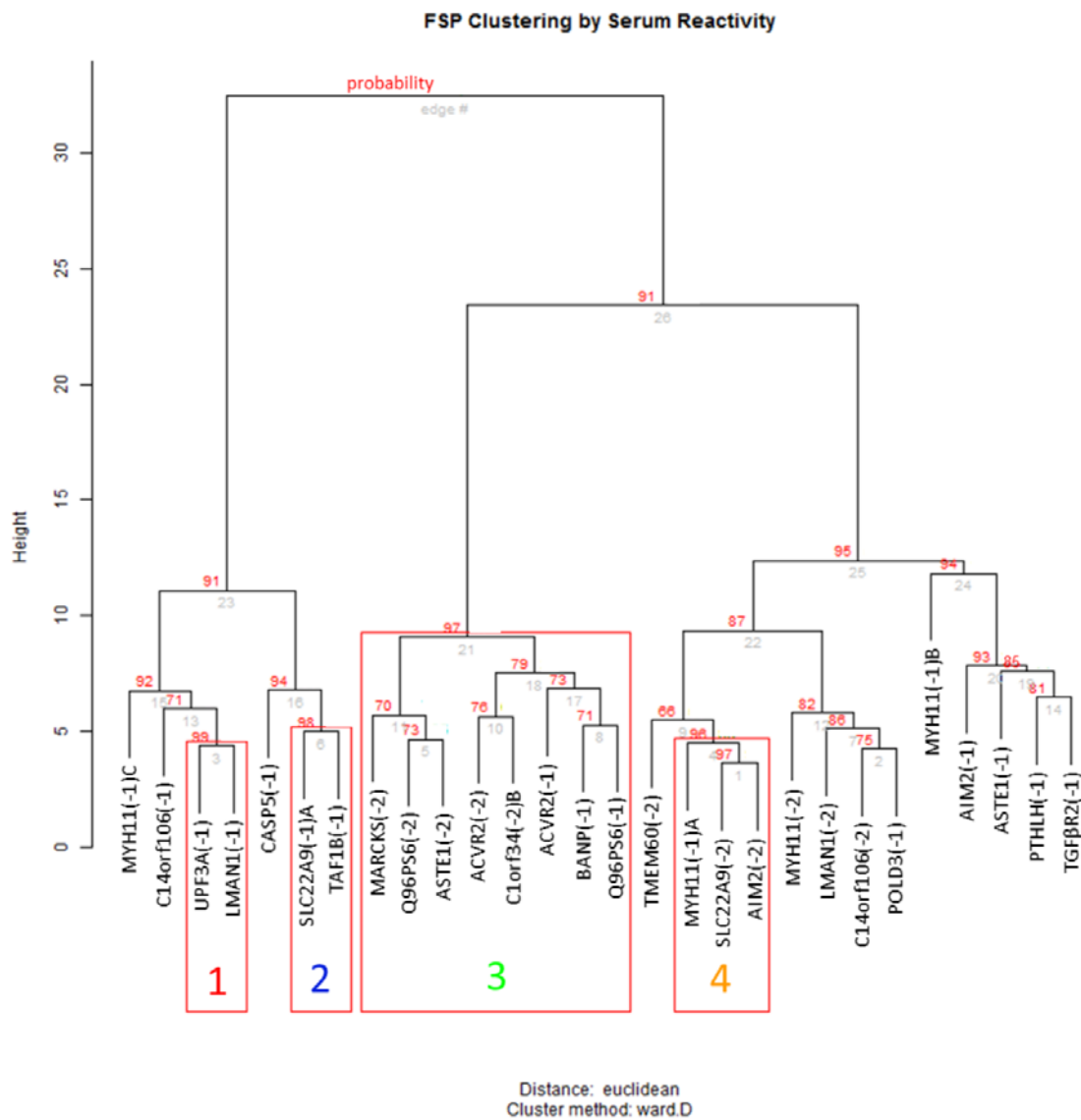


Figure 3.8: Frameshift peptide (FSP) clustering by serum reactivity. FSP serum reactivity was clustered using Ward's method and bootstrapping used to define statistically significant cluster groups ($p > 95\%$), with p values shown at each branch point in red. Red boxes highlight four significant cluster groups.

The presence of FSP clustering is an interesting finding, and biological explanations are discussed in Section 3.8. However, it is potentially a technical artefact of the method, but the resources and controls to explore this further were lacking. Irrespective of its source, the association between FSPs was informative for results interpretation.

3.7. A History of Colorectal Cancer is associated with Frameshift Peptide Serum Reactivity

In addition to cancer incidence, it was assumed that several patient variables may affect α FSP-Ab titres (see Section 3.3). To inform results interpretation, I looked for correlations between these patient variables before any analyses of association with FSP serum reactivity. No correlations were particularly striking or unexpected (Figure 3.9). Patient age was associated with a history of cancer, which is consistent with the increasing cumulative cancer risk for Lynch syndrome gene carriers as they age (Møller *et al*, 2017b). Patient age was also associated with surgical removal of colorectal or gynaecological tissues or whole organs, which is consistent with guidelines for extensive surgery in Lynch syndrome gene carriers for therapeutic or prophylactic reasons (Vasen *et al*, 2013). Patient age was associated with daily intake of aspirin, which is consistent with use of aspirin in cardiovascular disease (Calonge *et al*, 2009). Female sex was negatively correlated with a history of CRC and colorectal surgery, in agreement with an increased CRC rate in males with path_*MLH1* mutations relative to females and the presence of path_*MLH1* variants in approximately 30% of the patient cohort (Møller *et al*, 2017b), but positively correlated with colorectal or gynaecological surgery, consistent with prophylactic surgery to reduce risk of gynaecological cancers (Vasen *et al*, 2013). CRCs made up the majority of Lynch syndrome cancers diagnosed prior to blood draw and hence there was a strong, positive correlation between a history of CRC and Lynch spectrum cancers. A history of cancer was also associated with surgery, likely to be therapeutic according to guidelines (Vasen *et al*, 2013).

Due to the multiple covariates of interest that may affect FSP serum reactivity, a Multivariate Analysis of Covariance (MANCOVA) was used to analyse the data, with the patient variables as the independent variables and the serum reactivity of all 28 FSPs as the dependent variables. As this was an exploratory analysis, variables to be analysed further were identified by a p value < 0.05, with no correction for multiple testing. The MANCOVA found age (p = 0.04) and a history of CRC (p = 0.02) to be significantly associated with FSP serum reactivity (Table 3.3).

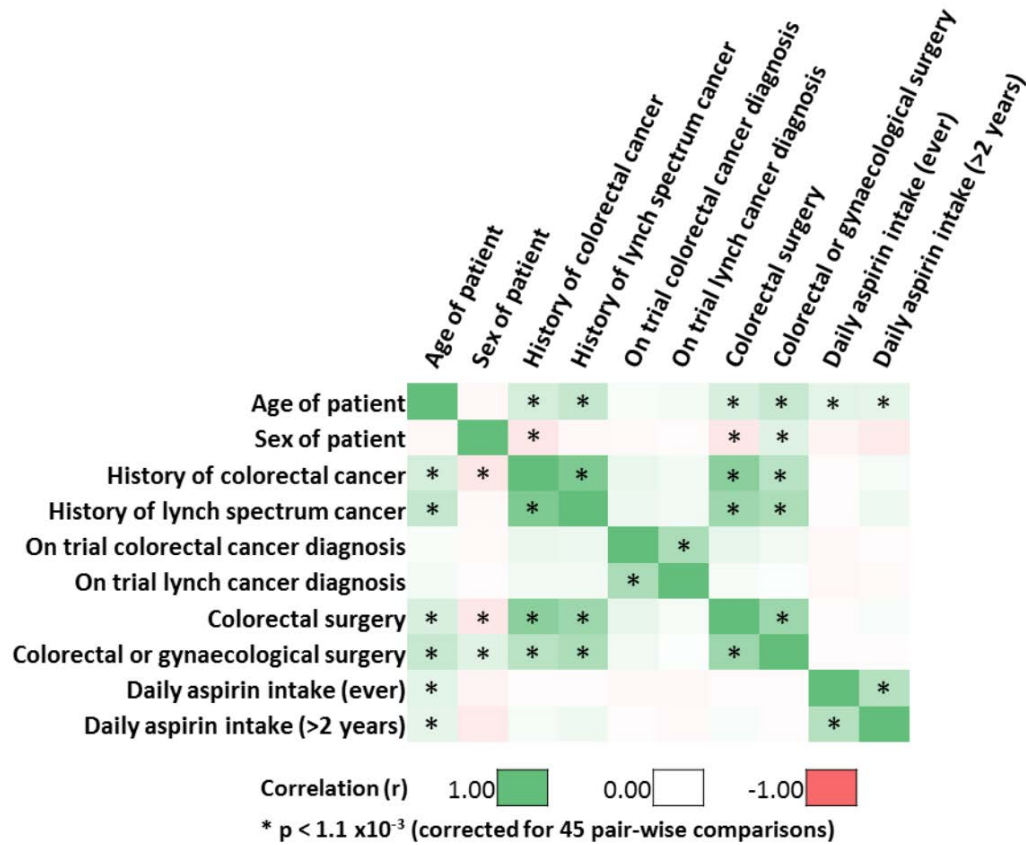


Figure 3.9: Correlation of patient variables. Pair-wise Pearson correlation was used to test associations between patient variables. The correlation r statistic (colour) and significant p values (*) are indicated.

Patient variable	p value
Age	0.040 *
Sex	0.195
Colorectal surgery	0.743
Colorectal or gynaecological surgery	0.073
Daily aspirin intake (>2years)	0.560
Daily aspirin intake (ever)	0.740
History of colorectal cancer	0.020 *
History of Lynch spectrum cancer	0.596
On trial colorectal cancer diagnosis	0.958
On trial Lynch spectrum cancer diagnosis	0.069
Global model	0.077

Table 3.3: Multivariate analysis of frameshift peptide (FSP) serum reactivity and patient variables. Patient variables with a statistically significant association with FSP serum reactivity are indicated by an asterisk (*).

The association of FSP serum reactivity with patient age and history of CRC was followed up by univariate analyses. In univariate analysis, patient age was not significantly associated with FSP serum reactivity (multiple linear regression, $p = 0.0542$), and none of the regression coefficients of individual FSPs, derived from the multiple regression model, were significant using a Bonferroni corrected threshold for multiple testing ($p > 0.0018$). In univariate analysis, patient history of CRC was again significantly associated with FSP serum reactivity (multiple logistic regression, $p = 0.0089$). Furthermore, two of the regression coefficients for individual FSP serum reactivity were significant using a Bonferroni corrected threshold for multiple testing ($p < 0.0018$, Table 3.4). The distributions of serum reactivity for these two FSPs, LMAN1(-2) and TAF1B(-1), were subsequently analysed with respect to a history of CRC. In LMAN1(-2), the significant difference in serum reactivity between patients with and without a history of CRC diagnosis was confirmed (Mann-Whitney U test, $p = 0.003$; Figure 3.10A). However, this was not so for TAF1B(-1) (Mann-Whitney U test, $p = 0.153$; Figure 3.10B), although a trend toward higher serum reactivity was evident in the patient group with a history of CRC. In the multiple regression that identified LMAN1(-2) and TAF1B(-1), the significance of each regression coefficient accounts for variance attributable to other variables. Interestingly, LMAN1(-2) is not part of a significant cluster group (Figure 3.8) and, therefore, will have less shared variance in serum reactivity with other FSPs than TAF1B(-1), leading to similar p values when analysed as a single variable or as part of a multiple regression. TAF1B(-1), however, is part of a significant cluster group (cluster 2; Figure 3.8), and therefore its shared variance with other FSPs in the same cluster will not contribute to calculations of significance in the multiple regression, producing a significant result which is not observed when analysed as a single variable.

Previously, Reuschenbach *et al* compared the highest optical density (a measure of antibody binding from ELISA, equivalent to serum reactivity) detected in MSI-high CRC patients and controls, and found a significant increase in the MSI-high CRC patients ($p = 0.036$) (Reuschenbach *et al*, 2010). A similar observation was made in our cohort of Lynch syndrome gene carriers, those patients with a history of CRC had a significant increase in their highest serum reactivity compared to those patients without a history of CRC (Mann-Whitney U test, $p = 0.012$). The time from CRC diagnosis to blood draw may also affect FSP serum reactivity as long term antibody production decreases with time (Dörner and Radbruch, 2007). Multiple regression against the serum reactivity of all FSPs showed that there was a trend associating serum reactivity with time since CRC diagnosis ($p = 0.072$), but

FSP	β	p value	FSP	β	p value
TAF1B(-1)	2.9167	0.0003 *	BANP(-1)	0.0153	0.9857
LMAN1(-1)	1.2530	0.1815	TGFBR2(-1)	-0.0815	0.8950
ASTE1(-1)	1.1966	0.0511	MYH11(-1)B	-0.1393	0.6729
SLC22A9(-2)	1.1528	0.2619	PTHLH(-1)	-0.1484	0.8643
POLD3(-1)	1.1229	0.3892	C1orf34(-2)B	-0.3996	0.5785
Q96PS6(-1)	1.0784	0.2231	C14orf106(-1)	-0.4733	0.4565
MYH11(-1)C	0.6382	0.2960	SLC22A9(-1)A	-0.5405	0.4737
ASTE1(-2)	0.4763	0.6378	MYH11(-1)A	-0.7110	0.4074
ACVR2(-2)	0.4568	0.6211	CASP5(-1)	-0.8659	0.1826
C14orf106(-2)	0.3867	0.5656	Q96PS6(-2)	-0.9560	0.2720
MYH11(-2)	0.3791	0.5546	AIM2(-1)	-1.1139	0.0687
ACVR2(-1)	0.3772	0.4611	AIM2(-2)	-1.2687	0.2657
TMEM60(-2)	0.2305	0.8152	UPF3A(-1)	-1.4690	0.1369
MARCKS(-2)	0.0905	0.8921	LMAN1(-2)	-4.2851	0.0009 *

Table 3.4: Regression coefficients (β) from multiple logistic regression of patient colorectal cancer (CRC) history and frameshift peptide (FSP) serum reactivity. FSPs with serum reactivity that are significantly associated with a history of CRC ($p < 0.0018$) are indicated by an asterisk (*).

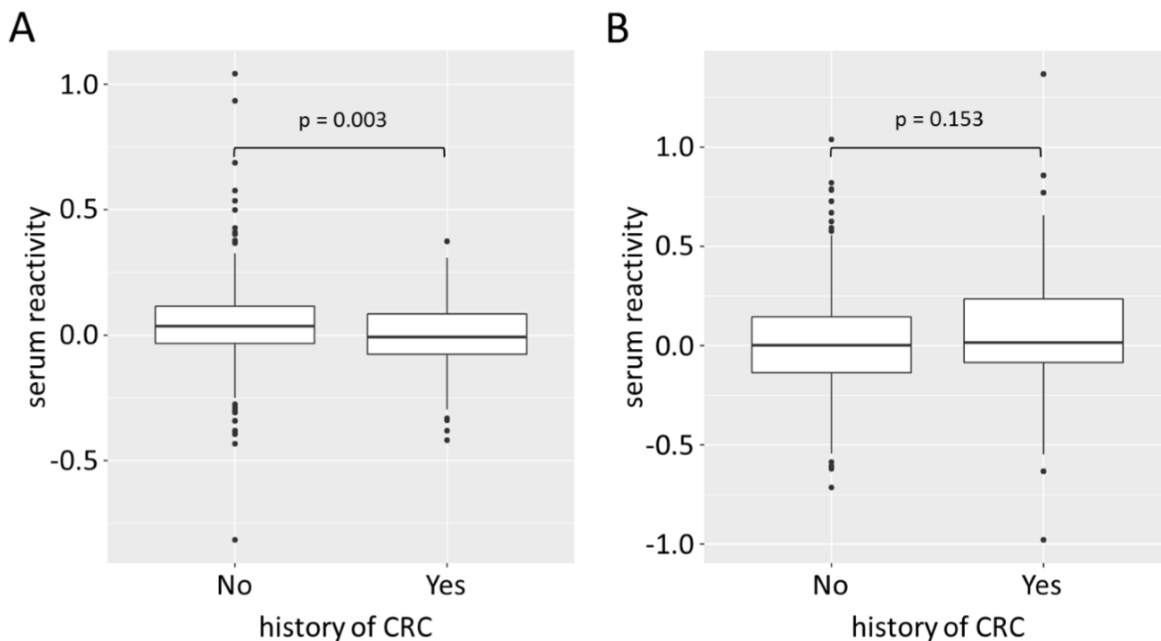


Figure 3.10: Distribution of serum reactivity in patients with or without a history of colorectal cancer (CRC). Shown for two FSPs that had significant associations between serum reactivity and a history of CRC by multiple regression, (A) LMAN1(-2), (B) TAF1B(-1).

the regression coefficients for individual FSPs were not significant, following Bonferroni correction for multiple testing ($p > 0.0018$). Finally, the three patients that were diagnosed with CRC on-trial (i.e. following blood draw, Table 3.1) were analysed. Serum reactivity for all FSPs was <0.4 for these three patients, giving no indication of serum reactivity beyond that observed in the majority of other Lynch syndrome gene carriers.

3.8. Discussion

The overall aim of this study was to assess the applicability of α FSP-Ab monitoring as a potential surveillance tool in Lynch syndrome gene carriers for early detection of CRC. To achieve this, a novel multiplex method based on antibody binding to beads coated with synthetic FSPs was used. The method had previously been explored in a relatively small cohort of 20 MSI-high CRC patients, with no selection by germline MMR gene mutation status (Reuschenbach *et al*, 2014). Therefore, a larger cohort of Lynch syndrome gene carriers was needed to address this aim. Sera taken at trial entry from patients participating in the CaPP3 clinical trial provided a cohort of 464 samples. However, due to its novelty, the technical aspects of the bead-based multiplex also needed to be assessed to optimise the method of normalisation.

Normalisation is required to account for technical or sample variables that may affect binding kinetics or quantification of fluorescence. In this method, such variables are accounted for by a FLAG-only control, and it was originally assumed that fluorescence from binding would follow an additive model. However, I found that a multiplicative model better represented the relationship between FSP MFI and control MFI, evident in the heteroscedasticity in variance for the raw data, which was removed by log transformation. Modelling the MFI signal versus control MFI by a multiplicative model fits with common assumptions of binding kinetics. Modelling serum protein P (specific and non-specific antibodies, and other molecules that contribute to PE fluorescence) and ligand L (FSP-bead conjugate) binding as a second order reaction:



the concentration of protein-ligand $[PL]$, which is equivalent in this case to the concentration of the fluorescent antibody-FSP-bead conjugate, can be calculated by:

$$[PL] = K_A[P][L]$$

where K_A represents the association constant of binding (Pollard, 2010). The sample variables that affect binding (specific and non-specific), such as the concentration of

peptides in the serum, or the concentration of salts, etc, will affect one of K_A or $[P]$. Likewise, technical variables, such as batch temperature, or the concentration of FSP-bead conjugates, etc, will affect one of K_A or $[L]$. These are multiplicative terms, and therefore this second order model of binding, for background and true signal, is consistent with the observation that an additive model was not appropriate for the relationship between FSP and control MFI.

When testing the method of normalisation, a regression-based and per patient correction was superior to normalisation by division (subtraction of log transformed data), giving an equivalent base-line serum reactivity between patients and between FSPs. Regression-based normalisation is a common tool. For example, it is frequently used to analyse microarray data in gene expression analyses to account for intensity variations from the quantity of input RNA, differences in detection or labelling efficiencies of different dyes, and spatial differences on the microarray surface (Quackenbush, 2002). However, as discussed within the results section, the validity of the regression and per patient normalisation needs to be confirmed and it is likely a superior normalisation method could be developed with additional experimentation. For example, whilst it was assumed for this study that the majority of Lynch syndrome gene carriers will not have α FSP-Abs based on the evidence of Reuschenbach *et al* (2010), it is feasible that their frequent history of MMRd CRC (Møller *et al*, 2017b), and increased rate of MMRd precancerous lesions, such as colorectal adenomas (de Jong *et al*, 2004a) and MMR-DCF (Kloor *et al*, 2012), produces a background of α FSP-Abs obscuring the normalisation procedure. Comorbidities and other disease could also have an effect on the immunological background of the patient, but this data was not available. These unaccounted variables could be a biological explanation for the large range in the median serum reactivity between different patients observed before per patient normalisation. However, here it was assumed that this variation was a technical artefact given there was no means to explore biological explanations, and because any signal would need to be detectable above such background for this method to be clinically useful for surveillance in Lynch syndrome gene carriers. Hence, samples from cancer-free and MMR mutation negative controls would be desirable to accurately describe the relationship between FSP and control MFI and further improve the normalisation method. Furthermore, the regression method assumes a linear model is appropriate, but additional data transformation may be needed for this to apply (Quackenbush, 2002). Again, negative controls would clarify this. Despite these caveats and limitations of the study cohort, the

regression and per patient normalisation showed a clear improvement from the previous method of normalisation, and therefore was used to generate serum reactivity data for all analyses.

The clustering of FSPs by serum reactivity was an interesting finding. Explanations could be biological or technical. Technical reasons could include similarities in preparation of FSP-bead conjugates, or non-specific features of FSP structure such as length, hydrophobicity, or formation of secondary structure. As was the case for optimisation of normalisation, such technical variables would ideally need a cohort of negative controls and a series of experiments to be tested. However, such additional work was beyond the resources of this study. The biological explanations for FSP clustering by serum reactivity come back to the influences of the different stages through which mutations in an MMRd tumour stimulate a humoral immune response. For example, cMNR mutations are likely drivers of MMRd tumorigenesis and would therefore be subject to strong selection pressures (Duval and Hamelin, 2002) and, hence, it is feasible that patterns of cMNR mutations arise depending on the other mutations within the tumour, the tumour microenvironment, and so on. Similarly, the HLA type of the patient will determine the affinity of MHC receptors for any FSPs expressed, leading to variation in the pattern of FSP antigen presentation to the immune system from patient to patient (Saeterdal *et al*, 2001). However, much of this is speculation and there is no evidence, to my knowledge, to support or refute these possibilities. Given time and resources, the method of Reuschenbach *et al* (2014) could be used to answer these hypotheses.

α FSP-Abs generated in response to MMRd CRCs could be used for surveillance of Lynch syndrome gene carriers as an alternative to the invasive procedure of colonoscopy, which cannot detect some precancerous lesions, evident in the frequent diagnosis of CRC during surveillance intervals (Seppälä *et al*, 2017), and has low rates of compliance (Stuckless *et al*, 2012). An association between FSP serum reactivity and a history of CRC was found in the cohort of Lynch syndrome gene carriers analysed. In addition, Lynch-spectrum cancers were not associated with FSP serum reactivity, consistent with selection of FSPs based on the frequency of the associated cMNR frameshift mutation in MMRd CRCs (Woerner *et al*, 2010) and the different frequencies of cMNR mutations in different cancer types (Kim *et al*, 2013). A history of CRC, rather than diagnosis of CRC after blood draw, was used due to the low number of on-trial cancers and based on the assumption that the production of antibodies will continue for a long time after disease. Previously, it has been shown that

long-lived plasma cells and memory B cells are responsible for antibody secretion over many years, potentially decades, following viral infection (Dörner and Radbruch, 2007), which is likely applicable to tumours. For example, B cell follicles (centres of antibody production and B cell activation and differentiation) form within the tumour microenvironment (Bindea *et al*, 2013), and both T and B cell populations develop anti-tumour immunological memory (Nielsen *et al*, 2012; Sarvaria *et al*, 2017; Amsen *et al*, 2018). In support of this, I found no association between FSP serum reactivity and the length of time from CRC diagnosis to blood draw despite the association between CRC history and FSP serum reactivity.

Other observations from this study include a weak association between age and FSP serum reactivity. This may be due to the association of age with a history of CRC (Pearson's correlation $r = 0.277$, $p < 10^{-9}$), and could also reflect humoral responses against a history of MMRd lesions that do not progress to cancer, such as MMRd colorectal adenomas or MMR-DCF, which have both been shown to contain cMNR frameshift mutation also found in CRC (Iino *et al*, 2000; Staffa *et al*, 2015). Also, when the association between a history of CRC and serum reactivity was analysed by multiple regression, it was found that regression coefficients of two FSPs were statistically significant. TAF1B(-1) showed the expected positive correlation between CRC history and serum reactivity, explicable by a humoral immune response against cancers containing cMNR frameshift mutations in the respective gene. However, LMAN1(-2) serum reactivity was negatively correlated with CRC history, which is difficult to explain biologically; could LMAN1min2 serum reactivity represent a shadow of immunological prevention of cancer? 26/28 FSPs analysed showed no significant association of serum reactivity with a history of CRC, which does not hold promise for these markers being used as early detection biomarkers of disease.

α FSP-Abs were hypothesised to be a good candidate biomarker for the early detection of MMRd CRC due to the high immunogenicity of FSPs (Kloor and von Knebel Döberitz, 2016). However, the individual insensitivity of an FSP to detect α FSP-Abs in the MMRd CRC patient sera could be explained by a lack of any of several required conditions within the patient tumour immune response. These include a lack of the frameshift mutation in the tumour, an incompatible HLA type of the patient, or the evolution of immune evasion by the tumour (Kloor and von Knebel Döberitz, 2016). The method used in this study was, therefore, chosen as it could multiplex many FSPs to improve the sensitivity of autoantibody detection (Robinson *et al*, 2002). Diagnostic accuracy of early detection using this method could not be assessed due to a lack of controls to validate the analysis method and the low

number of on-trial CRC diagnoses. However, FSP serum reactivity values observed in the three patients who developed on-trial CRCs did not appear distinct from background noise, with all values falling below a serum reactivity of 0.4, despite the high likelihood of latent malignancy at the time of blood draw (diagnosis was within 2-19 months of trial entry). Therefore, detection of α FSP-Abs by the described method is not a sensitive assay despite its multiplex analysis, and it is unlikely to have clinical utility for the surveillance of Lynch syndrome gene carriers. This lack of sensitivity could be due to technical or biological reasons. Technical reasons are less likely given the detection of α FSP-Abs against ASTE1(-1) and TAF1B(-1) in the positive control sera of patients vaccinated with these same FSPs. With respect to possible biological explanations, there is sufficient evidence to show that cMNR mutations lead to specific immune responses against the associated FSP antigen (Saeterdal *et al*, 2001; Schwitalle *et al*, 2008; Tougeron *et al*, 2009; Maby *et al*, 2015; Le *et al*, 2017), but these studies assessed cellular rather than humoral immunity, for example by stimulating peripheral and tumour infiltrating lymphocytes with FSPs, or T cell killing of MMRd cell lines. The publication of Reuschenbach *et al* in 2010 is the only study I am aware of, other than Resuschenbach *et al*, 2014, that has analysed humoral immunity against MMRd cancer using multiple sera and multiple FSPs. Reuschenbach *et al* (2010) used ELISA, a well-established technique for detection of serum antibodies, and showed serum reactivity in 20/69 (29%) MMRd CRC patients against at least one FSP in a panel of 8 derived from 6 cMNR frameshift mutations common to MMRd CRC. However, 9/31 (29%) healthy, Lynch syndrome gene carriers and 8/52 (15.4%) controls, respectively, also had serum reactivity against one FSP or more. This suggests that α FSP-Abs are only infrequently generated against MMRd CRCs and that non-specific binding to FSPs may account for the majority of signal observed. Although Reuschenbach *et al* (2010) used a much smaller panel of FSPs, their results are consistent with the observations of this study, suggesting that α FSP-Ab titres are likely to be poor biomarkers for the early detection of MMRd CRC.

3.9. Conclusions and Future Work

By using the multiplex method of Reuschenbach *et al* (2014), I showed that there is an association in Lynch syndrome gene carriers between a diagnosis of CRC and serum reactivity to FSPs, likely due to α FSP-Abs in the peripheral circulation. However, detection of α FSP-Abs by the described method is unlikely to be a useful biomarker test for early detection of MMRd CRC as the majority of patients with a CRC diagnosis (pre- or post-blood

draw) have serum reactivity values equivalent to patients without a CRC diagnosis. Furthermore, all but two of the FSPs analysed show no correlation with CRC incidence. It is feasible that optimisation of the method, particularly data normalisation, may improve its sensitivity for α FSP-Abs, but this would require significant investment to analyse cancer-free and MMR mutation-negative controls to validate normalisation and test technical variables. If the protocol can be optimised and shown to be robust to technical variables, additional data and samples collected as the CaPP3 clinical trial progresses will be available to, for example, re-assess the FSP serum reactivity of additional patients with on-trial CRC diagnoses, which would be a more direct evaluation of the analytical validity of the method for early detection of CRC in Lynch syndrome gene carriers. The longitudinal design of the CaPP3 study would also allow monitoring of patient FSP serum reactivity over time, which could be more informative than using a cross sectional study of a Lynch syndrome gene carrier population. Again, more detailed technical validation of the method would be the priority before any additional studies were carried out.

Whilst this study failed to develop a novel biomarker test for Lynch syndrome surveillance, extensive guidelines exist for alternative surveillance methods (Vasen *et al*, 2013). Colonoscopy, for example, remains an effective surveillance technique (Järvinen *et al*, 2000) even if it does not detect all colorectal lesions in Lynch syndrome gene carriers (Seppälä *et al*, 2017). Therefore, irrespective of these results, the identification of Lynch syndrome gene carriers is required for optimal patient management, and novel biomarker tests for MMR deficiency in CRC are needed to meet the demand for high throughput screening of all CRCs (Newland *et al*, 2017).

Chapter 4. Development of a Short Mononucleotide Repeat Sequencing Assay to Detect Microsatellite Instability in Colorectal Cancer

4.1. Introduction

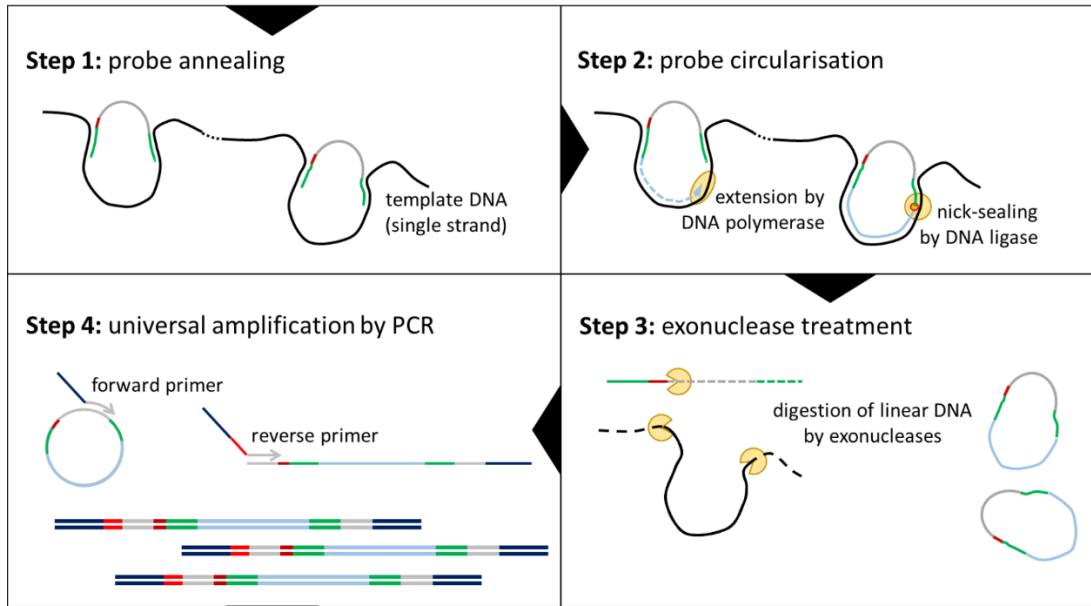
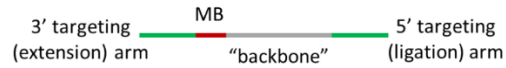
High throughput MMR deficiency testing is needed to meet two clinical needs: testing of all CRCs to screen for Lynch syndrome, and testing of any solid cancer to predict response to immune checkpoint blockade therapy. I aimed to develop an NGS-based MSI assay that facilitates automated and cheap MSI analysis, continuing the work of Redford *et al* (2018). In their study, Redford *et al* defined 120 informative microsatellite markers from whole genome sequence data from MSI-high and MSS CRCs, and matched normal tissue in The Cancer Genome Atlas. MNRs, rather than DNRs for example, were selected for their superior sensitivity and specificity for MSI status (Bacher *et al*, 2004), including sensitivity for MSH6 deficiency (You *et al*, 2010). The lengths of these MNRs ranged from 7-12bp, making them “short” relative to the longer markers employed by other assays. The MSI Analysis System (Promega), for example, uses 21-28bp MNRs (Bacher *et al*, 2004). Short MNRs are less likely to be polymorphic (Ananda *et al*, 2013), meaning that matched normal DNA was not needed for results interpretation. Furthermore they can be accurately sequenced due to low PCR and sequencing error (Fazekas *et al*, 2010). Finally, each selected marker had an associated SNP with a minor allele frequency >20%, within 30bp of the microsatellite, allowing the allelic origin of microsatellite length variants to be determined in heterozygotes.

A panel of 17 of these markers was selected based on sensitivity and specificity for MSI status from sequencing a discovery cohort of 6 MSI-high CRCs and 6 MSS CRCs. The method amplified the 17 marker panel in singleplex from each sample, and amplicons from each sample were pooled, purified and subject to a second PCR in which sample index sequences and sequencing adapters (for compatibility with Illumina sequencing platforms) were added. Sample-indexed amplicons were pooled into a sequencing library, sequenced in forward and reverse orientations on the MiSeq platform (Illumina), and fastq files were processed, as described in Section 2.12.1. A naïve Bayes approach (Section 2.12.2) was followed to develop an MSI classifier that would classify samples as MSI-high or MSS using the proportion of reads containing deletions and their allelic distribution. It was trained using a cohort of 67 MSI-high and 72 MSS CRCs, which had previously been typed by the MSI Analysis System (Promega). An independent cohort of 70 CRCs (36 MSI-high, 34 MSS, again

typed by MSI Analysis System) was used for validation and, across both cohorts, the assay achieved 98% sensitivity and 98% specificity (Redford *et al*, 2018), equivalent to FLA, IHC and NGS-based methods (Zhu *et al*, 2018). Deletions, rather than any variant in microsatellite length, were used as deletions are more frequent than insertions in MSI-high CRCs and cell lines, suggesting that any insertions detected are likely PCR or sequencing error (Sia *et al*, 1997; Lu *et al*, 2013). Increased discrimination of MSI status was achieved by analysing the allelic bias of reads containing deletions, as single deletion events will stochastically affect one allele and not the other. MSI-low samples were considered equivalent to MSS samples (Section 2.12.2; Halford *et al*, 2002; Laiho *et al*, 2002).

The turnaround time (TAT) of the assay was estimated to be 11 days and the cost was £26.20 per sample, assuming 96 samples being analysed per batch. Both TAT and cost are therefore inferior to the dominant MSI Analysis System (Promega), and this was largely due to the singleplex amplification of the 17 markers (Alhilal PhD Thesis, 2016). Protocol optimisation by multiplexing of the markers was needed. There is a plethora of programs available to design PCR primers suitable for multiplexing, but interactions between primers limit the number of markers in the multiplex, which is often difficult to predict, and differential amplification efficiency of different primer pairs produces unequal representation of each amplicon, which is not trivial to balance (Sint *et al*, 2012). Molecular inversion probes (MIPs) are an attractive technique for target enrichment due to robust multiplexing, with several thousand loci being amplified in one reaction (O’Roak *et al*, 2012), and a simple, automatable protocol (Hiatt *et al*, 2013; Neveling *et al*, 2017; Figure 4.1). To balance the number of sequencing reads detected from each locus the concentration of each MIP can be modified with ease (Niedzicka *et al*, 2016). Many of the 120 markers defined by Redford *et al* were untested and, therefore, potentially superior to those selected and so could be added into the multiplexed assay. Furthermore, detection of *BRAF* V600E mutations in MMRd CRCs indicates the tumour is of sporadic origin and is not Lynch syndrome (Domingo *et al*, 2004), hence simultaneous testing of *BRAF* reduces the Lynch syndrome screening pipeline to one step prior to germline genetic testing rather than the current practice of MMR deficiency testing followed by *BRAF* testing (Newland *et al*, 2017). *RAS* gene mutations predict CRC response to anti-EGFR therapy (De Roock *et al*, 2010a) and substitutions at *KRAS* G12 and G13 account for approximately 93% of all *RAS* gene mutations in CRC. Therefore, inclusion of other, relevant biomarkers in the multiplex, such as *BRAF* and *KRAS* mutation hotspots, would make the assay competitive with gene panel sequencing.

smMIP structure



sequencing-ready amplicons

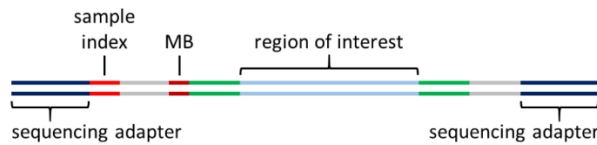


Figure 4.1: Protocol for multiplex loci capture, molecular barcode (MB) tagging, and amplification by single molecule molecular inversion probes (smMIPs). The protocol uses four steps. **1:** The two targeting arms of each probe anneal to marker loci in the template DNA. Each MIP molecule contains an MB as a unique identifier. **2:** probes circularise by polymerase extension from the 3' arm of the probe to gap-fill between the targeting arms, and nick sealing by ligation of the 3' end of the extension to the 5' targeting arm. **3:** exonuclease treatment removes linear DNAs, including template DNA and un-circularised probes. **4:** PCR amplification using universal primers adds sample index sequences and sequencing adapters to create sequencing-ready amplicons.

In addition to optimising the assay protocol, it was important to show the analytical validity of the assay with a view to its acceptance into clinical practice (Hayes, 2018). A collaboration between the Association for Molecular Pathology and the College of American Pathologists has defined guidelines for the validation of NGS-based oncology assays, covering the parameters that must be tested to support analytical validity and define assay limits (Jennings *et al*, 2017). These guidelines extensively cover the quality controls (QCs) for samples and sequencing to ensure reliable assay performance. Many of these QCs are independent of the assay. For example, it is recommended that, during sample preparation, equipment is thoroughly cleaned or new disposable consumables be used for each sample to

prevent contamination. For sequencing, it is recommended that cluster density, or total reads generated, be monitored to ensure they are within the expected range, as exceptionally densities or read depths suggest error in library preparation, or execution of sequencing. Some QCs, however, are dependent on the assay. For example, the lower limit of detection (LLoD) of an assay will determine the minimum tumour cell content required in a sample to generate reliable results.

The MSI classification method of Redford *et al* (2018) uses the probability that the observed proportion of reads containing microsatellite deletions and the observed distribution to deletions to different alleles belong to an MSI-high or MSS CRC. Sample variables that will affect results include:

1. sample composition, as a lower MSI-high content will reduce the signal to noise ratio
2. sample quality, as poor quality may introduce changes in the microsatellite sequence
3. sample quantity, as low library complexity (i.e. a low number of sample molecules sequenced) may skew representation of template DNA in sequencing reads

Sequencing variables that will affect results include:

1. read depth, as lower read counts will increase the confidence intervals of an observed proportion of reads (assuming read counting follows a binomial distribution)
2. base-call quality, as erroneous base-calls may change the detected length of a microsatellite

Therefore, each of these variables should be assessed during assay development to define assay QCs. For such assessments, the guidelines of Jennings *et al* recognise that it is infeasible to control for every sample variable. Therefore, based on calculations of non-parametric tolerance intervals, they recommend that a minimum of 59 independent samples, representative of the sample population the assay is intended for, should be used to test any of these parameters.

4.2. Aims

To optimise the MSI assay of Redford *et al* (2018), and to test its analytical validity, I aimed to:

1. Multiplex the 17 short MNRs, previously proven to be highly accurate for MSI status, together with additional microsatellite markers defined by Redford *et al* (2018) using smMIPs.

2. Determine the diagnostic accuracy (i.e. sensitivity and specificity) of the smMIP-based MSI assay, by re-training of the MSI classifier described by Redford *et al* (2018) and validation in an independent cohort of CRCs.
3. Include additional, clinically actionable biomarker loci, specifically *BRAF* V600 and *KRAS* G12 and G13 mutation hotspots, in the smMIP-based MSI assay.
4. Determine the limits of the smMIP-based MSI assay with respect to sample and sequencing variables, and define assay-specific QCs.

4.3. Multiplex Amplification of Microsatellites using Molecular Inversion Probes

Amplification and sequencing errors in microsatellite length can be reduced by use of high fidelity, Phusion polymerases, and the highest fidelity of these was found to be Herculase ii polymerase (Agilent; Fazekas *et al*, 2010). For this reason, Herculase ii polymerase was used in the work of Redford *et al* (2018). To show that the smMIP protocol could use Herculase ii polymerase instead of the Taq polymerase used in the original protocol, a positive control smMIP sequence was obtained from the study of Hiatt *et al* (2013) and shown to be amplifiable with Herculase ii polymerase in our laboratory (Figure 4.2A). Subsequently, smMIPs for the 17 short MNR loci analysed by Redford *et al* (2018) were designed using MIPgen (Section 2.8.2; Boyle *et al*, 2014a). These smMIPs were tested in singleplex (Figure 4.2B) and 15/17 produced visible amplicons. The two markers that failed to amplify (IM66 and LR20) were not taken forward. To verify the content of each smMIP amplicon, primers targeting the sequence internal to the smMIP targeting arms were designed, compatible with primers specific to the universal sequence of smMIP amplicons (Section 2.8.2). PCR amplification, using these primers and purified smMIP amplicon as template, produced secondary amplicons of the expected size, confirming that the smMIP amplicons contained the correct sequence (Figure 4.2C). The 15 MIPs were then pooled together, the markers were amplified in a 15plex reaction (Figure 4.2D), and the expected products were again verified by amplification of sequence internal to the amplicons (Figure 4.2E).

4.4. Amplicon Sequencing identifies Variants in Control Samples

The smMIP protocol produces amplicons containing sequencing adapters and sample index sequences (for read demultiplexing) for use with Illumina sequencing, a ubiquitous platform in healthcare services globally (Levy and Myers, 2016) and, hence, appropriate for high throughput assays. Four control samples were selected for sequencing, including DNAs

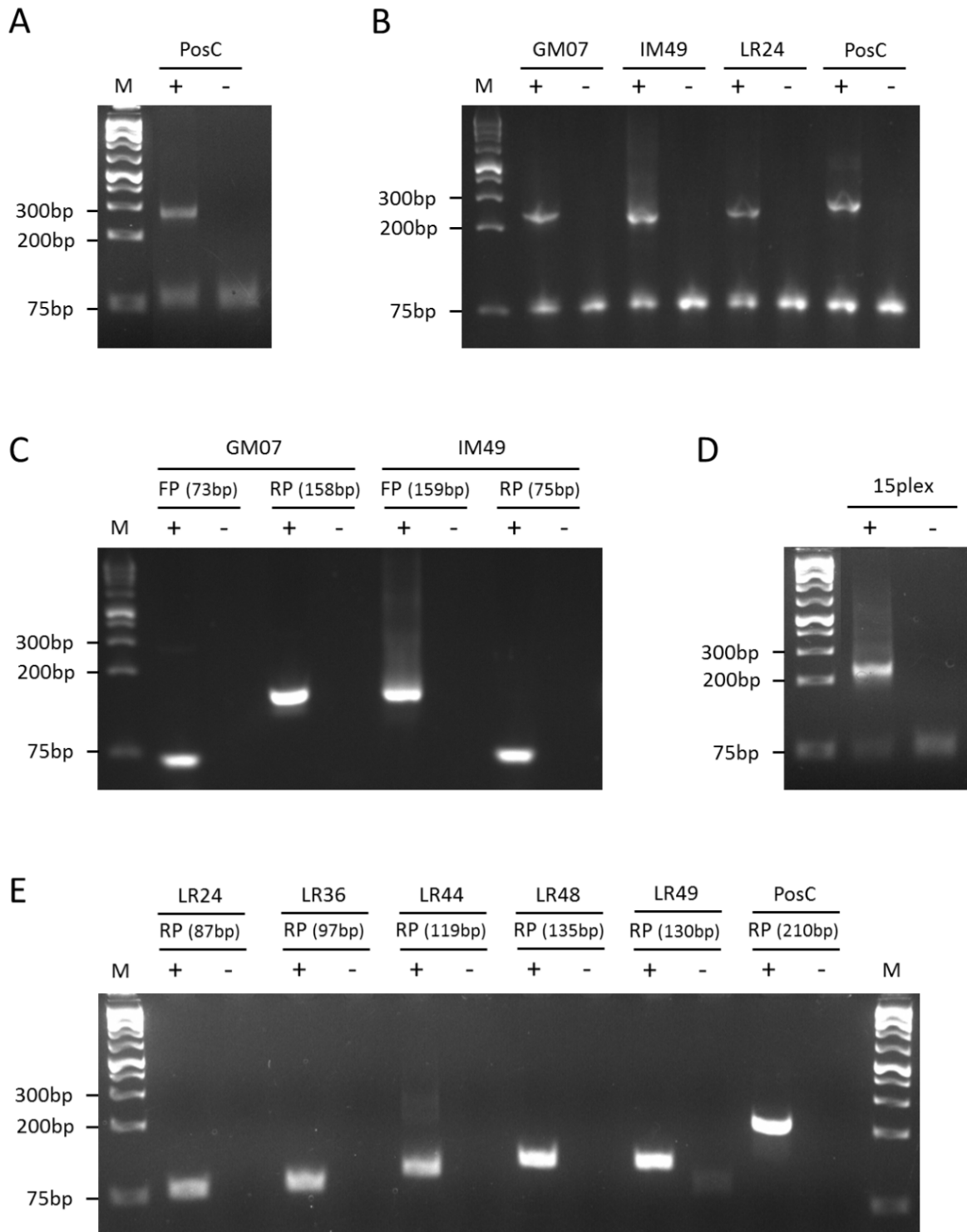


Figure 4.2: Amplification of marker loci using single molecule molecular inversion probes (smMIPs). **(A)** A positive control smMIP (PosC) (Hiatt *et al*, 2013) was used in a singleplex reaction to capture and amplify template DNA from K562 (+), alongside a template negative reaction (-). The expected smMIP amplicon size was 272bp. **(B)** Singleplex smMIPs were used to capture and amplify 17 microsatellite markers (Redford *et al*, 2018; three examples shown), with expected amplicon size of 240-270bp. **(C)** PCR verification of amplicons from B, showing examples from 2 markers using internal primers in both forward and reverse orientations (FP and RP). **(D)** 15 microsatellite markers were captured and amplified in one multiplexed-smMIP reaction. **(E)** PCR verification of amplicons from D, showing examples from 5 markers.

extracted from K562 (an MMRp chronic myeloid leukaemia cell line), HCT116 (an MMRd CRC cell line), and an MMRd CRC with a matched normal colorectal mucosa biopsy. Each sample was amplified using the 15plex smMIP reaction. The target read depth of sequencing was ≥ 5000 reads/marker/sample, requiring a minimum of 400,000 total reads (Section 2.10), assuming that the proportion of total reads generated would equal 0.75 x total read output, based on observations from Niedzicka *et al* (2016). Therefore, a MiSeq v2 Nano Kit (capacity of 1,000,000 reads, and the smallest kit available) was chosen. Amplicons from each sample were purified, quantified and diluted to 4nM, before pooling in equal volumes to give a final 4nM sequencing library. There was a 6-week delay between library preparation and MiSeq loading and the number of reads passing the MiSeq quality filter was only 301,367 (Table 4.1), much lower than anticipated, perhaps due to degradation of the library during the 6-week delay. Reads were aligned to reference genome hg19 and summarised in Marker Result tables that count the reads according to microsatellite length and SNP detected (Section 2.12.1). The proportion of reads passing filter aligned to the marker loci was 0.84, in agreement with the assumed 0.75 used in target read depth calculations. Reads were detected for each of the 15 markers, ranging from 706 to 12,553 (Table 4.2).

To expand the microsatellites analysed, in a separate analysis, smMIPs were designed for 9 additional markers from the original list of 120 (Redford *et al*, 2018), and redesigns for LR20 and IM66 were attempted. smMIPs for the *BRAF* V600 and *KRAS* G12 and G13 mutation-hotspots were also designed. Loci coordinates were compiled by Dr Harsh Sheth, I ran MIPgen, and subsequently Dr Sheth tested the smMIPs in singleplex and then multiplex based on protocols optimised in the first 17 markers. The smMIP for IM66 again failed to produce a visible amplicon, but smMIPs for LR20, the 9 additional microsatellites, and *BRAF* and *KRAS* were all successfully amplified in singleplex and multiplex. The same four samples, K562, HCT116, and the MMRd CRC with matched normal colorectal mucosa, were sequenced using the 12plex of additional markers, a 10pM loading concentration and a MiSeq v2 Nano Kit. In this sequencing run, a total of 1,444,882 reads were generated (Table 4.1), which was much higher than the previous run and kit capacity. At such high numbers of reads the 10pM loading concentration risked run failure due to over clustering of the MiSeq flow cell. Therefore, it was decided that DNA libraries should be prepared as near to MiSeq loading as possible to ensure library quality and sufficient read depth, and that 8pM should be loaded in future to avoid over clustering. In the 12plex sequencing run, all markers were covered, with read depth ranging from 6,388 to 49,795 reads/marker/amplicon (Table 4.2).

Parameters	15plex Library	12plex Library
Loading Concentration	8pM*	10pM
Reads passing filter	301,367 (95.0%)	1,444,882 (89.2%)
Reads aligned	253,692	1,272,430
Reads aligned/passing filter	0.84	0.88
Reads/marker/sample (mean±SD)	4,228 ±3,817	23,692 ±12,437
Base calls ≥Q30	78.00%	69.80%

*library was ~6 weeks old

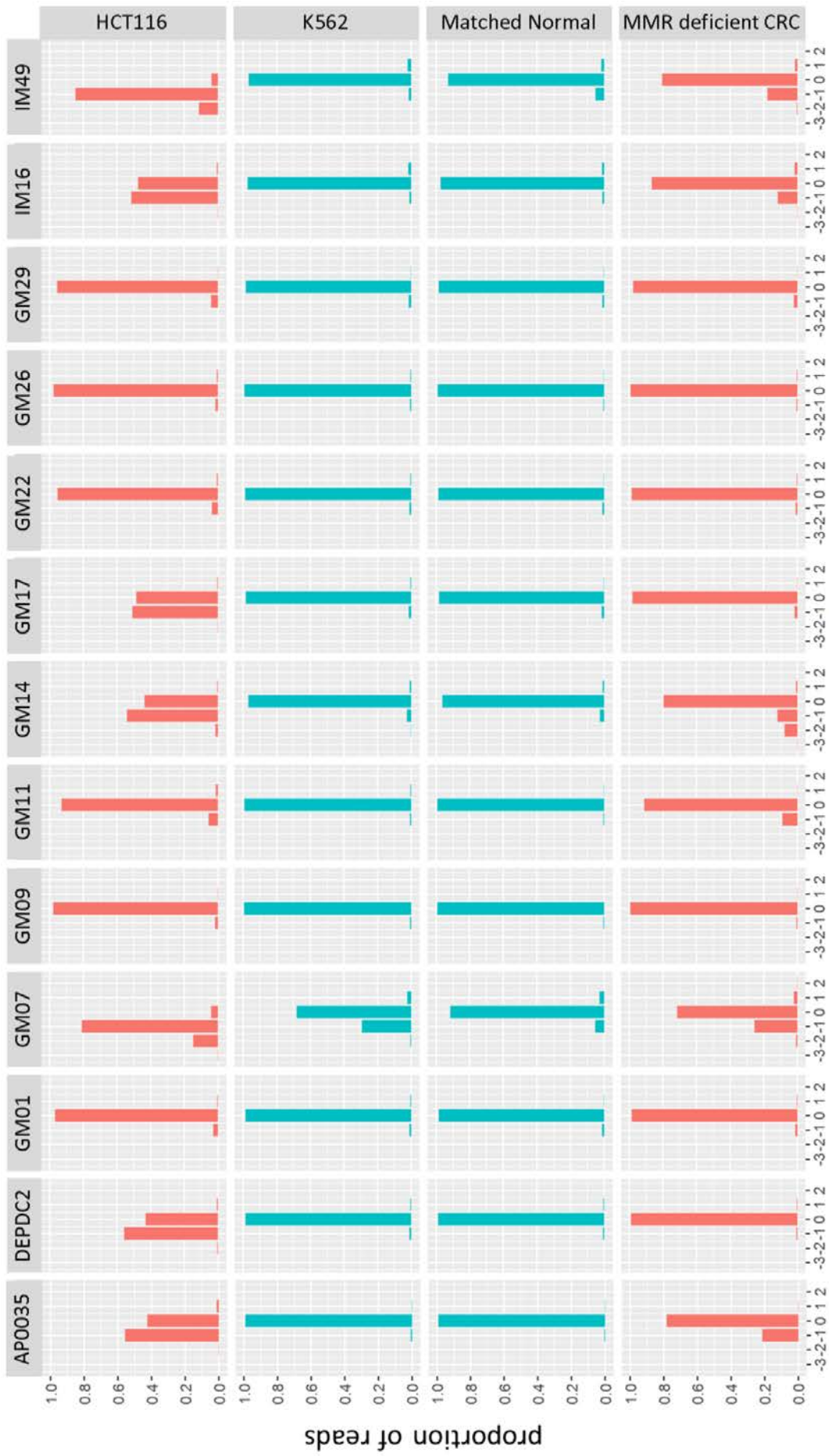
Table 4.1: Statistics from the two sequencing runs.

Marker	Marker Details	Reads Detected	Library
AP0035322	Microsatellite A(9)	1,538	15plex
<i>BRAF</i>	Hotspot p.V600	49,795	12plex
DEPDC2	Microsatellite G(8)	2,184	15plex
GM01	Microsatellite A(10)	35,515	12plex
GM07	Microsatellite A(11)	3,816	15plex
GM09	Microsatellite A(8)	3,557	15plex
GM11	Microsatellite A(9)	4,081	15plex
GM14	Microsatellite A(11)	6,684	15plex
GM17	Microsatellite A(9)	8,537	15plex
GM22	Microsatellite A(10)	46,054	12plex
GM26	Microsatellite A(10)	13,365	12plex
GM29	Microsatellite A(10)	22,352	12plex
IM16	Microsatellite A(9)	4,630	15plex
IM49	Microsatellite A(12)	706	15plex
<i>KRAS</i>	Hotspot p.G12,G13	31,391	12plex
LR10	Microsatellite A(10)	18,042	12plex
LR11	Microsatellite A(11)	3,795	15plex
LR17	Microsatellite A(10)	21,684	12plex
LR20	Microsatellite A(8)	6,388	12plex
LR24	Microsatellite A(9)	1,684	15plex
LR36	Microsatellite A(12)	4,170	15plex
LR40	Microsatellite A(9)	12,908	12plex
LR44	Microsatellite A(12)	1,359	15plex
LR46	Microsatellite A(8)	37,161	12plex
LR48	Microsatellite A(11)	12,553	15plex
LR49	Microsatellite A(7)	4,131	15plex
LR52	Microsatellite A(12)	23,455	12plex

Table 4.2: Mean reads detected per marker per sample from the two sequencing runs.

Marker Results tables from both sequencing runs were used to analyse the distribution of reads to different lengths of microsatellite in each marker. It was anticipated that the two MMRp samples (K562 and the matched normal colorectal mucosa) would show fewer variants in microsatellite length compared to the two MMRd samples (HCT116 and the MMRd CRC). The matched normal colorectal mucosa showed only low frequencies of length variants, likely due to the expected PCR and sequencing error (Figure 4.3). Similar observations were made for K562 except for -1 deletions in approximately one third of reads in markers GM07 and LR48 (Figure 4.3). This suggests that one allele of these markers had been mutated given that K562 is triploid (Klein *et al*, 1976). Mutation in 2 of 25 microsatellites is consistent with the MSI-low phenotype of some MMRp cancers (Halford *et al*, 2002; Laiho *et al*, 2002). HCT116 contained increased frequency of length variants in roughly 21-23/25 microsatellites compared to the MMRp samples, consistent with an MSI-high phenotype (Figure 4.3). For the majority of these markers, approximately one half of reads contained one length of microsatellite and the other half of reads contained another length, suggesting differential mutation in the two alleles of a clonal population. Markers that showed only slight increases in frequency of length variants could represent subclonal mutations. The MMRd CRC had increased frequency of length variants in 15/25 microsatellites, consistent with an MSI-high phenotype. However, reads were not evenly distributed between two different lengths as was seen in HCT116, which is consistent with the sample containing a mix of MMRd tumour cells and MMRp stromal cells (Figure 4.3 – spans 2 pages). For both MMRd samples, increases in length variants were due to deletions, as expected (Sia *et al*, 1997; Lu *et al*, 2013), which supports the use of deletion and not insertion frequency in the MSI classifier developed by Redford *et al* (2018).

Detection of allelic bias of microsatellite deletions requires a sample to be heterozygous at the associated SNP. Where >80% of reads were associated with the same SNP allele the sample was considered homozygous for the marker-associated SNP and therefore allelic bias of deletions could not be assessed, for example K562 at LR44_SNP1 (Figure 4.4). Where the multiple alleles were detected at the SNP locus and $\geq 20\%$ of reads were associated with the minor allele the sample was considered heterozygous for the marker-associated SNP, and allelic bias of deletions could be assessed. For example, the matched normal sample showed a similar proportion of reads assigned to C and T alleles at both WT (0) and deletion (-1) lengths of microsatellite in marker LR44, suggesting no allelic bias (Figure 4.4). However, for the MMRd CRC there is a greater proportion of reads



length of microsatellite relative to wild type

Figure 4.3: The distribution of reads to different lengths of microsatellite in the 25 markers in 4 samples. Red: MMR deficient. Blue: MMR proficient.

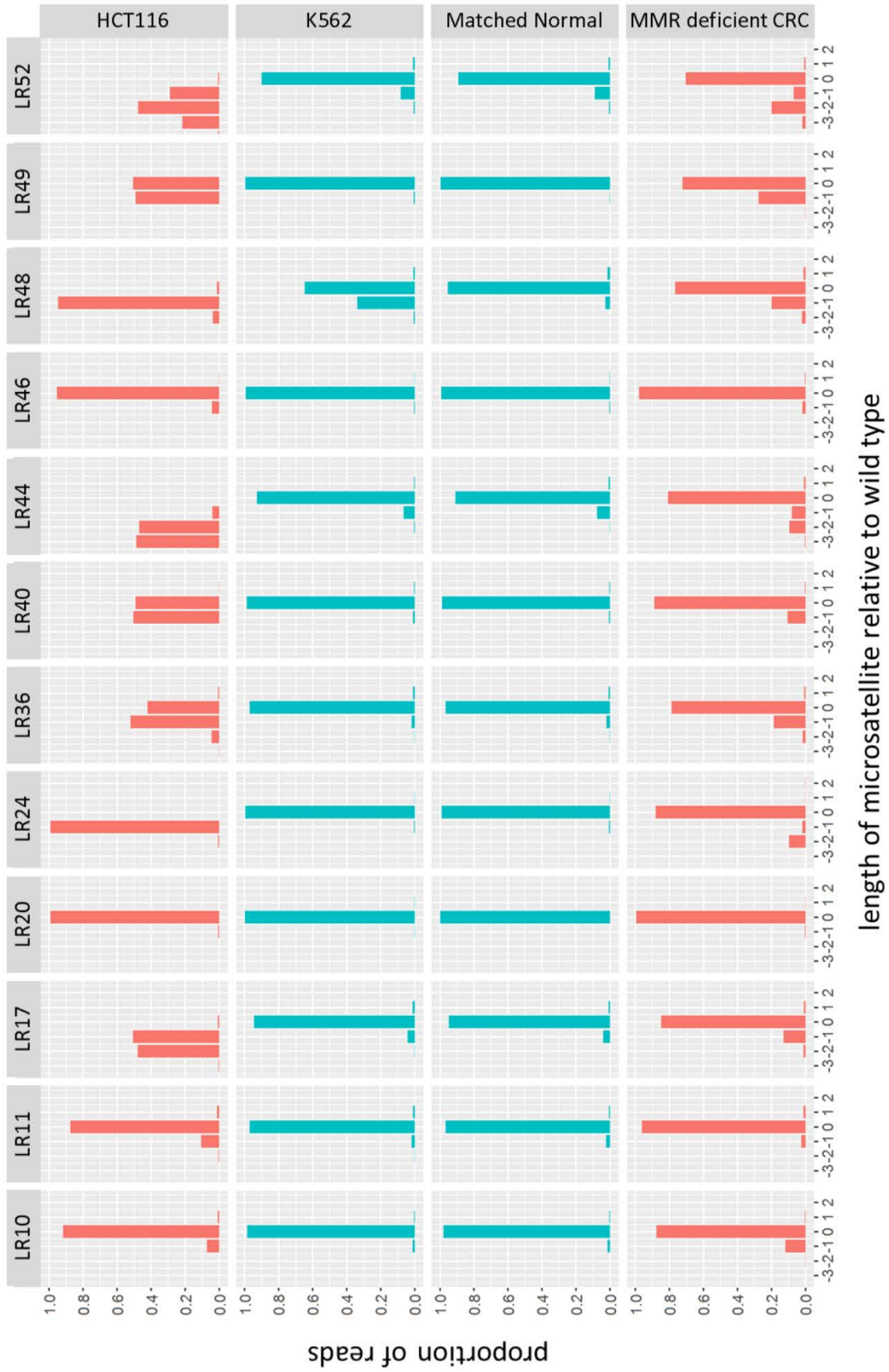


Figure 4.3: The distribution of reads to different lengths of microsatellite in the 25 markers in 4 samples. Red: MMR deficient. Blue: MMR proficient.

containing a -2 deletion assigned to the C allele than there are reads containing a -2 deletion assigned to the T allele, evidence of allelic bias of this deletion (Figure 4.4). The statistical significance of read count distribution to the different alleles was calculated by constructing a two-by-two table for the count of reads containing deletions versus the count of reads containing WT microsatellite length, as distributed between alleles, and using Fisher's Exact test. In the MMRd samples, the majority of markers (where the sample was heterozygous at the SNP locus) showed a significant difference in the deletion frequency between alleles, which was not the case for the MMRp samples (Table 4.3).

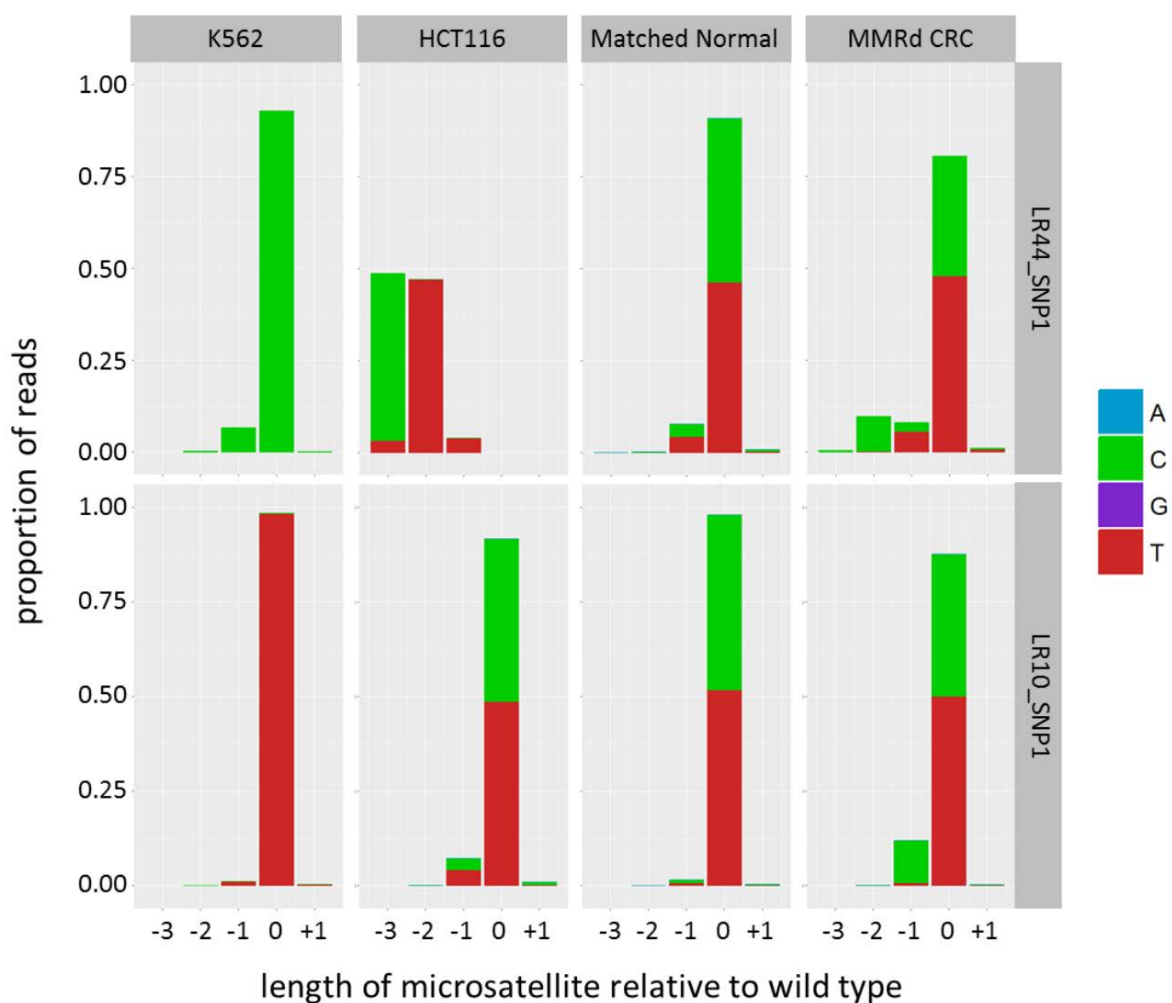


Figure 4.4: Distribution of reads by microsatellite length and allele. Examples are shown for two markers with associated SNPs, LR44_SNP1 and LR10_SNP1, in four samples: K562, a mismatch repair proficient (MMRp) cell line; HCT116, a mismatch repair deficient (MMRd) cell line; matched normal, a biopsy of MMRp colorectal mucosa from the same patient as the MMRd CRC; and a MMRd CRC.

Marker_SNP	K562	HCT116	Matched Normal	MMRd CRC
AP0035_SNP1	NA	NA	0.755	0.495
DEPDC2_SNP1	0.763	0.000	NA	NA
GM01_SNP1	0.373	NA	0.404	0.060
GM07_SNP1	NA	0.000	NA	NA
GM09_SNP1	NA	0.075	NA	NA
GM11_SNP1	NA	0.000	NA	NA
GM11_SNP2	NA	NA	NA	NA
GM14_SNP1	NA	NA	NA	NA
GM17_SNP1	NA	NA	NA	NA
GM22_SNP1	NA	NA	0.636	0.018
GM26_SNP1	NA	0.000	0.229	0.316
GM29_SNP1	NA	0.000	0.449	0.000
IM16_SNP1	0.013	NA	NA	NA
IM16_SNP2	NA	NA	NA	NA
IM16_SNP3	NA	NA	NA	NA
IM49_SNP1	NA	NA	NA	NA
LR10_SNP1	NA	0.001	0.042	0.000
LR10_SNP2	NA	0.001	NA	NA
LR11_SNP1	NA	0.095	NA	NA
LR11_SNP2	NA	0.062	NA	NA
LR17_SNP1	NA	NA	NA	NA
LR17_SNP2	NA	NA	NA	NA
LR17_SNP3	NA	NA	NA	NA
LR17_SNP4	NA	NA	NA	NA
LR20_SNP1	0.387	NA	0.737	0.670
LR24_SNP1	1.000	1.000*	NA	NA
LR36_SNP1	0.764	NA	NA	NA
LR40_SNP1	NA	NA	0.452	0.000
LR44_SNP1	NA	1.000*	0.539	0.000
LR44_SNP2	NA	1.000*	0.133	0.000
LR46_SNP1	NA	NA	0.828	0.000
LR48_SNP1	NA	NA	0.264	0.550
LR49_SNP1	0.774	NA	NA	NA
LR52_SNP1	0.012	NA	0.000	0.000
Sample Summary	Homozygous: 26 Heterozygous: 8 p < 0.05: 2	Homozygous: 21 Heterozygous: 13 p < 0.05: 7	Homozygous: 21 Heterozygous: 13 p < 0.05: 2	Homozygous: 21 Heterozygous: 13 p < 0.05: 9

Table 4.3: Detection of allelic bias of deletions in the microsatellite markers. Markers where the associated SNP is homozygous are denoted by NA. Markers where the associated SNP is heterozygous were assessed for statistical significance of the allelic bias of deletions in the microsatellite, using Fisher’s Exact test: p values are presented in the table and the data is summarised for each sample.

**Note: for HCT116 markers LR24 and LR44 there are very few or no WT reads (see Figure 4.3 for LR44 as an example) resulting in non-significant tests for allelic bias.*

Marker Result tables were also used to detect substitution mutations at the *BRAF* and *KRAS* mutation-hotspots. For each locus, the bases detected across the four samples were determined and then, for each sample, the proportion of reads containing each base was plotted (Figure 4.5). For *BRAF*, the matched normal colorectal mucosa, K562, and HCT116 all contained the WT adenine in nearly 100% of reads, with any variants likely due to error (present in <1% of reads). However, the MMRd CRC had 11% of reads assigned to thymine (Figure 4.5A). The observed A>T substitution represents the *BRAF* V600E mutation, suggesting that this MMRd CRC is of sporadic origin. For *KRAS*, there were four loci sequenced that are associated with substitution mutations in the G12 and G13 codons. At these loci, all samples contain the WT base in near to 100% of reads, except for HCT116 at chr12 25398281, where 52% of reads are assigned to thymine rather than the WT cytosine (Figure 4.5B). This C>T substitution is associated with *KRAS* G13D, a known mutation in HCT116 (Yun *et al*, 2009).

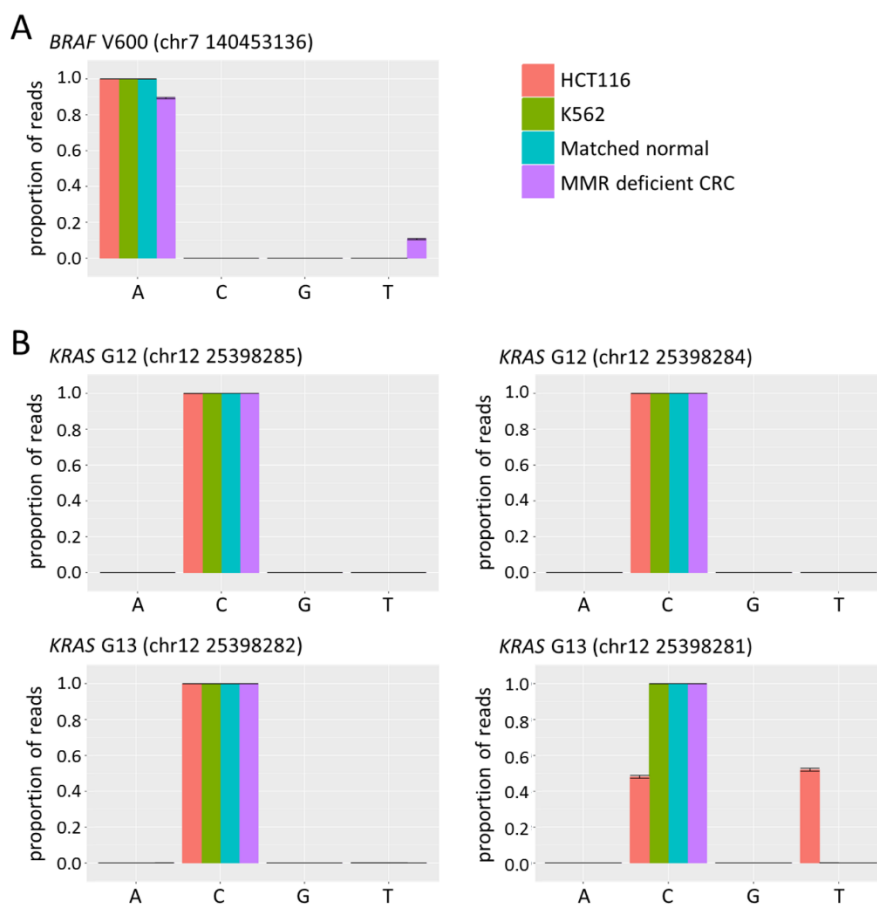


Figure 4.5: Detection of *RAS/RAF* mutations. The proportion of reads containing different bases at mutation hotspots (chromosomes and coordinates using reference genome hg 19) are shown for **(A)** *BRAF* and **(B)** *KRAS*. The A>T substitution at chr7 140453136 in the MMR deficient CRC represents *BRAF* V600E. The C>T substitution at chr12 25398281 in HCT116 represents *KRAS* G13D.

4.5. Training the MSI Classifier

The MSI classifier requires training in a cohort of samples to determine the probability that observations of microsatellite deletion frequencies and the allelic bias of deletions belong to either an MSI-high or MSS CRC population. Residual DNA samples from the work of Redford *et al* (2018) were compiled into a training cohort of 51 MSI-high and 47 MSS CRCs (*a priori* classification by MSI Analysis System, Promega). HCT116 and H9, an embryonic stem cell line, were included as MSI-high and MSS controls, giving a total of 100 samples. The training cohort was amplified using a multiplex of the 27 smMIPs, comprising the 25 short MNR markers, and *BRAF* and *KRAS* mutation hotspots. Amplicons were sequenced to a mean read depth (\pm SD) of 3,719 \pm 3,149 reads/marker/sample with 75.3% of base-calls \geq Q30. Read depth was lower than the target >5,000reads/marker/sample. Previously, MiSeq v2 kits were used and so it was assumed that v3 kits may need a higher DNA library loading concentration; subsequent sequencing used 8pM for v2 kits and 12pM for v3 kits and achieved target read depths. Unfortunately, no reads were detected for marker AP0035 in 87/98 samples. AP0035 was one of the less accurate markers in the work of Redford *et al* (2018) and, with the addition of 9 new markers not analysed by Redford *et al*, its inclusion was not necessary. Therefore, AP0035 was excluded from all further analyses.

The relative frequency of deletions in the microsatellite was determined for each marker across the 47 MSS CRCs and, from this empirical distribution, a threshold was set at the 95th percentile for each marker individually, as different markers have different, intrinsic error rates. Therefore, assuming the empirical distributions to be representative of the MSS CRC population, there would be 95% probability of an MSS CRC having a deletion frequency below the threshold, and 5% probability of an MSS CRC having a deletion frequency above the threshold. The probability of an MSI-high CRC having a deletion frequency above or below these thresholds could then be determined, again assuming the empirical distributions from the 51 MSI-high CRCs sequenced in the training cohort to be representative of the MSI-high CRC population. Considering two example markers, GM11 and IM49, the proportions of MSI-high samples falling above the threshold were 44/51 and 45/51, respectively. Therefore, for these two markers, the probabilities of observing a deletion frequency above the threshold in an MSI-high CRC would be 86.3% and 88.2% for GM11 and IM49, respectively. Conversely the probabilities of observing a deletion frequency below the threshold in an MSI-high CRC would be 13.7% and 11.8% for GM11 and IM49, respectively (Figure 4.6).

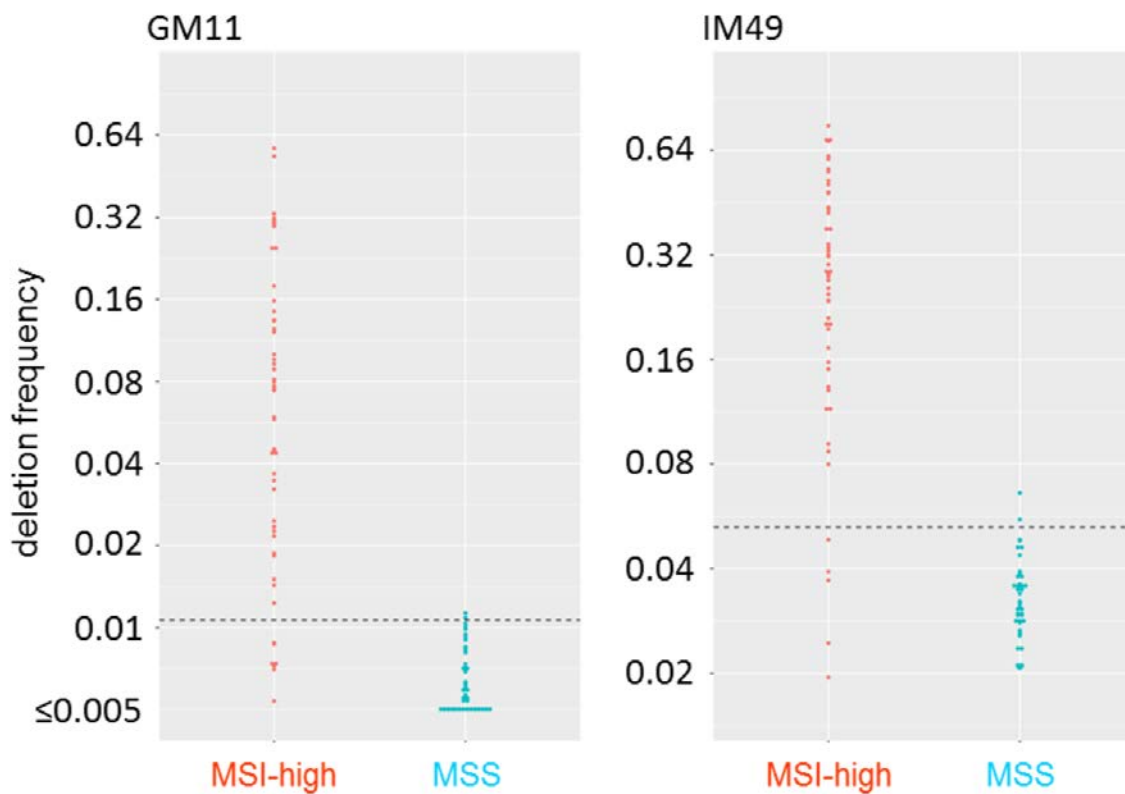


Figure 4.6: Distribution of training cohort samples by the relative frequency of microsatellite deletions. The proportion of reads containing a deletion in microsatellite length (deletion frequency) was determined for both MSI-high and MSS CRCs in the training cohort; the 0.95 quantile of deletion frequencies detected in the MSS samples was used as a threshold (dotted lines) to dichotomise the distributions. The proportion of MSI-high CRCs above and below the threshold can be used to calculate probabilities that an observed deletion frequency belongs to a MSI-high CRC population.

It is also possible to assess deletion frequency in different alleles when the sample is heterozygous at the neighbouring SNP. If allelic bias of deletion is present, it gives additional confidence that it is a true mutation rather than the result of error, which should affect both alleles equally. Assessment of allelic bias was possible in >30% of samples in all of the 24 microsatellite markers (Table 4.4). To establish the significance of allelic bias, two-by-two contingency tables were constructed, distributing reads according to length of microsatellite detected (deletion versus WT) and according to the allele detected, and Fisher's Exact tests were performed. A threshold was set at $p = 0.05$, and for each marker the probability of an observation from an MSI-high or MSS sample falling above or below this threshold was determined from the empirical distributions of p values from the 47 MSS and the 51 MSI-high CRCs of the training cohort. Again using GM11 and IM49 as example markers, for the MSI-high samples 23/28 and 19/26 samples fall below the threshold, and for MSS samples

Marker_SNP	Relative Frequency
DEPDC2_SNP1	0.316
GM01_SNP1	0.398
GM07_SNP1	0.541
GM09_SNP1	0.439
GM11_SNP1	0.510
GM11_SNP2	0.122
GM14_SNP1	0.449
GM17_SNP1	0.490
GM22_SNP1	0.367
GM26_SNP1	0.459
GM29_SNP1	0.378
IM16_SNP1	0.469
IM16_SNP2	0.245
IM16_SNP3	0.000
IM49_SNP1	0.490
LR10_SNP1	0.429
LR10_SNP2	0.418
LR11_SNP1	0.418
LR11_SNP2	0.408
LR17_SNP1	0.000
LR17_SNP2	0.000
LR17_SNP3	0.490
LR17_SNP4	0.490
LR20_SNP1	0.520
LR24_SNP1	0.469
LR36_SNP1	0.469
LR40_SNP1	0.194
LR44_SNP1	0.439
LR44_SNP2	0.439
LR46_SNP1	0.480
LR48_SNP1	0.327
LR49_SNP1	0.439
LR52_SNP1	0.510

Table 4.4: The proportion of samples in which allelic bias can be assessed for each marker. A total of 98 samples were analysed in the training cohort. If a sample was heterozygous for a marker-associated SNP the allelic bias of microsatellite deletions could be assessed. The table shows the proportion of the 98 samples that were heterozygous for each SNP associated with each of the markers.

Note: AP0035 has been excluded due to a lack of reads in Marker Result tables for 87/98 samples.

4/22 and 3/22 samples fall below the threshold, respectively (Figure 4.7). Therefore, for GM11, the probabilities of observing a p value for allelic bias below the 0.05 threshold would be 82.1% and 18.2% for an MSI-high and an MSS sample, respectively. Conversely, the probabilities of observing a p value for allelic bias above the 0.05 threshold in GM11 would be 17.9% and 81.8% for an MSI-high and an MSS sample, respectively. For IM49, the probabilities of observing a p value for allelic bias below the 0.05 threshold would be 73.1% and 13.6% for an MSI-high and an MSS sample, respectively. Conversely, the probabilities of observing a p value for allelic bias above the 0.05 threshold in IM49 would be 26.9% and 86.4% for an MSI-high and an MSS sample, respectively.

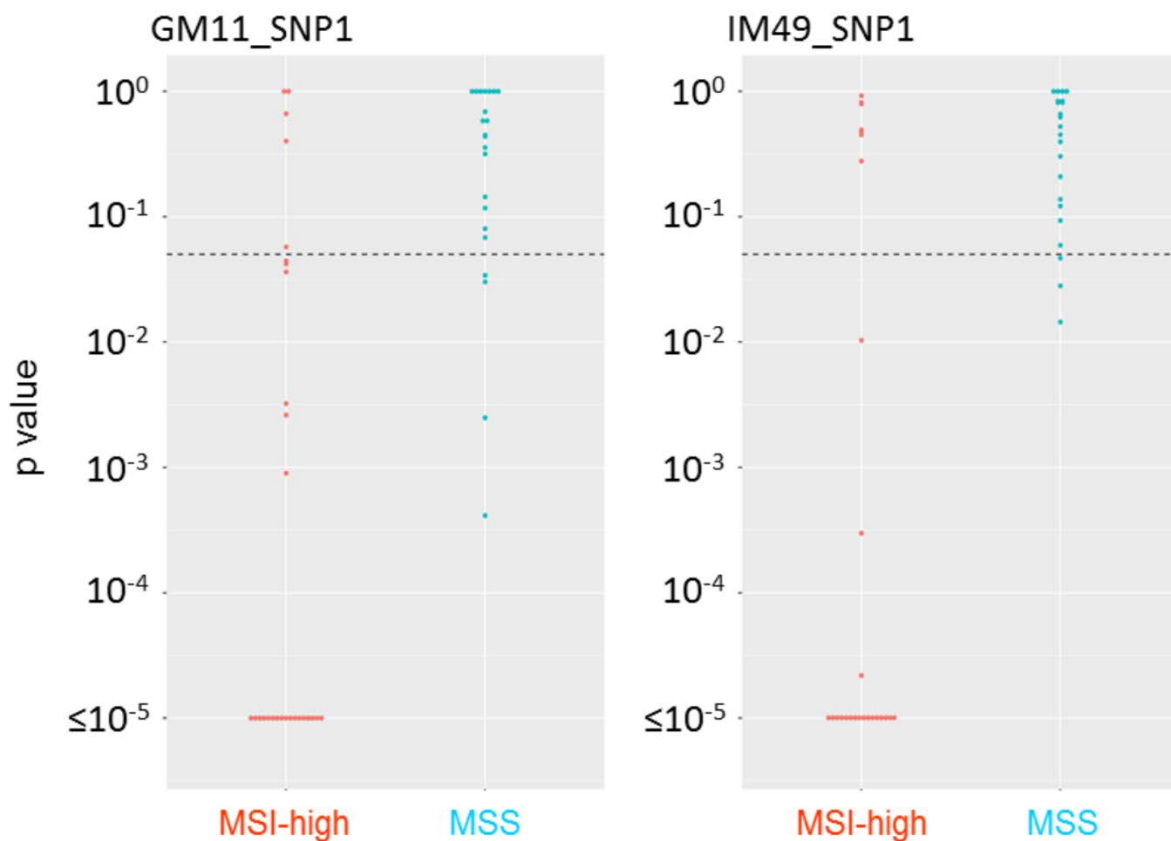


Figure 4.7: Distribution of samples relative to classifier thresholds. (A) The proportion of reads containing a deletion in microsatellite length (deletion frequency) was determined for both MSI-high and MSS CRCs in the training cohort; the 0.95 quantile of deletion frequencies detected in the MSS samples was used as a threshold (dotted lines). **(B)** The significance of allelic bias of deletions was represented by the p value from Fisher’s Exact tests for the count of reads distributed to different microsatellite lengths across alleles, with a threshold set at $p = 0.05$. MSI status (MSI-high or MSS) was determined *a priori* by MSI Analysis System (Promega).

Having trained the MSI classifier, for any marker an observed deletion frequency or observed p value of allelic bias could be converted into two probabilities: one probability of the observation O_i belonging to a MSI-high population $p(O_i|MSI)$, and a second probability of the observation O_i belonging to a MSS population $p(O_i|MSS)$. Where allelic bias could not be computed (e.g. the sample was homozygous at the associated SNP) the probabilities associated with allelic bias for that marker were set to 1. Also, where the deletion frequency was below the threshold for a marker, indicating no evidence for deletion at the microsatellite, the probabilities associated with allelic bias for that marker were set to 1. Relative probabilities from each marker could be condensed into one term by multiplication:

$$\frac{p(O|MSI)}{p(O|MSS)} = \prod_{i=1}^N \frac{p(O_i|MSI)}{p(O_i|MSS)}$$

A sample score (S) could then be calculated:

$$S = \log_{10} \frac{p(MSI)}{p(MSS)} \cdot \frac{p(O|MSI)}{p(O|MSS)}$$

where $p(MSI)$ and $p(MSS)$ are set to 0.15 and 0.85 as the *a priori* probability of sample being MSI-high or MSS, respectively. Samples with $S > 0$ are classified as MSI-high, indicating MMR deficiency (Redford *et al*, 2018).

The trained MSI classifier subsequently typed the training cohort with 100% sensitivity (95% CIs: 93.0-100.0%) and 100% specificity (95% CIs: 92.5-100.0%) (Figure 4.8: left-hand panel). In addition, control samples, HCT116 and H9, were correctly classified, with scores of 38.04 and -20.6, respectively. To determine if there was redundancy in the panel of microsatellites, the most discriminatory markers were defined by backward stepwise selection (performed by Dr Mauro Santibanez-Koref). Reducing the number of markers analysed would reduce assay costs (see Section 4.10) by increasing the number of samples that could be tested per sequencing run. In brief, the full panel of 24 markers was reduced by removing the least discriminatory marker at each step, until the classifier was no longer 100% accurate. Loss of accuracy occurred from 6 to 5 markers. From this 5 marker panel, each of the 19 markers that had been removed in previous steps was added individually to see which 6 marker panel gave the best separation between MSI-high and MSS samples. The most discriminatory panel comprised: GM07, GM11, GM14, LR36, LR44, and LR52. Naturally, classification by the 6 marker panel also achieved 100% accuracy (Figure 4.8: right-hand panel). Equivalent accuracy from one quarter of the markers shows redundancy in the microsatellites analysed.

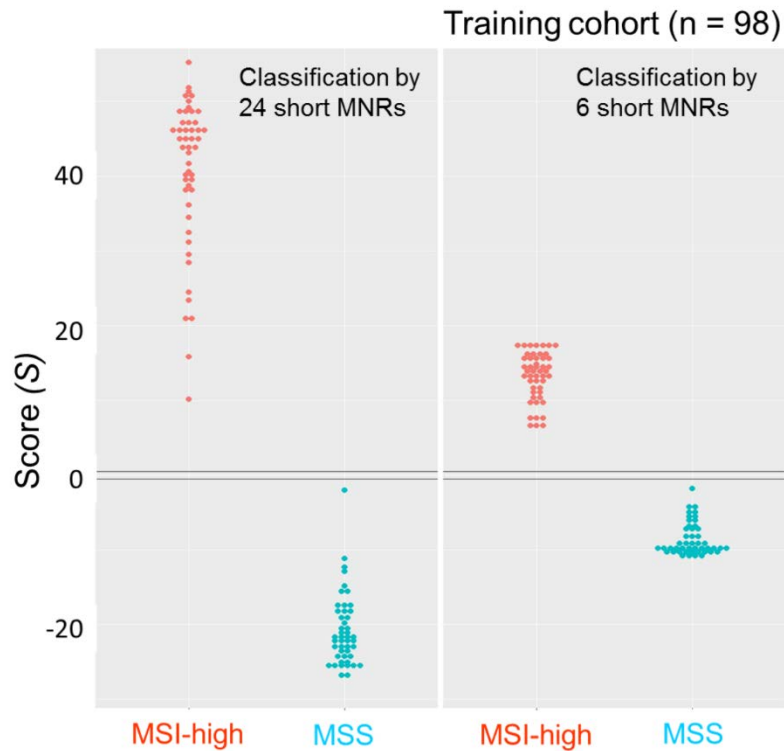


Figure 4.8: Self-classification of the training cohort. 98 CRCs with known MSI status were used to train an MSI classifier that analyses the relative frequency and allelic bias of deletions in a panel of short mononucleotide repeats (MNRs), according to the method described by Redford *et al* (2018). Classifier score (S) > 0 is MSI-high, and $S < 0$ is MSS. Self-classification of the training cohort achieved 100% sensitivity and 100% specificity, relative to typing by the Promega MSI Analysis System (colour), using either 24 markers (left-hand panel) or 6 of the most discriminatory markers (right-hand panel).

4.6. MSI Classification is Accurate and Reproducible

To balance the reads from each smMIP, the mean number of reads detected for each marker in the training cohort was calculated relative to the overall mean read depth per marker per sample, and a new multiplex pool of 26 smMIPs was made with the concentration of each smMIP inversely proportional to the relative number of reads detected for that marker (Appendix I). Using this read-balanced smMIP multiplex, 50 MSI-high CRCs and 49 MSS CRCs were then amplified as an independent validation cohort, with the assistance of Christine Hayes, and sequenced to a mean read depth (\pm SD) of $7,320 \pm 4,192$ reads/marker/sample with 57.2% of base-calls \geq Q30. Sequencing using the read-balanced smMIP multiplex had a much lower coefficient of variation in read depth between markers (35%) compared to the training cohort (68%). In the validation cohort, the MSI classifier again achieved 100% sensitivity (95% CIs: 92.9-100.0%) and 100% specificity (95% CIs: 92.8-100.0%) relative to typing by MSI Analysis System (Promega), using all 24 short MNRs (Figure 4.9: left-hand

panel). Furthermore, classification using the 6 most discriminatory short MNRs, as identified from the training cohort, was also 100% accurate (Figure 4.9: right-hand panel).

To assess reproducibility of the assay and classifier, 16 MSI-high and 16 MSS CRCs from the validation cohort were amplified a second time using a freshly prepared, read-balanced smMIP pool, again targeting the 24 short MNRs, and *BRAF* and *KRAS* mutation hotspots. These amplicons were sequenced to a mean read depth (\pm SD) of 5,408 \pm 2,160 reads/marker/sample with 85.4% of base-calls \geq Q30. Classification was 100% concordant with previous results and classifier scores were strongly correlated between sample repeats ($\beta = 0.97$, $p < 10^{-16}$, $R^2 = 0.97$).



Figure 4.9: MSI classifier validation. 99 CRCs, independent from the training cohort, with known MSI status were used to validate the MSI classifier. Classifier score (S) > 0 is MSI-high, and $S < 0$ is MSS. Classification of the validation cohort achieved 100% sensitivity and 100% specificity, relative to typing by the Promega MSI Analysis System (colour), using either 24 markers (left-hand panel) or 6 of the most discriminatory markers (right-hand panel).

4.7. MSI Classification is Robust to Low MSI-high Content

To assess the lower limit of detection (LLoD), defined here as the lowest proportion of MSI-high DNA within total template DNA at which a sample is classified as MSI-high, a DNA-

mixture series of 0.78-100% MSI-high DNA content (two-fold increments) was created in triplicate, by mixing HCT116 MSI-high DNA into control MSS DNA extracted from peripheral blood leukocytes (PBLs, Section 2.6). This triplicate series and control MSS DNAs were amplified using a read-balanced, 24 MNR smMIP pool. Amplicons were sequenced to a mean read depth (\pm SD) of $4,763 \pm 1,288$ reads/marker/sample with 84.7% of base-calls \geq Q30.

Increasing the MSI-high DNA content of the template DNA increased the proportion of reads containing insertion-deletion mutations in the microsatellite (Figure 4.10A). To confirm that the mixture series was accurate, I compared the observed proportion of reads containing variants in microsatellite length with the expected result. The expected proportion could be calculated by the following equation:

$$\text{propMUTreads}|P_{MSI} = \text{propMUTreads}_{MSS} + P_{MSI} * (\text{propMUTreads}_{MSI} - \text{propMUTreads}_{MSS})$$

where P_{MSI} is the proportion of MSI-high DNA content within the sample mixture, $\text{propMUTreads}_{MSS}$ is the proportion of mutant reads observed in sequencing pure MSS DNA, and $\text{propMUTreads}_{MSI}$ is the proportion of mutant reads observed in sequencing pure MSI-high DNA. The observed and the expected proportions were strongly correlated ($\beta = 1.009$, $p = 2 \times 10^{-16}$, $R^2 = 0.996$, Figure 4.10B), giving confidence in the accuracy of DNA mixing.

MSI classification of the DNA-mixture series was accurate from 3.13% or more MSI-high content in each replicate sample series (Figure 4.10C), approximating the LLoD to 3%. According to Jennings *et al* (2017), to specify a LLoD with confidence, 59 samples of the variant allele frequency (VAF) of interest (which, in this case, would be the MSI-high content of the sample DNA) should be tested. However, they recognise the difficulty in collecting 59 independent samples of equal VAF and suggested artificial samples could be used. To simulate additional samples a method of randomly mixing reads from two samples was designed by Dr Santibanez-Koref, creating *in silico* samples. 27 simulated sample series, again ranging from 0.78% to 100% MSI-high content (two-fold increments), were generated using reads from pure MSI-high and pure MSS samples, with reads/marker equal to the reads/marker of the MSI-high sample, and each sample was scored. Simulated sample scores were closely associated with the corresponding score from the mixing of template DNAs (Figure 4.10C), suggesting that *in silico* read mixing is a valid method for simulating additional samples. To simulate a large number of independent samples to analyse the robustness of the MSI classifier to low MSI-high content, reads from 50 of the MSI-high and 48 of the MSS CRCs from the validation cohort were mixed, generating 2400 simulated

sample series. These were scored and the proportion of samples classified as MSI-high calculated. Approximately 95% of simulated samples were correctly classified as MSI-high using 25% reads from an MSI-high CRC or more (Figure 4.10D). Due to the heterogeneous mixture of stromal and tumour cells in MSI-high CRCs, this cannot be used to estimate a LLoD, but supports the conclusion that the MSI classifier is robust to low MSI-high content.

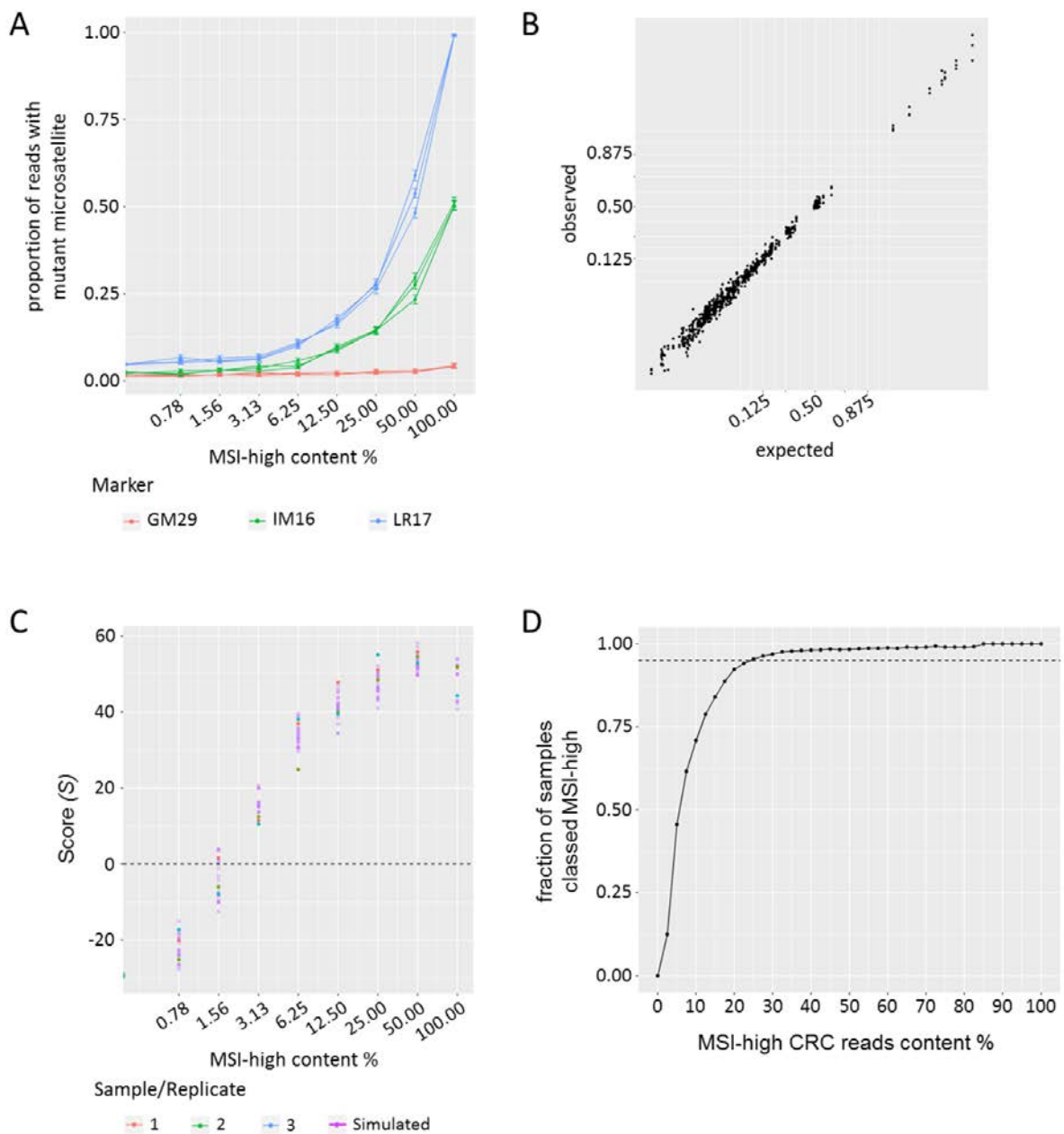


Figure 4.10: The robustness of the MSI classifier to low MSI-high content. (A) As the MSI-high content of a sample increases so does the frequency of reads containing variants in microsatellite length, which correlates closely with expected results **(B)**. **(C)** Scoring of samples of varying MSI-high content by mixing reads or template DNA from MSI-high and MSS controls generates comparable results. By both methods, the assay lower limit of detection can be approximated to 3% MSI-high content. **(D)** Samples simulated from read mixing of MSI-high and MSS CRCs shows the MSI classifier is robust (95% of samples correctly classified as MSI-high) down to 25% MSI-high CRC read content.

To directly compare the assay LLoD with that of FLA, replicates ranging from 1.56-12.5% MSI-high DNA content were independently classified using the MSI Analysis System (Promega), with the observer blinded to both sample content and experimental purpose. FLA reliably detected 6.25% MSI-high DNA content (Table 4.5).

MSI-high content (%)	Diagnosis	Unstable Markers	Uncertain Markers
1.56	MSS	0/5	0/5
1.56	MSS	0/5	0/5
1.56	MSS	0/5	0/5
3.13	MSI-high	3/5	5/5
3.13	MSI-high	2/5	5/5
3.13	MSI-high	2/5	5/5
6.25	MSI-high	5/5	2/5
6.25	MSI-high	5/5	0/5
6.25	MSI-high	5/5	0/5
12.5	MSI-high	5/5	0/5
12.5	MSI-high	5/5	0/5
12.5	MSI-high	5/5	0/5

Table 4.5: Microsatellite instability classification by fragment length analysis of DNA-mixtures of varying MSI-high DNA content. A series of samples with varying MSI-high DNA content tested by the smMIP-based MSI assay were also analysed using the MSI Analysis System (Promega). Fragment length analysis classified samples as MSI-high which contained $\geq 3.13\%$ MSI-high DNA. However, the pathologist was uncertain of the status of all 5 markers. Therefore, confident classification as MSI-high was only achieved in samples with $\geq 6.25\%$ MSI-high DNA content.

4.8. MSI Classification is Reliable from sequencing 75 Molecules per Marker

To establish the lowest quantity of template DNA required for accurate smMIP-based classification, 2-fold dilution series of 9 DNA samples, comprising 3 cell lines (HCT116, K562 and H9), 3 MMRd CRCs and 3 MMRp CRCs, were generated. CRC samples were selected based on availability of residual DNA. 0.78-100ng of each sample was amplified using a read-balanced 24 MNR smMIP pool. Production of smMIP amplicons of the expected 240-270bp was visually inspected using 3% agarose gel electrophoresis. smMIP amplicons were deemed visible between 3.13-100ng of template DNA across all samples (Figure 4.11) and so these reactions were sequenced to a mean read depth (\pm SD) of 243,073 \pm 64,485 reads/sample with 82.8% of base-calls \geq Q30. I have quoted read depth in units of reads/sample rather than reads/marker/sample as the proportion of reads aligned to the marker loci was

correlated with the input quantity of template DNA, with an increase in aligned reads as input quantity increased (Figure 4.12A). Due to the low quantity of amplicons from some of these samples, a template negative was also sequenced. For the template negative sample 127,756 total read pairs were detected, but only 152 reads (0.12%) were aligned to a marker, consistent with low frequency index mis-assignment observed on Illumina platforms (Illumina Inc., 2017).

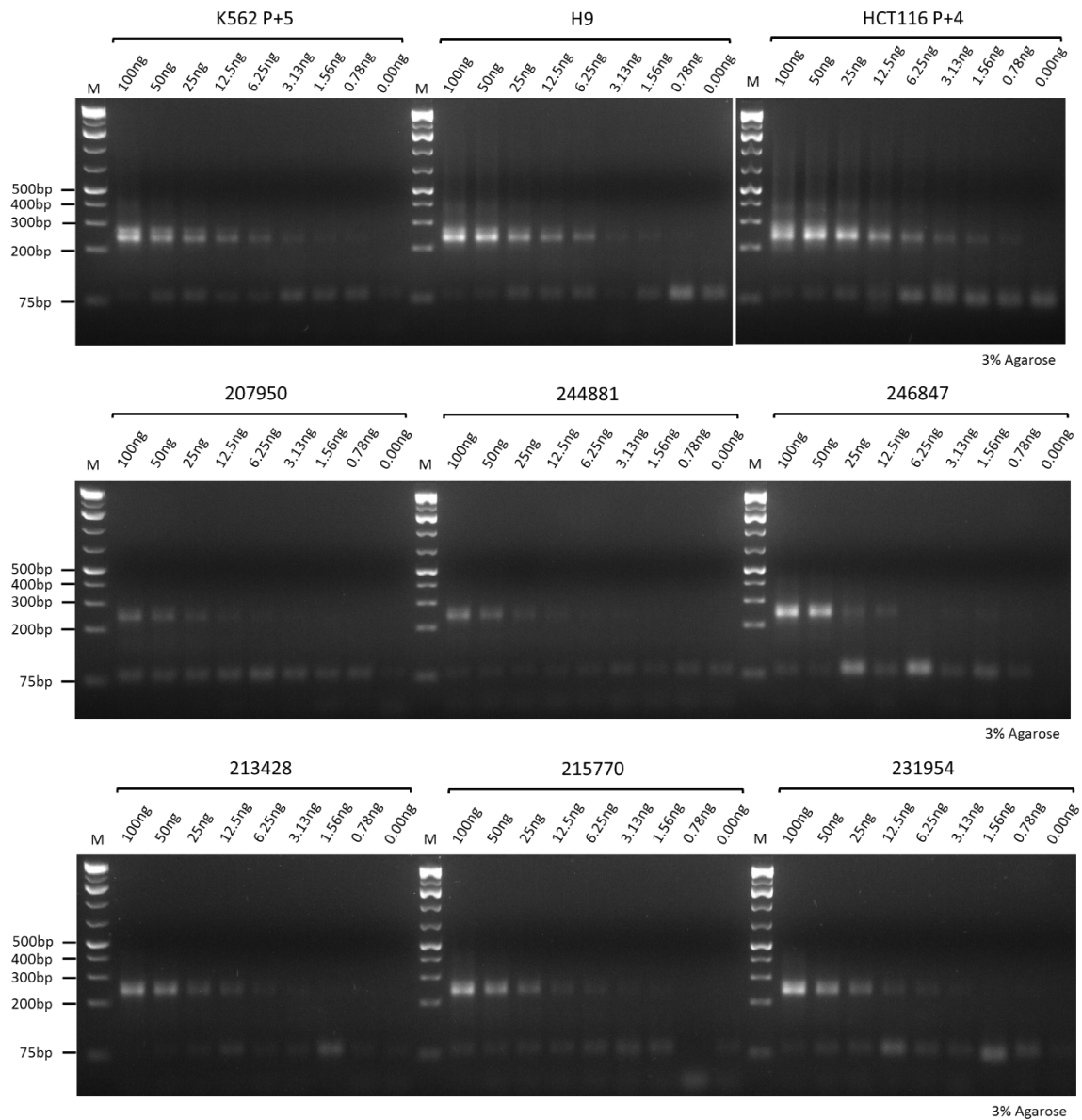


Figure 4.11: Visualisation of smMIP amplicons. Serial 2-fold dilutions of 9 samples were created, including 3 fresh sources (HCT116, H9, and K562) and 6 FFPE tissues (MMRd CRCs 207950, 244881 and 246847, and MMRp CRCs 213428, 215770 and 231954). Equal input volumes of the dilution series, such that template DNA ranged from 0.78-100ng for each sample, were amplified using the smMIP-based MSI assay. Primer dimers are visible at approximately 80bp.

To test the accuracy of the dilution series, the number of template molecules sequenced, as measured by the number of molecular barcodes (MBs) detected, was compared to the input quantity of template DNA for the 9 samples. The number of molecules sequenced and input quantity were closely correlated in each of the 9 samples ($\beta = 0.84-0.96$, $p < 10^{-3}$, $R^2 = 0.986-0.997$, Figure 4.12B), giving confidence in the dilution series. The effect of decreasing the quantity of template DNA on the detection of variants in microsatellite length was roughly assessed by looking at the absolute change in the frequency of reads containing length variants in samples where 3.13-50ng of template DNA was used, with change measured relative to results from 100ng of template DNA (Figure 4.12C). Notably the median, interquartile range, and total range of absolute change, all increased across the MSI-high samples as the quantity of template DNA decreased. However, no such effect was evident for the MSS samples and the absolute change in these samples remained small. As the MSI classifier uses the relative frequency of microsatellite deletions to classify samples, these observations suggested there would be increased error in classification when using low sample quantity, particularly for MSI-high samples.

To assess the effect of low quantities of template DNA on classification, each sample was scored for each input quantity sequenced. I noted that two of the MMRd CRC samples derived from FFPE tissue (207950 and 244881), had consistently much lower numbers of MBs detected (equivalent to template molecules sequenced) compared to the other samples (Figure 4.12B), suggesting that these samples were of a lower quality. Therefore, sample scores were compared to the mean template molecules sequenced per marker, to assess the effect of both template quantity and quality. This showed that, in these 9 samples, a minimum of 75 MB/marker is sufficient for reliable classification (Figure 4.12D). In summary, recommended QCs should include a minimum input of 25ng of sample DNA and a minimum of 75 MB/marker.

4.9. *BRAF* and *KRAS* Mutations are Reproducibly Detected

Within the validation cohort, 46 of the 50 MSI-high CRCs had been independently tested for *BRAF* V600E using high resolution melt curve analysis (HRM; Nikiforov *et al*, 2009). All of the 14 CRCs that tested positive for *BRAF* V600E by HRM had $\geq 5\%$ variant alleles associated with *BRAF* V600E mutation. Of the 32 CRCs that tested negative for *BRAF* V600E by HRM, 30 samples had *BRAF* V600E detected in $\leq 0.6\%$ of reads and 2 samples had *BRAF* V600E detected in 1.67% and 1.72% of reads. The error rate of NGS platforms is estimated to be 1-

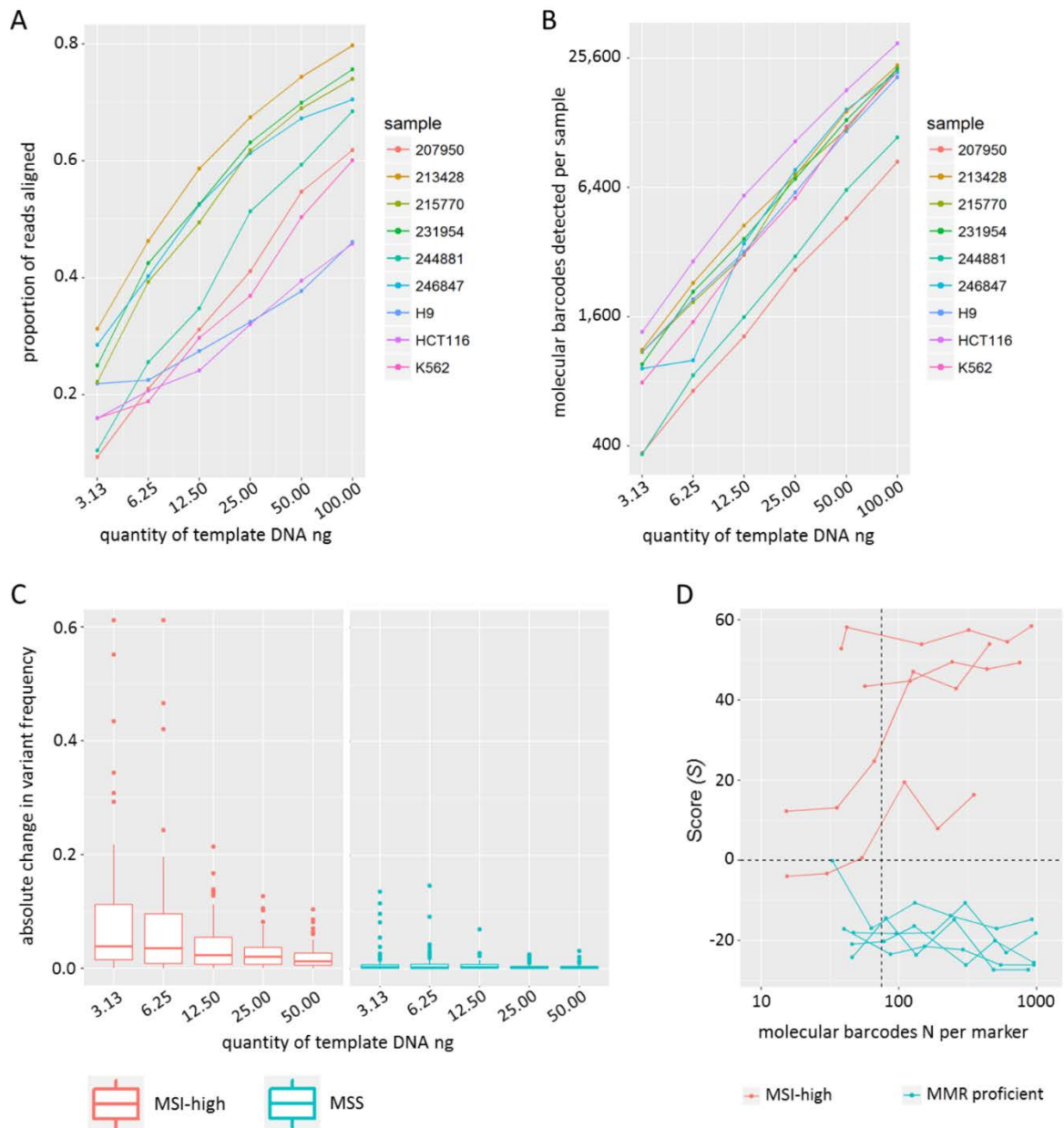


Figure 4.12: Sequencing results and classification of low quantity samples. (A)* Proportion of all sequencing reads aligning to markers at varying input quantities of template DNA of 9 samples. **(B)** Number of molecular barcodes (MBs) detected per sample at varying input quantities of template DNA of 9 samples. **(C)** The absolute change in microsatellite deletion frequency detected when varying input quantities of template DNA, as shown for MSI-high and MSS samples collectively. **(D)** MSI classifier scores from varying input quantities of template DNA of 9 samples, as measured by the mean number of MBs detected per marker. A 75 MBs minimum is shown by the vertical dotted line. **Note: 4 additional markers were included in this sequencing run, part of another piece of work not described in this thesis. The proportions of aligned reads shown are only for the 24 shortMNRs versus the total read output from the MiSeq, hence proportions are lower than the 0.75 used when calculating target read depth (Niedzicka et al, 2017; Table 4.1).*

1.5% (Shendure and Ji, 2008), suggesting these 2 samples may contain true mutations at very low VAFs not detectable by HRM.

Unfortunately, no samples were independently tested for *KRAS* mutations to determine the accuracy of *KRAS* genotyping by the smMIP-based assay. However, NGS is generally accepted to be as accurate as the long established technique of Sanger sequencing for variant detection (Beck *et al*, 2016). To compare our results to the literature, the frequencies of *BRAF* and *KRAS* mutations were analysed. Given the error rate of NGS platforms (Shendure and Ji, 2008), a $\geq 1.5\%$ VAF threshold was used for mutation calling. The observed frequencies of *BRAF* and *KRAS* mutations was in line with the literature (Table 4.6), and the expected association between *BRAF* V600E and an MSI-high phenotype (Muzny *et al*, 2012) was observed (OR: 5.56, 95% CIs: 2.56-12.5, $p < 10^{-5}$). Only one sample had both *BRAF* V600E and a *KRAS* G12 or G13 mutation, with VAFs of 1.67% for *BRAF* V600E and 11.6% for *KRAS* G13D. I also observed an over representation of *KRAS* G13D in MSI-high tumours relative to other *KRAS* mutations (OR: 7.69, 95% CIs: 2.27-25.0, $p < 10^{-3}$).

When assessing the reproducibility of the MSI classifier, *BRAF* and *KRAS* mutation hotspots were also sequenced (Section 4.6). Using the same 1.5% VAF threshold for mutation calling, results from the repeat testing of these 32 CRCs had 100% concordance for both *BRAF* V600E and *KRAS* G12 and G13 variants. For the *BRAF* V600 locus, there was a strong correlation between VAF in the validation cohort and repeat testing (Figure 4.13A; $\beta = 0.93$, $p < 10^{-16}$, $R^2 = 0.99$), and a similarly strong correlation was found for *KRAS* (Figure 4.13B; $\beta = 1.06$, $p < 10^{-16}$, $R^2 = 0.97$), suggesting hotspot mutation detection is reproducible.

Mutation	Observed		Literature	
	MSI-high (n=99)	MSS (n=98)	MSI-high	MSS
<i>BRAF</i> V600E	36.4% (27.6-46.2%)	9.2% (4.9-16.5%)	31%	7%
<i>KRAS</i> G12, G13 variants*	21.2% (14.3-30.3%)	38.8% (29.7-48.7%)	43%*	59%*

Table 4.6: Frequency of *BRAF* and *KRAS* mutations in colorectal cancers (CRCs). CRC samples from training and validation cohorts were combined and the frequency of mutations determined, with 95% confidence intervals. Frequencies observed in the literature are taken from Rajagopalan *et al* (2002).

*Note: Rajagopalan *et al* (2002) analysed *KRAS* codons 59 and 61 in addition to codons 12 and 13.

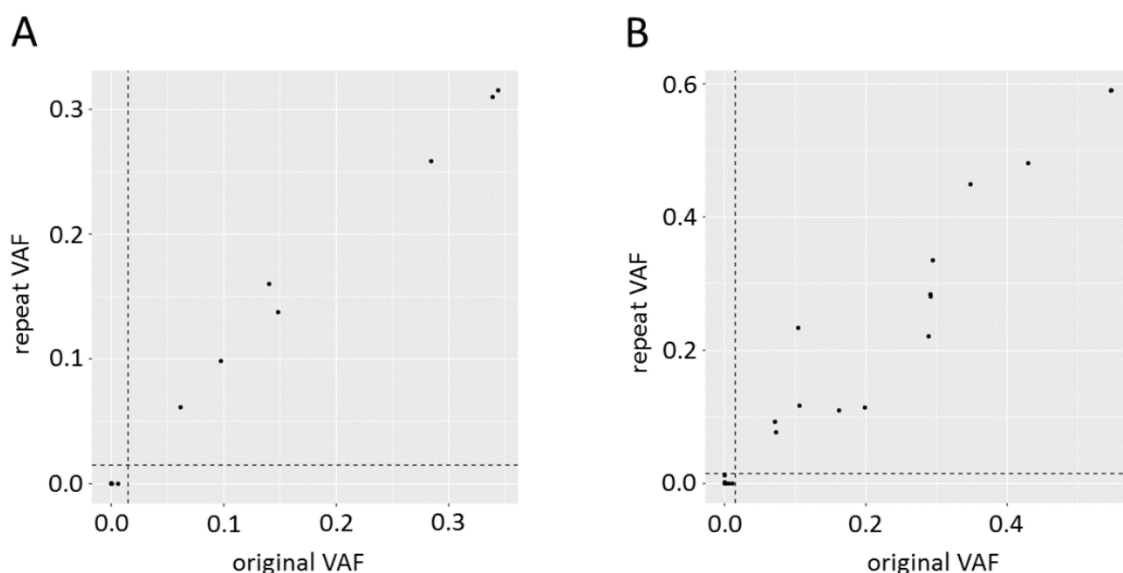


Figure 4.13: Variant allele frequency compared between repeat testing of 32 CRCs... (A) at the *BRAF* V600 mutation hotspot, and (B) at the *KRAS* G12, G13 mutation hotspot. 1.5% thresholds used for mutation calling are shown by dotted lines.

4.10. Assay Cost and Turnaround Time are Superior to Established Methods

The cost and TAT of an assay are significant determinants of its aptness for high throughput diagnostics. The cost of reagents per sample for the smMIP-based MSI assay, including charges for use of the MiSeq instrument, and the TAT were calculated and compared to Promega MSI Analysis System v1.2 (Table 4.7). A full breakdown of reagent costs is available in Appendix J. This comparison shows the smMIP-based MSI assay is superior in cost to Promega MSI Analysis System v1.2, especially at the highest throughput using a panel of 10 markers. With respect to TAT, the Promega MSI Analysis System v1.2 is faster per batch, but a reliance on marker by marker, sample by sample, results interpretation limits the number of samples per batch. Therefore, at scale the smMIP-based MSI assay is superior in TAT. Additionally, for Lynch syndrome screening, the smMIP-based MSI assay includes *BRAF* V600E testing for no additional cost or time, reducing the recommended two test screening pipeline (Newland *et al*, 2017) to one test, further improving its cost and TAT in relation to current methods. A potential disadvantage of the smMIP protocol is the multiple steps required. However, for each step the hands on time is brief and is fully automatable (Neveling *et al*, 2017).

	smMIP-based MSI assay		MSI Analysis System v1.2, Promega	
Cost	25 markers	£8.19-12.70/sample	Tumour, normal	£16.06/sample*
	10 markers	£5.94-7.75/sample	Tumour only	£8.03/sample*
Batch	96+ samples		24 samples	
TAT	4-5 days/batch		2-3 days/batch	

Table 4.7: Summary of cost analysis and turnaround time (TAT). Costs for the smMIP-based MSI assay assumes a read depth of 2000 reads/marker/sample, and the range covers sequencing on a MiSeq v2 Micro, v2, or v3 kit (Appendix J). 25 markers includes 24 short MNRs and the *BRAF* V600 mutation hotspot. *Costs for MSI Analysis System v1.2 are calculated from Promega list price for product MD1641, which includes 100 reactions, and does not include costs of capillary electrophoresis. TAT for both assays includes DNA extraction and sample preparation, through to result. Details of batch size and TAT for the MSI Analysis System v1.2 were provided by the Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust.

4.11. Discussion

I aimed to continue the development of a sequencing-based MSI assay applicable to high throughput cancer diagnostics, as a screening tool for Lynch syndrome and to inform use immunotherapy. To do so, smMIP technology was used to multiplex and expand the marker panel of a singleplex, 17 marker, MSI assay previously developed by our research group (Redford *et al*, 2018). Only one marker, IM66, failed to be amplified using singleplex smMIPs, despite multiple designs, suggesting that MIPgen (Boyle *et al*, 2014a) is a reliable tool for creating MIP-based assays. Furthermore, MIPs are known to be robust to multiplexing (O’Roak *et al*, 2012), and in this project I found that pooling several smMIPs and adjusting their concentrations (to balance reads from each marker) was simple and effective. The marker panel was also expanded to include an additional 9 short MNRs, and *BRAF* and *KRAS* mutation hotspots relevant to CRC diagnostics. The modularity of a smMIP-based diagnostic assay is appealing for the ease with which it could be adapted to other cancer types, for example by inclusion or exclusion of clinically actionable biomarkers as appropriate. The success of the use of smMIPs is evident in the cost reduction and TAT of the assay, whilst maintaining high sensitivity and specificity. Previously, the singleplex assay was estimated to cost £26.20/sample, with a TAT of 11 days (Alhilal PhD Thesis, 2016), whereas the assay now costs £5.94-12.70/sample with a TAT of 4-5days for batches of 96 or more samples.

The smMIP-based MSI assay uses monomorphic and short (7-12bp) MNRs that have significantly lower PCR and sequencing error rates compared to longer markers (Fazekas *et al*, 2010). This low error rate allowed thresholds in the frequency of microsatellite deletions to be defined and used for MSI classification. Each marker is also associated with a SNP of minor allele frequency >20%, allowing the allelic bias of deletions to be assessed, giving further discriminatory power. An algorithm using these thresholds and the relative frequency and allelic bias of deletions in the 24 microsatellite markers, has been trained to calculate the relative probability that a sample belonged to an MSI-high or MSS phenotype, summarising the result in one score (Redford *et al*, 2018). Scoring is an automated process and hence the assay does not require expert, case-by-case, marker-by-marker result interpretation. The monomorphism of the markers also removes the need for matched normal DNA. When applied to both training and validation cohorts, the assay classified 197 CRCs with 100% sensitivity and 100% specificity, relative to FLA using the Promega MSI Analysis System. Classifier performance was therefore found to be equivalent using either the singleplex MSI assay or smMIP-based MSI assay. The 197 samples used were all DNA from FFPE tissues derived from pathology services, representing the spectrum of samples the assay is designed for, and in excess of the minimum 59 independent samples recommended for analytical validation of NGS-based assays (Jennings *et al*, 2017).

The smMIP-based MSI assay fulfils other requirements of an ideal diagnostic test. For example, Jennings *et al* (2017) suggest repeat testing of three samples using new batches of all reagents to show that the assay is reproducible. Here, I used a freshly prepared smMIP multiplex pool, new reagents and a distinct sequencing run, and showed 100% classification concordance in repeat testing of 32 CRC samples. The linear correlation between original and repeat scores was very strong, with low variation (linear regression $R^2 = 0.97$). Despite the reproducibility of the assay in my hands, ideally it should also be run by an independent operator in a different laboratory to confirm this.

Diagnostic tests must also be robust to sample variables. The most critical variables with respect to tumour samples are tumour cell content, and the quantity and quality of DNA. For the smMIP-based MSI assay, I simulated variation in the MMRd cell content of a sample by mixing different quantities of MSI-high DNA, from MMRd CRC cell line HCT116, and MSS DNA, from MMRp PBLs. The assay correctly classified the samples generated down to 3.13% MSI-high content, suggesting a LLoD of approximately 3%. This was superior to FLA by the MSI Analysis System, which correctly classified the same DNA mixtures at $\geq 6.25\%$

MSI-high content. At 3.13%, classification by FLA was uncertain in every marker. Jennings *et al* (2017) recommend testing 59, independent sample containing the VAF of interest to reliably define the LLoD, but recognise this is not feasible and that artificial substitutes may be used. In this study, with the help of Dr Santibanez-Koref, I opted for an *in silico* method of sequencing-read mixing, which was shown to give comparable results to DNA mixing in 27 simulated series. To further test the robustness to MSI-high content, read mixture series were created between the MSI-high and MSS CRCs from the validation cohort, generating 2400 simulated series. 95% of these simulated samples were classified as MSI-high when only 25% of reads originated from an MSI-high CRC. Given that the MSI-high CRCs will not consist of purely MMR tumour cells and are therefore already diluted, this 25% actually represents a much lower MSI-high content, potentially as low as 1.25% in some samples assuming a minimum of 5% tumour cell content in these samples. Whilst this *in silico* method of mixing reads from MSI-high CRCs and MSS CRCs cannot help define a LLoD due to the “impurity” of the MSI-high CRC read samples, it supports that the smMIP-based MSI assay is robust to low MSI-high content.

The CRC DNA samples used were extracted from FFPE tissues, meaning they would have a poorer quality than the control DNAs extracted from fresh cell lines. However, the assay amplified these FFPE-derived DNAs and MSI classification was 100% accurate. smMIPs incorporate MBs into reads allowing the number of sample DNA molecules sequenced to be quantified. As discussed by Jennings *et al* (2017), the number of template molecules sequenced, or library complexity, is a critical QC metric for any NGS-based diagnostic test. By diluting 9 samples, including low quality FFPE-derived DNAs and cell line controls, I showed that classification was reliable when more than a mean 75 MB/marker was detected.

Additional sequencing metrics that should be accounted for by a diagnostic assay include read depth and the percentage of base-calls above or equal to a quality score of Q30 (Jennings *et al*, 2017). Target read depths were calculated to be >5000reads/marker/sample, which was not achieved in all runs. Notably, the training cohort read depth was only 3,719reads/marker/sample as the capacity of the MiSeq flow cell was not used. Early experiments used MiSeq v2 kits and defined 8pM as the optimal DNA library loading concentration. However, due to the number of samples and need for a higher capacity kit, v3 kits were used for subsequent experiments, including sequencing of the training cohort. The lower read depth of the training cohort may be due to the loading of an 8pM library on a

MiSeq v3 flow cell – subsequently 12pM libraries were loaded for v3 kits, which gave expected read depths, and an 8pM loading concentration continued to be used for v2 kits.

Another consequence of reduced quantity of template DNA was reduced read depth, as fewer reads aligned to marker loci. Whilst the reason for this has not been confirmed, I speculate that non-specific amplicons and primer dimers carried through purification constitute a greater proportion of reaction product as template DNA decreases (see Figure 4.11). This is supported by the 127,756 sequencing reads generated from a template negative reaction, of which only 52 reads (0.12%) were aligned to markers. Those few reads that were aligned to a marker are possibly due to sample index mis-assignment (Illumina Inc., 2017). The consequence of this observation is that target read depth calculations should consider the quantity of template DNA being used in reactions. For example, the 0.75 adjustment factor for off-target reads should be decreased if using less than 100ng of sample. An alternative is to optimise purification to remove non-specific products, in particular primer dimers. However, this was not pursued during this work due to time constraints.

For the majority of sequencing runs, more than 75% of base-calls were \geq Q30, which is within the expected range of Illumina sequencing platforms. However, the validation cohort sequencing run only had 57.2% of base-calls \geq Q30, but classification was 100% accurate. Therefore, whilst it is desirable for sequencing to have $>75\%$ of base-calls \geq Q30, it appears that the smMIP-based MSI assay and classifier can tolerate lower. To formally test this would require excessive investment due to the cost of each sequencing run.

It is recommended that *BRAF* V600E or *MLH1* promoter methylation testing are carried out following MMR deficiency testing in all MMRd CRCs to improve screening specificity by identification and exclusion of sporadic cases. The inclusion of *BRAF* V600E testing in the smMIP-based MSI assay streamlines the LS screening pipeline, requiring only one tumour test prior to germline testing of MMR genes, equivalent to tumour-sequencing (Hampel *et al*, 2018). The assay was able to detect low VAF in *BRAF* down to 1.7%, with improved sensitivity compared to HRM analysis, which has an estimated LLoD of 10% (Nikiforov *et al*, 2009). The alternative test for *MLH1* promoter methylation has a higher specificity than *BRAF* V600E when screening for Lynch syndrome (Pérez-Carbonell *et al*, 2010). However, testing both markers is redundant due to their association (Pérez-Carbonell *et al*, 2010). Arguably, *BRAF* V600E is also the superior marker as *MLH1* methylation occurs as a second hit in the CRCs of approximately 55% of *MLH1* mutation carriers (Young *et al*,

2001; Kaz *et al*, 2007), and therefore *MLH1* methylation testing has a lower sensitivity for Lynch syndrome screening than *BRAF* V600E testing (sensitivity 84.2% versus 100%, respectively) (Moreira *et al*, 2015). In addition, germline epimutations in *MLH1* cause Lynch syndrome, and these too would be excluded by *MLH1* methylation testing (Suter *et al*, 2004). This lower sensitivity of *MLH1* methylation testing of Lynch syndrome screening was also observed by Hampel *et al* (2018), relative to tumour-sequencing that analysed *BRAF* V600 only.

One smMIP targeting *KRAS* G12 and G13 mutation hotspots was also included in the assay multiplex, as a proof of principle that the assay can be expanded to other, clinically actionable, biomarkers beyond MSI and *BRAF* V600. Using a $\geq 1.5\%$ mutant read threshold, the frequencies of *BRAF* and *KRAS* mutations detected were similar to frequencies previously observed (Rajagopalan *et al*, 2002). However, the 95% CIs quoted show that significantly fewer *KRAS* mutations were detected in both MSI-high and MSS CRCs. This can be explained by the slightly higher frequency of *BRAF* mutations, which are considered mutually exclusive with *KRAS* mutation (De Roock *et al*, 2010a), and the inclusion of mutations in *KRAS* codons 59 and 61 in the reference study (Rajagopalan *et al*, 2002). Therefore, for our smMIP-based assay to comprehensively cover *RAS* gene mutations, additional smMIPs would be needed. The smMIP-based assay detected only one sample with both *BRAF* and *KRAS* mutations, which could be an extremely rare, sub-clonal co-occurrence (Sahin *et al*, 2013), or perhaps sequencing error as the *BRAF* VAF in this sample was only 1.67% given estimates of sequencing error of 1.0-1.5% on NGS platforms (Shendure and Ji, 2008). A significant predominance of *KRAS* G13D mutations was found in the MSI-high versus MSS CRCs, consistent with the findings of others (Oliveira *et al*, 2004; Phipps *et al*, 2013). Whilst selection pressures between the different functions of G12 and G13 mutations may be the cause of this (De Roock *et al*, 2010b), it is interesting that the specific C>T substitution responsible for the G13D mutation is prevalent in mutational signature 6 (the pattern of random mutations throughout a tumour genome), which is associated with MMRd CRCs (Alexandrov *et al*, 2013). This suggests that loss of MMR may influence the specific driver mutations that are acquired during tumorigenesis (Ahadova *et al*, 2018).

The cost of a diagnostic assay is a significant factor in its clinical uptake. Tumour-sequencing for example, has an estimated cost of 607 \pm 207€ per sample (Marino *et al*, 2018), which may inhibit its uptake. Whilst the cost estimates of Marino *et al* (2018) include overheads, personnel costs, etc, the reagent and consumables costs for target enrichment

and sequencing were estimated to be 291€ per sample, significantly more than a targeted assay such as our smMIP-based MSI assay. Indeed, we found that our assay has an equivalent reagent cost to FLA when using 24 microsatellite plus *BRAF* markers, ranging from £8.19-£12.70 depending on the capacity of the MiSeq kit used. However, 6 microsatellites were sufficient for accurate MSI classification so these costs can be reduced by decreasing the number of markers to increase the number of samples per sequencing run. Furthermore, amplicons were purified per sample in the protocol of this study, but it is feasible to pool amplicons prior to purification, saving additional cost and time. The smMIP protocol is also fully automatable (Neveling *et al*, 2017), which would again reduce cost and handling. Finally, *BRAF* V600E detection is included within the assay, avoiding expenditure on additional tests for Lynch syndrome screening. The modularity of smMIPs means it would be trivial to incorporate other clinically actionable markers for negligible extra cost. These advantages make the described MSI assay particularly suited to high throughput diagnostics, for example in large testing laboratories where hundreds to several thousand CRCs may be assessed each year, but does not preclude use of IHC or FLA in smaller scale laboratories given the long established efficacy of these methods. However, given the clinical guidelines for MMR deficiency testing of CRCs (Newland *et al*, 2017), and the strong likelihood that these will be expanded to other cancer types due to the pan-cancer efficacy of immune checkpoint blockade therapy (Le *et al*, 2017), it is likely that clinical service will become more reliant on centralised diagnostic services to meet demand.

4.12. Conclusions and Future Work

The smMIP-based MSI assay developed here, which has built upon the work of Redford *et al* (2018), is highly sensitive and specific for MSI status in CRCs, simultaneously detects *BRAF* V600E, is reproducible, and is robust to sample variables given the specified QCs (Table 4.8). The automation of laboratory workflow and results interpretation removes the need for expert personnel and provides a cheap, scalable assay. Combined, these factors suggest that a high throughput smMIP-based MSI assay is a suitable companion diagnostic for immune checkpoint blockade therapy and is applicable to two-step Lynch syndrome screening strategies.

From here, our research group intends to commercialise the assay, which may require further protocol optimisation, most notably removing redundant markers from the panel and selection of an optimal set to further reduce cost. We will also deploy the assay

into the Northern Genetics Service, who have been our collaborators throughout this study, for clinical validation (whereby the assay will be ran in parallel with standard diagnostic procedures on clinical samples in real time) and formal testing of assay reproducibility. Finally, we have been working with a commercial partner, NimaGen, to transfer the assay into pre-aliquoted plates, to further reduce sample and reagent handling. NimaGen currently market smMIP assays for *BRCA* gene sequencing (Neveling *et al*, 2017).

Having developed the MSI assay for CRC diagnostics, it was of interest to explore its application to the detection of low-level MSI in normal tissues as a biomarker of CMMRD.

Variable or Parameter		Quality Control Criteria
DNA sample	tumour cell content	≥3%
	input quantity	≥25ng
Sequencing	base-call quality	75% ≥Q30
	molecules sequenced	≥75 MB/marker

Table 4.8: Quality controls for the smMIP-based MSI assay for reliable classification.

Chapter 5. Accurate Detection of Constitutional Mismatch Repair Deficiency by a Sequencing-based Microsatellite Instability Assay

5.1. Introduction

CMMRD is a highly penetrant cancer-predisposition syndrome that manifests in childhood to early adolescence, caused by germline mutation in both alleles of an MMR gene (Wimmer *et al*, 2017). Guidelines for the management of this condition recommend surveillance and altered treatment (Vasen *et al*, 2014). Identification of CMMRD uses clinical features, such as diagnosis of malignancy, family history and non-neoplastic features, according to published guidelines (Wimmer *et al*, 2014). However, many of the clinical features of CMMRD overlap with other syndromes, and hence genetic diagnosis by germline sequencing of MMR genes is required (Wimmer *et al*, 2014). Genetic diagnosis can be confounded by variants of unknown significance (VUS) in MMR genes and the multiple pseudogenes of *PMS2* (De Vos *et al*, 2004), which accounts for approximately 60% of CMMRD (Wimmer *et al*, 2017).

Low-level MSI occurs in the non-neoplastic tissues of CMMRD patients and is detectable by highly sensitive MSI assays (Ingham *et al*, 2013; Bodo *et al*, 2015). Such diagnostic tests can be used to clarify uncertain genetic diagnoses. However, current assays are limited by insensitivity for biallelic, germline mutation of *MSH6* by analysis of DNRs (Ingham *et al*, 2013), or require laborious and expensive methodology (Bodo *et al*, 2015), which restrict clinical utility. In chapter 4, I presented a smMIP and sequencing-based MSI assay, applicable to high throughput cancer diagnostics. The smMIPs utilise molecular barcodes to count the number of template DNA molecules sequenced as an assay QC, based on the assumption that each molecular barcode corresponds to a single template molecule of DNA. However, an alternative use of molecular barcodes is to group reads that share the same barcode to summarise the sequence content of the majority of reads within the group in a single molecule sequence (smSequence), representing the sequence of the original template molecule. Analysing smSequences rather than all reads equalises representation of template DNA molecules and reduces PCR and sequencing errors, which will only be present in the minority of reads within a molecular barcode group. By reducing noise, smSequences therefore facilitate detection of low frequency variants (Casbon *et al*, 2011). Hence, whilst use of smSequences was not necessary to assess MSI in CRCs, it was of interest to see if our assay would be sensitive to the low-level MSI in non-neoplastic tissues of CMMRD patients

by adoption of an alternative, smSequence-based analysis of microsatellite length. The analysis of MNRs suggested it would also be sensitive for constitutional MSH6 deficiency (Bodo *et al*, 2015).

As well as being suitable for routine confirmation of genetic diagnosis in suspected CMMRD patients, the simplicity and low cost of the smMIP and sequencing-based MSI assay would make it ideal for screening larger cohorts of patients. For example, there are currently no estimates of the frequency of CMMRD in paediatric haematological malignancy. Whilst it is recognised that germline genetic testing for causative mutations is required in this population, diagnosis is challenging due to the number of genes to screen and frequent lack of family history or other distinguishing clinical features (Furutani and Shimamura, 2017). A CMMRD screening tool to streamline these diagnostic pathways could be an invaluable addition to the clinical management of haematological malignancies, and similarly for other childhood malignancies. Furthermore, the full phenotypic spectrum and prevalence of CMMRD are not known (Durno *et al*, 2017); cheap diagnostic assays would facilitate research efforts to identify CMMRD in populations with related conditions.

5.2. Aims

To apply the smMIP and sequencing-based MSI assay, developed in Chapter 4, to CMMRD diagnostics, I aimed to:

1. Assess the use of molecular barcodes to reduce PCR and sequencing error in the 24 short MNRs analysed by our MSI assay.
2. Develop an automatable method to detect low frequency variants in microsatellite length, and determine the ability of the assay to detect CMMRD.

5.3. Study Samples and Method

To address the study aims, the study was split into two parts; here I will summarise the division of samples for clarity.

A pilot cohort of 40 control, germline DNA samples extracted from the PBLs of anonymised patients, and 5 CMMRD, germline DNA samples extracted from PBLs of genetically confirmed CMMRD patients, was sequenced using the smMIP-based MSI assay across three sequencing runs, to exclude batch effects. The 40 pilot control samples were used to determine the reduction in PCR and sequencing error by use of molecular barcodes. The pilot CMMRD samples were then analysed to see if they could be distinguished from the

controls. A blinded cohort of 31 CMMRD and 54 control samples was analysed across three more sequencing runs. The spread of these samples across multiple runs was due to the incremental collection of samples for analysis, as explained in Section 5.6. All CMMRD patient samples, from pilot and blinded cohorts, are summarised in Appendix B.

5.4. Single Molecule Reads reduce Error in Microsatellite Length Variant Detection

A pilot cohort of 5 CMMRD samples and 40 anonymised control DNAs extracted from peripheral blood leukocytes (PBLs) were amplified and sequenced across three sequencing runs, using the same smMIP-based method developed in Chapter 4 and described in Section 2.9, to a mean (\pm SD) read depth of 2,735 \pm 1,120 reads/marker/sample, with a mean 83.4% of base calls of quality >Q30. The reads were processed according to Section 2.12.1, utilising molecular barcodes. The frequency distribution of molecular barcode groups, according to the number of reads they contain, showed that the vast majority of reads share a molecular barcode with at least one other read and therefore, whilst groups containing only one read are the most frequent, the majority of molecular barcode groups contain ≥ 2 reads. The assignment of multiple reads to the majority of barcode groups showed that error correction by use of smSequences would be testable. As an illustrative example, a distribution is shown for marker GM07 in one of the control samples (Sample ID: 40) (Figure 5.1).

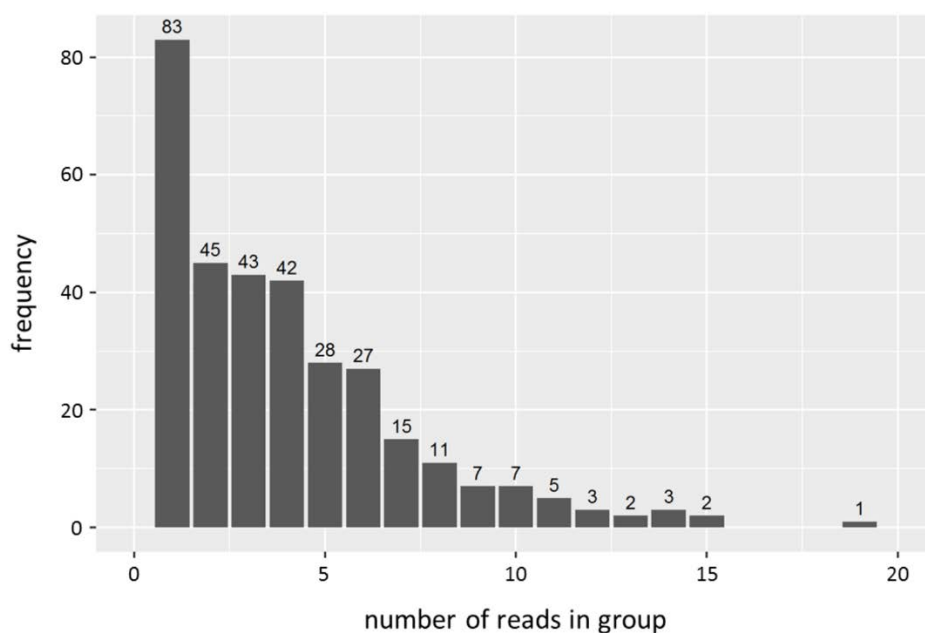


Figure 5.1: The frequency distribution of molecular barcode groups by the number of reads within each group. Sequencing reads from microsatellite marker GM07 in a control sample (Sample ID: 40) were grouped according to molecular barcode. Each group was classed by the number of reads within it, and the frequency of groups containing different numbers of reads was determined.

There are two considerations to define criteria for smSequence generation. First, a minimum of two reads per group are needed to allow correction of PCR and sequencing errors as correction relies on comparison of sequences between reads (Figure 5.2). Increasing the minimum number of reads required per group could lead to large numbers of reads being discarded, potentially counteracting any benefit from error correction. Second, length variants can originate from true mutations, or PCR and sequencing errors. By grouping reads by molecular barcode, erroneous microsatellite lengths that do not represent the original template molecule can be identified as they will only occur in the minority of reads in a group. Where there is only one read in a group or different lengths are equally represented, PCR and sequencing errors cannot be identified. Therefore, the length of the microsatellite in the smSequence must be found in the majority of reads in the group to be confident that it is the true length in the original, template DNA molecule (Figure 5.2).

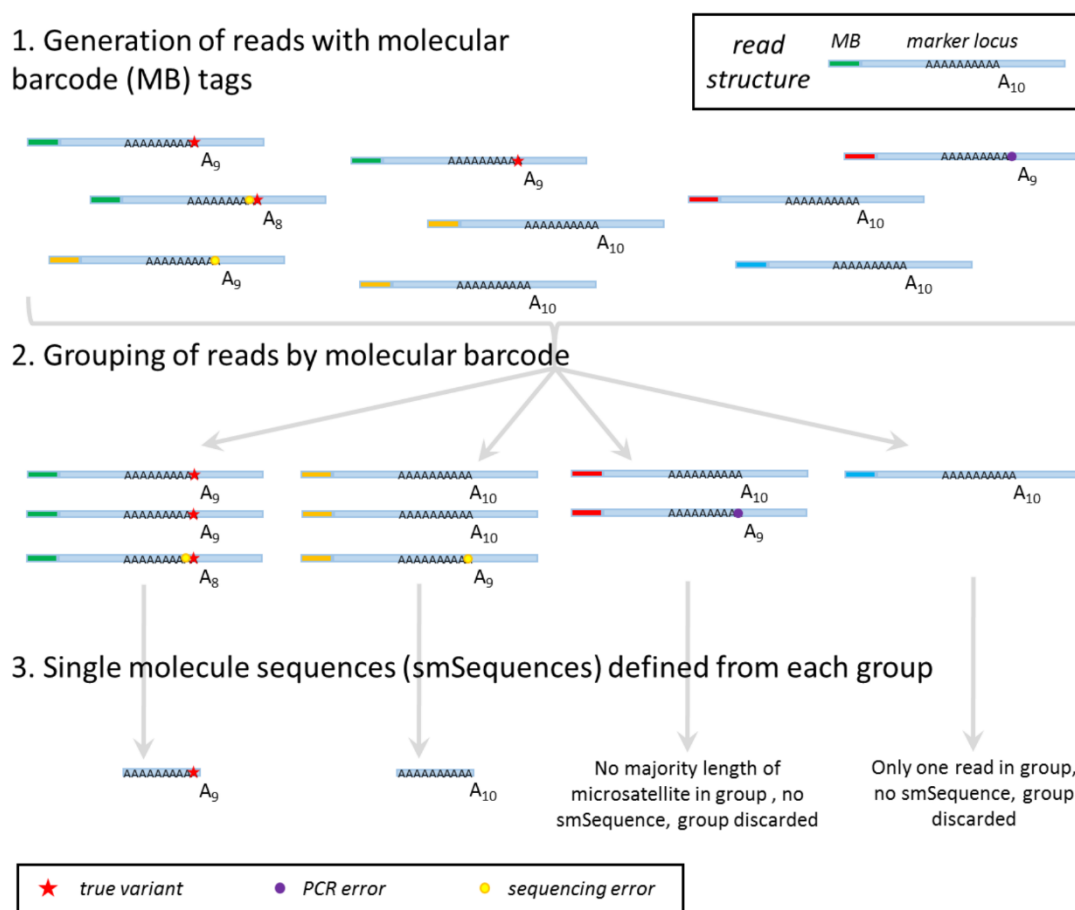


Figure 5.2: Definition of single molecule sequences (smSequences). There are three steps to define smSequences. One, reads are generated with molecular barcodes (MB, unique sequences identified by colour). Two, reads are grouped by molecular barcode, assuming that all reads in a group represent the same template DNA molecule. Three, molecular barcode groups are discarded if one length of microsatellite cannot be found in the majority of reads within the group, and where there is only one read in the group.

I aimed to use smSequences to reduce the noise in detection of variants in microsatellite length so that low frequency, true variants would be detectable. Based on the assumption that template DNA from MMRp PBLs should not be affected by MSI, variants in microsatellite length from control samples were classed as errors of PCR or sequencing. To determine whether or not use of smSequences was able to reduce error rate, the relative frequency of variants in microsatellite length in the 40 controls was, therefore, used as a quantitative measure of assay error rate in each marker. Again using marker GM07 in one of the control samples (Sample ID: 40) as an illustrative example, the relative frequency of variants detected reduces from 7.07% (98/1387) for all reads to 0.87% (2/230) for smSequences, which is equivalent to an 8-fold reduction in error rate (Figure 5.3A). However, by modelling read counts as a binomial distribution, changes in the relative frequency of variants may be a result of reduced count number; statistical analysis was needed to confirm that any change in error rate was significant. Two-by-two tables were constructed that included the counts containing WT or variant microsatellite lengths for all reads and for smSequences (Figure 5.3B), and Fisher’s exact test was used to compare count distribution. In the example, the reduction in error rate is significant ($p < 0.05$).

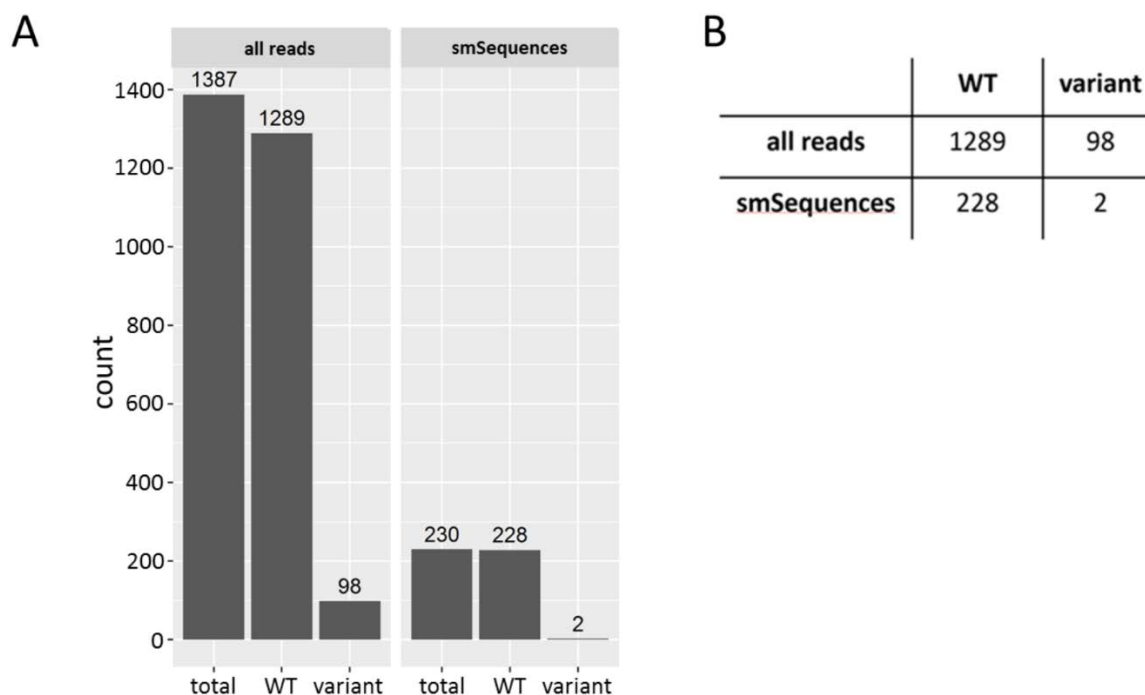


Figure 5.3: Count of reads with different microsatellite lengths, using either all reads irrespective of molecular barcode, or single molecule sequences (smSequences). Sequencing reads from microsatellite marker GM07 in an anonymised control sample (Sample ID: 40) were counted, using either reads irrespective of molecular barcode, or reads grouped by molecular barcode and summarised in one smSequence. Counts include wild type (WT) and variant microsatellite length, and are shown as **(A)** a graph or as **(B)** a two-by-two contingency table suitable for Fisher’s exact test.

The results of error rate analysis in all 24 microsatellites in the 40 control samples are shown in Figure 5.4. To prevent germline variants from affecting analyses, observations where variants in microsatellite length had a relative frequency > 0.4 (indicating the sample is heterozygous or homozygous for a novel length of microsatellite) were excluded from further analyses. Across the 960 observations (24 microsatellite markers in 40 controls) only 6 germline variants were detected and excluded. Using all reads, it was evident that different markers had different, base error rates (Figure 5.4, top panel). To quantify the change in error rate when smSequences were used rather than all reads, I used the equation:

$$\text{fold. change} = \frac{\text{error rate}_{\text{smSequences}}}{\text{error rate}_{\text{all reads}}}$$

Fold-change above 1 means smSequences have increased error rate of microsatellite length detection, and fold-change below 1 means smSequences have decreased error rate of microsatellite length detection, relative to analysis of all reads irrespective of molecular barcode. All markers show a reduction in error rate in all samples when smSequences are used except for GM09 in four samples, with the vast majority showing a two-fold or greater reduction in error (fold-change in error rate < 0.5) (Figure 5.4, middle panel). The four samples that showed an increase in error rate for GM09 had very low counts of smSequences (15-25) detected at GM09, despite counts of smSequences for other markers (166-596) and read depth in GM09 (1999-3010) equivalent to other samples. Indeed very low counts of smSequences in GM09 were observed for several other samples, but it was uncertain what the cause of this was (see Section 5.8), so I chose to keep GM09 in the marker panel. In some observations, use of smSequences removed all error in microsatellite length detection such that all smSequences contained a wild type (WT) microsatellite length, giving an infinite-fold reduction in error (fold-change in error rate = 0) (Figure 5.4, middle panel); this was more frequent in less error prone markers where the base error rate from all reads was already very low (Figure 5.4, compare middle and top panels). Excluding those samples where fold-change = 0, the magnitude of fold-change is correlated with the base error rate of the marker ($r_s = -0.29$, $p < 10^{-10}$), showing that smSequences facilitate a greater reduction of error in more error-prone markers. For 15/24 markers, this fold-change was significant ($p < 0.05$) in the majority of the control samples analysed (Figure 5.4, bottom panel). In summary, using smSequences significantly reduced the error in detection of variants in microsatellite length, and for some markers the majority of samples contained no false variants in smSequences. smSequences would, therefore, improve detection of true,

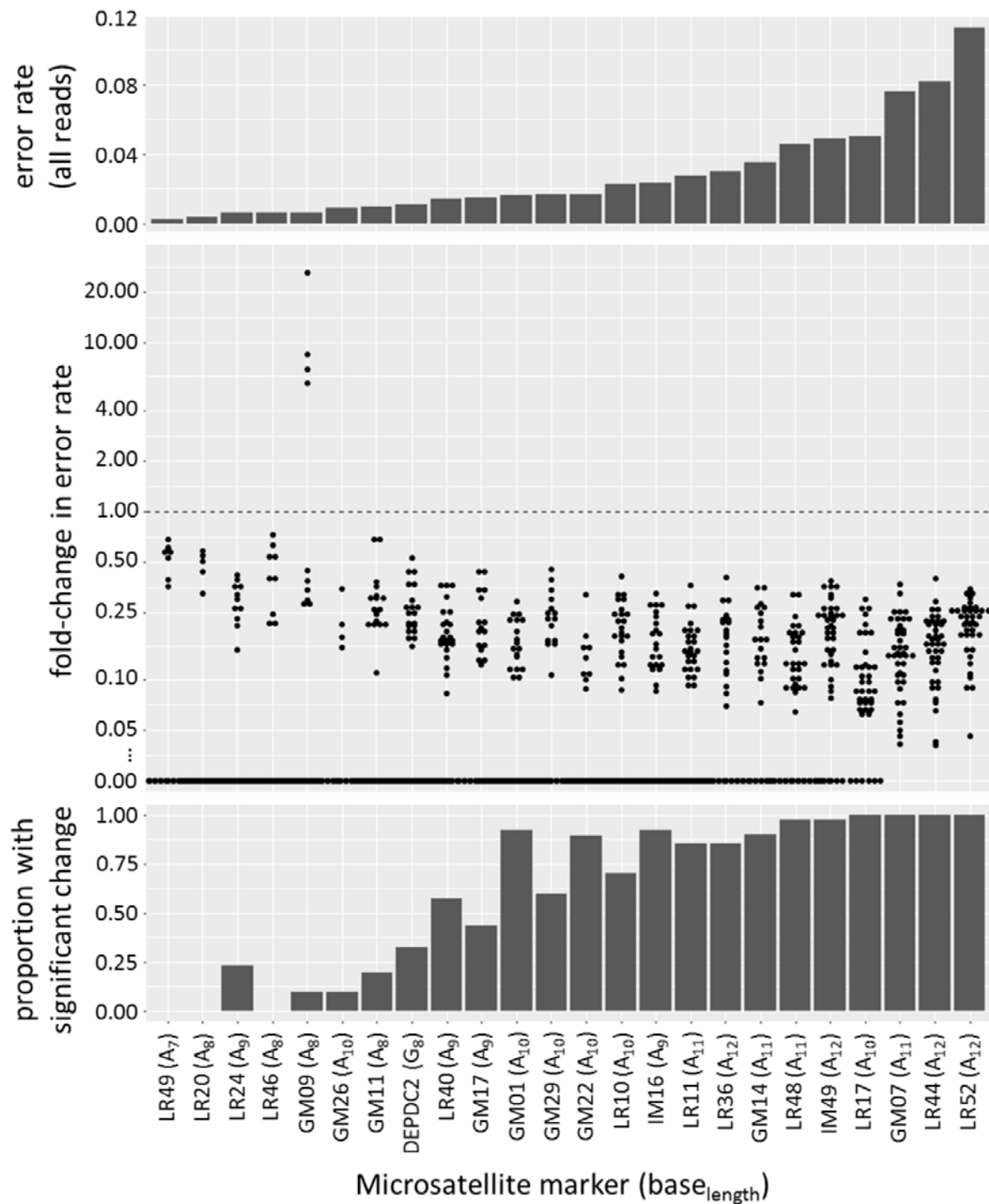


Figure 5.4: Using single molecule sequences (smSequences) reduces the error in detection of variants in microsatellite length. Top panel: microsatellites are listed from left to right in order of increasing error rate, as measured by the relative frequency of microsatellite length variants from all reads, averaged (mean) across the 40 control samples analysed (Sample IDs: 1-40). **Middle panel:** when smSequences were analysed compared to when all reads were analysed irrespective of molecular barcode, the change in error rate was determined (fold-change <1 represents a reduction in error rate). **Bottom panel:** the proportion of fold-changes in error rate that are statistically significant was determined for each marker using Fisher’s exact test.

low frequency variants and it was decided that smSequences should be used in the analysis of CMMRD samples. In addition, 6 germline variants were detected among the 40 samples. Hence, any automated method of CMMRD classification using the smMIP and sequencing-based MSI assay would need to identify and exclude rare germline variants so that they cannot influence result interpretation.

5.5. CMMRD Samples are identifiable by Deviation in Microsatellite Lengths from Controls

As a preliminary assessment of whether or not the assay could detect low frequency variants in microsatellite length associated with CMMRD, 5 CMMRD samples (Sample ID: A-E) had been included in the pilot cohort. These included samples from two patients with homozygous *PMS2* mutation, one patient with homozygous *MSH6* mutation, one patient with compound heterozygous *MSH6* mutation, and one patient with compound heterozygous *MLH1* mutation. Again, sequencing of the pilot cohort was spread across three different runs to prevent batch effects from obscuring analyses. Read data was summarised as the relative frequency of smSequences that contained a wild type length of microsatellite (prWT). I hypothesised that there would be a decrease in prWT in the 24 markers in CMMRD samples due to presence of low frequency variants in microsatellite length. Indeed, prWT was significantly lower in 22/24 markers ($p < 0.05$) from the 5 CMMRD patients.

Having observed this difference in prWT in the majority of markers, I aimed to develop a simple method of sample classification based on the prWT. When analysing CRCs, the distribution of microsatellite deletion frequency was modelled in both MSI-high and MSS cancers to define marker specific thresholds, therefore allowing the MSI classifier to determine the relative probability a sample was either MSI-high or MSS (Chapter 4). Due to the rarity of CMMRD, with approximately 200 known cases globally (Wimmer *et al*, 2017), using such a classifier is not feasible as samples are not readily available to model distributions in CMMRD or validate thresholds. Therefore, for CMMRD classification a scoring method was devised that would quantify deviation of a sample from a control distribution. For each marker, the Beta distribution was used to model the prWT in a control, non-CMMRD population using the smSequences from the 40 control samples (Figure 5.5; Appendix K), with exclusion of germline length variants (prWT < 0.6). Using these distributions, the probability of an observed prWT being smaller than expected of a control population was determined, and for each sample a single score (Table 5.1) was calculated by

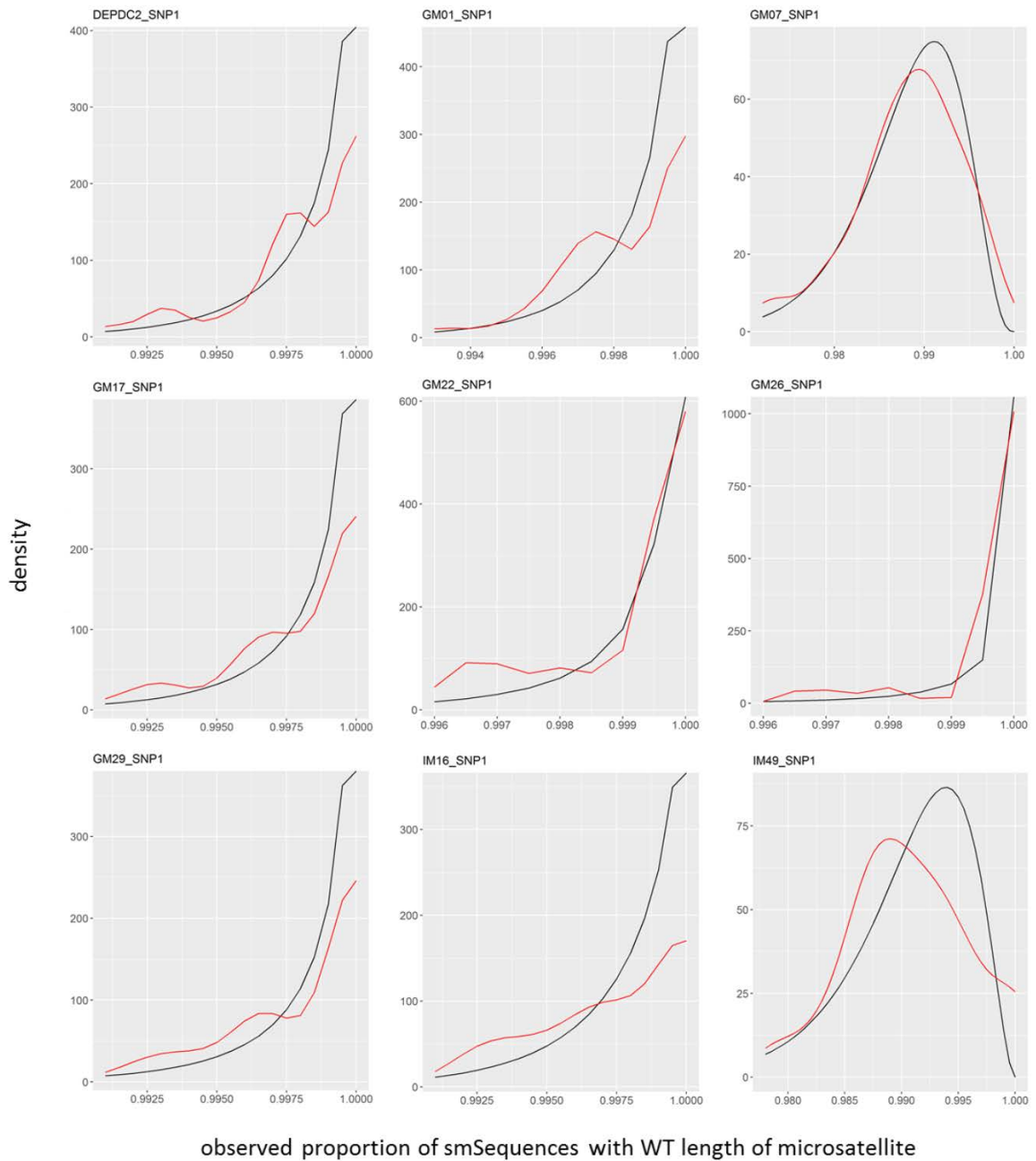


Figure 5.5: Modelling the distribution of the proportion of smSequences containing wild type (WT) length of microsatellite (prWT). The prWT was determined in each marker across 40 control samples (Sample IDs: 1-40), and for each marker the prWT were modelled by a Beta distribution, excluding samples with germline length variants (prWT < 0.6). A comparison of Beta (black line) and empirical (red line) distributions is shown for 9 markers. Graphs for all markers are shown in Appendix K.

combining the probabilities from the 24 markers, again excluding germline variants (prWT < 0.6), using Fisher's method and the following equation:

$$score = -\log_{10}(combined\ probability)$$

The 40 control sample scores ranged from 0.00 to 1.47 and the 5 CMMRD sample scores ranged from 10.02 to 27.34, showing that these CMMRD samples have a minimal probability

of belonging to a control, non-CMMRD population based on the observed prWT in the 24 microsatellite markers.

Marker	Sample ID: 16 (control)		Sample ID: C (CMMRD)	
	prWT	probability	prWT	probability
DEPDC2	1.000	1.000	0.994	0.055
GM01	0.997	0.155	0.997	0.168
GM07	0.992	0.711	0.936	0.000
GM09	0.996	0.205	0.999	0.259
GM11	0.997	0.166	0.997	0.143
GM14	1.000	1.000	0.966	0.001
GM17	1.000	1.000	0.997	0.160
GM22	0.997	0.044	0.998	0.122
GM26	1.000	1.000	0.989	0.000
GM29	1.000	1.000	0.991	0.025
IM16	0.997	0.286	0.989	0.014
IM49	0.986	0.182	0.973	0.008
LR10	0.996	0.276	0.981	0.002
LR11	0.996	0.302	0.987	0.015
LR17	0.995	0.458	0.966	0.000
LR20	1.000	1.000	0.997	0.027
LR24	0.529	NA	0.996	0.030
LR36	1.000	1.000	0.989	0.036
LR40	0.997	0.189	0.990	0.003
LR44	0.989	0.590	0.951	0.001
LR46	0.996	0.048	0.995	0.030
LR48	0.989	0.100	0.977	0.007
LR49	1.000	1.000	1.000	1.000
LR52	0.962	0.086	0.897	0.000
	combined p: 0.32		combined p: 4.57 x10⁻²⁸	
	score: 0.49		score: 27.34	

Table 5.1: Conversion of observed proportion of smSequences containing a wild type microsatellite length (prWT) to a probability and per sample score. Using the Beta distribution of prWT for each marker in 40 controls (Sample IDs: 1-40), observed prWT can be converted to a probability. This probability therefore represents the probability that an observation is less than would be expected in a control, non-CMMRD population. Observed prWT and the associated probabilities are shown for one control and one CMMRD sample (Sample IDs: 16 and C, respectively). Note the presence of a germline length variant in sample 16, marker LR24; such germline length variants were excluded from scoring. Probabilities were combined by Fisher’s method and sample score is equal to $-\log_{10}(\text{combined probability})$; the higher the score the greater the deviation of the sample from a control, non-CMMRD population.

5.6. CMMRD Samples are Identifiable with High Accuracy

A blinded cohort of 56 samples was amplified and sequenced using the smMIP-based MSI assay to a mean (\pm SD) read depth of 4,539 \pm 1,320 reads/marker/sample, with a mean 84.5% of base calls of quality $>Q30$. The blinded cohort included 16 samples from 15 genetically confirmed CMMRD patients, covering biallelic mutation of each of the four MMR genes implicated in CMMRD, and 40 control samples. All samples were independent from those analysed in the pilot cohort. Each sample was scored according to its observed prWT from the 24 markers and using the method described in Section 5.5. Again, all samples, clinical details and scores are summarised in Appendix B.

The 16 CMMRD samples scored from 1.59 to 53.72, and the 40 control samples scored from 0.00 to 1.08. This establishes that the method can fully separate CMMRD samples from the controls with high accuracy. One CMMRD sample (Patient ID: 8, Sample ID: 99, score = 1.59) scored low relative to the other CMMRD samples (scores = 5.71-53.72) and much lower than an affected sibling (Patient ID: 9, Sample ID: 82, score = 19.09) who shares the same compound heterozygous mutation of *MSH6*. This score of 1.59 is equivalent to a 2.6% probability the sample comes from a control population, and therefore raised suspicion that the sample was DNA from another individual. Sanger sequencing was used to confirm sample identity by detection of the c.3557-1G>C splice site mutation at the 5' end of exon 7 that affects one allele of *MSH6* in this patient. Their second allele contains an intragenic deletion covering exons 3-7, hence it was expected that Sanger sequencing would detect only the substitution at the affected locus. In parallel with sample 99, sample 82 (from the affected sibling) was sequenced as a positive control, and sample 95 was sequenced as a negative control. Sanger sequencing confirmed the identity of sample 99 (Figure 5.6). Interestingly, patient 8 was aplastic at the time of blood draw for sample 99 due to chemotherapy for T cell lymphoma, and it is feasible that the low leukocyte count or therapy might have influenced the frequency of microsatellite length variants in PBLs, and hence sample score.

To see if the low score from patient 8 was reproducible, 2 additional samples (one from an independent blood draw taken at a time similar to sample 99, and the other blood draw taken 8 weeks later once the patient had recovered from aplasia) were collected. This was also an opportunity to assess additional CMMRD samples, including samples from patients homozygous for a hypomorphic mutation in *PMS2*. These patients have residual MMR activity and an attenuated phenotype, with much later onset of malignancy than is

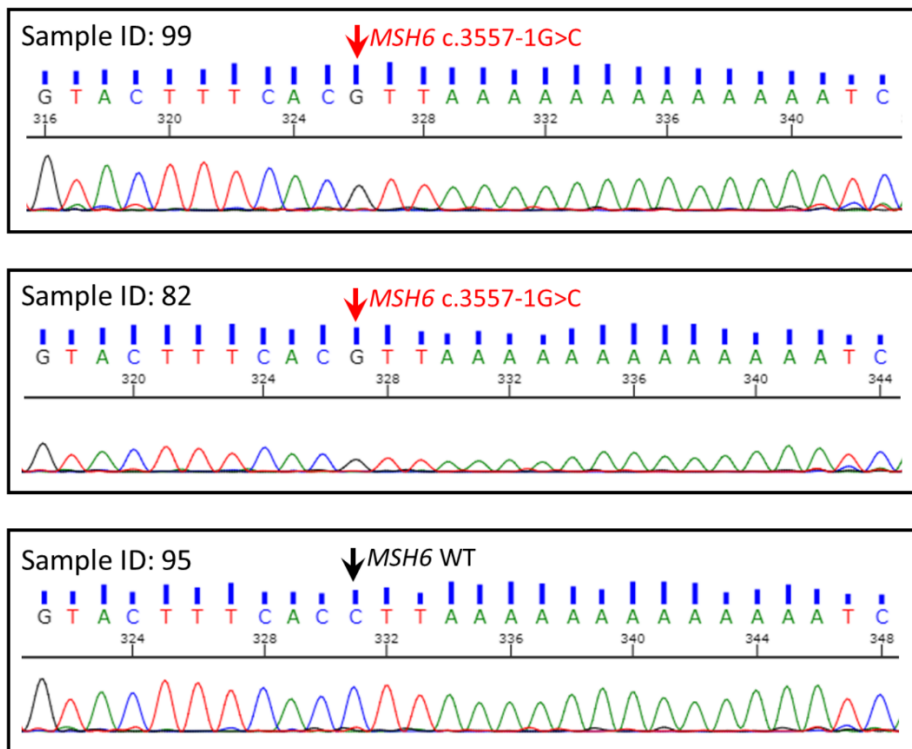


Figure 5.6: Confirmation of the identity of sample 99. The patient that sample 99 was extracted from has a causative splice site mutation in *MSH6*, specifically c.3557-1G>C, and an intragenic deletion in the second allele of *MSH6*, spanning the same locus. Sanger sequencing of the *MSH6* splice site confirmed presence of the hemizygous c.3557-1G>C mutation in sample 99 and a positive control sample (Sample ID: 82) from an affected sibling. Sample 95, a negative control, did not contain the mutation. Note the sequence shown is that of the reverse strand.

typical of CMMRD (Li *et al*, 2015) and, therefore, it was expected that their PBLs would have lower frequencies of variants in microsatellite length than other CMMRD samples. A second, blinded cohort was assembled, including 10 samples from 9 genetically confirmed CMMRD patients, 14 control samples, the two new samples from patient 8, and the 3 samples from the 3 patients homozygous for the hypomorphic *PMS2* mutation. The cohort was amplified and sequenced using the smMIP-based MSI assay to a mean (\pm SD) read depth of 3,288 \pm 1,898 reads/marker/sample, with a mean 84.7% of base calls of quality >Q30. Each sample was scored by the described method. Scores from the 10 CMMRD samples ranged from 3.54 to 54.55, and the 14 control samples scored from 0.00 to 1.14, again showing that CMMRD samples are separable from controls. The two samples from patient 8 again scored relatively low, including scores of 2.08 and 4.62, but remained distinguishable from controls. The samples from patients homozygous for the hypomorphic *PMS2* mutation scored 2.76, 4.28, and 5.90, showing that they were also separable from controls but had lower scores than the majority of CMMRD patients, consistent with their residual MMR activity (Li *et al*, 2015).

Finally, to ensure the method of CMMRD identification would not pick up Lynch syndrome, DNAs extracted from the PBLs of Lynch syndrome gene carriers participating in the CaPP3 clinical trial (n = 40, see Section 2.2.3) was analysed. These samples covered 9 *MLH1*, 21 *MSH2*, 8 *MSH6*, and 1 *PMS2* mutation carriers. One patient did not disclose which MMR gene was affected. Samples were sequenced by the smMIP-based MSI assay to a mean (\pm SD) read depth of 2,681 \pm 985 reads/marker/sample, with a mean 83.2% of base calls of quality >Q30, and scored. Scores ranged from 0.00 to 0.92, meaning that Lynch syndrome gene carriers are indistinguishable from non-CMMRD controls.

A collective analysis of all 36 CMMRD samples, covering 32 patients, and 94 control samples, excluding any repeats of the same sample (see Section 5.7), showed that the assay was capable of separating all of the analysed CMMRD samples from the controls (Figure 5.7; Table 5.3). As well as having genetic confirmation of mutations in both alleles of one of the MMR genes in the CMMRD patients, all samples were analysed with the gMSI assay (Ingham *et al*, 2013) by Barbara Mühlegger at the Division of Human Genetics, Medical University of Innsbruck. gMSI accurately detected CMMRD, with no false positives, except for 15 CMMRD samples from patients with biallelic germline mutation of *MSH6* (Table 5.2).

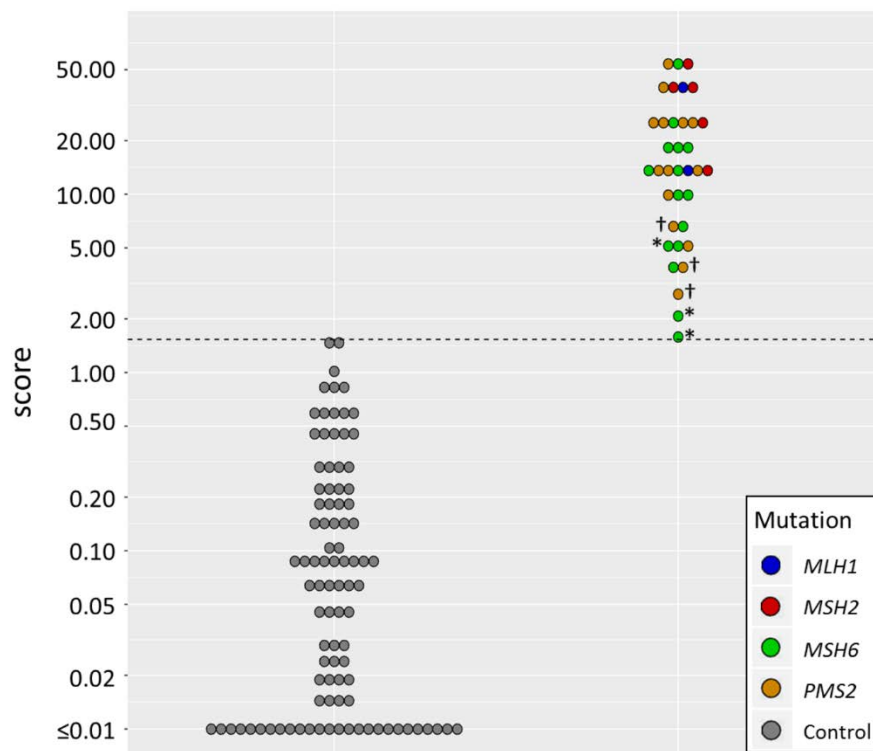


Figure 5.7: Score distribution of constitutional mismatch repair deficiency (CMMRD) and control samples. DNA samples from peripheral blood leukocytes of 32 CMMRD and 94 control patients were scored, and can be separated by an *a posteriori* threshold (score >1.53, dotted line). * scores for patient 8. † scores for patients homozygous for hypomorphic *PMS2* mutation.

Patient ID	Genotype	Sample ID	Score	gMSI A	gMSI B	gMSI C
25	<i>MSH2</i> hom	132	54.55	0.155	0.146	0.396
17	<i>MSH6</i> comp het	43	53.72	-0.062	-0.007	-0.053
7	<i>PMS2</i> hom	54	53.59	0.037	0.233	0.294
24	<i>MSH2</i> hom	116	43.10	0.065	0.038	0.197
12	<i>MLH1</i> hom	49	42.98	0.055	0.11	-0.009
23	<i>MSH2</i> hom	104	42.52	0.048	0.082	0.259
15	<i>PMS2</i> hom	98	36.97	0.067	0.085	0.482
20	<i>MSH2</i> hom	87	27.67	0.047	0.079	0.232
1	<i>PMS2</i> hom	C	27.34	0.300	0.067	0.061
10	<i>PMS2</i> hom	56	25.52	0.179	0.082	0.328
3	<i>MSH6</i> hom	D	24.88	-0.038	-0.022	-0.048
7	<i>PMS2</i> hom	93	23.20	-0.025	0.182	0.27
2	<i>PMS2</i> hom	A	23.03	0.093	0.287	0.394
9	<i>MSH6</i> comp het	82	19.09	-0.065	-0.049	-0.079
14	<i>MSH6</i> hom	76	18.07	-0.052	-0.026	-0.053
21	<i>MSH6</i> hom	101	17.61	-0.036	-0.058	-0.061
19	<i>MSH2</i> hom	58	14.85	0.068	0.06	0.259
18	<i>PMS2</i> hom	71	14.49	-0.016	0.096	0.29
5	<i>MLH1</i> comp het	E	14.43	0.022	0.051	0.062
16	<i>MSH6</i> comp het	83	13.70	-0.074	-0.05	-0.07
26	<i>PMS2</i> comp het	113	13.08	-0.045	0.056	0.216
28	<i>PMS2</i> hom	130	12.82	0.24	0.138	0.323
6	<i>MSH6</i> hom	65	12.47	-0.066	-0.059	-0.054
4	<i>MSH6</i> comp het	B	10.02	-0.050	-0.040	-0.024
11	<i>MSH6</i> hom	91	9.85	-0.066	-0.059	-0.064
27	<i>PMS2</i> comp het	124	9.83	0.175	0.152	0.249
22	<i>MSH6</i> hom	109	7.39	-0.04	-0.06	-0.111
31 †	<i>PMS2</i> hom	115	5.90	0.049	0.064	0.065
13	<i>PMS2</i> hom	51	5.71	-0.009	0.151	0.294
32	<i>MSH6</i> hom	128	4.78	0.012	-0.035	-0.044
8 *	<i>MSH6</i> comp het	102	4.62	-0.066	-0.055	-0.062
30 †	<i>PMS2</i> hom	120	4.28	-0.039	0.061	0.145
22	<i>MSH6</i> hom	107	3.54	-0.04	-0.061	-0.105
29 †	<i>PMS2</i> hom	125	2.76	0.028	-0.03	0.053
8 *	<i>MSH6</i> comp het	105	2.08	-0.067	-0.053	-0.065
8 *	<i>MSH6</i> comp het	99	1.59	-0.07	-0.047	-0.083

Table 5.2: Genotype, sample scores and gMSI results of the 32 patients with constitutional mismatch repair deficiency. Each sample was scored by the described method. gMSI ratios for three markers (A, D2S123; B, D17S250; C D17S791) were calculated with the Peak Heights software (Ingham *et al*, 2013). Marker ratios presented here are the observed ratio minus the marker-specific threshold; positive values represent ratios above the threshold. If two or more of the gMSI markers are above the threshold the sample is classified as CMMRD. Thresholds were calculated as per Ingham *et al* (2013), using the same 40 controls as were used for the control distributions for score calculation (see Section 5.5). * patient 8, blood drawn whilst aplastic or recently recovered from aplasia. † patients homozygous for hypomorphic *PMS2* mutation.

5.7. Identification of Contamination in a Control Sample

Three control samples (selected by availability of DNA, Sample IDs: 6, 7, and 9) were included as repeats in sequencing runs 1, 2, 3, and 5 (unfortunately I neglected to include them on sequencing run 4 and 6). The repeats were scored as described before, and all but one fell within the range expected of controls (score = 0.00-0.87); sample 7 on sequencing run 5 (second, blinded cohort) scored 6.00 (Table 5.3). Whilst this repeat did have the lowest read depth, it is not exceptionally lower than other repeats, and the minimum marker depth was 746 reads. Furthermore, the observed prWT of each marker and the probability it was from a control population was compared between sequencing run 1 and sequencing run 5, and showed that multiple markers contribute to the unexpected high score in sample 7 (Table 5.4). This confirmed this wasn't an erroneous case due to low coverage of one or two markers.

To clarify the reason for this discordant result I wanted to confirm sample identity. The only genetic data available for sample 7 were the microsatellite markers and associated SNPs sequenced. Polymorphisms in our microsatellites are rare, and sample 7 contained none. However, the 33 SNPs of MAF >20% are suitable for sample identification due to the high likelihood that any two samples will have different alleles. The different bases detected

Sample ID	Sequencing Run	Read depth (per marker)	Score
6	1	3422	0.87
	2	2712	0.49
	3	1240	0.01
	5	1300	0.01
7	1	3043	0.04
	2	3610	0.18
	3	1650	0.52
	5	1102	6.00
9	1	4030	0.21
	2	4203	0.00
	3	1954	0.00
	5	1381	0.03

Table 5.3: Repeat testing of three control samples as a quality control. Three control samples were sequenced on four of the sequencing runs and scored as a rough indication of whether or not score may be affected by batch. Read depth is the mean number of reads detected per marker.

Marker	Sample ID: 7 (run 1)		Sample ID: 7 (run 5)	
	prWT	probability	prWT	probability
DEPDC2	0.998	0.292	1.000	1.000
GM01	1.000	1.000	0.995	0.036
GM07	0.983	0.189	0.956	0.000
GM09	1.000	1.000	1.000	1.000
GM11	0.998	0.198	1.000	1.000
GM14	0.991	0.097	1.000	1.000
GM17	1.000	1.000	1.000	1.000
GM22	1.000	1.000	0.993	0.006
GM26	1.000	1.000	1.000	1.000
GM29	1.000	1.000	1.000	1.000
IM16	0.995	0.119	1.000	1.000
IM49	0.979	0.039	1.000	1.000
LR10	1.000	1.000	0.990	0.030
LR11	1.000	1.000	0.990	0.036
LR17	0.990	0.139	0.972	0.001
LR20	1.000	1.000	1.000	1.000
LR24	1.000	1.000	0.996	0.018
LR36	1.000	1.000	0.986	0.018
LR40	0.998	0.266	1.000	1.000
LR44	0.984	0.313	0.987	0.510
LR46	1.000	1.000	1.000	1.000
LR48	0.996	0.478	0.967	0.001
LR49	1.000	1.000	0.992	0.003
LR52	0.981	0.702	0.964	0.108
	score: 0.04		score: 6.00	

Table 5.4: Summary of read data from sample 7 in two sequencing runs. The observed proportion of smSequences containing a WT length of microsatellite (prWT) and the probability the observation belongs to a control population in each microsatellite marker for sample 7 in two sequencing runs.

across the four sequencing runs for each of the SNP loci were summarised, and subsequently, the proportion of reads assigned to each base was determined for each run. Read proportions by allele were plotted and analysed to see if sample 7 had a distinct profile in sequencing run 5 compared to the previous runs. Indeed, it was evident that reads were assigned to the same bases in all runs except for sample 7 in sequencing run 5, which had a minority of reads assigned to novel alleles or significantly different proportions of reads

assigned to the same allele (Figure 5.8). This suggested the reaction was contaminated with another DNA. There were two possible points of contamination, either in the template DNA or in the sample indexing reverse primer, as all other samples gave scores as expected and template negative reactions were blank, suggesting common reagents were not the source of contamination. However, determining the exact source of contamination is not relevant to the work described in this thesis. Given that the scores from the other repeats and controls fell within the expected range of controls, equivalent to 13-100% probability the samples belong to a control population, and given that the one exceptional score is explained by contamination, this data suggests that the assay is stable. Due to the limited availability of CMMRD samples, a statistically robust analysis of assay reproducibility was not feasible.

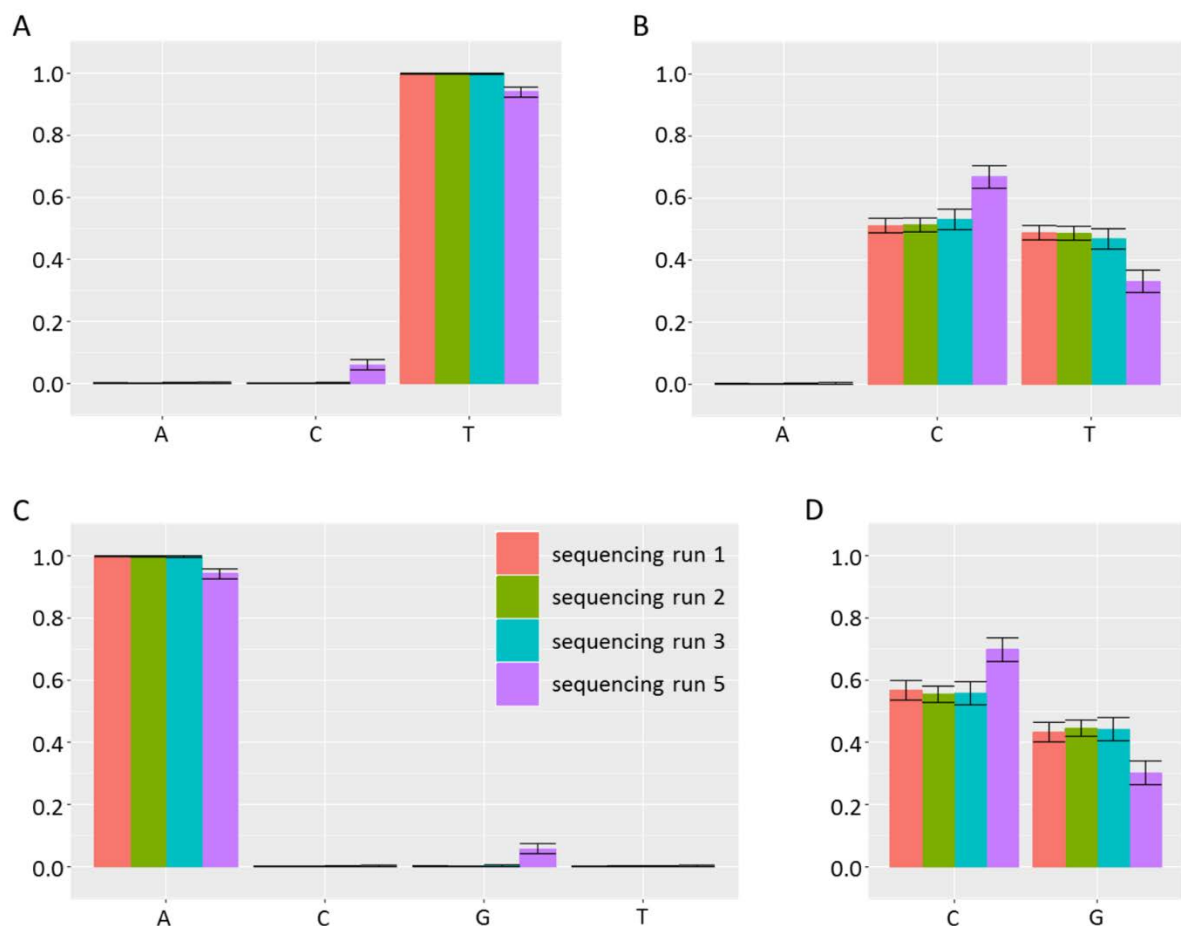


Figure 5.8: The proportion of reads assigned to different SNPs in sample 7 repeats. Sample 7 was sequenced across 4 different sequencing runs (numbered 1, 2, 3, and 5). In sequencing run 5, the second blinded cohort, sample 7 shows significant differences in the proportions of reads assigned to different SNPs in several of the microsatellite markers, compared to the runs 1, 2, and 3. Four markers are shown as illustrative examples, **(A)** DEPDC2, **(B)** LR48, **(C)** IM16, and **(D)** IM49. Error bars show 99% confidence intervals from the binomial distribution of read counts.

5.8. Discussion

Here, I have presented an MSI assay capable of detecting CMMRD with 100% accuracy by detection of low frequency variants in microsatellite length in PBL DNA samples. The assay is therefore applicable as a companion to genetic testing for CMMRD in suspected patients, such as those identified by the criteria of the C4CMMRD consortium guidelines (Wimmer *et al*, 2014). Also, the assay is a development of the smMIP-based MSI assay for CRC diagnostics, and hence has the benefits of simplicity, scalability and low cost, as discussed in section 4.11. Therefore, it is an appropriate screening tool in larger cohorts of patients. This is particularly pertinent as there are, approximately, only 200 known cases of CMMRD (Wimmer *et al*, 2017), yet it is estimated that up to one per 370-1000 of the population carry a heterozygous MMR mutation (Aaltonen *et al*, 1998; Hampel and de la Chapelle, 2011), implying that CMMRD may be more prevalent than the 200 known cases suggests. For example, adoption of germline genetic testing is required in childhood haematological malignancies as it is likely many are associated with unidentified germline mutations (Furutani and Shimamura, 2017). However, sequencing the many potential genes would incur a prohibitive cost, and therefore screening tools to guide differential diagnosis, such as the assay described here, are desirable. Unselected screening of patients with other childhood malignancies related to CMMRD is equally applicable.

Our current understanding of the CMMRD phenotype may be skewed by the ascertainment bias of current clinical guidelines. For example, the effect of ascertainment bias on the estimates of phenotype has been observed in Lynch syndrome, whereby analysis of affected family members of probands, rather than the probands themselves, showed a much later median age of disease onset (Hampel *et al*, 2005b). This can have a downstream effect on screening strategies, with early guidelines recommending screening in CRC patients under 50 years of age despite 70% of Lynch syndrome CRCs being diagnosed in patients over 50 years of age when screening of CRCs by MMR deficiency testing was applied (van Lier *et al*, 2012). Indeed, alongside guidelines for surveillance and management of CMMRD, the US Multi-Society Task Force on Colorectal Cancer raised the gaps in our knowledge that require further research, including the prevalence, disease spectrum, and genotype-phenotype correlations of CMMRD (Durno *et al*, 2017). Therefore, despite its rarity, identification of additional cases by large scale screening would begin to fill in these gaps. In summary, due to its scalability, the assay presented here has broad applicability to both clinical practice and research with respect to CMMRD.

The error prone amplification and sequencing of microsatellites (Fazekas *et al*, 2010) will obscure accurate detection of low frequency variants, such as those in microsatellite length found in the normal tissues of CMMRD patients (Bodo *et al*, 2015). Therefore, to address the difficulty of detecting low frequency variants in microsatellites I used molecular barcodes to group reads originating from the same template molecule so that PCR and sequencing errors could be identified and a hypothetical smSequence, assumed to represent the sequence content of a single template DNA molecule, used in variant detection (Casbon *et al*, 2011). Very recently Waalkes *et al* (2018) used smMIPs to sequence a panel of 111 long (16-40bp) microsatellites for MSI testing in cancer diagnostics. They used molecular barcodes in a very similar method, according to Carlson *et al* (2015), to show a reduction in the number of length variants detected in microsatellites, suggesting a reduction in error. Here, I have confirmed that use of smSequences significantly reduces the error of detecting variants in microsatellite length using 40 control samples which can be assumed should not contain detectable microsatellite mutations. In combination with our selective use of short (7-12bp) markers known to have lower error rates (Fazekas *et al*, 2010), the use of smSequences frequently removed all PCR and sequencing error. However, marker GM09 showed an increased error rate in four of the control samples. These samples, and others, had very low counts of smSequences in GM09 despite unremarkable counts for total reads in GM09 and counts for smSequences in other markers. Therefore, this result is difficult to explain. For example, if there was a variable in these samples interfering with target capture or amplification of GM09 it would be expected that this would affect other markers as well or would be detectable in the read depth of GM09. An alternative is that there was an error in probe synthesis or the bioinformatic pipeline that makes detection of molecular barcodes for this marker unreliable. However, this is speculation and would require further research. Regardless, use of the full marker panel, including GM09, was able to accurately detect CMMRD patients. However, GM09 could be excluded in future analyses, and the modified panel validated in a set of independent samples.

By characterising the proportion of smSequences with a WT microsatellite length in 40 control samples, I generated distributions reflective of the control population, against which samples could be scored. This removed the need to characterise the distribution in the CMMRD population, which would be limited by the scarcity of CMMRD samples, and heterogeneity of germline mutations. To assess assay performance, I analysed samples from 32 genetically-confirmed CMMRD patients, which is a relatively large cohort given that only

~200 patients have been published so far (Wimmer *et al*, 2017). This cohort covers biallelic mutation of each of the four MMR genes to address genetic heterogeneity of this syndrome, and the assay was able to separate all CMMRD samples from controls irrespective of mutation (Figure 5.7). The probability a sample was not from a control population was converted into an easily described score such that higher scores indicated increased MSI and increased likelihood of CMMRD. Using a score threshold of >1.53 (>97% probability the sample is not from a control population), the assay detected CMMRD with 100% accuracy, irrespective of which MMR gene was mutated (Figure 1). Naturally, a score threshold should be picked *a priori* and validated, but the scarcity of CMMRD samples prevented threshold validation. Instead, standard thresholds used for statistical probabilities can be applied. For example, a score threshold of >1.30, equivalent to >95% probability, would have generated two false positives (100% sensitivity, 98% specificity). A more conservative score threshold of >2.00, equivalent to >99% probability, would have missed only one CMMRD sample (97% sensitivity, 100% specificity, Sample ID: 99). However, this sample was collected during exceptional circumstances, discussed below, and hence a score threshold of >2.00 ensures high specificity, which is particularly important if used as a screening tool in larger cohorts.

Included within the cohort was a variety of samples, which allowed a limited analysis of variables that may affect score, such as hypomorphic MMR mutation, patient age, and clinical history. Samples from patients 29, 30 and 31, homozygous for a hypomorphic *PMS2* variant, all scored relatively low (score = 2.76-5.90; Table 5.2), which may be indicative of their residual MMR activity and attenuated phenotype (Li *et al*, 2015). Therefore, it may be worth testing if assay score has any prognostic value, for example by indicating the penetrance of germline mutations.

Three samples from patient 8 (score = 1.59-4.62) scored much lower than the sample from an affected sibling, patient 9 (score = 19.09; Table 5.2). At the time of blood draws, patient 8 was aplastic, or just recovered from aplasia, due to chemotherapy for T cell lymphoma. A decreased frequency of microsatellite length variants in repopulating PBLs in CMMRD patients is consistent with the observation in mice that hematopoietic stem cells with a higher burden of microsatellite mutation are associated with defective repopulation (Reese *et al*, 2003). Whilst this is speculative, it may be pertinent to avoid using blood samples drawn from aplastic patients, or to interpret negative results as inconclusive if such samples are unavoidable. In the future, a time course of samples taken during patient recovery from aplasia may be informative.

We also observed relatively lower scores from patient 22 (score = 3.54-7.39) compared to an affected sibling, patient 21 (score = 17.61; Table 5.2), who is 8-9 years older. In this case, patient 22 was only 13 and 15 months old at blood draws, and had not presented with cancer. An association between age and frequency of microsatellite length variants has been observed in the general population (Coolbaugh-Murphy *et al*, 2005) and may be applicable to CMMRD patients. However, formal analysis of the effect of age, and other, variables on score would require larger numbers of samples and patients, and is beyond the scope of this work.

Due to the limitation in the number of CMMRD samples and quantity of each sample provided for this work, it was not feasible to assess other parameters of assay performance, despite the recommendations from the Association for Molecular Pathology and the College of American Pathologists for the analytical validation of NGS-based diagnostic tests (Jennings *et al*, 2017). For example, assessment of assay reproducibility would require repeat testing of the majority of samples, yet only a handful of the CMMRD samples have sufficient DNA remaining to do so. Given more time and access to additional samples these assay parameters could be assessed. One issue raised by these guidelines is the sensitivity of NGS-based diagnostic tests to contamination and the need for quality checks in sample handling. In the repeat testing of three control samples, it was found that a repeat of sample 7 gave an unexpected high score. In this case, the SNPs associated with the microsatellite markers were able to detect contaminating DNA by detection of novel or changed representation of bases at the SNP loci in the affected sequencing run compared to previous runs. It is feasible that this use of SNPs could be explored further to give an additional assay QC for contamination, for both CMMRD and cancer diagnostics. Furthermore, even small quantities of contamination, depending on the source, could be critical to detection of CMMRD given that assay score is sensitive to very small changes in variant frequency (Table 5.1). Whilst assessment of contamination would not be of general utility to clinical diagnostic services, in which appropriate controls will be in place to prevent contamination, the analysis of the SNPs provides a means to explore unexpected or inconsistent results.

5.9. Conclusions and Future Work

In conclusion, a smMIP and sequencing-based assay that utilises molecular barcodes can detect low frequency variants in microsatellite length in PBLs of CMMRD patients with high accuracy, including CMMRD caused by *MSH6* mutation. The laboratory workflow is simple

and analysis is automated, applicable to rapid turnaround times within clinical decision windows. Therefore, the assay could be deployed immediately to compliment genetic testing. Being cheap and designed for high throughput diagnostics, the assay could also be used for screening in larger cohorts of patients, for example in cases of childhood, haematological malignancy. Such unselected screening of relevant childhood malignancies will improve our understanding of CMMRD, including its frequency, phenotype and disease spectrum. Given additional samples and a more thorough understanding of the distribution of scores in the CMMRD population, a more robust classifier could be defined. Finally, it is apparent that several factors, including age, mutation pathogenicity, and leukocyte repopulation, may affect the frequency of variants in microsatellite length in PBLs, and therefore sample score. The assay could be used to explore these biological mechanisms in more detail in the context of CMMRD, which may improve score interpretation and provide novel insights into disease mechanisms.

Chapter 6. General Discussion and Future Work

6.1. The Clinical Utility and Analytical Validity of Mismatch Repair Deficiency Biomarkers and Tests in Cancer Diagnostics

The overall utility of biomarker tests can be assessed by their analytical validity and clinical utility (Hayes, 2018). As discussed in Section 1.7, several factors inform analytical validity, including accuracy of results, concordance of repeat testing, robustness to sample and technical variables, and clinical validation. Clinical utility can be judged by assay influence on clinical decisions, its adverse effects to the patient, its cost and practicality, and generation of measurable improvement in healthcare practice and patient outcomes (Ray *et al*, 2010; Henry and Hayes, 2012).

MMR deficiency affects approximately 1 in 6 of all CRCs (Thibodeau *et al*, 1998) and is present in 90-100% of CRCs diagnosed in the context of Lynch syndrome (Leach *et al*, 1996; Liu *et al*, 1996; Thibodeau *et al*, 1996; Hampel *et al*, 2008). Furthermore, all tissues of CMMRD patients, normal and neoplastic, are MMRd (Wimmer *et al*, 2008). In the broadest sense, detection of MMR deficiency is important to clinicians and patients as it can be used to identify associated cancer-predisposition syndromes and informs disease management due to the unique properties of MMRd relative to MMRp tumours (Bodo *et al*, 2015; Le *et al*, 2015; Newland *et al*, 2017). Here, I will summarise the clinical utility and analytical validity of each of the biomarkers and assays of MMR deficiency investigated or developed during this work.

6.1.1. *Anti-frameshift peptide antibodies as a liquid biopsy biomarker of colorectal cancer*

The “holy grail” of cancer diagnostics is considered to be early detection due to the greater variety of treatment options available to, and the improved survival of, patients with lower stage disease (Etzioni *et al*, 2003). MMRd CRCs are a distinct molecular subtype (Guinney *et al*, 2015) and the immunogenicity of these tumours provides a potential source of antibody-based biomarkers for early detection of disease (Reuschenbach *et al*, 2010). My study of α FSP-Abs titres (Chapter 3) was aiming to find a novel biomarker of MMRd CRC using a recently published and novel technique. Had initial findings been promising it would have been feasible to develop the method into a biomarker assay, applicable to clinical surveillance in patients at high risk of MMRd CRC, such as Lynch syndrome gene carriers. Given the invasiveness and limitations of colonoscopy in detection and prevention of some

Lynch syndrome colorectal tumours (Stuckless *et al*, 2012; Seppälä *et al*, 2017), the clinical utility of such a surveillance assay could have been high.

FSP serum reactivity was found to be associated with a history of CRC in the cohort of Lynch syndrome gene carriers analysed, suggesting that α FSP-Abs are detectable and in agreement with previous studies that have used the alternative technique of ELISA to quantify α FSP-Ab titres in MSI-high CRC patients (Ishikawa *et al*, 2003; Reuschenbach *et al*, 2010). However, my results suggest the individual and collective sensitivity of FSP serum reactivity for CRC is low as the majority of values from patients with a history of CRC fell within the background noise, as was also observed for three samples where CRC was diagnosed shortly after blood draw. It is conceivable that the sensitivity could be improved by optimising the method, but cancer-free and MMR mutation-negative controls were lacking, and to address this was not possible due to financial and logistical constraints. However, given the clear signals from positive control samples (patients vaccinated with synthetic FSPs) that were not observed in the CRC patients, I believe that α FSP-Abs are insensitive biomarkers for early detection of CRC possibly due to the lack, or instability, of a humoral immune response to cMNR frameshift mutations. In conclusion, α FSP-Abs are likely to have low clinical utility, but additional research is needed to confirm this, starting with technical validation of the method (Section 3.9). The analytical validity of the method was not formally assessed due to limitations of the cohort analysed, such as the low number of on-trial CRC diagnose and lack of controls. Alternative biomarkers for liquid biopsy-based surveillance of Lynch syndrome gene carriers should be considered (see Section 6.2.1).

6.1.2. A sequencing-based microsatellite instability assay for colorectal cancer diagnostics

NICE Diagnostic Guidance 27 states that all CRCs should be tested for MMR deficiency to screen for Lynch syndrome, either by loss of MMR protein expression or detection of high levels of MSI in the tumour, such that patients and affected family members can benefit from altered treatment and surveillance (Newland *et al*, 2017). These guidelines are based on the high accuracy and cost-effectiveness of unselected molecular screening, using the established techniques of IHC or FLA (Pérez-Carbonell *et al*, 2012; Snowsill *et al*, 2014). MMR deficiency also informs use of immune checkpoint blockade, with pembrolizumab recently receiving FDA-approval as a second-line therapy in MSI-high solid cancers, irrespective of the tissue of origin (MERCK & Co. Inc, 2017). Despite the accepted analytical validity and clinical utility of these biomarker assays, they are not suitable for high throughput MMR deficiency

testing due to reliance on case-by-case and marker-by-marker results interpretation. Furthermore, to fulfil Lynch syndrome screening guidelines separate tests for *BRAF* V600E or *MLH1* promoter methylation are needed.

I developed a sequencing-based MSI assay from the assay of short MNRs described by Redford *et al* (2018), adopting smMIP technology to multiplex the markers, and expanding the panel to include additional short MNRs and the *BRAF* V600 locus (Chapter 4). The assay classifier was trained and validated, achieving 100% accuracy relative to FLA by MSI Analysis System (Promega), and giving 100% results concordance on repeat testing. The MSI assay was also robust to sample variables, including low quality template DNA from FFPE tissue, low MSI-high content, and low quantity template DNA. Accurate classification was achieved from as few as 75 molecular barcodes/marker. Following the guidelines of Jennings *et al* (2017) for the analytical validation of NGS-based oncology assays, assay QCs were also defined. *BRAF* V600E testing was included in the test to reduce the Lynch syndrome screening pipeline to one screening test, to be followed by germline genetic testing of MMR genes. The cost and TAT of the assay is equivalent to, and arguably better than, the dominant MSI Analysis System (Promega), and the laboratory workflow and analysis are both fully automatable. These results demonstrate the analytical validity and clinical utility of the smMIP and sequencing-based MSI assay, and in several aspects the assay is superior to those used in current practice, most notably in its scalability. Hence, it is suitable for deployment and clinical validation. Alongside clinical validation, additional improvements can be made to the assay, such as reduction in the number of markers analysed to reduce sequencing costs, and transfer of the assay into a kit to further streamline the protocol (Section 4.12).

6.1.3. A sequencing-based microsatellite instability assay to detect constitutional mismatch repair deficiency

Identification of CMMRD is critical for the appropriate clinical management of the patient, with patients benefitting from surveillance and altered treatment (Vasen *et al*, 2014). Due to a pleiotropic phenotype, genetic diagnosis by detection of pathogenic mutations affecting both alleles of the same MMR gene is the gold standard. However, MMR VUS and *PMS2* pseudogenes can confound genetic diagnosis and so companion diagnostic tools are needed (Wimmer *et al*, 2017). CMMRD can be identified by MSI testing of normal, non-neoplastic

tissues using highly sensitive techniques, but current assays are laborious or insensitive to MSH6 deficiency (Ingham *et al*, 2013; Bodo *et al*, 2015).

The smMIP and sequencing-based MSI assay was repurposed, using a novel analysis method capable of detecting low level MSI in the PBLs of CMMRD patients (Chapter 5). It incorporates molecular barcodes into reads to facilitate reduction of PCR and sequencing error in microsatellites for accurate detection of low frequency length variants. Other methods to detect low frequency variants have been developed, such as Safe-SeqS (Kinde *et al*, 2011), Duplex Sequencing (Schmitt *et al*, 2012; Kennedy *et al*, 2014) and CypherSeq (Gregory *et al*, 2016). However, the original methods are not optimised to target specific loci and instead shear the template DNA, ligate molecular barcodes to the non-specific DNA fragments, universally amplify the fragments, and then sequence the amplicons. Adapting these protocols to targeted sequencing is limited. SafeSeqS, for example, was further adapted to target specific loci using an initial two-round PCR with primers to introduce the molecular barcodes, followed by universal amplification (Kinde *et al*, 2011). However, the reliance on PCR to initially capture the targets limits multiplexing, and multiplexing was not demonstrated by Kinde *et al* (2011), only discussed. Similar, more recent, PCR-based methods for introducing molecular barcodes to amplicons can multiplex tens of markers (Ståhlberg *et al*, 2016), whereas smMIPs can be multiplexed in their thousands (Hiatt *et al*, 2013). The concentration of each smMIP in a pool can also be balanced to equalise read depth between markers with ease, as shown in this work and by others (Niedzicka *et al*, 2016). The smMIP protocol is also simple and has been shown to be fully automatable (Neveling *et al*, 2017).

The assay uses short MNRs that are sensitive to MSH6 deficiency so that it is informative irrespective of the affected MMR gene. The clinical utility of the assay was evident in the high accuracy that was achieved, with perfect separation of CMMRD samples from controls, making it a suitable complementary test for uncertain genetic diagnoses. In addition, its low cost and scalability could facilitate screening of large cohorts of patients, such as cases of childhood leukaemia (Furutani and Shimamura, 2017). With respect to analytical validity, it was difficult to formally test the assay in the context of CMMRD diagnostics given the scarcity of CMMRD samples; the 32 patients analysed are a relatively large cohort with respect to the total patient population, accounting for approximately 15% of all known cases in the literature (Wimmer *et al*, 2017). However, the assay is evidently highly sensitive and specific, uses high quality and quantity DNA extracted from PBLs, and

uses the same robust protocols as used in the MSI assay for cancer diagnostics, suggesting it is appropriate for clinical deployment.

6.2. The Future Direction of Biomarker Tests for Mismatch Repair Deficiency

The work presented in this thesis has an impact on MMR deficiency testing in cancer diagnostics beyond the interpretation of clinical utility and analytical validity of the individual biomarkers and tests described.

6.2.1. Surveillance for mismatch repair deficient cancers in high risk populations

α FSP-Abs were insensitive for the early detection of on-trial CRCs in three patients, yet the gold standard technique for the early detection of CRC in at-risk populations remains colonoscopy, a relatively invasive procedure that fails to detect 7-20% of colorectal tumours in Lynch syndrome (Stuckless *et al*, 2012; Ahadova *et al*, 2015). α FSP-Abs are derived from the frequent cMNR frameshift mutations found in MMRd cancers. Therefore, as generation of α FSP-Abs against an MMRd cancer appears to be an infrequent or undetectable event, detection of the frameshift mutations in cfDNA, indicative of ctDNA from an MMRd cancer, could be a more tractable biomarker for Lynch syndrome surveillance. As discussed in Section 1.8, ctDNA can be used to monitor cancer progression and relapse (Diaz *et al*, 2012; Taly *et al*, 2013; Siravegna *et al*, 2015; Schøler *et al*, 2017), and several studies have reported fractions of ctDNA within cfDNA ranging between 0.01% and 93% by detection of cancer-associated variants (Jahr *et al*, 2001; Diehl *et al*, 2005). The lowest ctDNA fractions are found in early stage disease (0.01-1.7%) (Jahr *et al*, 2001). Hence, for the early detection of cancer, highly sensitive but costly and laborious techniques are needed to detect low frequency variants (<1%) from ctDNA in a background of WT cfDNA, such as picodroplet digital PCR (Taly *et al*, 2013; Bettegowda *et al*, 2014) and high read depth NGS (Shu *et al*, 2017).

Here, I have described the use of smMIPs and molecular barcodes to reduce PCR and sequencing error such that CMMRD samples can be separated from controls by detection of low frequency microsatellite length variants present in genomic DNA from PBLs. Based on this experience, it would be of interest to see if an smMIP-based assay can detect cMNR frameshift mutations in cfDNA for potential use as a liquid biopsy-based biomarker of MMRd cancer. Sequencing and molecular barcoding of reads has previously been used to detect mutations at VAFs as low as 0.1% in cfDNA extracted from blood plasma (Ståhlberg *et al*, 2016). Similar sensitivity was achieved when smMIPs and their molecular barcodes were

used to analyse 33 mutations across clinically informative cancer genes in cell line and tumour samples, although cfDNA was not tested (Hiatt *et al*, 2013). Typically, cMNRs range between 5-13bp (Woerner *et al*, 2010), which is comparable to the short MNRs analysed for CMMRD detection. Hence, despite the high error rate in sequencing of microsatellites (Fazekas *et al*, 2010), it is reasonable to assume that molecular barcodes will also allow error reduction and detection of low frequency variants in cMNRs. Furthermore, due to their robustness to multiplexing (O’Roak *et al*, 2012), smMIPs can analyse many cancer-associated loci simultaneously providing a relatively cheap screening assay, as was achieved with our panel of 24 short MNRs. The high frequency of cMNR frameshift mutations in MMRd CRCs, with many occurring in >60% (Woerner *et al*, 2010), means a reliable panel could be constructed for cancer detection without prior knowledge of mutation status. As cMNRs frameshifts are driver mutations, it is also more likely that they will be detectable compared to the marker of short MNRs used in the MSI test, and, due to the differences in cMNR frameshifts between different cancer types (Woerner *et al*, 2010), they may provide information on the type of cancer. However, it is also possible that the current assay of short MNRs can detect MSI in cfDNA as a biomarker of MMRd cancer. Finally, as was hypothesised of α FSP-Abs, detection of MSI in cfDNA may additionally give a measure of the mutational burden of a cancer and therefore its likely response to immunotherapy (Topalian *et al*, 2016; Yarchoan *et al*, 2017).

There are several disadvantages of analysing cfDNA rather than antibody titres. Whilst antibodies are relatively stable in serum (Anderson and LaBaer, 2005), serum cfDNA can be contaminated by lysis of PBLs and hence rapid processing of the blood sample or use of stabilising reagents is required (Norton *et al*, 2013). However, this can be addressed by setting up the appropriate clinical pipeline for sample workup. cfDNA also tends to be low quantity and highly fragmented (Jahr *et al*, 2001). By assessing the robustness of the MSI test for cancer diagnostics, I showed that amplicons were visible from as little as 3.13ng of template DNA, a quantity of sample that is achievable from the median 184ng of cfDNA extracted from 1ml of patient serum (Fong *et al*, 2009). Also, the size of the smMIP annealing site (120-150bp) is less than the average 180-220bp fragment size of cfDNA (Jahr *et al*, 2001). In conclusion, it would be pertinent to assess the applicability of the smMIP, sequencing and data analysis methods presented here to the detection of cMNR frameshift mutations, or short MNR instability, in cfDNA, with a view to developing an assay for surveillance in Lynch syndrome gene carriers.

6.2.2. Microsatellite instability testing in cancer diagnostics

A plethora of methods have been presented in recent years with respect to MMR deficiency testing of cancers to screen for Lynch syndrome and, more recently, with a view to companion diagnostics for immune checkpoint blockade therapy. Beyond the established techniques of IHC and FLA, and the growing application of NGS-based methods, some research groups have developed alternative assays. For example, N_LysT is a PCR and high resolution melting curve (HRM)-based assay that detects microsatellite length variants by differential melting temperatures of amplicons. The key advantage of this approach is the rapid TAT as both PCR amplification and HRM analysis of amplicons are conducted in one tube using one thermocycler program (Susanti *et al*, 2018). However, the use of HRM means that only one marker can be analysed at a time, and whilst individual markers can show high (>95%) sensitivity for the MSI-high phenotype (Findeisen *et al*, 2005), multiple markers must be analysed due to the continuum of the MSI spectrum from MSS and MSI-low phenotypes in MMRp cancers to the MSI-high phenotype in MMRd cancers (Boland *et al*, 1997).

Parallel analysis of multiple microsatellite markers in many samples requires techniques capable of generating large quantities of data, such as NGS. Hence novel NGS-based methods of assessing MSI have been appearing in the literature for 5 years or more, particularly with a focus on software that analyses microsatellites captured in whole genome, whole exome, or gene panel sequencing (Lu *et al*, 2013; Niu *et al*, 2014; Salipante *et al*, 2014; Gray *et al*, 2018; Hampel *et al*, 2018; Zhu *et al*, 2018). However, the high cost of such approaches (Marino *et al*, 2018) is prohibitive to their use for screening the 41,000 CRCs diagnosed each year in the UK (Cancer Research UK Statistics, 2015), as recommended by NICE and others (Balmana *et al*, 2013; Stoffel *et al*, 2015; Newland *et al*, 2017). Hence cheaper MSI assays are required, that can also benefit from the advantages of NGS. Here, I have shown that the smMIP and sequencing-based MSI assay has a superior cost per sample than gene panel, whole exome, or whole genome sequencing. Like other NGS-based assays, it uses the Illumina sequencing platforms that are ubiquitous in research and clinical laboratories around the world (Levy and Myers, 2016), and multiplexes multiple microsatellites and other clinically actionable biomarkers to optimise clinical testing.

The low cost and scalability of the smMIP and sequencing-based MSI assay make it a competitive option for routine diagnostics. However, an argument against targeted sequencing assays is that, if gene panel sequencing is to become routine in cancer health care, why not deploy extensive sequencing assays now? For example, it is estimated that

9.9% of all CRCs are due to mutations associated with hereditary cancer syndromes, and only a third of these are attributable to Lynch syndrome (Yurgelun *et al*, 2017). Therefore, germline gene panel sequencing of CRC patients may be more appropriate than screening for Lynch syndrome only. The costs and QALYs-gained from gene panel testing of germline DNA in patients with >5% probability of hereditary cancer (based on a predictive, clinical algorithm) have recently been modelled, using two approaches, either germline gene panel testing, or MMR IHC of the tumour, followed by germline MMR gene testing according to Lynch syndrome screening recommendations. When immediate germline gene panel testing was applied, 8,076 (29.1%) patients were found to have a hereditary cancer syndrome, with an additional 5,984 affected, first degree relatives identified through cascade testing. When Lynch syndrome screening was applied, 2,584 (9.3%) patients were found to have Lynch syndrome, with an additional 1,915 first degree relatives also testing positive for MMR gene mutation. From the appropriate clinical management of identified patients, it was estimated that gene panel sequencing had a cost of \$1,543 per QALY-gained, whilst Lynch syndrome screening had a cost of \$1,882 per QALY-gained, suggesting that germline, gene panel sequencing of CRC patients suspected of hereditary disease is more cost-effective than screening for Lynch syndrome-only in the same patient cohort (Gu *et al*, 2018). However, these analyses were performed in preselected patients, which will reduce costs per QALY-gained, but will also reduce screening sensitivity. For Lynch syndrome screening, it has previously been shown that Bethesda criteria, which rely on age at diagnosis, family history of Lynch spectrum cancers, etc, are less sensitive and less specific for identification of Lynch syndrome gene carriers than molecular screening for MMR deficiency in CRC patients (Pérez-Carbonell *et al*, 2012). Therefore, such analyses of cost-effectiveness can be skewed in favour of gene panel sequencing as they do not account for the cases missed by pre-selection of patients based on clinical criteria. Finally, the more genes that are sequenced without guidance as to which genes are likely to contain the causative mutation, the more time will be required to interpret VUS. For example, in their study of germline, gene panel sequencing of 1112 CRC patients suspected of hereditary disease, Yurgelun *et al* (2015) identified at least one VUS in 479 patients. It is, therefore, perhaps more appropriate to consider economic models that use multiple diagnostic pathways for the application of NGS to clinical oncology rather than a “one-size-fits-all” approach. In such an economic model of the CRC diagnostic pipeline, cheap, high throughput screening tools, such as the smMIP and sequencing-based MSI assay presented in this work, would be the first line of reference for

the clinician, ensuring rapid identification of those patients with germline, pathogenic mutations where clinical guidelines exist for altered treatment. Such rapid screening is not mutually exclusive with germline gene panel sequencing of additional cases of interest, for example where clinical criteria elicit suspicion of hereditary disease.

The health technology assessment by Snowsill *et al* (2014) concluded that unselected screening for Lynch syndrome by MSI testing of all CRCs diagnosed under the age of 70 years, followed by *BRAF* V600E testing of MSI-high CRCs to remove sporadic cases, and, finally, germline MMR gene testing, had an incremental cost-effectiveness ratio of £5,491 per QALY-gained relative to no testing (diagnosis of Lynch syndrome based on clinical criteria), which was well below the NICE cost-effectiveness threshold of £20,000 per QALY-gained. In their diagnostic guidance, NICE expanded these recommendations to include all CRC diagnoses irrespective of age (Newland *et al*, 2017). However, Lynch syndrome gene carriers are also at high risk of multiple, other cancer types, most prominently endometrial, ovarian, upper GI, and urinary tract cancers (Møller *et al*, 2017b), and it appears that the majority of these cancer types are affected by MMR deficiency (Gurin *et al*, 1999; Simpson *et al*, 2001; Hampel *et al*, 2006; Gylling *et al*, 2007; Pal *et al*, 2008; van der Post *et al*, 2010). The most frequent cancer in female Lynch syndrome gene carriers is EC, not CRC (Hampel *et al*, 2006). Strategies to identify Lynch syndrome based on EC result in very similar conclusions as strategies using CRC diagnoses. Primarily, unselected molecular screening of tumours for MMR deficiency followed by germline genetic testing reveal that 1.8% (95% CI: 0.9-3.5%) of ECs are due to Lynch syndrome, similar to rates observed in CRC (Hampel *et al*, 2006), and inclusion of *MLH1* methylation testing can exclude sporadic cases to improve screening specificity (Leenen *et al*, 2012). Furthermore, 41% of Lynch syndrome patients identified by molecular screening had no indicators of Lynch syndrome based on Bethesda criteria or other clinical features (Mills *et al*, 2014), showing MMR deficiency testing to be the superior screening strategy in terms of sensitivity and specificity. Finally, cost-effectiveness analyses have shown that screening for Lynch syndrome by MMR deficiency testing of ECs diagnosed under the age of 70 is cost-effective, with incremental cost-effectiveness ratio of €6,668 per life year gained relative to screening by Bethesda guidelines, which had only 43% sensitivity in the cohort tested in parallel to cost analyses (Goverde *et al*, 2016); cost effectiveness is therefore equivalent to MMR deficiency testing of CRCs (Snowsill *et al*, 2014). Whilst clinical guidelines do not exist yet for MMR deficiency testing of extracolonic cancers to screen for Lynch syndrome, the cheap, automatable and scalable MSI test described in this thesis

would be able to meet the increasing demand. Indeed, as the number of samples to be analysed increases, the per sample cost of the assay decreases, as higher capacity sequencing kits can be used (Appendix J).

Another application of NGS gene panel, whole exome, or whole genome sequencing is the identification of somatic mutations in the tumour that inform therapeutic choice. In CRC, gain of function mutations in *BRAF*, *KRAS*, *NRAS*, or *PIK3CA* are clinically informative as they predict response to anti-EGFR therapy (Lièvre *et al*, 2006; De Roock *et al*, 2010a). From the plethora of evidence in the literature, NICE Technology Appraisal 439 states that cetuximab and panitumumab (monoclonal antibodies that block EGFR signalling) should be used as first line therapy only in *RAS* wild type metastatic CRC (Cooper *et al*, 2017). As discussed in Section 4.11, a key advantage of an smMIP-based assay is that it is modular, with it being relatively trivial to add smMIPs targeting additional biomarkers into the multiplex. This was shown using a smMIP targeting the *KRAS* G12 and G13 mutation hotspot. MSI testing informs use of pembrolizumab, or other immune checkpoint blockade therapies, following FDA approval of pembrolizumab as a second line therapy for any MSI-high solid cancer irrespective of the tissue of origin (MERCK & Co. Inc, 2017). With ongoing clinical trials to confirm the efficacy of immune checkpoint blockade in MMRd cancers (Cummings and Garon, 2017), its use as a first line therapy will likely increase in coming years, further fuelling the demand for high throughput MSI testing in a broad spectrum of cancer types. Finally, the score generated by the assay covers a broad scale that allows quantification of the MSI signal, rather than the tripartite classification using FLA. Whilst this was not explored here due to a lack of clinical or pathology data on the CRC patients and their tumour, it would be possible to correlate classifier score with a variety of disease phenotypes, such as patient age, prognosis, tumour stage, genetic background, and so on. Specifically, it would be interesting to see if assay score within MSI-high samples correlates with tumour response to immune checkpoint blockade, as has been shown for tumour mutational burden (Yarchoan *et al*, 2017).

Due to its modularity, an smMIP-based MSI test could also be tailored to different cancer types to maximise the number of clinically relevant biomarkers analysed for each tumour tested, making it competitive with gene panel, whole exome, and whole genome sequencing. It is also worth considering that actionable mutations, such as *RAS* gene mutations, occur in hotspots. For example, *KRAS* G12 and G13 mutations account for more than 90% of *RAS* gene mutations in CRC (Rajagopalan *et al*, 2002), and therefore select loci

could be included with minimal additional smMIPs. By targeting hotspot loci, an smMIP-based assay would have a much lower price point than more extensive sequencing methods, and the 3-5 day TAT would allow rapid profiling of tumours to extract the most relevant information within treatment decision windows. Again, it is worth considering that cheap, targeted assays are not mutually exclusive with gene panel, whole exome, or whole genome sequencing – it is justifiable to reserve more expensive but comprehensive sequencing to cases of interest rather than apply these as the front line diagnostic tool. A final argument often used in favour of gene panel, whole exome, or whole genome sequencing is the ever falling cost of NGS (Horak *et al*, 2016). However, the MSI assay presented is also an NGS-based method and hence would benefit from these cost reductions; given that it is already cheaper than the dominant MSI Analysis System (Promega), this only argues in favour of targeted sequencing.

Ultimately, the development of this assay was driven by a need to improve the uptake of MMR deficiency testing, and the true measure of an assays clinical utility is in the improved outcomes for patients (Ray *et al*, 2010; Henry and Hayes, 2012). Current estimates of the rate of clinical uptake of MMR deficiency testing and the number of known Lynch syndrome gene carriers are becoming outdated. For example, the most recent estimates of clinical uptake are based on data from 2010-2012 (Shaikh *et al*, 2018), and estimates of what percentage of Lynch syndrome gene carriers are known have not been updated since 2011 to my knowledge (Hampel and de la Chapelle, 2011). Therefore, with deployment of the smMIP and sequencing-based MSI assay into local clinical practice, it would be pertinent to audit the rate of MMR deficiency testing in recent years to formally assess the effect of both the guidelines that recommend MMR deficiency testing, such as NICE DG27 (Newland *et al*, 2017), and the advances in available technologies.

6.2.3. Microsatellite instability testing of non-neoplastic tissues

MMR deficiency affects all tissues of CMMRD patients and the low level MSI in the non-neoplastic tissues is a biomarker by which CMMRD can be identified. Advantages of a blood-based assay are that it is minimally invasive and does not rely on the excision of tumour tissue, and therefore a diagnosis can be determined prior to surgery and other treatment decision. By applying the smMIP and sequencing-based assay to screening large cohorts of patients, for example those affected by childhood haematological malignancy (Furutani and Shimamura, 2017), the phenotypic spectrum of CMMRD will be better understood (Durno *et*

al, 2017) and the assay refined. One point of assay refinement is the panel of markers used. As *PMS2* is the predominantly affected gene in CMMRD (Wimmer *et al*, 2017) and is the arguably the most difficult to interpret, due to pseudogenes (Nakagawa *et al*, 2004) and being the poorest annotated of the MMR genes in variant databases (Blount and Prakash, 2017), it would be appropriate to analyse a wider spectrum of microsatellite markers to find those that are most sensitive and specific for *PMS2* deficiency. For example, although insensitive for *MSH6* deficiency, the DNRs used in the gMSI assay of Ingham *et al* (2013) achieve high accuracy for CMMRD detection using a simple PCR-based protocol and analysis. Additional, longer MNRs, and di-, tri- and tetra-nucleotide repeats are all candidate markers, and a reduction in the error of sequencing these different types of microsatellite, by use of molecular barcodes, has been shown in the literature (Carlson *et al*, 2015; Waalkes *et al*, 2018), suggesting they too could be used to detect low-level MSI. Furthermore, the short MNRs used in the panel of the smMIP and sequencing-based MSI assay were selected for their instability in MSI-high CRCs, and it should not be assumed that markers sensitive to MMR deficiency in cancer are the most sensitive to MMR deficiency in normal tissues.

Differences between cancers and normal tissue are also apparent in the level of MSI detected. Comparing the results in Chapters 4 and 5, it is evident that the rate of indel mutations in microsatellites is greatly increased in MMRd CRCs compared to MMRd non-neoplastic PBLs. A likely explanation for this is the mono- or oligo-clonality of cancers, whereby the majority of tumour cells originate from one dominant clone and are, therefore, genetically homogeneous compared to precursor lesions, as shown by whole exome sequencing of CRCs and colorectal adenomas (Cross *et al*, 2018). In contrast, the PBL population is derived from a heterogeneous population of hematopoietic stem cells, especially in young individuals, as modelled *in silico* and in mouse models (Roeder *et al*, 2008) and more recently shown by whole exome sequencing of 12,380 patients (Genovese *et al*, 2014). Indeed, clonal (rather than polyclonal) haematopoiesis is seen as an aberration of age and is associated with risk of haematological malignancy (Genovese *et al*, 2014). Therefore, variants in microsatellite length may be common to CMMRD PBLs, but they will not be represented throughout the PBL population and hence individual variants occur at a low frequency and are difficult to detect. Interestingly, ultra-hypermutated glioblastomas diagnosed from either a sporadic or CMMRD background are MSS by conventional MSI testing, despite MMR deficiency and loss of polymerase proof reading leading to complete ablation of replication error correction. This surprising observation was explained by the

rapid mutation rate of these tumours (approximately 600 mutations acquired per cell division) leading to rapid progression of a highly heterogeneous tumour population, with no dominant clone ever becoming established (Shlien *et al*, 2015). Therefore, the evolutionary landscape of a tumour or tissue as well as its MMR deficiency determines the strength of an MSI signal.

These considerations are also relevant to another application of MSI testing of non-neoplastic tissues: screening for Lynch syndrome. The presence of MMR-DCF in the normal colorectal mucosa of Lynch syndrome gene carriers shows that MMR deficiency can strike in their non-neoplastic tissues (Kloor *et al*, 2012). It follows that MSI may be detectable in the non-neoplastic tissues of Lynch syndrome gene carriers as a biomarker of germline, heterozygous MMR gene mutation. Indeed, increased frequency of variants in microsatellite length have been detected in PBLs and buccal cells of Lynch syndrome gene carriers using small pool PCR of three DNRs D2S123, D5S346, and D17S518 (Coolbaugh-Murphy *et al*, 2010; Hu *et al*, 2011), two of which are used by the gMSI assay for CMMRD diagnosis (Ingham *et al*, 2013). Alternatively, bacterial vectors have been used for high fidelity replication of single copies of BAT26 (an A₂₆ MNR) initially amplified from patient PBLs to show increased microsatellite deletions in Lynch syndrome gene carriers (Alazzouzi *et al*, 2005). Both techniques require dilution of template DNA to single copies, to facilitate detection of microsatellite length variants, and hundreds of PCRs per sample, but were able to separate the Lynch syndrome gene carriers analysed (n = 6, Alazzouzi *et al*, 2005; n = 7, Coolbaugh-Murphy *et al*, 2010; n = 8, Hu *et al*, 2011) from controls. Furthermore, Coolbaugh-Murphy *et al* (2010) and Hu *et al* (2011) found a correlation between Lynch syndrome gene carrier age and the frequency of microsatellite length variants, in agreement with previous studies using the same method in the general population (Coolbaugh-Murphy *et al*, 2005).

The limited number of samples analysed in these studies, due to the laborious methods used, restrict the conclusions that can be drawn from their results, however it is an intriguing possibility that normal tissues could be screened to identify Lynch syndrome gene carriers by detection of low-level MSI, as was achieved for CMMRD. In this study, DNAs extracted from the PBLs of a small cohort of Lynch syndrome gene carriers (n = 40) was analysed using the same method as described for CMMRD detection (Section 5.6). Scores ranged from 0.00 to 0.92, meaning that, by this method, Lynch syndrome gene carriers are indistinguishable from controls. However, as discussed above, the panel of short MNRs may

not be the most sensitive markers for MSI in non-neoplastic tissues, supported by the use of a long MNR by Alazzouzi *et al* (2005), and the use of DNRs by Coolbaugh-Murphy *et al* (2010) and Hu *et al* (2011). Also, the analysis method, quantifying instability by the proportion of smSequences containing a WT length of microsatellite, is relatively simplistic and could be developed, for example, to look at the allelic distribution of microsatellite lengths detected as was shown to be effective in cancer diagnostics. In addition, protocol optimisation may be required, such as use of higher sequencing depth.

As discussed in Chapter 5, the assay score for CMMRD could be affected by a number of factors, including patient age, their clinical history, and their genetic background, particularly the penetrance of their MMR variants. However, the rarity of the syndrome may make answering these research questions difficult. Alternatively, Lynch syndrome gene carriers may be as common as 1 in 300 of the general population (Hampel and de la Chapelle, 2011; Win *et al*, 2017), and should the assay be adapted to detect Lynch syndrome gene carriers as suggested, these questions could be answered in the context of monoallelic path_MMR gene variants. The different penetrance of MMR variants is particularly interesting – it is known that in some cancer-predisposition syndrome that different pathogenic variants within the same gene can be associated with greatly different disease phenotypes, including age of onset and tumour spectrum; in Li-Fraumeni syndrome, for example, dominant negative mutations in *TP53* are associated with osteosarcomas, adrenocortical carcinomas, CNS tumours, and soft tissue sarcomas diagnosed in childhood, whereas non-dominant negative variants are associated with adult age-of-onset and a predominance of breast cancer (Bougeard *et al*, 2015).

An smMIP and sequencing-based assay for the detection of Lynch syndrome gene carriers from analysis of low level MSI in PBLs would remove the requirement for tumour tissue, and therefore bring Lynch syndrome screening and diagnosis forward in the clinical management of cancer patients. This is particularly relevant for the first cancer diagnosis in a Lynch syndrome patient, where early diagnosis of germline MMR defects is informative for patient treatment (Vasen *et al*, 2013). For example, the high risk for metachronous CRCs in Lynch syndrome (Aarnio *et al*, 1995) dictates that patients should be offered more extensive surgery, with the risk of subsequent CRC being reduced by 31% for every additional 10cm of colorectum resected (Parry *et al*, 2011), but this is not possible if diagnosis follows surgical resection of the tumour for testing. Another strategy to identify Lynch syndrome gene carriers is to screen the general population irrespective of cancer diagnosis, which is feasible

given estimates that there may be 3-4 carriers per 1000 in the general population (Hampel and de la Chapelle, 2011; Win *et al*, 2017). Population screening has been modelled previously using familial risk criteria to screen those aged 25 years or older in a population representative of the USA. Individuals with a $\geq 5\%$ risk of being from a Lynch syndrome family were tested for pathogenic MMR gene mutations, and this general population screening model showed that colorectal and endometrial cancer incidence in the Lynch syndrome gene carrier population would fall by 12.4% and 8.8%, respectively, at a cost of \$26,000 per QALY-gained relative to no testing (Dinh *et al*, 2011). Again, it has been shown multiple times that familial criteria have a poor sensitivity relative to MMR deficiency testing (Pérez-Carbonell *et al*, 2012) and, with appropriate modifications to the low cost and scalable, smMIP and sequencing-based MSI assay presented, molecular screening of the general population could be used to detect low level MSI in normal tissues as a biomarker of heterozygous MMR gene mutation. General population screening strategies must have an exceptionally high specificity to reduce the number of false positives detected (Hartwell *et al*, 2006) and the detection of MSI in the non-neoplastic tissues of Lynch syndrome gene carriers needs to be proven with larger numbers of patients; however, the tools developed in this study provide a means to explore this idea further.

6.3. Concluding Remarks

MMR deficiency defines a distinct subtype of CRC (Guinney *et al*, 2015) and is associated with cancer predisposition syndromes, Lynch syndrome and CMMRD (Wimmer *et al*, 2008; Lynch *et al*, 2009). The unique features of MMR deficiency, for example the immunological interactions of affected tumours (Kloor and von Knebel Döberitz, 2016), and the high risks for multiple cancers in associated syndromes (Møller *et al*, 2017b; Wimmer *et al*, 2017), make its identification critical for disease and patient management. The importance of such personalised medicine has been recognised for over a decade, particularly with respect to therapeutic response (Schilsky, 2010), and will become increasingly available to health care services as we further understand the heterogeneity of cancer. In their short review titled “Personalised medicine in oncology: questions for the next 20 years”, Blay *et al* (2012) highlight the need to transition knowledge with respect to heterogeneity in genetic and clinical characteristics into practice, and the need to address the cost burden of novel tests and treatments. Alongside personalised medicine is “precision prevention and early detection”, in which at-risk populations are identified by their mechanistic association with

disease, such that appropriate interventions can reduce cancer risk and increase detection rates (Rebbeck *et al*, 2018). It is my hope that, by continuing the development of a novel MSI test, I have contributed to personalised medicine by providing a cheap and accurate biomarker test for MMR deficiency, an appropriate companion diagnostic for immune checkpoint blockade therapy. The novel MSI test can also be used for the identification of Lynch syndrome gene carriers and CMMRD patients by MSI testing of CRC and non-neoplastic tissues, respectively, facilitating precision prevention and early detection strategies (Vasen *et al*, 2013; Vasen *et al*, 2014). Whilst additional research is needed to determine whether or not α FSP-Abs are an appropriate biomarker for the early detection of MMRd CRC, I believe this study has provided alternative ideas for the pursuit of cancer care's holy grail (Etzioni *et al*, 2003).

Chapter 7. Appendices

7.1. Appendix A: Colorectal Cancer Sample Data and Source

Sample	MSI Status	BRAF V600E	Source	Sample	MSI Status	BRAF V600E	Source
D206487	MSI-high	NA	Newcastle	L0226	MSI-high	NA	Pamplona
D227036	MSI-high	NA	Newcastle	L0247	MSS	NA	Pamplona
D248097	MSS	NA	Newcastle	L0255	MSS	NA	Pamplona
D250194	MSS	NA	Newcastle	L0261	MSS	NA	Pamplona
D250997	MSI-high	NA	Newcastle	L0275	MSI-high	NA	Pamplona
D251725	MSS	NA	Newcastle	L0284	MSI-high	NA	Pamplona
E08	MSS	NA	Edinburgh	L0287	MSI-high	NA	Pamplona
E43	MSI-high	NA	Edinburgh	L0288	MSS	NA	Pamplona
E44	MSI-high	NA	Edinburgh	L0290	MSS	NA	Pamplona
E49	MSI-high	NA	Edinburgh	L0300	MSS	NA	Pamplona
E55	MSI-high	NA	Edinburgh	L0303	MSI-high	NA	Pamplona
E57	MSI-high	NA	Edinburgh	L0376	MSI-high	NA	Pamplona
E60	MSS	NA	Edinburgh	L0379	MSI-high	NA	Pamplona
E65	MSI-high	NA	Edinburgh	L0400	MSI-high	NA	Pamplona
E67	MSS	NA	Edinburgh	L0406	MSI-high	NA	Pamplona
E74	MSS	NA	Edinburgh	L0408	MSI-high	NA	Pamplona
E76	MSS	NA	Edinburgh	L0409	MSI-high	NA	Pamplona
E81	MSS	NA	Edinburgh	L0421	MSI-high	NA	Pamplona
E83	MSS	NA	Edinburgh	L0444	MSS	NA	Pamplona
E85	MSS	NA	Edinburgh	L0455	MSI-high	NA	Pamplona
E89	MSI-high	NA	Edinburgh	L0481	MSI-high	NA	Pamplona
E90	MSI-high	NA	Edinburgh	L0489	MSI-high	NA	Pamplona
E91	MSI-high	NA	Edinburgh	L0497	MSI-high	NA	Pamplona
E93	MSS	NA	Edinburgh	L0515	MSI-high	NA	Pamplona
E97	MSI-high	NA	Edinburgh	L0525	MSI-high	NA	Pamplona
L0006	MSI-high	NA	Pamplona	L0526	MSI-high	NA	Pamplona
L0029	MSI-high	NA	Pamplona	L0531	MSS	NA	Pamplona
L0054	MSS	NA	Pamplona	L0533	MSS	NA	Pamplona
L0080	MSS	NA	Pamplona	L0535	MSS	NA	Pamplona
L0086	MSS	NA	Pamplona	L0536	MSS	NA	Pamplona
L0091	MSI-high	NA	Pamplona	L0552	MSS	NA	Pamplona
L0093	MSS	NA	Pamplona	L0576	MSS	NA	Pamplona
L0100	MSS	NA	Pamplona	L0584	MSS	NA	Pamplona
L0104	MSI-high	NA	Pamplona	L0650	MSI-high	NA	Pamplona
L0106	MSI-high	NA	Pamplona	L0688	MSI-high	NA	Pamplona
L0113	MSI-high	NA	Pamplona	L0718	MSI-high	NA	Pamplona
L0142	MSS	NA	Pamplona	L0811	MSI-high	NA	Pamplona
L0143	MSI-high	NA	Pamplona	L0812	MSI-high	NA	Pamplona
L0146	MSS	NA	Pamplona	L0817	MSS	NA	Pamplona
L0149	MSI-high	NA	Pamplona	L0819	MSS	NA	Pamplona
L0150	MSS	NA	Pamplona	L0863	MSS	NA	Pamplona
L0153	MSS	NA	Pamplona	L0897	MSS	NA	Pamplona
L0179	MSI-high	NA	Pamplona	L0899	MSS	NA	Pamplona
L0203	MSI-high	NA	Pamplona	L0914	MSS	NA	Pamplona
L0210	MSI-high	NA	Pamplona	L0924	MSS	NA	Pamplona
L0211	MSI-high	NA	Pamplona	L0928	MSS	NA	Pamplona
L0213	MSS	NA	Pamplona	L0953	MSS	NA	Pamplona
L0214	MSI-high	NA	Pamplona	L0954	MSI-high	NA	Pamplona
L0218	MSS	NA	Pamplona	L0956	MSS	NA	Pamplona

Table 7.1: CRCs in the classifier training cohort. MSI status assessed by the Promega MSI Analysis System v1.2. BRAF V600E assessed by high resolution melt curve analysis (Nikiforov *et al*, 2009).

Sample	MSI status	BRAF V600E	Source	Sample	MSI status	BRAF V600E	Source
155063	MSS	NA	Newcastle	233715	MSS	NA	Newcastle
155087	MSS	NA	Newcastle	234543	MSI-high	pos	Newcastle
155088	MSS	NA	Newcastle	237260	MSI-high	neg	Newcastle
155089	MSS	NA	Newcastle	237780	MSI-high	neg	Newcastle
155090	MSS	NA	Newcastle	238659	MSI-high	pos	Newcastle
155501	MSS	NA	Newcastle	239222	MSI-high	neg	Newcastle
155502	MSS	NA	Newcastle	239405	MSI-high	neg	Newcastle
155694	MSS	NA	Newcastle	239970	MSI-high	pos	Newcastle
155695	MSS	NA	Newcastle	241981	MSI-high	pos	Newcastle
156188	MSS	NA	Newcastle	242117	MSI-high	pos	Newcastle
167234	MSS	NA	Newcastle	244031	MSI-high	neg	Newcastle
168888	MSS	NA	Newcastle	244881	MSI-high	neg	Newcastle
204448	MSI-high	neg	Newcastle	245457	MSI-high	NA	Newcastle
205882	MSI-high	neg	Newcastle	245836	MSI-high	pos	Newcastle
207950	MSI-high	neg	Newcastle	245838	MSI-high	pos	Newcastle
210173	MSS	NA	Newcastle	246656	MSI-high	neg	Newcastle
210177	MSS	NA	Newcastle	246847	MSI-high	pos	Newcastle
210178	MSS	NA	Newcastle	246849	MSI-high	neg	Newcastle
210179	MSS	NA	Newcastle	247641	MSI-high	neg	Newcastle
210180	MSI-high	NA	Newcastle	249555	MSI-high	neg	Newcastle
210386	MSS	NA	Newcastle	249985	MSI-high	neg	Newcastle
210777	MSS	NA	Newcastle	250505	MSI-high	neg	Newcastle
210778	MSS	NA	Newcastle	250512	MSI-high	neg	Newcastle
211727	MSS	NA	Newcastle	251058	MSI-high	pos	Newcastle
212963	MSS	NA	Newcastle	252045	MSI-high	neg	Newcastle
213233	MSI-high	neg	Newcastle	252048	MSI-high	neg	Newcastle
213428	MSS	NA	Newcastle	252782	MSI-high	pos	Newcastle
213520	MSI-high	neg	Newcastle	253580	MSI-high	neg	Newcastle
215118	MSS	NA	Newcastle	253977	MSI-high	neg	Newcastle
215770	MSS	NA	Newcastle	254175	MSI-high	neg	Newcastle
216379	MSS	NA	Newcastle	254340	MSS	NA	Newcastle
220045	MSS	NA	Newcastle	254574	MSS	NA	Newcastle
220926	MSS	NA	Newcastle	255075	MSS	NA	Newcastle
223129	MSS	NA	Newcastle	255078	MSI-high	pos	Newcastle
223962	MSS	NA	Newcastle	255809	MSI-high	neg	Newcastle
225162	MSS	NA	Newcastle	255810	MSI-high	neg	Newcastle
225729	MSS	NA	Newcastle	255811	MSI-high	NA	Newcastle
226491	MSS	NA	Newcastle	256265	MSS	NA	Newcastle
226724	MSS	NA	Newcastle	256267	MSI-high	neg	Newcastle
227175	MSS	NA	Newcastle	256271	MSI-high	NA	Newcastle
228082	MSS	NA	Newcastle	257349	MSI-high	neg	Newcastle
228417	MSS	NA	Newcastle	D222913	MSI-high	neg	Newcastle
228418	MSS	NA	Newcastle	D223305	MSI-high	neg	Newcastle
229072	MSS	NA	Newcastle	D227036	MSI-high	pos	Newcastle
229073	MSS	NA	Newcastle	D229104	MSI-high	neg	Newcastle
229291	MSS	NA	Newcastle	D229113	MSI-high	pos	Newcastle
229618	MSS	NA	Newcastle	D234036	MSI-high	neg	Newcastle
229619	MSS	NA	Newcastle	D238498	MSI-high	neg	Newcastle
229995	MSS	NA	Newcastle	D238861	MSI-high	pos	Newcastle
231954	MSS	NA	Newcastle				

Table 7.2: CRCs in the assay validation cohort. MSI status assessed by the Promega MSI Analysis System v1.2. BRAF V600E assessed by high resolution melt curve analysis (Nikiforov *et al*, 2009).

7.2. Appendix B: Constitutional Mismatch Repair Deficiency and Control Patient Samples

Patient ID	Genotype	Age First Malignancy	Malignancy Type	Referring Physician	Sample ID	Age at Blood Draw	Score	gMSI A	gMSI B	gMSI C
1	<i>PMS2</i> hom	13years	B-cell Burkitt lymphoma	Iman Ragab	C	13years	27.34	0.30	0.07	0.06
2	<i>PMS2</i> hom	5years	glioblastoma	Christian Kratz, Tim Ripperger	A	5years	23.03	0.09	0.29	0.39
3	<i>MSH6</i> hom	13years	colorectal cancer	Iman Ragab	D	20years	24.88	-0.04	-0.02	-0.05
4	<i>MSH6 comp het</i>	6years	medulloblastoma	Amedeo Azizi	B	6years	10.02	-0.05	-0.04	-0.02
5	<i>MLH1 comp het</i>	6years	T-NHL	Michaela Nathrath	E	21years	14.43	0.02	0.05	0.06
6	<i>MSH6</i> hom	11years	glioblastoma	Christian Kratz, Andreas Beilken	65	11years	12.47	-0.07	-0.06	-0.05
7	<i>PMS2</i> hom	9years	glioblastoma	Claudia Blattmann, Hans-Jürgen Pander	54 93	≤9years 9years	53.59 23.20	0.04 -0.03	0.23 0.18	0.29 0.27
8 * (F1)	<i>MSH6 comp het</i>	10years	T-cell lymphoma	Stephan Lobitz	99 102 105	12years 12years 12years	1.59 4.62 2.08	-0.07 -0.07 -0.07	-0.05 -0.06 -0.05	-0.08 -0.06 -0.07
9 (F1)	<i>MSH6 comp het</i>	No Tumour	NA	Stephan Lobitz	82	9years	19.09	-0.07	-0.05	-0.08
10	<i>PMS2</i> hom	ND	ND	Manon Suerink	56	ND	25.52	0.18	0.08	0.33
11	<i>MSH6</i> hom	5years	Wilms tumour	Daniel Rueda	91	ND	9.85	-0.07	-0.06	-0.06
12	<i>MLH1</i> hom	7months	T-cell lymphoma	Daniel Rueda	49	7months	42.98	0.06	0.11	-0.01
13	<i>PMS2</i> hom	2years	ALL	Daniel Rueda	51	ND	5.71	-0.01	0.15	0.29
14	<i>MSH6</i> hom	13months	medulloblastoma	Julia Täubner	76	ND	18.07	-0.05	-0.03	-0.05
15	<i>PMS2</i> hom	20years	colorectal cancer	Hagit Baris	98	26years	36.97	0.07	0.09	0.48
16	<i>MSH6 comp het</i>	3years	B-ALL	Danuta Lewandowska	83	8years	13.70	-0.07	-0.05	-0.07
17	<i>MSH6 comp het</i>	4years	medulloblastoma	Thorsten Rosenbaum	43	4years	53.72	-0.06	-0.01	-0.05
18	<i>PMS2</i> hom	7years	sPNET	Benoit Florquin	71	10years	14.49	-0.02	0.10	0.29
19 (F2)	<i>MSH2</i> hom	9years	colorectal cancer	Karin Dahan	58	13years	14.85	0.07	0.06	0.26
20 (F2)	<i>MSH2</i> hom	17years	glioblastoma	Karin Dahan	87	17years	27.67	0.05	0.08	0.23
21 (F3)	<i>MSH6</i> hom	9years	glioblastoma	Imschweiler	101	9years	17.61	-0.04	-0.06	-0.06

22 (F3)	<i>MSH6 hom</i>	No Tumour	NA	Imschweiler	107	13months	3.54	-0.04	-0.06	-0.11
					109	15months	7.39	-0.04	-0.06	-0.11
23 (F2)	<i>MSH2 hom</i>	9years	colorectal cancer	Karin Dahan	104	19years	42.52	0.05	0.08	0.26
24	<i>MSH2 hom</i>	2years, 4months	medulloepithelioma	Demirsoy	116	2years, 11months	43.10	0.07	0.04	0.20
25	<i>MSH2 hom</i>	ND	T-cell lymphoma	Demirsoy	132	6years	54.55	0.16	0.15	0.40
26	<i>PMS2 comp het</i>	4years	sPNET	Aretz	113	14years	13.08	-0.05	0.06	0.22
27	<i>PMS2 comp het</i>	9years	B-cell lymphoma	Aretz	124	9years	9.83	0.18	0.15	0.25
28	<i>PMS2 hom</i>	ND	T-cell NHL	Aretz	130	2years	12.82	0.24	0.14	0.32
29 †	<i>PMS2 hom</i>	24years	brain tumor	George Chong, William Foulkes	125	24years	2.76	0.03	-0.03	0.05
30 †	<i>PMS2 hom</i>	3years	medulloblastoma	George Chong, William Foulkes	120	18years	4.28	-0.04	0.06	0.15
31 †	<i>PMS2 hom</i>	No Tumour	NA	George Chong, William Foulkes	115	21years	5.90	0.05	0.06	0.07
32	<i>MSH6 hom</i>	10years	colorectal cancer	George Chong, William Foulkes	128	23years	4.78	0.01	-0.04	-0.04
Control01	NA	NA	NA	NA	1	NA	0.002	-0.06	-0.02	-0.06
Control02	NA	NA	NA	NA	2	NA	0.829	-0.05	-0.04	-0.05
Control03	NA	NA	NA	NA	3	NA	0.533	-0.07	-0.05	-0.06
Control04	NA	NA	NA	NA	4	NA	0.424	-0.04	-0.08	-0.08
Control05	NA	NA	NA	NA	5	NA	0.046	-0.06	-0.04	-0.06
Control06	NA	NA	NA	NA	6	NA	0.868	-0.07	-0.05	-0.06
Control07	NA	NA	NA	NA	7	NA	0.042	-0.06	-0.05	-0.05
Control08	NA	NA	NA	NA	8	NA	0.32	-0.08	-0.02	-0.08
Control09	NA	NA	NA	NA	9	NA	0.205	-0.05	-0.04	-0.05
Control10	NA	NA	NA	NA	10	NA	0.145	-0.05	-0.03	-0.07
Control11	NA	NA	NA	NA	11	NA	0.28	-0.04	-0.05	-0.11
Control12	NA	NA	NA	NA	12	NA	0.03	-0.04	-0.02	-0.02
Control13	NA	NA	NA	NA	13	NA	0.10	-0.05	-0.06	-0.04
Control14	NA	NA	NA	NA	14	NA	0.00	0.00	-0.06	-0.04
Control15	NA	NA	NA	NA	15	NA	0.42	-0.08	-0.06	-0.07

Control16	NA	NA	NA	NA	16	NA	0.49	-0.07	-0.07	-0.04
Control17	NA	NA	NA	NA	17	NA	0.00	-0.06	-0.04	-0.06
Control18	NA	NA	NA	NA	18	NA	0.06	0.00	-0.06	-0.06
Control19	NA	NA	NA	NA	19	NA	0.05	-0.08	-0.05	-0.07
Control20	NA	NA	NA	NA	20	NA	0.03	-0.08	0.00	-0.02
Control21	NA	NA	NA	NA	21	NA	0.01	-0.06	-0.03	0.00
Control22	NA	NA	NA	NA	22	NA	0.02	-0.07	-0.05	-0.05
Control23	NA	NA	NA	NA	23	NA	0.00	-0.07	-0.07	-0.07
Control24	NA	NA	NA	NA	24	NA	0.06	-0.07	-0.07	-0.08
Control25	NA	NA	NA	NA	25	NA	0.78	-0.03	-0.07	-0.07
Control26	NA	NA	NA	NA	26	NA	0.16	-0.07	-0.02	-0.08
Control27	NA	NA	NA	NA	27	NA	0.00	-0.07	-0.06	-0.08
Control28	NA	NA	NA	NA	28	NA	0.00	-0.04	-0.06	-0.03
Control29	NA	NA	NA	NA	29	NA	0.48	-0.05	-0.06	-0.09
Control30	NA	NA	NA	NA	30	NA	0.00	-0.07	-0.05	-0.07
Control31	NA	NA	NA	NA	31	NA	0.16	-0.08	-0.07	-0.06
Control32	NA	NA	NA	NA	32	NA	1.47	-0.08	-0.06	-0.08
Control33	NA	NA	NA	NA	33	NA	0.02	-0.08	-0.06	-0.08
Control34	NA	NA	NA	NA	34	NA	0.06	-0.08	-0.05	-0.05
Control35	NA	NA	NA	NA	35	NA	1.46	-0.07	-0.07	-0.07
Control36	NA	NA	NA	NA	36	NA	0.00	-0.07	-0.06	-0.07
Control37	NA	NA	NA	NA	37	NA	0.14	-0.05	-0.07	-0.07
Control38	NA	NA	NA	NA	38	NA	0.07	-0.08	-0.04	-0.07
Control39	NA	NA	NA	NA	39	NA	0.57	-0.08	-0.07	-0.08
Control40	NA	NA	NA	NA	40	NA	0.01	-0.06	-0.04	-0.09
Control41	NA	NA	NA	NA	41	NA	0.66	-0.07	-0.06	-0.09
Control42	NA	NA	NA	NA	42	NA	0.28	-0.08	-0.05	-0.08
Control43	NA	NA	NA	NA	44	NA	0.01	-0.08	-0.05	-0.09
Control44	NA	NA	NA	NA	45	NA	0.46	-0.05	-0.06	-0.07
Control45	NA	NA	NA	NA	46	NA	0.21	-0.08	-0.05	-0.08

Control46	NA	NA	NA	NA	47	NA	0.00	-0.07	-0.05	-0.07
Control47	NA	NA	NA	NA	48	NA	0.03	-0.02	-0.05	-0.04
Control48	NA	NA	NA	NA	50	NA	0.03	-0.05	-0.04	-0.07
Control49	NA	NA	NA	NA	53	NA	0.05	-0.08	-0.05	-0.08
Control50	NA	NA	NA	NA	55	NA	0.20	-0.08	-0.07	-0.09
Control51	NA	NA	NA	NA	59	NA	0.09	-0.08	-0.04	-0.09
Control52	NA	NA	NA	NA	60	NA	0.03	-0.08	-0.06	-0.07
Control53	NA	NA	NA	NA	61	NA	1.02	-0.08	-0.05	-0.08
Control54	NA	NA	NA	NA	62	NA	0.09	-0.08	-0.05	-0.09
Control55	NA	NA	NA	NA	63	NA	0.13	-0.08	-0.06	-0.09
Control56	NA	NA	NA	NA	64	NA	0.13	-0.06	-0.05	-0.09
Control57	NA	NA	NA	NA	66	NA	0.00	-0.07	-0.05	-0.07
Control58	NA	NA	NA	NA	68	NA	0.10	-0.06	-0.06	-0.09
Control59	NA	NA	NA	NA	69	NA	0.60	-0.07	-0.05	-0.08
Control60	NA	NA	NA	NA	70	NA	0.00	-0.08	-0.06	-0.05
Control61	NA	NA	NA	NA	72	NA	0.58	-0.07	-0.04	-0.09
Control62	NA	NA	NA	NA	73	NA	0.22	-0.07	-0.06	-0.10
Control63	NA	NA	NA	NA	74	NA	0.02	-0.08	-0.07	-0.05
Control64	NA	NA	NA	NA	75	NA	0.09	-0.07	-0.06	-0.08
Control65	NA	NA	NA	NA	77	NA	0.11	-0.08	-0.05	-0.08
Control66	NA	NA	NA	NA	78	NA	0.01	-0.06	-0.09	-0.08
Control67	NA	NA	NA	NA	79	NA	0.00	-0.07	-0.07	-0.07
Control68	NA	NA	NA	NA	80	NA	0.10	-0.08	-0.07	-0.07
Control69	NA	NA	NA	NA	81	NA	0.00	-0.06	-0.05	-0.09
Control70	NA	NA	NA	NA	84	NA	0.00	-0.08	-0.06	-0.09
Control71	NA	NA	NA	NA	85	NA	0.24	-0.08	-0.03	-0.09
Control72	NA	NA	NA	NA	86	NA	0.06	-0.07	-0.07	-0.07
Control73	NA	NA	NA	NA	88	NA	0.10	-0.07	-0.10	-0.06
Control74	NA	NA	NA	NA	89	NA	0.01	-0.08	-0.07	-0.07
Control75	NA	NA	NA	NA	90	NA	0.20	-0.08	-0.05	-0.05

Control76	NA	NA	NA	NA	92	NA	0.22	-0.08	-0.06	-0.10
Control77	NA	NA	NA	NA	94	NA	0.02	-0.08	-0.03	-0.08
Control78	NA	NA	NA	NA	95	NA	0.01	-0.09	-0.05	-0.06
Control79	NA	NA	NA	NA	96	NA	0.08	-0.08	-0.05	-0.09
Control80	NA	NA	NA	NA	97	NA	0.00	-0.08	-0.05	-0.09
Control81	NA	NA	NA	NA	111	NA	0.01	-0.08	-0.04	-0.10
Control82	NA	NA	NA	NA	112	NA	0.02	-0.08	-0.03	-0.08
Control83	NA	NA	NA	NA	114	NA	0.10	-0.07	-0.05	-0.07
Control84	NA	NA	NA	NA	117	NA	0.02	-0.09	-0.05	-0.07
Control85	NA	NA	NA	NA	119	NA	0.27	-0.08	-0.06	-0.09
Control86	NA	NA	NA	NA	121	NA	0.02	-0.07	-0.06	-0.10
Control87	NA	NA	NA	NA	122	NA	0.01	-0.08	-0.05	-0.04
Control88	NA	NA	NA	NA	123	NA	0.00	-0.07	-0.04	-0.06
Control89	NA	NA	NA	NA	126	NA	0.01	-0.07	-0.05	-0.11
Control90	NA	NA	NA	NA	127	NA	0.02	-0.06	-0.07	-0.06
Control91	NA	NA	NA	NA	129	NA	0.00	-0.07	-0.03	-0.06
Control92	NA	NA	NA	NA	131	NA	0.07	-0.06	-0.07	-0.08
Control93	NA	NA	NA	NA	133	NA	0.08	-0.07	-0.06	-0.11
Control94	NA	NA	NA	NA	134	NA	0.00	-0.08	-0.06	-0.08

Table 7.3: Clinical details and test results of constitutional mismatch repair deficiency and control samples. 36 DNA samples from 32 genetically-confirmed CMMRD patients were sourced from a number of referring physicians. 94 anonymised control DNA samples, extracted from peripheral blood leukocytes, were acquired from patients consulted for non-cancer related conditions at the Division of Human Genetics, Medical University of Innsbruck, Innsbruck, Austria. All control patients had consented for use of residual DNA samples in assay development. Assay Score for each sample was calculated as described in Chapter 5. Higher score indicates increased MSI and therefore increased likelihood of CMMRD. gMSI ratios for three markers (A, D2S123; B, D17S250; C D17S791) were calculated with the Peak Heights software (Ingham *et al*, 2013). Marker ratios presented here are the observed ratio minus the marker-specific threshold; positive values represent ratios above the threshold. If two or more of the gMSI markers are above the threshold the sample is classified as CMMRD. Thresholds were calculated as per Ingham *et al* (2013), using the same 40 controls as were used for the control distributions for score calculation.

ND: Not Disclosed. NA: Not Applicable.

Patients from three families (F1), (F2), (F3) are indicated in Patient ID. * Patient 8 was aplastic when blood samples were collected, or just recovered from aplasia.

† Patients homozygous for hypomorphic *PMS2* mutation.

7.3. Appendix C: Frameshift Peptides analysed for Serum Reactivity

FSP	Amino Acid Sequence
ACVR2(-1)	VVHKKRGLFDYKDDDDK
ACVR2(-2)	VHKKEACFKRLLAETCWNGNALDYKDDDDK
AIM2(-1)	KAKKKHREVKRTNSSQLVDYKDDDDK
AIM2(-2)	IKAKKNIEKDYKDDDDK
ASTE1(-1)	NSKKKGRRNRIPAVLRTEGEPLHTPSVGMRETTGLGCDYKDDDDK
ASTE1(-2)	NSKKKAEETEQYQLFDYKDDDDK
BANP(-1)	FFPFFCSVGADYKDDDDK
C14orf106(-1)	RVEKKNCSTIPTYVKRRTTNHSSQMTVHDYKDDDDK
C14orf106(-2)	RVEKKIAAYLPMDYKDDDDK
C1orf34(-2)B	RAAWEDKGGGGICGAWDFWEIDYKDDDDK
CASP5(-1)	NHKKKQLRCWNTWAKMFFMVFLIHWQNTMFDYKDDDDK
LMAN1(-1)	LDKKKRNSRRATPTSKGSLRRKYLRVDYKDDDDK
LMAN1(-2)	ELDKKRGIPGPPRPPRAACGGNI DYKDDDDK
MARCKS(-2)	TPKKKEALFLQEVFQAERLLLQEEQEGWRRRDYKDDDDK
MYH11(-1)A	LRGPPHRKLRSDAPGEETRPLSFLLEGLEDVELLMQMVLDYKDDDDK
MYH11(-1)B	LLMQMVLRKRRTLETQTSMEPRPVNKQLSTVLHHDYKDDDDK
MYH11(-1)C	QLSTVLHGHGKTKNQNKQTKKTQQPRTKQNPADCTDYKDDDDK
MYH11(-2)	LRGPPTGNFVAMHQARKRDLFRSFDYKDDDDK
POLD3(-1)	QKEKKGSEDYKDDDDK
PTHLH(-1)	GLKKKRKTTEEHIICNDYKDDDDK
Q96PS6(-1)	IFFFFKDGVLSSHLDYKDDDDK
Q96PS6(-2)	PIFFFFSKMESYSLTDYKDDDDK
SLC22A9(-1)A	AAQKKNLLCVKCSCTPTYVKGSPSCPLRDLQTLWPILADYKDDDDK
SLC22A9(-2)	AAQKKTFSVDYKDDDDK
TAF1B(-1)	GLKKKTILKKAGIGMCVKVSSIFFINKQKPDYKDDDDK
TGFBR2(-1)	MKEKKSIVRLSSCVPVALMSAMTTSSSQKNITPAILTCCDYKDDDDK
TMEM60(-2)	HNIKKSIVPHCNVTDYKDDDDK
UPF3A(-1)	RCKKKRQINRRKLQRKDYKDDDDK

Table 7.4: Synthetic frameshift peptides (FSP) and their amino acid sequence. FSPs are denoted by “Gene(deletion length)”. The amino acid sequence is determined by the translation product of the gene following deletion in its coding mononucleotide repeat, with the N-terminus to the left and C-terminus to the right. Amino acids in black represent the 5 amino acids upstream of the frameshift mutation, which will be found in the wild type protein. Amino acids in blue represent the novel amino acid sequence downstream of the frameshift mutation, generated by the change in reading frame. Where the novel sequence is >35 amino acids in length it is split between multiple FSPs, for example see MYH11(-1)A, MYH11(-1)B, and MYH11(-1)C. Amino acids in red are the FLAG octapeptide that is tagged to the C-terminus of all FSPs and is used as a control for non-specific binding.

7.4. Appendix D: Marker Loci for the smMIP and Sequencing-based MSI Assay

Marker	Chromosome	MNR	MNR Start	MNR End	Variant	Variant Position
<i>BRAF</i>	chr7	-	-	-	V600(E)	140453136
AP0035_SNP1	chr11	A9	127625067	127625075	rs10893736	127625130
DEPDC2_SNP1	chr8	G8	68926683	68926690	rs4610727	68926700
GM01_SNP1	chr11	A10	28894429	28894438	rs7951012	28894411
GM07_SNP1	chr7	A11	93085748	93085758	rs2283006	93085722
GM09_SNP1	chr20	A8	6836977	6836984	rs6038623	6836952
GM11_SNP1	chr5	A9	166099891	166099899	rs347435	166099902
GM11_SNP2	chr5	A9	166099891	166099899	rs72817807	166099948
GM14_SNP1	chr3	A11	177328818	177328828	rs6804861	177328829
GM17_SNP1	chr11	A9	95551111	95551119	rs666398	95551136
GM22_SNP1	chr14	A10	43401010	43401019	rs17113692	43400964
GM26_SNP1	chr14	A10	49584751	49584760	rs11628435	49584720
GM29_SNP1	chr3	A10	70905560	70905569	rs2687195	70905581
IM16_SNP1	chr18	A9	1108767	1108775	rs4392141	1108738
IM16_SNP2	chr18	A9	1108767	1108775	rs59912715	1108746
IM16_SNP3	chr18	A9	1108767	1108775	rs73367791	1108784
IM49_SNP1	chr3	A12	56682066	56682077	rs7642389	56682093
IM66_SNP1	chr17	G7	48433967	48433973	rs147847688	48433971
IM66_SNP2	chr17	G7	48433967	48433973	rs141474571	48433973
IM66_SNP3	chr17	G7	48433967	48433973	rs4794136	48433958
IM66_SNP4	chr17	G7	48433967	48433973	rs143225448	48433979
IM66_SNP5	chr17	G7	48433967	48433973	rs140457310	48433950
<i>KRAS</i> _1	chr12	-	-	-	G12(R/C/S)	25398285
<i>KRAS</i> _2	chr12	-	-	-	G12(V/A/D)	25398284
<i>KRAS</i> _3	chr12	-	-	-	G13(C)	25398282
<i>KRAS</i> _4	chr12	-	-	-	G13(D)	25398281
LR10_SNP1	chr1	A10	81591388	81591397	rs1768398	81591398
LR10_SNP2	chr1	A10	81591388	81591397	rs1768397	81591415
LR11_SNP1	chr2	A11	217217871	217217881	rs13011054	217217857
LR11_SNP2	chr2	A11	217217871	217217881	rs16855951	217217913
LR17_SNP1	chr14	A10	55603031	55603040	rs79618905	55603041
LR17_SNP2	chr14	A10	55603031	55603040	rs77482253	55603042
LR17_SNP3	chr14	A10	55603031	55603040	rs1009978	55603061
LR17_SNP4	chr14	A10	55603031	55603040	rs1009977	55603002
LR20_SNP1	chr1	A8	64029634	64029641	rs217474	64029606
LR24_SNP1	chr1	A9	153779429	153779437	rs1127091	153779412
LR36_SNP1	chr4	A12	98999723	98999734	rs17550217	98999699
LR40_SNP1	chr2	A9	13447470	13447478	rs6432372	13447484
LR44_SNP1	chr10	A12	99898286	99898297	rs7905384	99898268
LR44_SNP2	chr10	A12	99898286	99898297	rs7905388	99898281
LR46_SNP1	chr20	A8	10660085	10660092	rs6040079	10660063
LR48_SNP1	chr12	A11	77988097	77988107	rs11105832	77988123
LR49_SNP1	chr15	A7	93619048	93619054	rs12903384	93619037
LR52_SNP1	chr16	A12	63861441	63861452	rs2434849	63861437

Table 7.5: Marker loci for the smMIP and sequencing-based MSI assay. Loci are specified by chromosomal coordinates using reference genome hg 19. Mononucleotide repeats (MNRs) are specified by sequence content and the chromosomal coordinate at which the MNR starts and ends. Variant can refer to a somatic mutation known to be a driver in colorectal tumorigenesis, or can refer to a germline single nucleotide polymorphism.

7.5. Appendix E: Sequences of Molecular Inversion Probes, PCR Primers, and Sequencing Primers

Oligonucleotide	Description	Oligonucleotide Sequence
AP0035_2.0001_MIP	Molecular inversion probe	GCACATTATGTTGTAGTCAAGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGTTTATTGGCCATTTGTATATATT
BRAF_E_0007_MIP	Molecular inversion probe	CCATCAGTTTGAACAGTTGTCTGGATCCACTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTCTACTGTTTTCTTT
DEPDC2_0039_MIP	Molecular inversion probe	GTCTTTGACTCACCTGTGTAGTGTCTGCACTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNATGTTACACACATGC
GM01_0004_MIP	Molecular inversion probe	GGCTGTTACCAACTAAATCTTACCCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCCTTTTAGAATGATCAGATT
GM07_0036_MIP	Molecular inversion probe	CCAAACCCCATATGTGTGGTTGCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTGGGCCCTTTTAGGCATATAG
GM09_0026_MIP	Molecular inversion probe	GCATAAGGCTAGGATCATTTTCATTCAAGACTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCACAAAAATCAATGCT
GM11_0005_MIP	Molecular inversion probe	GAATACTTAGATACGTAGGTGATACTGAACTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCAAAAAAGTACAGTGG
GM14_0030_MIP	Molecular inversion probe	CAATGTTTATCCTTTGCTGGAATCAATTCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNAATGACTTCCAGGCT
GM17_0009_MIP	Molecular inversion probe	GCAAGGGCCTGCATTGTGGTAAGTTTGTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGCTATAAATATCCAGTG
GM22_0007_MIP	Molecular inversion probe	CATCTTTCTTCAAGGTGGTGCTCTTGGTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCTATATCCCCCAAAG
GM26_D_0002_MIP	Molecular inversion probe	GTTCTGCTCCCGCTTGCAGATCAAAGGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGGATTTAGAATCCAGCTC
GM29_0020_MIP	Molecular inversion probe	CTCAGGGCTGAGGAGACTTTTTGTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTTTCAGTGTGCCTTCTGAG
IM16_0021_MIP	Molecular inversion probe	TTTTGAAGATGCTTGCATAGCTATCTACCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGCTGAGTAATATATGGG
IM49_0028_MIP	Molecular inversion probe	GCACGCCTGTAATCCCAAGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGGATCGCTTCAGGCCAGGAGTCAA
IM66_1.0019_MIP	Molecular inversion probe	CACGCCAGCCCTCAAGGCCTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTCTCCAGACCCACCTTCTCGCCC
KRAS_0007_MIP	Molecular inversion probe	GTGACTATATTAGAACATGTCACACCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTCGTCAAGGCACTCTGCCT
LR10_0001_MIP	Molecular inversion probe	CACTGTGAAGCAACTGCGCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGTAGTACTGGTTGAGTCTATTTTT
LR11_0003_MIP	Molecular inversion probe	CCTCACATTTTATAAAGACTTTCAACAATCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCATTTCCTGTGCCTTT
LR17_0011_MIP	Molecular inversion probe	CTCCAACAGCACCTTTCCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTTTACCTTAGTTTGTGTACTGCCAAA
LR20_D.0001_MIP	Molecular inversion probe	GCAACTATTCAATTACAGTATATAGGGGCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTATCATGAAATTCTAT
LR24_0004_MIP	Molecular inversion probe	GTGGGAAAATACTTATTCCAGGGAGAGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTTTTAAAGGGGAAAGGA
LR36_0032_MIP	Molecular inversion probe	AGAGTGCAAAGATAAATGTGCCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNAGTGGCTGGCACTTGTGGT
LR40_0017_MIP	Molecular inversion probe	CAGTTATATATGAAGAAGCTTGGATACCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTCAGTTCAGTTGACTG
LR44_0006_MIP	Molecular inversion probe	CACTTTTGTTCCTTGACTGTTTTTACTCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNCTGAGGTAGGCTCATT
LR46_0012_MIP	Molecular inversion probe	GTGAGTCGTCTGTTCTTGTGAATGGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNGAGTTCAGTCTTTTCAGGGA
LR48_0014_MIP	Molecular inversion probe	GCCCAATTATTTCAACCAGTTTCCACTGACTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNAGAAGATTCACTCAA
LR49_0016_MIP	Molecular inversion probe	GGAGAAATGTCTGAGGCTGAATTTGGCTTCAGCTTCCCGATATCCGACGGTAGTGTNNNNNNTGGCTGCCTTTTAGGAGG

LR52_A_0010_MIP	Molecular inversion probe	GCATGTAGAACTGTTCTCTAGTAGTCTCCTTCAGCTTCCCGATATCCGACGGTAGTGTTTTNNNNNAGGCAATCTTTAAAAC
MIP_PosCon_MIP	Molecular inversion probe	GCAGTCTTCTACCTGTGTCTCTTCAGCTTCCCGATCCGACGGTAGTGTTTTNNNNNATTACCTCATAGTAGAGCA
MIP_Ampli_FP	MIP universal amplification forward primer	AATGATACGGCGACCACCGAGATCTACACATACGAGATCCGTAATCGGGAAGCTGAAG
MIP_Ampli_RP	MIP universal amplification reverse primer	CAAGCAGAAGACGGCATAACGAGATXXXXXXXXACACGCACGATCCGACGGTAGTGT
AP0035_2.0001_intRP	For amplicon validation	TGTATGGAAGGAACACAAGAGT
BRAF_E_0007_intRP	For amplicon validation	TGTCTGGATCCATTTTGTTCATGA
DEPDC2_1.0039_intFP	For amplicon validation	GCAAGCTGAAAGATCCTCGG
DEPDC2_1.0039_intRP	For amplicon validation	CCGAGGATCTTTCAGCTTGC
GM01_0004_intRP	For amplicon validation	TCGGGAAGCTGAAGAATCTGA
GM07_1.0036_intFP	For amplicon validation	TGACCAAACCCCATATGTGTG
GM07_1.0036_intRP	For amplicon validation	CACATATGGGGTTTGGTCACA
GM09_1.0026_intRP	For amplicon validation	CCTGGAATACGGAGCATAAGG
GM11_2.0005_intRP	For amplicon validation	TCTGAACCATTCTTAATTGCCT
GM14_0022_intRP	For amplicon validation	AGCTGGGAAGTCATTGAGTCT
GM14_1.0030_intRP	For amplicon validation	TGGTCTTTTAGCCTGGGAAGT
GM17_1.0009_intRP	For amplicon validation	ACACATGCACTGACTTCTGC
GM22_0007_intRP	For amplicon validation	CCAGAGCTTTATAACCAAGAGCA
GM26_0002_intRP	For amplicon validation	TACTAAAGTCCAATCGAGAGCC
GM29_0020_intRP	For amplicon validation	CAGGAAGGCACACTGAAACA
IM16_1.0021_intRP	For amplicon validation	GGTATGAACACTGCTGATTCCA
IM49_1.0028_intFP	For amplicon validation	TCAGGCCAGGAGTTCAAGAA
IM49_1.0028_intRP	For amplicon validation	TGTTCTTGAACCTCTGGCCT
IM66_1.0019_intRP	For amplicon validation	GAAGAAGGTGGGTCTGGAGA
KRAS_0007_intRP	For amplicon validation	GGGAAGCTGAAGAGGCAAGA
LR10_0001_intRP	For amplicon validation	ATGTATAACAATTTGGACTTAGCGC
LR11_2.0003_intRP	For amplicon validation	TGAAGTTAGGCTCCGTGGTT
LR17_0011_intRP	For amplicon validation	GGGGAGCTGAAGTTTGTATGT
LR20_1.0001_intRP	For amplicon validation	GGGGCAAACTAAACATGTAAGT

LR24_1.0004_intFP	For amplicon validation	GGTAACCAAAGCAGGAAAACAT
LR24_1.0004_intRP	For amplicon validation	TGGTTACCTTTCCTTCCCCT
LR36_1.0032_intFP	For amplicon validation	TGTGGTGACCCTGAACGTTA
LR36_1.0032_intRP	For amplicon validation	TCATTAACGTTCAGGGTCACC
LR40_0017_intRP	For amplicon validation	CTGAACTGATGAATGTATAAGCCAC
LR44_1.0006_intFP	For amplicon validation	GCCAAGAGTTCAAGACCAGC
LR44_1.0006_intRP	For amplicon validation	GTCTCACTTTGTTGCCCAGG
LR46_0012_intRP	For amplicon validation	CCTGAAAAGACTGAACTCTGTATCA
LR48_2.0014_intRP	For amplicon validation	TGGAAGGAGGGCTAAACTGA
LR49_1.0016_intRP	For amplicon validation	CTTTTGTGCCCTTTCCCAA
LR52_0010_intRP	For amplicon validation	GGCTTCTTGTAACCTTTTCTCAAAA
MIP_bbFP	For amplicon validation	ACGAGATCTCTAGCAACACG
MIP_bbRP	For amplicon validation	GACCACCGAGATCTACACATAC
MIP_PosCon_intFP	For amplicon validation	TCCTCCAAATGTAGAATCTTCACC
MIP_PosCon_intRP	For amplicon validation	ACACAGGTAGAAGACTGCACT
MIP_Index_Seq_Primer	Custom sequencing primer	ACACTACCGTCGGATCGTGCGTGT
MIP_Read1_Seq_Primer	Custom sequencing primer	CATACGAGATCCGTAATCGGGAAGCTGAAG
MIP_Read2_Seq_Primer	Custom sequencing primer	ACACGCACGATCCGACGGTAGTGT

Table 7.6: Oligonucleotide sequences of all probes and primers used in the development of the smMIP and sequencing-based MSI assay. See Chapter 4 for details of how each oligonucleotide was used. The MIP universal amplification reverse primer contains a sequence specified as “XXXXXXXX”. This is the sample index sequence, and circularised smMIPs from each sample are amplified with a unique sample index sequence to facilitate read de-multiplexing. Note: the reverse complement of the sample index sequence specified in the primer is used in the sample sheet for MiSeq loading (see Appendix F).

7.6. Appendix F: Example Sample Sheet for MiSeq Loading

```
[Header]
IEMFileVersion      4
Investigator Name   Richard Gallon
Experiment Name
Date                25/01/2017
Workflow            GenerateFASTQ
Application         FASTQ Only
Assay              Nextera XT
Description
Chemistry           Default
```

```
[Reads]
157
151
```

```
[Settings]
CustomRead1PrimerMix C1
CustomIndexPrimerMix C2
CustomRead2PrimerMix C3
ReverseComplement    0
```

```
[Data]
Sample_ID           index
K562                TGCTAGAG
HCT116              TGAGAGCT
PR32516Normal      ATAAGCGT
PR32516Tumour      GTCACTCA
```

Table 7.7: Example sample sheet for MiSeq loading. The sample sheet specifies the workflow to be used, the length of reads in forward and reverse orientations, and the use of custom sequencing primers. Samples and sample index sequences are specified for read de-multiplexing. This example is the sample sheet used for the first sequencing run described in Section 4.4.

7.7. Appendix G: PCR Primer Sequences for *MSH6* c.3557-1G>C

Oligonucleotide	Oligonucleotide Sequence
<i>MSH6</i> _7F	CCAATATGTGTAGCTCATGATAGC
<i>MSH6</i> _7R	TATTAGTGTTCTCATCCCCGTAG

Table 7.8: Primer sequences for amplification of the *MSH6* c.3557-1G>C locus. Used to confirm the identity of sample 99 from patient 8, as describe in Section 5.6.

7.8. Appendix H: Distribution of Frameshift Peptide versus FLAG-only Control Median Fluorescence Intensity

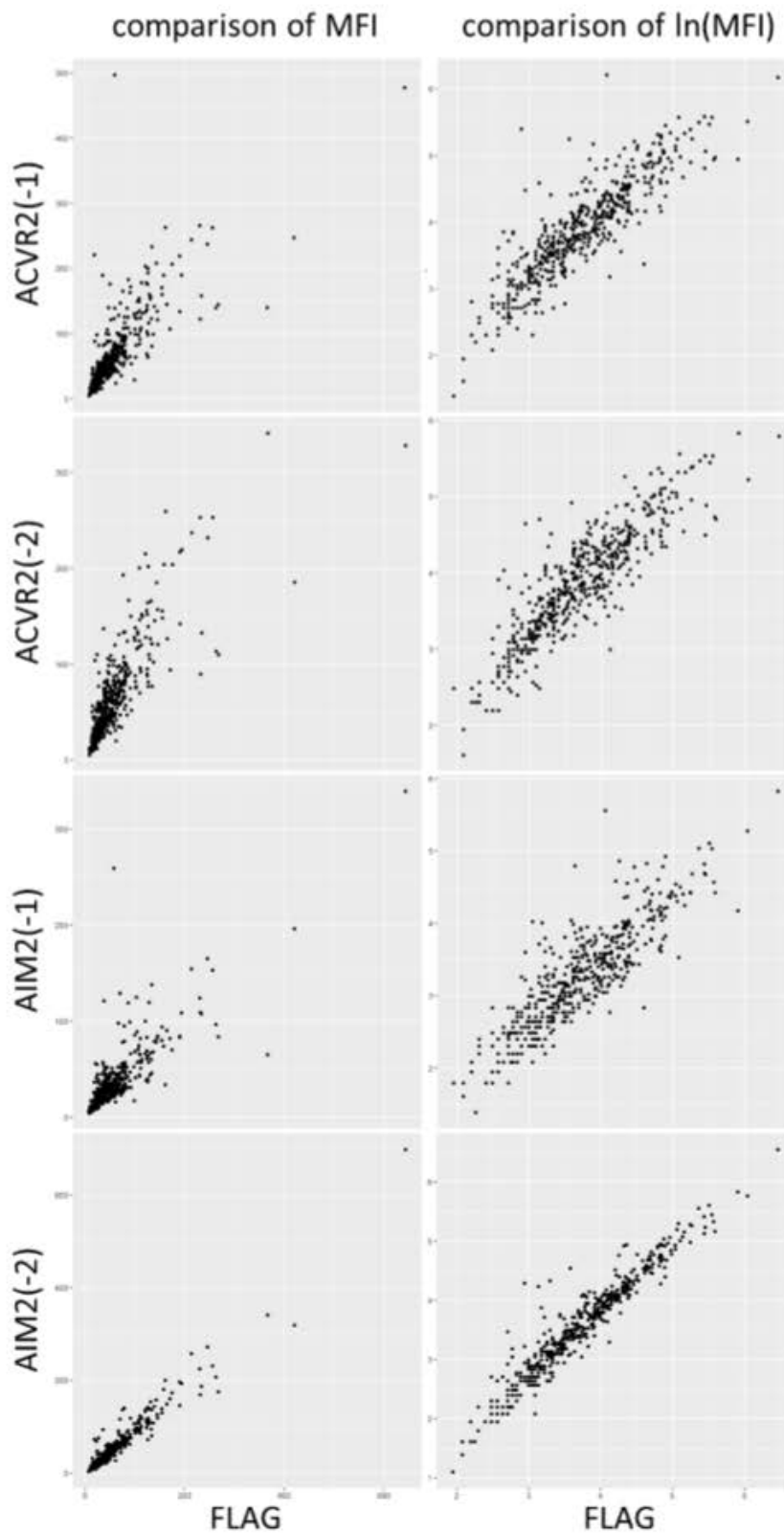


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

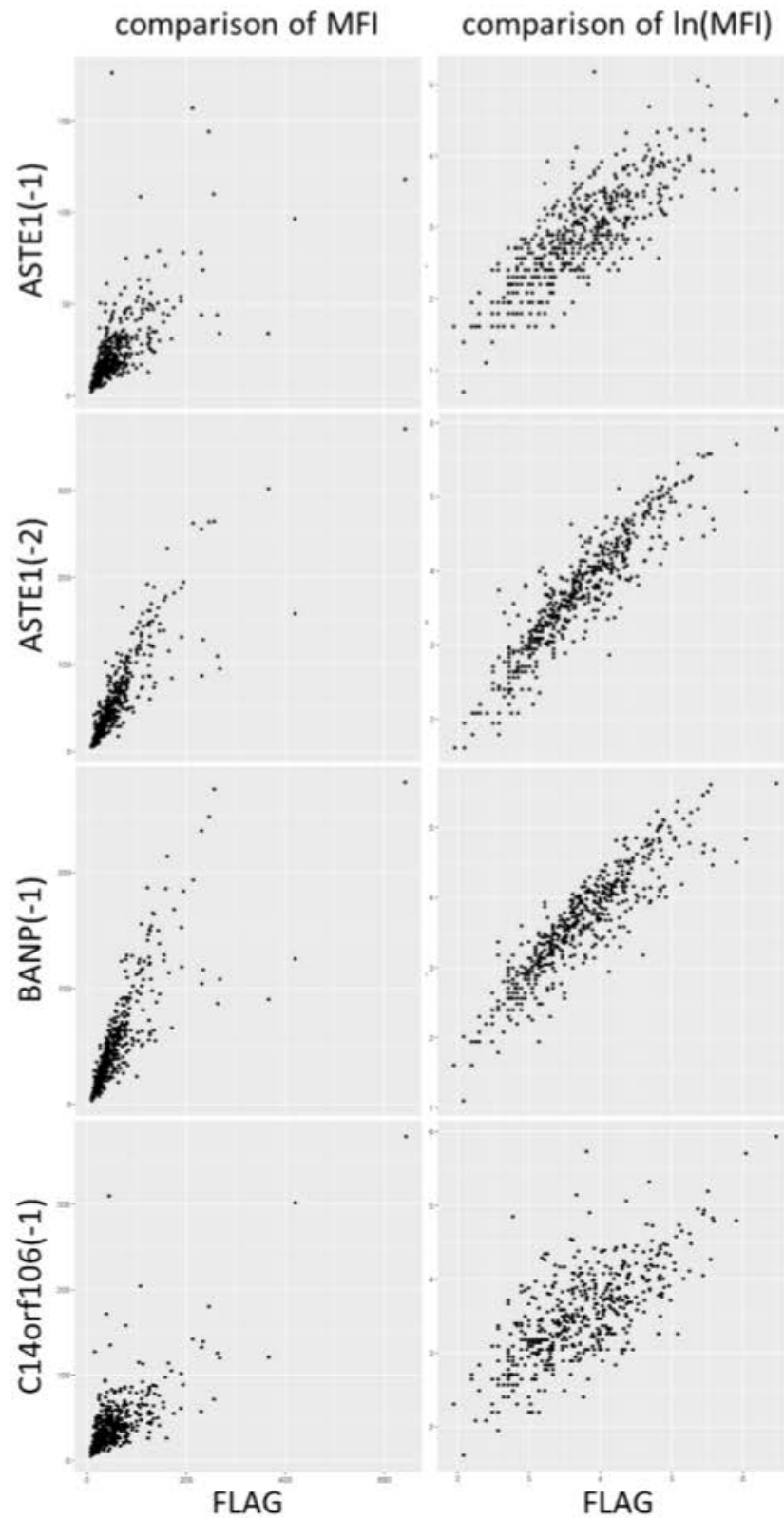


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

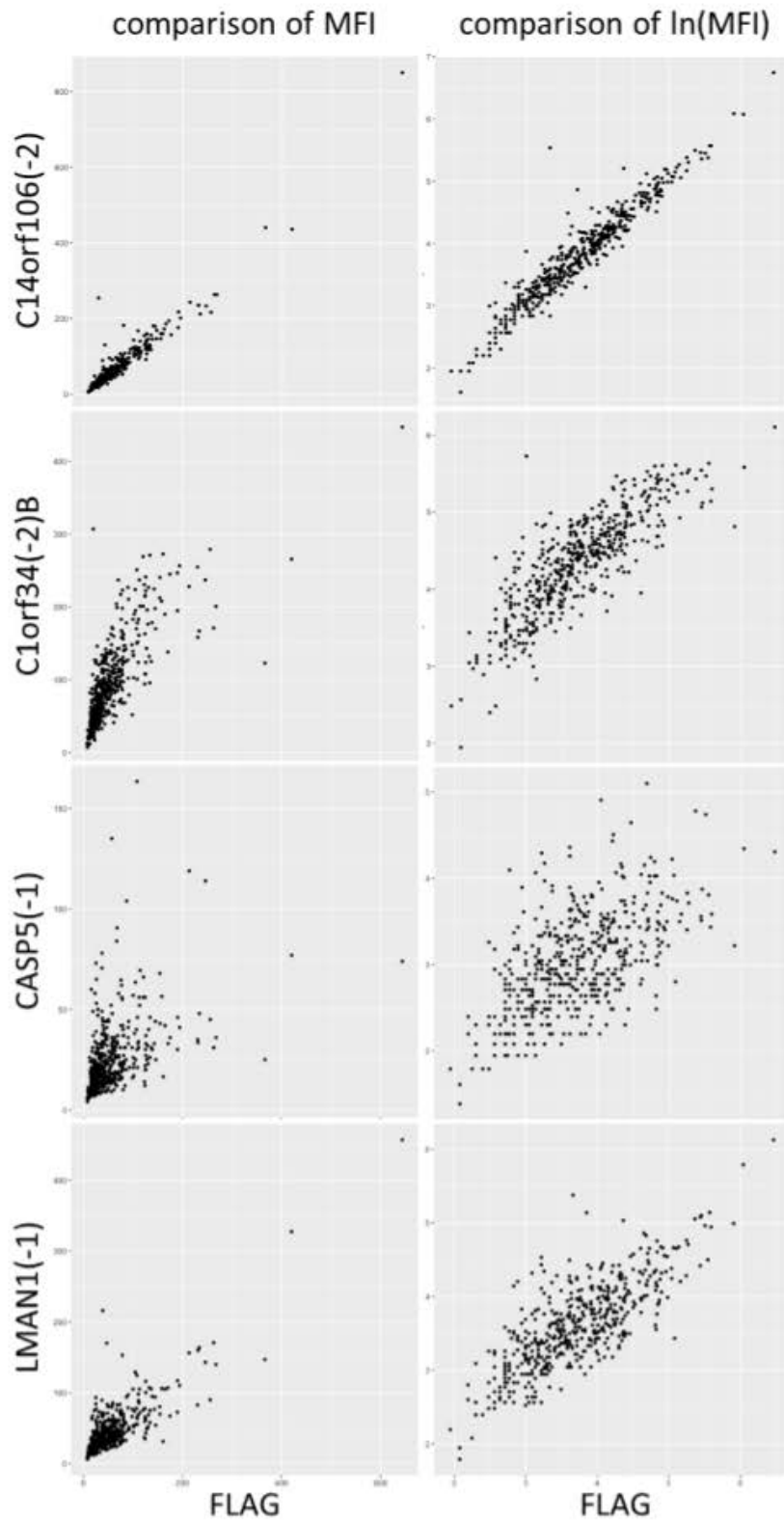


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

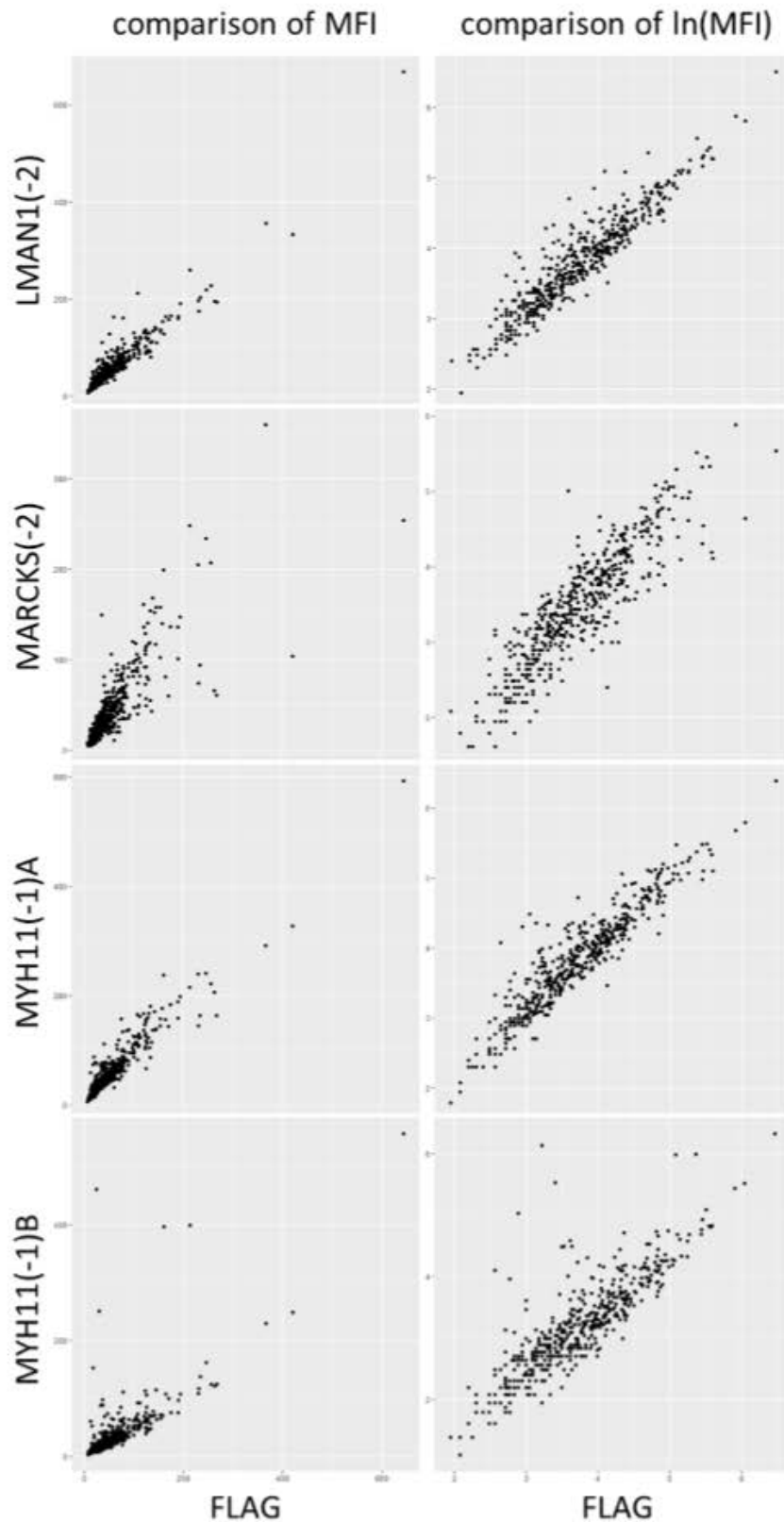


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

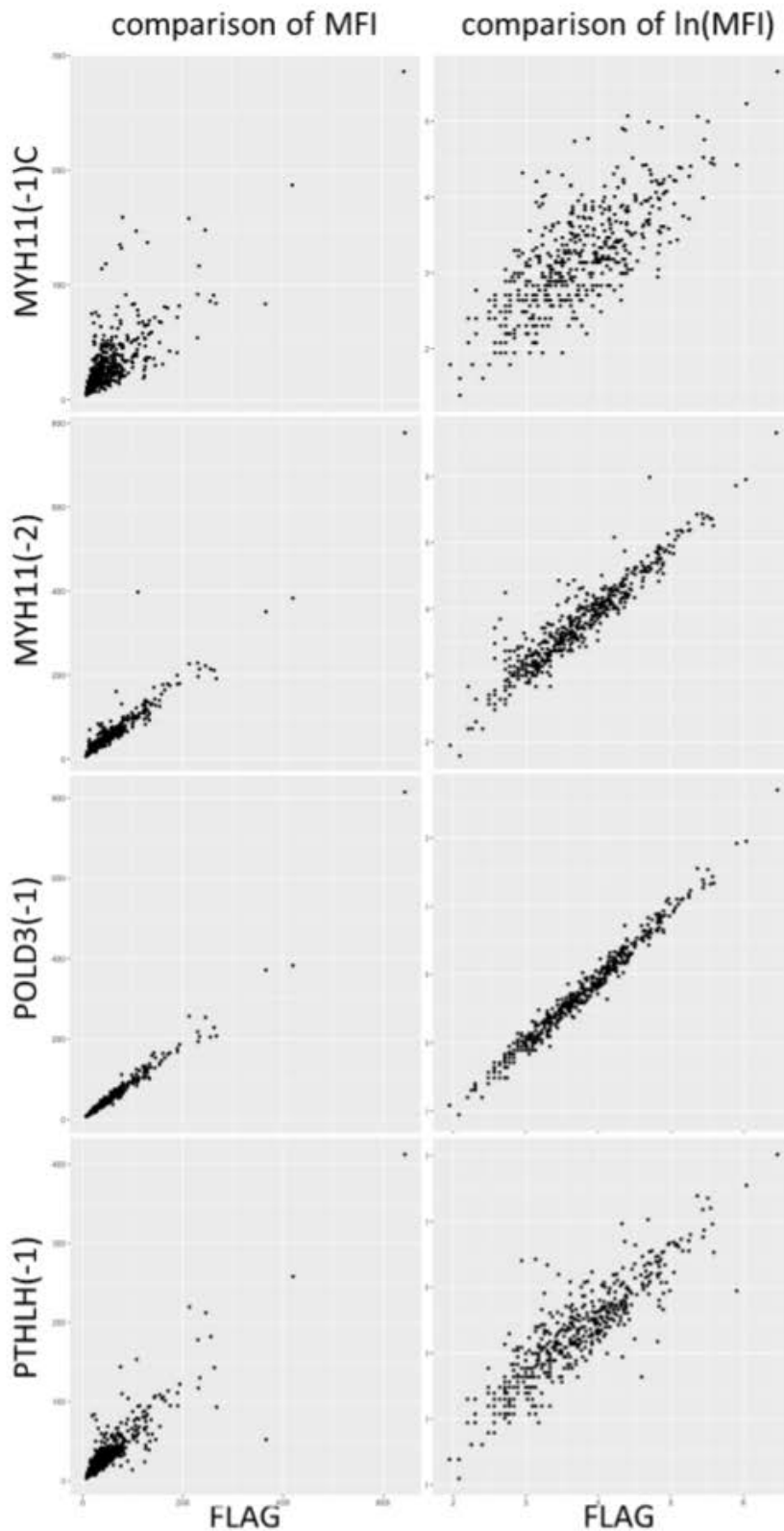


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

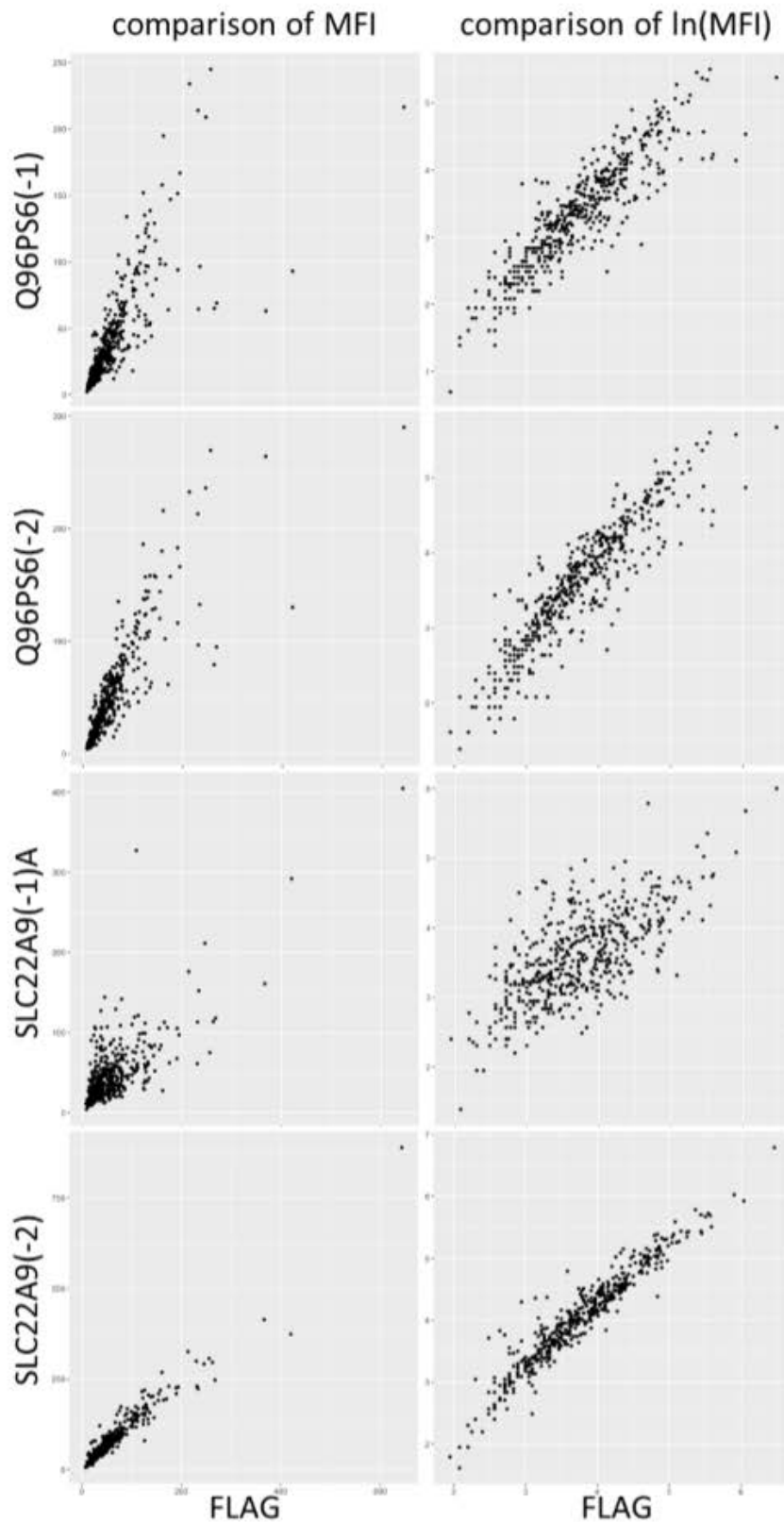


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

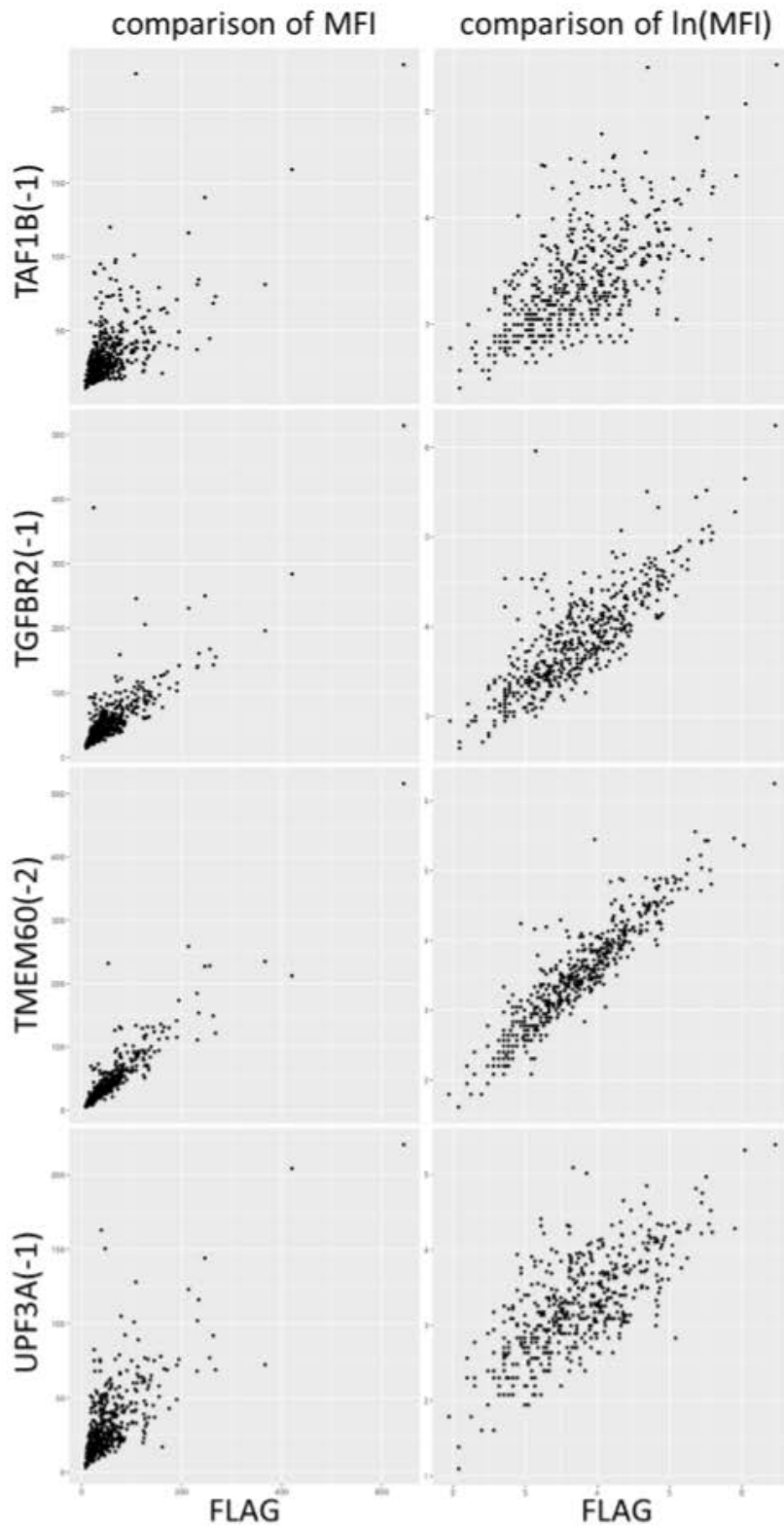


Figure 7.1: Distribution of frameshift peptide (FSP) versus FLAG-only control median fluorescence intensity (MFI). Both raw MFI data and $\ln(\text{MFI})$ data are shown. As discussed in Section 3.4, log transformation reduces heteroscedasticity in the variance between FSP and control MFI.

7.9. Appendix I: Read-balancing the Multiplex Pool of Molecular Inversion Probes

Marker	Reads Assigned/ Mean Reads	Read Balanced Volume μ l
<i>BRAF</i>	0.80	3.73
DEPDC2	0.59	5.06
GM01	1.57	1.91
GM07	1.20	2.50
GM09	0.58	5.15
GM11	0.81	3.71
GM14	1.65	1.82
GM17	1.40	2.14
GM22	3.14	0.96
GM26	0.64	4.65
GM29	1.44	2.09
IM16	1.00	2.99
IM49	0.24	12.55
<i>KRAS</i>	0.30	9.87
LR10	0.62	4.81
LR11	0.72	4.19
LR17	1.26	2.37
LR20	0.34	8.84
LR24	0.22	13.83
LR36	1.13	2.66
LR40	0.71	4.23
LR44	0.27	11.14
LR46	2.36	1.27
LR48	2.22	1.35
LR49	0.94	3.21
LR52	1.22	2.46
Mean MIP μ l	Total MIP μ l	Total pool μ l
4.60	119.50	460

Table 7.9: Read-balancing the multiplex pool of molecular inversion probes (MIP). For each marker, the number of reads assigned across all samples in the training cohort was divided by the mean number of reads detected across all markers. This gave a “correction factor”: markers with a correction > 1 were over represented in the read data, and markers with a correction factor < 1 were under represented in the read data. To balance the number of reads detected for each marker, this correction factor was applied to the volume of each MIP added into the multiplex pool, where the balanced volume was equal to 3μ l divided by the correction factor. The mean volume of 10nM MIPs added into the pool was calculated to determine the total pool volume required to dilute each MIP 100-fold to reach a mean working stock concentration of 0.1nM for each MIP.

7.10. Appendix J: Reagent Costs of the smMIP and Sequencing-based MSI Assay

smMIP-MSI assay on Illumina v2 Micro Kit (4 million read capacity)				
Item	Cost per item (GBP)	No. of samples	Cost per sample (GBP)	
MIPs (25 MIPs per reaction)	500	500000	0.00	*
T4 polynucleotide kinase (2500 Units, 10U/ul)	170	500000	0.00	*
T4 DNA ligase buffer (6ml)	16	500000	0.00	*
Ampligase DNA ligase (1000 Units, 5U/ul)	142	200	0.71	
Herculase DNA polymerase 2 (400 reaction kit)	290	225	1.29	†
Exonuclease I (15,000 Units, 20U/ul)	207	750	0.28	
Exonuclease III (25,000 Units, 100U/ul)	182	250	0.73	
MIP sample indexing reverse primers (60 oligos)	510	192000	0.00	*
MIP forward primer	8.5	3200	0.00	*
QIAxcel screening kit (2400 samples)	542	2400	0.23	
AMPure XP DNA cleanup kit (60 ml)	1146	1333	0.86	
Qubit dsDNA HS kit (500 reaction kit)	175	500	0.35	
Custom sequencing primers (3 oligos)	15	3000	0.01	
Illumina v2 Micro Kit (4 million read capacity)	365	60	6.08	‡
Illumina MiSeq machine run (Genomics Core charge)	130	60	2.17	
TOTAL			12.70	

Table 7.10: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v2 Micro Kit and 25 markers.

* Once smMIPs and primers are purchased the cost per sample is negligible.

† 0.89ul polymerase used per sample, equivalent to 1.78 kit reactions.

‡ Using 2000 reads/marker/sample and 25 markers/sample and assuming 0.75 on target reads.

smMIP-MSI assay on Illumina v3 Kit (25 million read capacity)				
Item	Cost per item (GBP)	No. of samples	Cost per sample (GBP)	
MIPs (25 MIPs per reaction)	500	500000	0.00	*
T4 polynucleotide kinase (2500 Units, 10U/ul)	170	500000	0.00	*
T4 DNA ligase buffer (6ml)	16	500000	0.00	*
Ampligase DNA ligase (1000 Units, 5U/ul)	142	200	0.71	
Herculase DNA polymerase 2 (400 reaction kit)	290	225	1.29	†
Exonuclease I (15,000 Units, 20U/ul)	207	750	0.28	
Exonuclease III (25,000 Units, 100U/ul)	182	250	0.73	
MIP sample indexing reverse primers (375 oligos)	3187.5	1200000	0.00	*
MIP forward primer	8.5	3200	0.00	*
QIAxcel screening kit (2400 samples)	542	2400	0.23	
AMPure XP DNA cleanup kit (60 ml)	1146	1333	0.86	
Qubit dsDNA HS kit (500 reaction kit)	175	500	0.35	
Custom sequencing primers (3 oligos)	15	18750	0.00	
Illumina v3 Kit (25 million read capacity)	1273	375	3.39	‡
Illumina MiSeq machine run (Genomics Core charge)	130	375	0.35	
TOTAL			8.19	

Table 7.11: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v3 Kit and 25 markers.

* Once smMIPs and primers are purchased the cost per sample is negligible.

† 0.89ul polymerase used per sample, equivalent to 1.78 kit reactions.

‡ Using 2000 reads/marker/sample and 25 markers/sample and assuming 0.75 on target reads.

smMIP-MSI assay on Illumina v2 Micro Kit (4 million read capacity)				
Item	Cost per item (GBP)	No. of samples	Cost per sample (GBP)	
MIPs (10 MIPs per reaction)	200	500000	0.00	*
T4 polynucleotide kinase (2500 Units, 10U/ul)	170	500000	0.00	*
T4 DNA ligase buffer (6ml)	16	500000	0.00	*
Ampligase DNA ligase (1000 Units, 5U/ul)	142	200	0.71	
Herculase DNA polymerase 2 (400 reaction kit)	290	225	1.29	†
Exonuclease I (15,000 Units, 20U/ul)	207	750	0.28	
Exonuclease III (25,000 Units, 100U/ul)	182	250	0.73	
MIP sample indexing reverse primers (60 oligos)	510	192000	0.00	*
MIP forward primer	8.5	3200	0.00	*
QIAxcel screening kit (2400 samples)	542	2400	0.23	
AMPure XP DNA cleanup kit (60 ml)	1146	1333	0.86	
Qubit dsDNA HS kit (500 reaction kit)	175	500	0.35	
Custom sequencing primers (3 oligos)	15	3000	0.01	
Illumina v2 Micro Kit (4 million read capacity)	365	150	2.43	‡
Illumina MiSeq machine run (Genomics Core charge)	130	150	0.87	
TOTAL			7.75	

Table 7.12: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v2 Micro Kit and 10 markers.

* Once smMIPs and primers are purchased the cost per sample is negligible.

† 0.89ul polymerase used per sample, equivalent to 1.78 kit reactions.

‡ Using 2000 reads/marker/sample and 10 markers/sample and assuming 0.75 on target reads.

smMIP-MSI assay on Illumina v3 Kit (25 million read capacity)				
Item	Cost per item (GBP)	No. of samples	Cost per sample (GBP)	
MIPs (10 MIPs per reaction)	200	500000	0.00	*
T4 polynucleotide kinase (2500 Units, 10U/ul)	170	500000	0.00	*
T4 DNA ligase buffer (6ml)	16	500000	0.00	*
Ampligase DNA ligase (1000 Units, 5U/ul)	142	200	0.71	
Herculase DNA polymerase 2 (400 reaction kit)	290	225	1.29	†
Exonuclease I (15,000 Units, 20U/ul)	207	750	0.28	
Exonuclease III (25,000 Units, 100U/ul)	182	250	0.73	
MIP sample indexing reverse primers (375 oligos)	3187.5	1200000	0.00	*
MIP forward primer	8.5	3200	0.00	*
QIAxcel screening kit (2400 samples)	542	2400	0.23	
AMPure XP DNA cleanup kit (60 ml)	1146	1333	0.86	
Qubit dsDNA HS kit (500 reaction kit)	175	500	0.35	
Custom sequencing primers (3 oligos)	15	18750	0.00	
Illumina v3 Kit (25 million read capacity)	1273	937	1.36	‡
Illumina MiSeq machine run (Genomics Core charge)	130	937	0.14	
TOTAL			5.94	

Table 7.13: Reagent costs for the smMIP and sequencing-based MSI assay, using a MiSeq v3 Kit and 10 markers.

* Once smMIPs and primers are purchased the cost per sample is negligible.

† 0.89ul polymerase used per sample, equivalent to 1.78 kit reactions.

‡ Using 2000 reads/marker/sample and 10 markers/sample and assuming 0.75 on target reads.

7.11. Appendix K: Modelling the Proportion of smSequences with WT Microsatellite Length by the Beta Distribution

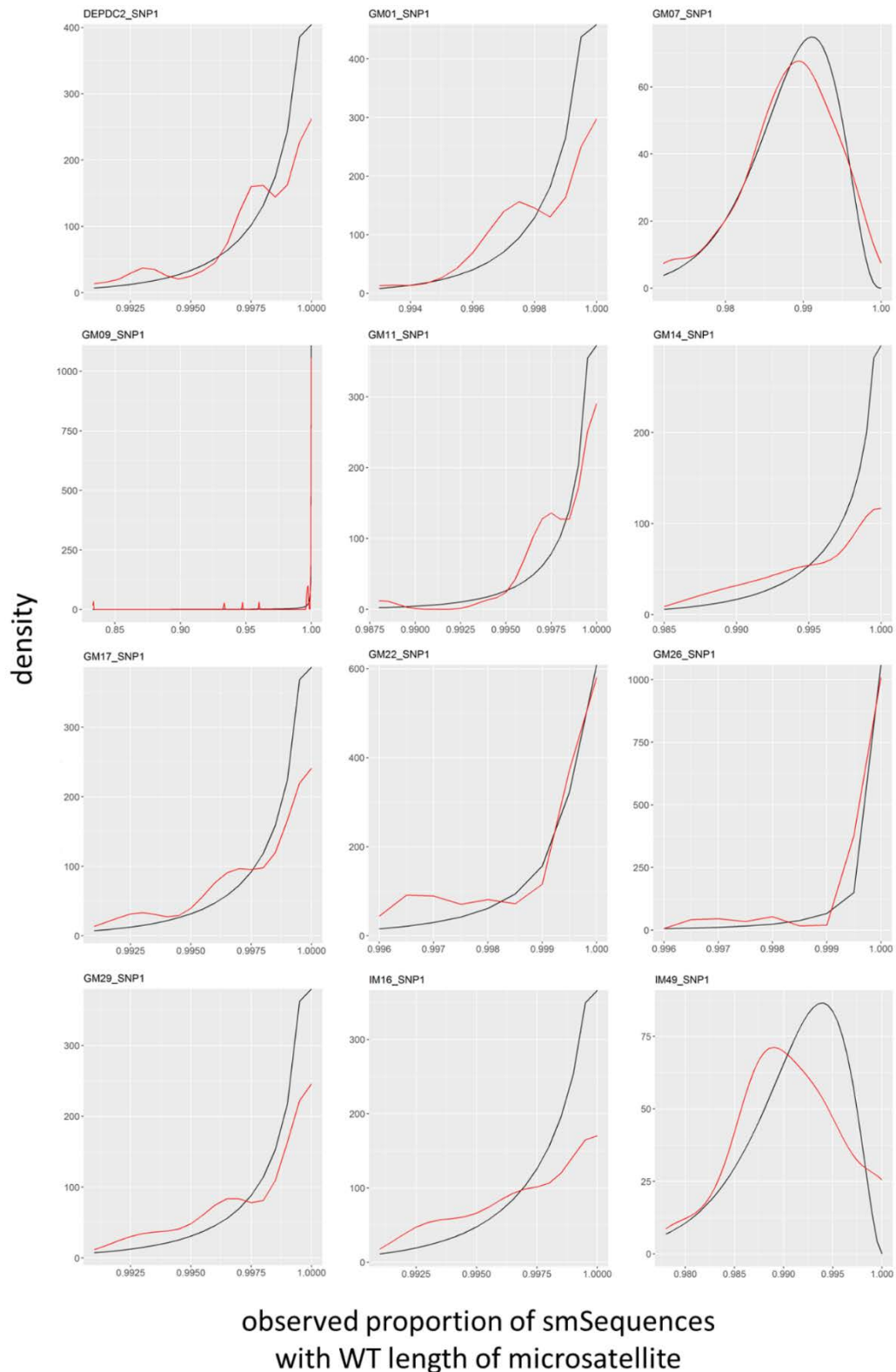


Figure 7.2: Comparison of Beta (black line) and empirical (red line) distributions. 40 non-CMMRD control samples were sequenced using the smMIP and sequencing-based MSI assay, and the proportion of smSequences containing WT reads (prWT) determined for each marker. The distribution of prWT in a control population was modelled by a Beta distribution for each marker (Section 5.5), which is compared here to the empirical distribution.

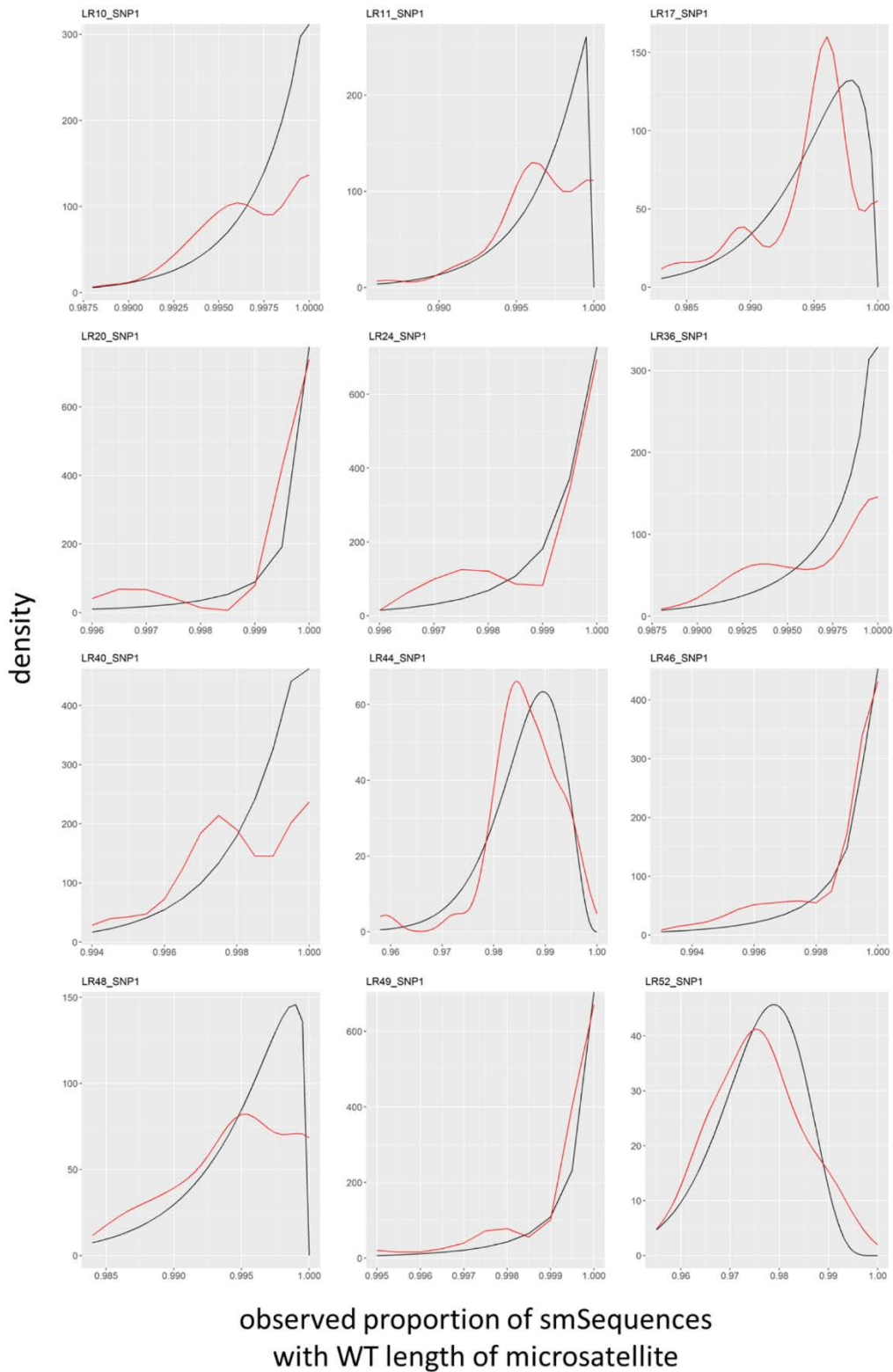




Figure 7.2: Comparison of Beta (black line) and empirical (red line) distributions of microsatellite length variants in controls. 40 non-CMMRD control samples were sequenced using the smMIP and sequencing-based MSI assay, and the proportion of smSequences containing WT reads (prWT) determined for each marker. The distribution of prWT in a control population was modelled by a Beta distribution for each marker (Section 5.5), which is compared here to the empirical distribution.

7.12. Appendix L: Three molecular pathways model colorectal carcinogenesis in Lynch syndrome. Ahadova, Gallon *et al*, 2018.

As part of the work described in Section 2.7.1 and Chapter 3, I had a one month placement at the Department of Applied Tumour Biology, Heidelberg University Hospital, Heidelberg, Germany. During this time, I worked with collaborators (Dr Aysel Ahadova and Dr Matthias Kloor) to study a cohort of adenomas collected from Lynch syndrome gene carriers participating in the CAPP2 clinical trial (Burn *et al*, 2011). In summary, 21 dysplastic adenomas were analysed by IHC for loss of MMR protein according to the germline affected MMR gene. Of these, 19 (90.5%) were MMRd, and we found 4 samples where it appeared that MMR deficiency had occurred in normal colorectal mucosa and progressed to an MMRd dysplastic adenoma. This was clearest in the sample shown in Figure 1.7. In addition, I performed statistical analyses comparing the observed frequency of mutations in MMRd and MMRp CRCs with mutational signatures (Alexandrov *et al*, 2013). I am joint first author for the respective publication (Ahadova *et al*, 2018). However I felt that this work was too distinct from the other studies presented in this thesis, and hence it has not been included. The manuscript is appended to this section; as a distinct publication it does not follow the same scheme of page numbering.

Three molecular pathways model colorectal carcinogenesis in Lynch syndrome

Aysel Ahadova ^{1,2,3†}, Richard Gallon^{4†}, Johannes Gebert^{1,2,3}, Alexej Ballhausen^{1,2,3}, Volker Endris⁵, Martina Kirchner⁵, Albrecht Stenzinger⁵, John Burn⁴, Magnus von Knebel Doeberitz ^{1,2,3}, Hendrik Bläker⁶ and Matthias Kloor^{1,2,3}

¹ Department of Applied Tumor Biology, Institute of Pathology, University Hospital Heidelberg Im Neuenheimer Feld 224, 69120 Heidelberg, Germany

² Clinical Cooperation Unit Applied Tumor Biology, German Cancer Research Center (DKFZ), Im Neuenheimer Feld 280, 69120 Heidelberg, Germany

³ Molecular Medicine Partnership Unit (MMPU), University Hospital Heidelberg, Heidelberg, Germany

⁴ Institute of Genetic Medicine, Newcastle University, International Centre for Life, Central Parkway, Newcastle upon Tyne, United Kingdom

⁵ Department of General Pathology, Institute of Pathology, University Hospital Heidelberg Im Neuenheimer Feld 224, Heidelberg 69120, Germany

⁶ Department of General Pathology, University Hospital Charité, Charitéplatz 1, Berlin 10117, Germany

Lynch syndrome is caused by germline mutations of DNA mismatch repair (MMR) genes. MMR deficiency has long been regarded as a secondary event in the pathogenesis of Lynch syndrome colorectal cancers. Recently, this concept has been challenged by the discovery of MMR-deficient crypt foci in the normal mucosa. We aimed to reconstruct colorectal carcinogenesis in Lynch syndrome by collecting molecular and histology evidence from Lynch syndrome adenomas and carcinomas. We determined the frequency of MMR deficiency in adenomas from Lynch syndrome mutation carriers by immunohistochemistry and by systematic literature analysis. To trace back the pathways of pathogenesis, histological growth patterns and mutational signatures were analyzed in Lynch syndrome colorectal cancers. Literature and immunohistochemistry analysis demonstrated MMR deficiency in 491 (76.7%) out of 640 adenomas (95% CI: 73.3% to 79.8%) from Lynch syndrome mutation carriers. Histologically normal MMR-deficient crypts were found directly adjacent to dysplastic adenoma tissue, proving their role as tumor precursors in Lynch syndrome. Accordingly, mutation signature analysis in Lynch colorectal cancers revealed that *KRAS* and *APC* mutations commonly occur after the onset of MMR deficiency. Tumors lacking evidence of polypous growth frequently presented with *CTNNB1* and *TP53* mutations. Our findings demonstrate that Lynch syndrome colorectal cancers can develop through three pathways, with MMR deficiency commonly representing an early and possibly initiating event. This underlines that targeting MMR-deficient cells by chemoprevention or vaccines against MMR deficiency-induced frameshift peptide neoantigens holds promise for tumor prevention in Lynch syndrome.

Lynch syndrome and familial adenomatous polyposis (FAP) are two major inherited tumor syndromes predisposing to colorectal cancer.¹ FAP is inherited through a germline

Key words: colorectal cancer, Lynch syndrome, microsatellite instability, mismatch repair deficiency

Abbreviations: CI: confidence interval; FAP: familial adenomatous polyposis; MMR: mismatch repair; MMR-DCF: mismatch repair-deficient crypt focus; MSI: microsatellite unstable; MSI-H: high level microsatellite instability; MSI-L: low level microsatellite instability; MSS: microsatellite stable

Additional Supporting Information may be found in the online version of this article.

†A.A. and R.G. contributed equally to the manuscript.

Grant sponsor: German Research Foundation (Deutsche Forschungsgemeinschaft, DFG); **Grant numbers:** KFO227, KL2354;

Grant sponsor: Wilhelm Sander Foundation (2016.056.1)

DOI: 10.1002/ijc.31300

History: Received 24 Oct 2017; Accepted 1 Feb 2018; Online 9 Feb 2018

Correspondence to: Matthias Kloor, Im Neuenheimer Feld 224, 69120 Heidelberg, Germany, E-mail: matthias.kloor@med.uni-heidelberg.de, Tel.: +49-6221-565210, Fax: +49-6221-565981

mutation in the *APC* gene, which, upon a second somatic hit, results in the formation of hundreds to thousands of adenomatous polyps (polyposis) in the colonic mucosa of mutation carriers.² The multiple clinically detectable lesions illustrate that *APC* germline mutations lead to a strong increase of the adenoma initiation rate in the colorectum.^{2,3}

In contrast to FAP, polyposis is absent in Lynch syndrome, the most common hereditary colorectal cancer syndrome in adults, which is caused by germline mutations of DNA mismatch repair (MMR) genes.^{4,5} Although some studies found an increased adenoma incidence in Lynch syndrome mutation carriers, adenomatous polyps in Lynch syndrome were only slightly more prevalent than in the unaffected population.^{6,7} This observation has suggested that Lynch syndrome-causing MMR gene germline mutations do not increase the adenoma initiation rate, but rather accelerate the progression of preformed adenomas, which have developed independently from MMR deficiency, into invasive cancer.^{6,8} Therefore, Lynch syndrome was long regarded as a prime example of an inherited tumor predisposition that does not act through enhanced tumor initiation.^{6,8} MMR deficiency, accordingly, has commonly been believed to be a

What's new?

Whether mutations in mismatch repair (MMR) genes play an initiating or a secondary role in colorectal carcinogenesis in Lynch syndrome is unclear. To better understand the pathogenic process, the authors of this study developed a Lynch syndrome model delineating three molecular pathways of colorectal cancer formation. Some colorectal cancers were found to grow from MMR-proficient adenomas after secondary inactivation of the MMR system. However, most colorectal cancers developed from MMR-deficient precursor lesions, either via an adenomatous phase or as nonpolypous lesions. The findings underline the importance of prevention measures targeting MMR-deficient cells in the clinical management of Lynch syndrome.

secondary event, and somatic mutations of MMR genes were thought to occur after the formation of polyps that had been caused by *APC* mutations or other events occurring independently from MMR deficiency.

Various observations of Lynch syndrome pathogenesis have been interpreted as supportive of this concept: these include the existence of polyps with retained or partial expression of MMR proteins⁹ found in some Lynch syndrome patients.^{10–12} Moreover, correlation of MMR deficiency with the higher grade and bigger size of the adenomas seemed to further corroborate the role of MMR deficiency as a “noninitiating” event in Lynch syndrome-associated colorectal carcinomas.^{7,9,11,13–15}

In recent years, however, the classical view of Lynch syndrome as an “accelerating” disease has been challenged, most importantly by the discovery of MMR-deficient crypt foci (MMR-DCF), colonic crypts presenting with a normal histological appearance but already lacking the expression of MMR proteins.^{10,16} Although it has not been clear whether these MMR-DCF had the potential of being true cancer precursors, it demonstrated that MMR deficiency can strike in phenotypically normal cells and it in fact does so very frequently during the life of Lynch syndrome mutation carriers.¹⁶ This observation opened up the possibility that MMR-deficient colorectal cancers in Lynch syndrome, at least to a certain proportion, may also develop from such MMR-DCF. This concept has also been supported by several observations in path_ *MMR* gene variant carriers: The majority of adenomas show complete and homogeneous lack of MMR protein expression in all dysplastic cells,¹⁷ 50% of MMR-deficient adenomas smaller than 5 mm have high grade dysplasia,¹⁴ and some MMR-DCF have aberrant histology,¹⁸ pointing to their potential role as cancer precursors. Very recently, molecular studies on Lynch syndrome-associated colorectal cancers have provided further support for MMR deficiency as an event commonly preceding adenoma formation.¹⁹

These conflicting observations could be explained by the existence of multiple, common pathways of colorectal tumorigenesis in Lynch syndrome whereby MMR deficiency can either precede or follow adenoma formation. Furthermore, an entirely different pathway of carcinogenesis has been described that may bypass the formation of polyps and lead to the formation of invasive cancers from MMR-DCF through a nonpolypous progression pathway.²⁰ The existence

of MMR-deficient and nonpolypous lesions destined for either rapid polypous growth or direct tissue invasion would have wide ranging clinical implications, as such lesions would escape colonoscopic detection and polypectomy,²⁰ the recommended surveillance and prevention method in Lynch syndrome patients. Indeed, whilst colonic surveillance and polypectomy of Lynch syndrome patients reduces colorectal cancer-associated mortality,²¹ the high frequency of interval cancers in patients under regular colonoscopic surveillance with polypectomy suggests that a significant proportion of colorectal lesions are undetectable by colonoscopy and subsequently manifest as cancer within surveillance intervals.²² This illustrates that the sequence of mutational events in Lynch syndrome carcinogenesis, which still represents a highly controversial topic, is of crucial relevance to the optimal clinical management of these patients.

In addition to the single putative pathway in which MMR deficiency is a secondary, accelerating event, we hypothesized that two more common pathways can contribute to Lynch colorectal pathogenesis. These pathways are both initiated by nonpolypous and MMR-deficient precursor lesions that can either develop into polypous adenocarcinoma or invade directly into the colorectal wall. Therefore, we aimed to reconstruct the sequence of somatic mutational events in Lynch syndrome carcinogenesis from two perspectives. First, we performed a systematic literature review to determine the proportion of MMR-deficient adenomas in Lynch syndrome mutation carriers, complemented by the analysis of our own collection of adenoma samples. Second, we evaluated next-generation sequencing data of Lynch syndrome-associated colorectal cancers to detect mutational signatures reflecting the sequence of mutational events²³ and to identify fingerprints indicative of polypous or nonpolypous growth.

Materials and Methods**Patients and tumor specimens**

Formalin-fixed paraffin-embedded tissue sections of 21 dysplastic adenomas from 15 Lynch syndrome mutation carriers participating in the CAPP2 trial (colorectal adenoma/carcinoma prevention programme 2, path_ *MLH1* variant carriers, $n = 10$; path_ *MSH2* variant carriers, $n = 5$) were retrieved and available for the analysis of MMR protein expression. The collection of tumors has been described in a previous study reporting the prevalence of polyps in Lynch syndrome mutation carriers.²⁴

Formalin-fixed, paraffin-embedded archival tissue blocks from 21 carcinomas from Lynch syndrome mutation carriers (path_*MLH1*, *n* = 9, path_*MSH2*, *n* = 8, path_*MSH6*, *n* = 3, path_*PMS2*, *n* = 1) were obtained from the Department of Applied Tumor Biology, Institute of Pathology, University Hospital Heidelberg. All patients provided their informed and written consent in frame of the German HNPCC Consortium, which was approved by the Institutional Ethics Committee.

Literature survey

The systematic literature search of studies listed in NCBI Pubmed by May 15, 2017 was performed using the following keywords: {mismatch repair deficiency} OR {mismatch repair protein expression} OR {mismatch repair gene} OR {microsatellite instability} OR {microsatellite unstable} OR {MMR deficiency} OR {MMR gene} OR {MMR loss} OR {MMR protein expression} OR {MSI} OR {MSI-H}) AND ({adenoma} OR {adenomatous} OR {colorectal adenoma} OR {colorectal polyp} OR {dysplasia} OR {dysplastic lesions} OR {polyp} OR {precancerous} OR {precursor}) AND ({hereditary non-polyposis colorectal cancer syndrome} OR {HNPCC} OR {Lynch} OR {Lynch patients} OR {Lynch syndrome} OR {mutation carriers}). All studies written in English and analyzing MMR deficiency in (1) adenomas (2) from Lynch syndrome patients (3) using immunohistochemical staining of MMR proteins and/or PCR fragment length analysis of mononucleotide microsatellite markers were collected and used for integrated data analysis. Adenomas were included as “Lynch syndrome adenomas” if the described patients fulfilled the following criteria: proven path_*MMR gene* variant carrier OR history of a tumor showing loss of MMR protein expression plus proven path_*MMR gene* variant carrier among first- or second-degree relatives. Only adenomas classified as “dysplastic” were included in the calculation of the proportion of MMR-deficient lesions. The lesions were considered as MMR-deficient if they presented with MMR protein expression loss, instability of >30% of tested markers or both. Studies analyzing <20 lesions from Lynch syndrome patients were excluded. Collection of articles and extraction of data was performed by one author (AA) and verified by a second author (MK).

Immunohistochemistry (IHC)

For detection of MMR protein expression, paraffin blocks were cut into 3- μ m-thick sections. Deparaffinization and tissue staining were performed according to standard protocols published previously.²⁵ The following primary antibodies were used: anti-*MLH1* (clone G168-15, dilution 1:300, BD Pharmingen, Heidelberg, Germany) or anti-*MSH2* (clone FE11, dilution 1:100, Calbiochem, Darmstadt, Germany), depending on the germline mutation status of the respective patient. Staining was visualized using the Vectastain elite ABC detection system (Vector, Burlingame, Calif., USA) and using 3,3'-Diaminobenzidine (DAB) (Dako) as chromogen.

Hematoxylin was used for blue counterstaining of cell nuclei. Cells with nuclear staining signals in MMR protein staining were assessed as MMR-proficient.

Mutational analysis of TCGA and DFCI databases

TCGA²⁶ and DFCI²⁷ databases were used to determine somatic mutation patterns of commonly mutated colon cancer genes *APC* and *KRAS* in microsatellite-unstable (MSI) cancer samples (www.cbioportal.org, status: January 31, 2017)).^{28,29} Mutational and clinical data (including MSI typing and *CIMP/MLH1* methylation status) were downloaded and used for stratification of MSI-H, MSI-L and MSS cancers. For the present study, MSI-L cancers were grouped together with MSS cancers and comprised together the “MSS” group, whereas “MSI-H” cancers are further referred to as “MSI” throughout the manuscript. MSI cancers with a negative *CIMP/MLH1* methylation status were classified as “MSI Lynch” and evaluated together with our own cohort of 21 tumors from Lynch syndrome patients, from which *APC* and *KRAS* mutation data were obtained through next generation panel sequencing. MSI cancers with positive *CIMP/MLH1* methylation status were classified as “MSI sporadic” and evaluated separately. Samples without mutational data or without clinical data containing results of MSI typing were excluded.

The order of mutational events in MSI Lynch colorectal cancers was explored using the relative frequency of mutations in *APC* and *KRAS* and the association of these mutations with MMR deficiency, including insertion and deletion mutations at homopolymers and substitutions. Substitutions were typed according to the flanking nucleotide bases as well as the base transition or transversion, according to the scheme used by Alexandrov *et al.*²³ The substitution probabilities of mutational signature 6 (COSMIC)²³ were used to define expected frequencies of different substitutions in MMR-deficient colorectal cancer. To estimate the proportion of Lynch syndrome colorectal cancers in which MMR deficiency preceded *APC* mutation, *APC* mutations were classified as “MMR deficiency-related” if they were single nucleotide insertions or deletions affecting homopolymer sequences (mononucleotide repeats) or if they were C>T mutations occurring in a CpG sequence context, following a simplified approach based on mutational signatures.²³ Other mutations were classified as “MMR deficiency-unrelated.”

Library preparation and semiconductor sequencing

Targeted next generation sequencing of 21 Lynch syndrome-associated colorectal cancers was performed on IonTorrent PGM and Proton sequencers using a custom 180 amplicon panel (CRC panel) encompassing mutation HotSpot regions in 30 genes³⁰ (Supporting Information, Table 2).

Briefly, amplicon library preparation was performed with the Ion AmpliSeq Library Kit v2.0 using approximately 10 ng of DNA. The DNA was mixed with the primer pool, containing all primers for generating the 180 amplicons and the

AmpliSeq HiFi Master Mix and transferred to a PCR cycler (BioRad, Munich, Germany). After the end of the PCR reaction, primer end sequences were partially digested using FuPa reagent, followed by the ligation of barcoded sequencing adapters (Ion Xpress Barcode Adapters, Life Technologies). The final library was purified using AMPure XP magnetic beads (Beckman Coulter, Krefeld, Germany) and quantified using qPCR (Ion Library Quantitation Kit, Thermo Fisher Scientific, Waltham, USA) on a StepOnePlus qPCR machine (Thermo Fisher Scientific, Waltham, USA). The individual libraries were diluted to a final concentration of 100 pM, pooled and processed to library amplification on an Ion OneTouch2. Unenriched libraries were quality-controlled using Ion Sphere quality control measurement on a QuBit instrument. After library enrichment (Ion OneTouch ES), the library was processed for sequencing using the Ion Torrent 200 bp HiQ sequencing chemistry and the barcoded libraries were loaded onto 318 or PI chips.

Variant calling and annotation

Data analysis was performed using the Ion Torrent Suite Software (version 5.0). After base calling, the reads were aligned against the human genome (hg19) using the TMAP algorithm within the Torrent Suite. Variant calling was performed with the variant caller plugin within the Torrent Suite Software and the IonReporter package using a corresponding bed-file containing the coordinates of the amplified regions. Only variants with an allele frequency >5% and minimum coverage >100 reads were taken into account. Variant annotation was performed using Annovar.³¹ Annotations included information about nucleotide and amino acid changes of RefSeq annotated genes, COSMIC and dbSNP entries as well as detection of possible splice site mutations. For data interpretation and verification, the aligned reads were visualised using the IGV browser (Broad Institute).³²

Statistical analysis

The calculation of 95% confidence intervals (CI) for the observed frequencies of MMR deficiency in Lynch syndrome adenomas was performed using the modified Wald method and GraphPad Prism software (Version 6.02). Fisher's exact test was performed to test for significant differences of mutation frequencies between groups, using GraphPad Prism software (Version 6.02). Median regression and Kernel density estimation of residuals was used in R (Version 3.3.1) to analyse the relationship between the relative frequencies of observed substitutions and their relative probabilities in mutational signature 6.²³

Results

MMR deficiency in Lynch syndrome adenomas

To determine the frequency of DNA mismatch repair deficiency in adenomas from Lynch syndrome mutation carriers, we performed a systematic literature analysis. The initial search of the NCBI Pubmed database resulted in 545 records.

Five hundred and nine publications that evidently did not address the research topic of interest were excluded based on the title. Out of the remaining 36 records, 22 were excluded because they either did not address the research question or information on MMR deficiency could not be extracted systematically for the analyzed samples. An additional 2 studies were excluded because the number of the analyzed samples was too low (Fig. 1a). The 12 studies that fulfilled the inclusion criteria reported information about MMR deficiency in dysplastic adenomas, with the number of samples studied ranging from 25 to 134. In total, 619 Lynch syndrome-associated dysplastic adenomas were included in the quantitative data synthesis for the calculation of the mean frequency of MMR deficiency in Lynch syndrome-associated adenomas. 7 out of 12 published studies used both IHC and PCR methods to determine MMR deficiency, 4 studies used only IHC and 1 study used only PCR-based methods (Fig. 1b). In studies that used both methods, lesions were classified as MMR-deficient if at least one of the methods showed evidence of MMR deficiency. In total, evidence of MMR deficiency was detected in 472 out of 619 adenomas, with a slightly higher proportion of MMR-deficient lesions in cohorts analyzed by IHC compared to PCR-based methods (76.41% vs 69.02%, $p = 0.017$).

In addition to the published series of dysplastic adenomas, we performed immunohistochemical staining of MMR proteins in 21 dysplastic adenoma specimens obtained during the CAPP2 trial (Supporting Information, Table 1, Fig. 2). No significant difference in the proportion of MMR-deficient lesions was observed among lesions from patients taking aspirin and those taking placebo (8 of 10 in the aspirin vs 11/11 in the placebo group, $p = 0.214$). These adenomas were then included in the quantitative data synthesis, constituting a final set of 640 dysplastic adenomas with confirmed MMR deficiency status. Overall, 491 (76.7%) out of these 640 examined lesions showed evidence of MMR deficiency (95% CI: 73.3–79.8%) (Fig. 1c Forest Plot).

Information about heterogeneity of MMR protein expression in Lynch syndrome adenomas, that is, the existence of MMR-proficient dysplastic crypts, can provide important information about the sequence of mutation events in the respective lesion. Among the identified publications, explicit information about heterogeneity was available for 219 dysplastic adenomas that presented with MMR protein loss.^{11,12,17} In addition, all 19 MMR-deficient adenomas from the CAPP2 cohort were analyzed for heterogeneity. Altogether, eight (3.3%) of these 238 adenomas showed MMR-proficient dysplastic crypts. Single adenomas with a heterogeneous MMR protein expression pattern have also been reported in manuscripts not included in the quantitative data synthesis.^{10,33}

Whereas onset of MMR deficiency occurring in an already existing adenoma may manifest as heterogeneity among dysplastic crypts, MMR deficiency preceding adenoma formation may be detectable through the presence of adjacent,

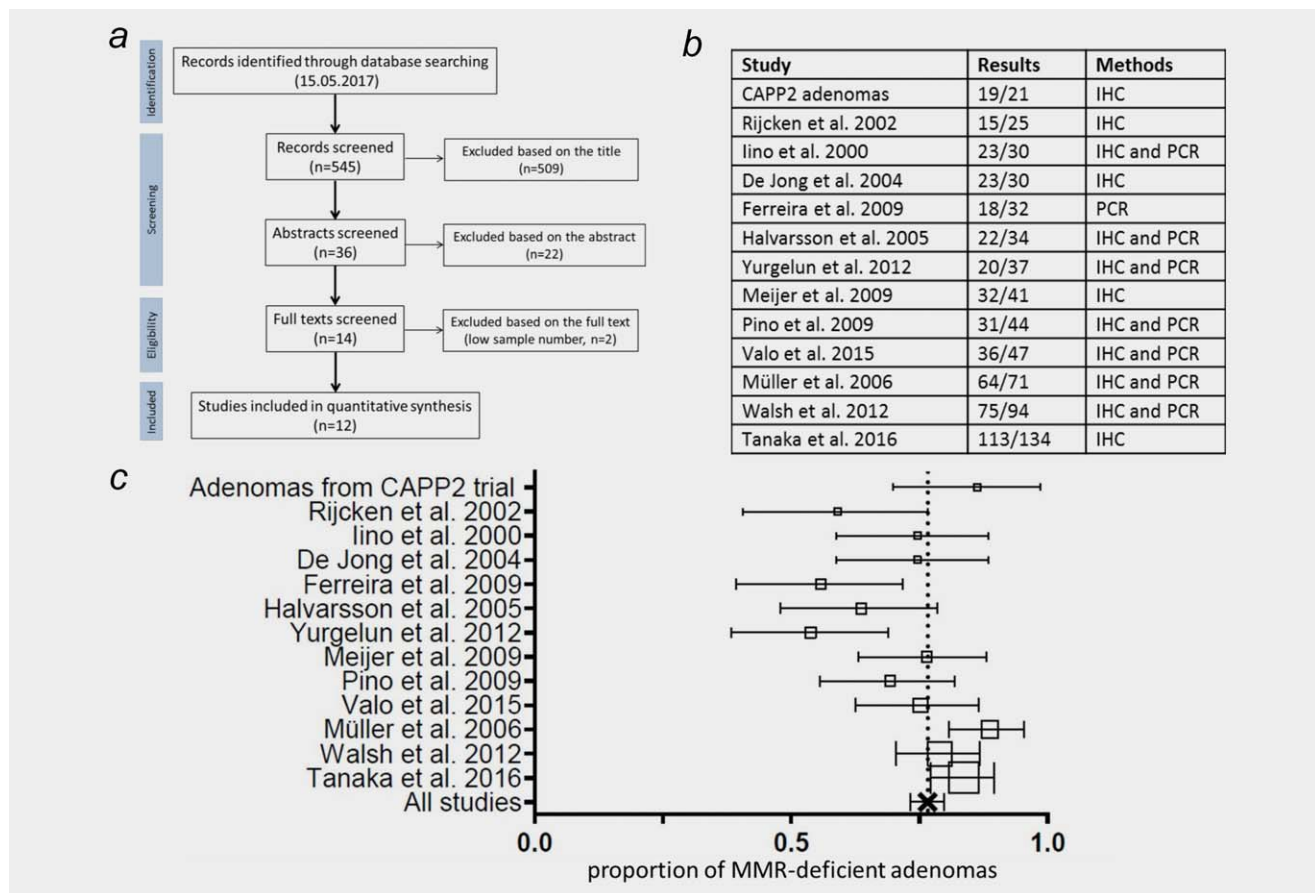


Figure 1. Systematic literature analysis of MSI frequency in Lynch syndrome adenomas. (a) Flow diagram illustrating the numbers of studies screened, assessed for eligibility, and included in the review according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations. From an initial set of 545 records screened, 12 studies were eligible and included in the quantitative synthesis. (b) Tabellary view of the included studies. (c) Forest plot of the results. Together with our own data, MMR deficiency/MSI was detected in 491 (76.7%) out of 640 adenomas from Lynch syndrome mutation carriers. [Color figure can be viewed at wileyonlinelibrary.com]

nondysplastic MMR-DCF.¹⁶ In fact, one of the MMR-deficient adenomas analyzed showed such MMR-deficient nondysplastic crypts (Fig. 2a), indicating that MMR deficiency in this lesion occurred prior to adenoma formation.

Mutation patterns in Lynch syndrome cancers

Although giving a useful hint regarding the possible sequence of mutational events, examination of adenomas does not provide any information about tumor progression to cancer. To estimate the timing of MMR deficiency in Lynch syndrome associated colorectal cancers, we analyzed the frequencies of known driver mutations that are likely related to MMR deficiency to reconstruct the sequence of mutational events in relation to the onset of MMR deficiency.

The focus of this analysis was on mutations of genes playing a key role in the adenoma-carcinoma model of colorectal cancer, *APC*, *KRAS* and *TP53*.^{8,34} The analyzed mutations included small insertions and deletions and base substitutions. Mutational signature 6 from the COSMIC database,²³ which is associated with MMR deficient colorectal cancer, was used as a reference. We hypothesized that the observed

mutations would be associated with this signature if MMR deficiency preceded their occurrence. We supplemented the next generation sequencing data of our collection of Lynch syndrome colorectal cancers ($n = 21$) with data available in the TCGA and DFCI databases, from which mutation data of a total of 752 colorectal cancers samples (“MSI Lynch,” $n = 26$) could be included, resulting in a series of 47 colorectal cancers classified as “MSI Lynch” (Supporting Information, Fig. 1).

For *TP53*, which is known to be rarely mutated in Lynch syndrome colorectal cancers,³⁵ the number of mutation events in Lynch syndrome cancers was too low for a reliable analysis. Therefore we first examined mutations of the *KRAS* gene, which are considered a late event commonly occurring after the initiation of carcinogenesis³⁶ and therefore likely to occur after the onset of MMR deficiency. Oncogenic *KRAS* mutations most commonly affect codons 12 and 13 leading to loss of GTPase activity and in turn constitutive activation of the *KRAS* protein.³⁷ In total, 8 different types of substitution (determined by the base substitution and the context of flanking nucleotides) were identified in the analyzed

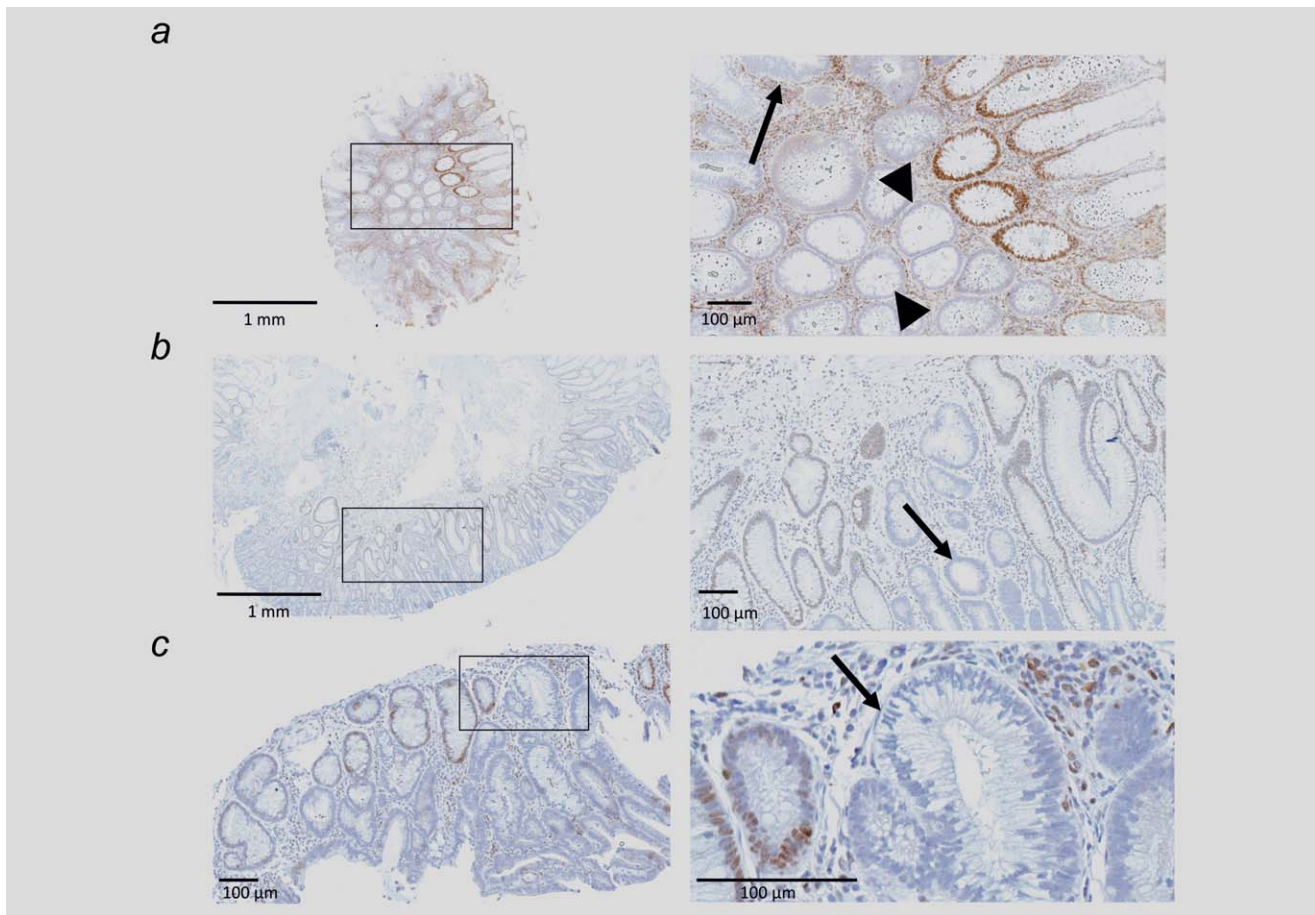


Figure 2. Immunohistochemical MMR protein staining of Lynch syndrome adenomas. Adenomas from the CAPP2 trial were stained for the MMR protein corresponding to the reported germline mutation (*a*: MSH2; *b*, *c*: MLH1). Loss of MMR protein expression is seen in dysplastic adenoma crypts (arrows, *a–c*). Interestingly, nondysplastic crypts demonstrating loss of MMR protein expression were detected in direct vicinity of one adenomatous lesion (arrowhead, *a*). MMR proficiency is indicated by brown nuclear staining, MMR deficient cells have blue nuclei (hematoxylin counterstaining).

colorectal cancers (Fig. 3a). The relative frequencies of point mutation types in the *KRAS* gene observed in MSI Lynch colorectal cancers were linearly related to their relative probabilities in mutational signature 6 using quantile regression ($\beta = 0.752$, $p = 0.033$). However, *KRAS* mutations in MSS colorectal cancers were not linearly related to their relative probabilities in signature 6 ($\beta = 0.325$, $p = 0.674$). These findings suggest that *KRAS* gene mutations in Lynch syndrome cancers commonly occur after the onset of MMR deficiency (Fig. 3a).

In contrast to the codon restriction of *KRAS* mutations, mutations of the *APC* tumor suppressor gene are more widespread over the entire gene sequence. Small insertion/deletion mutations at homopolymers and substitutions were observed in both, MSI Lynch and MSS colorectal cancers, but in different proportions. Single base pair insertion/deletions at homopolymers are strongly associated with MMR deficiency and mutational signature 6, and accounted for 16.7% of *APC* mutations in Lynch syndrome colorectal cancers, but only 5.0% of *APC* mutations in MSS colorectal cancers (Fisher's

exact test, $p = 0.012$). As for *KRAS*, the relative frequency of *APC* substitutions were linearly related to their relative probabilities from mutational signature 6 in MSI Lynch colorectal cancers ($\beta = 1.16$, $p < 0.001$) but not in MSS colorectal cancers ($\beta = 0.494$, $p = 0.336$, Fig. 3b).

As a rough approximation of the proportion of tumors that acquire MMR deficiency before *APC* mutation, the observed mutations were grouped into MMR deficiency-related and MMR deficiency-unrelated. Among Lynch syndrome cancers ($n = 47$), 27 (75%) out of the total 36 *APC* mutation events were considered as MMR deficiency-related, being C > T mutations at CpG sites ($n = 21$) or single nucleotide insertion/deletions at homopolymer sequences ($n = 6$), whereas this was the case only in 209 (35%) out of 603 *APC* mutations in MSS cancers (Table 1). Using the proportion of such mutations in MSS cancers as a background, we predicted the proportion of *APC* mutations occurring after the onset of MMR deficiency in hereditary MSI cancers to be 61% (95% CI: 33% to 80%) (Fig. 3c), which is consistent with the proportion of adenomas with MMR deficiency.

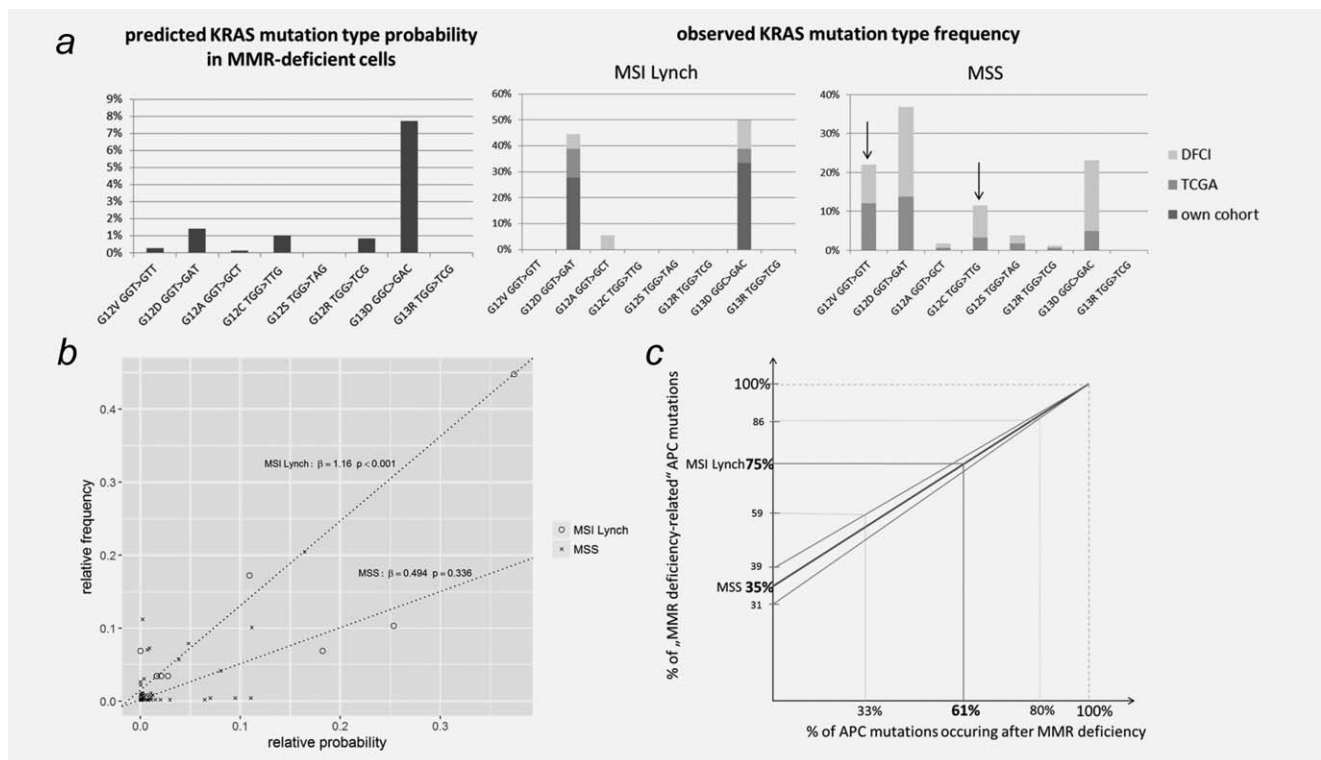


Figure 3. Mutation signature of MMR deficiency in *KRAS* and *APC* mutations. (a) *KRAS* mutation patterns differed significantly between MSI and MSS colorectal cancers, with a clear predominance of G13D and G12D mutations in MSI cancers, in line with mutation signatures of MMR deficiency.²³ (b) Quantile regression between the relative frequencies of *APC* substitutions and their probabilities according to the COSMIC mutation signature of MMR-deficient colorectal cancer (signature 6). A significant relationship was observed for MSI Lynch cancers (open circles represent types of substitution, $\beta = 1.16$, $p < 0.001$), whereas no such relationship was detected for MSS cancers (crosses represent types of substitutions, $\beta = 0.494$, $p = 0.336$). (c) Estimation of the proportion of *APC* mutations occurring after MMR-deficiency based on a simplified approach using mutational signatures: MMR-deficiency-related *APC* mutations were detected in 27 out of 36 (75%) *APC*-mutant Lynch syndrome colorectal cancers. Assuming that only MMR-deficiency related mutations would be observed if all cancers had MMR deficiency preceding *APC* mutation and accounting for a background of 35% such mutations also occurring in MSS cancers, this number corresponds to 61% (95% CI: 33–80%) of MSI Lynch cancers having MMR loss precede *APC* mutation.

Table 1. Type and number of mutations in the *APC* gene of *APC*-mutant cancers

Group	Number of cancers analyzed	Number of mutations	Number of C>T mut.	Number of C>T mut. at CpG	Number of ins/del mut. at MS	Number of other mut.	Mutations with MMR deficiency Si:nature (%)
MSI Lynch	47	36	24	21	6	6	75
MSI sporadic	93	34	15	10	8	11	53
Mss	633	603	271	179	30	302	35

Nonpolypous cancers in Lynch syndrome: the third pathway

Previously, we have provided evidence that some Lynch syndrome-associated colorectal cancers develop through an adenoma-independent, nonpolypous pathway of progression. Such nonpolypous colorectal cancers were previously shown to frequently harbor *CTNNB1* mutations as activators of Wnt signaling.²⁰ We now performed next generation sequencing to obtain a broader mutational pattern of such proposed nonpolypous Lynch syndrome colorectal cancers, identified by the absence of tumor-adjacent adenoma formations, and

compared the mutation profile to cancers with evident adenoma history. As previously published,²⁰ *CTNNB1* mutations predominantly occurred in cancers lacking evidence of an adenomatous precursor stage (5 out of 10 cancers, Fig. 4), but they were also found in one out of five cancers that developed in an adenoma. Significant differences between polypous and nonpolypous cancers with regard to mutations in colorectal cancer genes were only found for *TP53* mutations ($p = 0.044$, Fig. 4), which were restricted to cancers lacking evidence of polypous growth. The number of observed *CTNNB1* mutations was too low to determine the

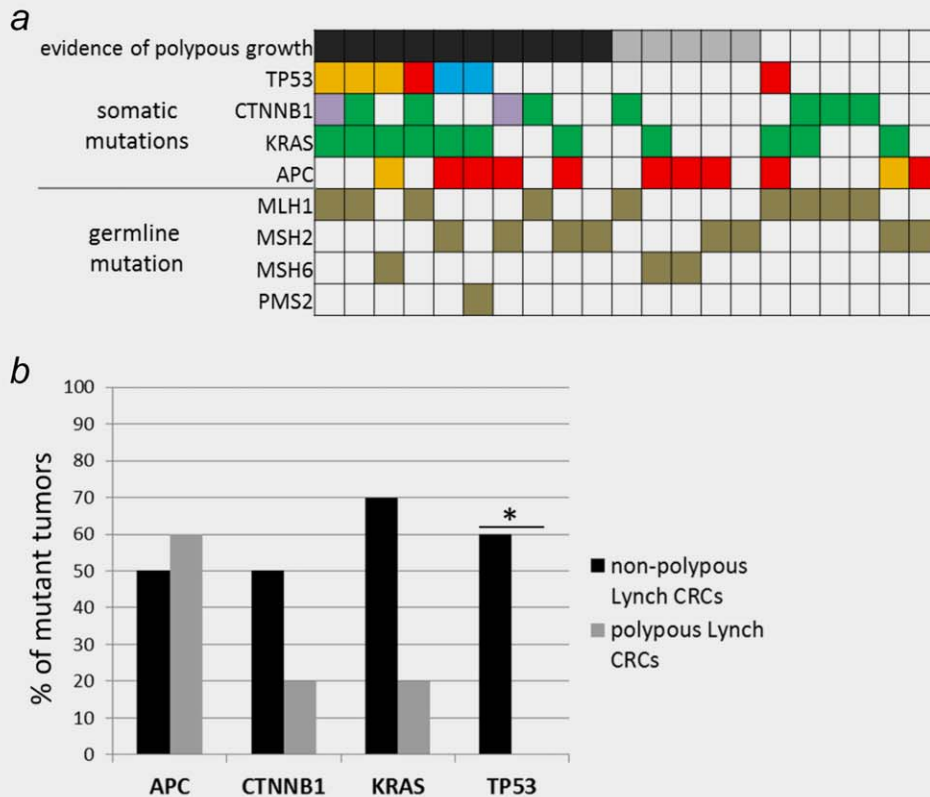


Figure 4. Mutation status of key colorectal genes in Lynch syndrome cancers. Colorectal cancers were grouped according to the evidence of polypous growth (no, black; yes, grey; not assessable, white). (a) Patient-wise color-coded mutation data (red, stop mutation; orange, frameshift insertion/deletion mutation; purple, in-frame deletion; green, amino acid exchange activating oncogene; blue, amino acid exchange inactivating tumor suppressor gene). (b) Comparison of mutation frequencies between Lynch syndrome-associated colorectal cancers with and without evidence of polypous growth. A significantly higher mutation frequency was observed in nonpolypous than polypous cancers for *TP53* (* $p = 0.044$).

timing of MMR deficiency. Two tumors lacking both, *APC* and *CTNNB1* mutations harbored frameshift mutations of homopolymers in the *RNF43* gene (not shown), an alternative activator of Wnt signaling.

Discussion

This study was initiated to comprehensively address the question of distinct pathogenesis pathways in Lynch syndrome on different levels, by analyzing colorectal lesions of different progression stages using literature data, molecular profiling and cancer genome databases.

First, we examined the overall frequency of MMR-proficient lesions among adenomas in Lynch syndrome mutation carriers by a systematic literature review, providing the largest dataset available so far. Literature data together with MMR deficiency data obtained in our own cohort revealed that MMR-proficient adenomas represent less than a quarter of all adenomas in Lynch syndrome (23.3%), indicating that the majority of Lynch syndrome adenomas are MMR-deficient. Naturally, most of the adenomas included in the literature analysis were from path_*MLH1* and path_*MSH2* carriers, and future analyses will have to show whether the

proportion of MMR-deficient adenomas may be different in path_*MSH6* and path_*PMS2* carriers.

Overall, adenomas analyzed by IHC had a slightly, but significantly higher frequency of MMR deficiency than those analyzed by PCR. This is in contrast to the situation in MMR-deficient colorectal cancers, in which PCR-based methods have been shown to have a higher sensitivity.³⁸ This discrepancy likely reflects the fact that loss of protein expression occurs synchronously with the onset of MMR deficiency, whereas microsatellite instability gradually increases with the progression of MMR-deficient lesions, hence being less pronounced in adenomas compared to clinically diagnosed cancers.³⁹ In the present analysis, we therefore used IHC to detect MMR deficiency assuming it to have higher sensitivity than PCR-based techniques and to explore the spatial heterogeneity of MMR protein expression.

The vast majority of analyzed adenomas presented with a complete loss of the respective MMR protein, whereas adenomas presenting with heterogeneous MMR protein expression patterns in dysplastic crypts were absent in our own cohort, and reported only very rarely in the literature (3.3%). This observation can either be explained by MMR-deficient

dysplastic cells very rapidly overgrowing MMR-proficient dysplastic cells, or alternatively by MMR deficiency preceding dysplasia formation. Compatible with the latter possibility, we detected nondysplastic MMR-deficient crypts directly adjacent to one dysplastic MMR-deficient adenoma from our own cohort, proving that MMR deficiency can be an initiating event in adenoma formation and that MMR-DCF are bona fide tumor precursors in Lynch syndrome.

To obtain a better estimate of the proportion of MSI colorectal cancer initiated by MMR deficiency, that is, MMR-DCF, and those originating from MMR-proficient adenomas, we reconstructed the sequence of events by determining whether *APC* and *KRAS* mutations in manifest cancers showed fingerprints of pre-existing MMR deficiency. It has recently been published that *APC* mutations in Lynch syndrome-associated colorectal cancers are more frequently insertion/deletion mutations at repetitive sequences than in microsatellite-stable colorectal cancers,¹⁹ which agrees with our observations in this study. We extended the mutation signature analysis to include substitutions and showed that the relative frequencies of the observed *APC* substitutions were different in MSI Lynch and MSS colorectal cancers. Furthermore, the relative frequencies of each substitution in MSI Lynch colorectal cancers were similar to what would be expected if caused by MMR-deficiency, based on the probabilities determined by mutational signature 6, while this was not the case for MSS colorectal cancers,²³ adding further evidence that MMR deficiency frequently precedes *APC* mutation in Lynch syndrome-associated colorectal carcinogenesis. By classifying C > T transitions at CpG sites and insertion/deletion mutations at repetitive sequences as “MMR deficiency-related,” we predict that approximately 61% of *APC* mutations in Lynch syndrome-associated cancers occur after the onset of MMR deficiency. This observation is compatible with the hypothesis that the majority of adenomas ultimately developing into colorectal cancer in Lynch syndrome are initiated by MMR deficiency, in agreement with the high rate of MMR deficiency in adenomas, the existence of MMR-DCF and our unique finding of a MMR-deficient adenoma outgrowing from a MMR-DCF. Accounting for the fact that part of Lynch syndrome adenomas, instead of *APC* mutations, harbor MMR deficiency-induced *RNF43* mutations as activators of Wnt/beta-catenin signaling, the percentage of adenomas initiated by MMR deficiency may even be higher than predicted from *APC* mutation signatures alone.^{19,40} Similarly the analysis of *KRAS* mutation patterns in MSI Lynch colorectal cancers revealed an association with the MMR-deficiency related mutational signature 6; as for *APC*, our data suggests a sequence of events in which MMR deficiency commonly precedes activating *KRAS* mutations. In addition, MMR deficiency being the cause of subsequent driver mutations is a likely explanation for the predominance of G13D mutations in MMR-deficient colorectal cancers as previously reported,⁴¹ in spite of its potentially lower oncogenic effect compared to codon 12 mutations.⁴² Our data

also demonstrate the applicability of mutational signature analysis for tracking sporadic MSI cancer development. Also in these tumors, an elevated proportion of MMR deficiency-related *APC* mutations was observed. However, these results need to be integrated in carcinogenesis models that account for the different mutational events and precursor lesions, such as sessile serrated adenomas.⁴³

Our approach has limitations. First, we did an overall analysis of *APC* and *KRAS* mutations occurring in all tumors together, not dissecting individual tumors for the occurrence of multiple mutations. Second, due to lack of information about path_ *MMR* gene variant status in tumors from DFCI/TCGA databases, 26 tumors were assigned into the “MSI Lynch” group due to negative MLH1 promoter methylation/CIMP status. This approach may lead to the inclusion of a small number of actually sporadic tumors, as two somatic mutations affecting both alleles of one MMR gene can lead to MMR deficiency in ~20% of CIMP-negative MSI tumors.⁴⁴ We also assume a linear carcinogenesis model, not referring to the possibility of independent pathways of transformation developing in parts of the same tumor or precancerous lesion. Moreover, we cannot formally exclude the possibility that the increased proportion of MMR deficiency-related mutations observed in Lynch syndrome-associated cancers merely reflects an increased number of passenger mutations accumulating after transformation and the onset of MMR deficiency. However, the functional impact of the observed mutations and the fact that no increased load of randomly distributed, functionally irrelevant mutations was observed in oncogenes such as *KRAS* argues against the passenger mutation assumption and supports the validity of our conclusions.

Recent prospective studies on CRC incidence in Lynch syndrome patients under colonoscopic surveillance show that colorectal cancers occurred despite regular colonoscopy with polypectomy.^{22,45} Based on this observation, Møller *et al.*²² raised the question whether CRC in Lynch syndrome necessarily has to always emerge from a macroscopically visible lesion. Using panel sequencing,^{46,47} we examined the possibility that there is a distinct pathway reflected by a distinct molecular profile of the manifest cancers, which may support the concept that part of Lynch syndrome colorectal cancers may emerge from nonpolypous, that is, “invisible” precursor lesions.²⁰

Interestingly, a subset of MSI Lynch tumors presented with mutations of the *TP53* gene, which are otherwise rare in MSI cancers.³⁵ All *TP53*-mutant tumors presented with a nonpolypous histology, suggesting that *TP53* mutations, alongside *CTNNB1* mutations,²⁰ may represent drivers of nonpolypous cancer formation in Lynch syndrome. On a mechanistic level, this observation is very well compatible with recent evidence that mutant gain-of-function variants of *TP53* are associated with the formation of flat lesions and an inflammatory phenotype favoring invasive growth in murine models.⁴⁸ Clinically, our data support the existence of a

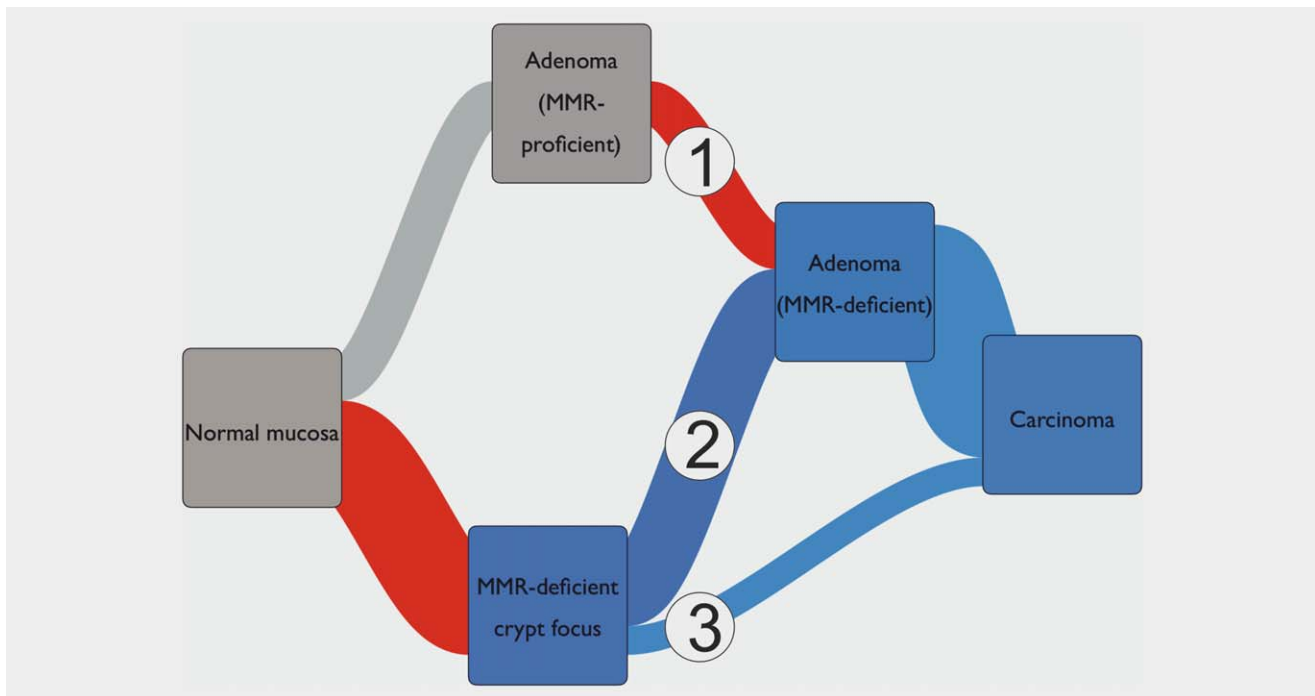


Figure 5. Integrative model of colorectal carcinogenesis in Lynch syndrome (Sankey diagram). Lynch syndrome colorectal cancer development follows three distinct routes. A subset of Lynch syndrome colorectal cancers develops through (1) MMR deficiency-independent adenoma formation with secondary MMR inactivation. Most commonly, however, tumor formation follows or is initiated by MMR deficiency, which can either lead to (2) MMR-deficient adenoma formation, or to (3) entirely nonpolypous progression into invasive cancer. The relative contribution of the three pathways is predicted to vary between populations and will depend on factors such as availability of colonoscopy screening and screening intervals. For better visibility, pre-malignant lesions that do not develop into cancer are not included in the diagram, because their number greatly exceeds the number of carcinomas. [Color figure can be viewed at wileyonlinelibrary.com]

distinct group of Lynch syndrome colorectal cancers that may manifest as interval cancers because they are not detectable even by high quality colonoscopy.^{22,45} Due to the limited number of tumors analyzed by panel sequencing, confirmation in independent tumor collections is strongly encouraged.

The contribution of nonpolypous cancers to the overall colorectal cancer burden in Lynch syndrome remains to be determined and most likely will vary between populations. In this context, it has to be kept in mind that the surveillance scheme applied in management of Lynch syndrome, which differs between countries,^{22,49} will most likely influence the picture significantly, because, as mentioned above, colonoscopy will, with a much higher likelihood detect, polypous adenomas compared to nonpolypous precursor lesions. Therefore, compared to a population not under regular colonoscopy, carcinomas developing through the nonpolypous tumorigenesis pathway will be much more frequent in populations participating in, for example, annual colonoscopy screening programs.

Due to low number of *CTNNB1* and *TP53* mutations, their timing with respect to MMR deficiency could not be reliably determined. However, *CTNNB1* mutations are known to be associated with MSI cancers,⁵⁰ in particular with hereditary MSI CRC.^{20,51} It has also been shown that in colonic tissue *CTNNB1* mutations alone are unable to drive activation of

Wnt signaling.⁵² Taken together with the high number of MMR-DCF in the normal colonic mucosa of Lynch syndrome patients, the scenario in which *CTNNB1* strikes in an already MMR-deficient cell and leads to nonpolypous progression is much more likely than the reversed order of the events. However, studies need to be performed to determine the timing of mutations associated with nonpolypous growth in Lynch CRC.

In summary, the conflicting models of colorectal Lynch syndrome pathogenesis—MMR deficiency as a late event^{3,8,53} versus MMR deficiency as an early event^{17,19,20}—can only be reconciled by a unifying model that accepts the existence of distinct pathways of colorectal carcinogenesis in Lynch syndrome (Fig. 5). Indeed, our study provides histological and molecular evidence that Lynch syndrome-associated colorectal cancers do not follow one single pathway, but three pathways separated from each other by the type and timing of key mutation events: colorectal cancers in Lynch syndrome can in fact grow out from MMR-proficient adenomas after secondary inactivation of the MMR system (pathway 1). However, a larger part of cancers appear to develop from precursor lesions in which MMR deficiency is an early event, likely to include MMR-DCF, either through an adenomatous phase (pathway 2) or as nonpolypous lesions with immediate invasive growth (pathway 3). Future studies will have to

assess the relative contribution of each of these pathways to the colorectal cancer burden in Lynch syndrome, the effectiveness of screening programs to prevent each of these cancer types and the impact of the pathogenesis on patients' outcome and survival.

Acknowledgements

Part of the results is based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>.

The excellent technical support provided by Beate Kuchenbuch, Lena Ehret and Petra Hoefler is gratefully acknowledged.

References

- Jasperson KW, Tuohy TM, Neklason DW, et al. Hereditary and familial colon cancer. *Gastroenterology* 2010;138:2044–58.
- Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. *Orphanet J Rare Dis* 2009;4:22.
- Jass JR. Colorectal adenoma progression and genetic change: is there a link? *Ann Med* 1995;27:301–6.
- Boland CR, Goel A. Microsatellite instability in colorectal cancer. *Gastroenterology* 2010;138:2073–87 e3.
- Mecklin JP, Aarnio M, Laara E, et al. Development of colorectal tumors in colonoscopic surveillance in Lynch syndrome. *Gastroenterology* 2007;133:1093–8.
- Lynch HT, Snyder CL, Shaw TG, et al. Milestones of Lynch syndrome: 1895–2015. *Nat Rev Cancer* 2015;15:181–94.
- De Jong AE, Morreau H, Van Puijenbroek M, et al. The role of mismatch repair gene defects in the development of adenomas in patients with HNPCC. *Gastroenterology* 2004;126:42–8.
- Fearon ER. Molecular genetics of colorectal cancer. *Annu Rev Pathol* 2011;6:479–507.
- Halvarsson B, Lindblom A, Johansson L, et al. Loss of mismatch repair protein immunostaining in colorectal adenomas from patients with hereditary nonpolyposis colorectal cancer. *Mod Pathol* 2005;18:1095–101.
- Shia J, Stadler ZK, Weiser MR, et al. Mismatch repair deficient-crypts in non-neoplastic colonic mucosa in Lynch syndrome: insights from an illustrative case. *Familial Cancer* 2015;14:61–8.
- Meijer TW, Hoogerbrugge N, Nagengast FM, et al. In Lynch syndrome adenomas, loss of mismatch repair proteins is related to an enhanced lymphocytic response. *Histopathology* 2009;55:414–22.
- Walsh MD, Buchanan DD, Pearson SA, et al. Immunohistochemical testing of conventional adenomas for loss of expression of mismatch repair proteins in Lynch syndrome mutation carriers: a case series from the Australasian site of the colon cancer family registry. *Mod Pathol* 2012;25:722–30.
- Rijcken FE, Hollema H, Kleibeuker JH. Proximal adenomas in hereditary non-polyposis colorectal cancer are prone to rapid malignant transformation. *Gut* 2002;50:382–6.
- Iino H, Simms L, Young J, et al. DNA microsatellite instability and mismatch repair protein loss in adenomas presenting in hereditary nonpolyposis colorectal cancer. *Gut* 2000;47:37–42.
- Jacoby RF, Marshall DJ, Kailas S, et al. Genetic instability associated with adenoma to carcinoma progression in hereditary nonpolyposis colon cancer. *Gastroenterology* 1995;109:73–82.
- Kloor M, Huth C, Voigt AY, et al. Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study. *Lancet Oncol* 2012;13:598–606.
- Tanaka M, Nakajima T, Sugano K, et al. Mismatch repair deficiency in Lynch syndrome-associated colorectal adenomas is more prevalent in older patients. *Histopathology* 2016;69:322–8.
- Pedroni M, Sala E, Scarselli A, et al. Microsatellite instability and mismatch-repair protein expression in hereditary and sporadic colorectal carcinogenesis. *Cancer Res* 2001;61:896–9.
- Sekine S, Mori T, Ogawa R, et al. Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis. *Mod Pathol* 2017.
- Ahadova A, von Knebel Doeberitz M, Blaker H, et al. CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome. *Familial Cancer* 2016;15:579–86.
- Jarvinen HJ, Aarnio M, Mustonen H, et al. Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer. *Gastroenterology* 2000;118:829–34.
- Moller P, Seppala T, Bernstein I, et al. Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database. *Gut* 2017;66:464–72.
- Alexandrov LB, Nik-Zainal S, Wedge DC, et al. Signatures of mutational processes in human cancer. *Nature* 2013;500:415–21.
- Liljegren A, Barker G, Elliott F, et al. Prevalence of adenomas and hyperplastic polyps in mismatch repair mutation carriers among CAPP2 participants: report by the colorectal adenoma/carcinoma prevention programme 2. *JCO* 2008;26:3434–9.
- Kloor M, Sutter C, Wentzensen N, et al. A large MSH2 Alu insertion mutation causes HNPCC in a German kindred. *Hum Genet* 2004;115:432–8.
- Cancer Genome Atlas N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* 2012;487:330–7.
- Giannakis M, Mu XJ, Shukla SA, et al. Genomic correlates of immune-cell infiltrates in colorectal carcinoma. *Cell Rep* 2016.
- Gao J, Aksoy BA, Dogrusoz U, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Signaling* 2013;6:p11.
- Cerami E, Gao J, Dogrusoz U, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Disc* 2012;2:401–4.
- Jesinghaus M, Pfarr N, Endris V, et al. Genotyping of colorectal cancer for cancer precision medicine: results from the IPH Center for Molecular Pathology. *Genes Chromosomes Cancer* 2016;55:505–21.
- Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 2010;38:e164.
- Robinson JT, Thorvaldsdottir H, Winckler W, et al. Integrative genomics viewer. *Nat Biotechnol* 2011;29:24–6.
- Giuffre G, Muller A, Brodegger T, German Hnpcc Consortium GCA, et al. Microsatellite analysis of hereditary nonpolyposis colorectal cancer-associated colorectal adenomas by laser-assisted microdissection: correlation with mismatch repair protein expression provides new insights in early steps of tumorigenesis. *J Mol Diagn* 2005;7:160–70.
- Vogelstein B, Kinzler KW. The multistep nature of cancer. *Trends Genet* 1993;9:138–41.
- Kloth M, Ruessler V, Engel C, et al. Activating ERBB2/HER2 mutations indicate susceptibility to pan-HER inhibitors in Lynch and Lynch-like colorectal cancer. *Gut* 2016;65:1296–305.
- Rajagopalan H, Bardelli A, Lengauer C, et al. Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status. *Nature* 2002;418:934.
- Capella D, Cronauer-Mitra S, Pienado MA, et al. Frequency and spectrum of mutations at codons 12 and 13 of the c-K-ras gene in human tumors. *Environ Health Perspect* 1991;93:125–31.
- Lindor NM, Burgart LJ, Leontovich O, et al. Immunohistochemistry versus microsatellite instability testing in phenotyping colorectal tumors. *JCO* 2002;20:1043–8.
- Shibata D. When does MMR loss occur during HNPCC progression? *CBM* 2006;2:29–35.
- Giannakis M, Hodis E, Jasmine Mu X, et al. RNF43 is frequently mutated in colorectal and endometrial cancers. *Nat Genet* 2014;46:1264–6.
- Oliveira C, Westra JL, Arango D, et al. Distinct patterns of KRAS mutations in colorectal carcinomas according to germline mismatch repair defects and hMLH1 methylation status. *Hum Mol Genet* 2004;13:2303–11.

42. Margonis GA, Kim Y, Spolverato G, et al. Association between specific mutations in KRAS codon 12 and colorectal liver metastasis. *JAMA Surg* 2015;150:722–9.
43. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet* 2014;383:1490–502.
44. Mensenkamp AR, Vogelaar IP, van Zelst-Stams WA, et al. Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology* 2014;146:643–6e8.
45. Toni Seppälä KP, Evans DG, Järvinen H, et al. Colorectal cancer incidence in path_MLH1 carriers subjected to different follow-up protocols: a Prospective Lynch Syndrome Database report. *Heredit Cancer Clin Pract* 2017;15.
46. Pfarr N, Penzel R, Klauschen F, et al. Copy number changes of clinically actionable genes in melanoma, non-small cell lung cancer and colorectal cancer—A survey across 822 routine diagnostic cases. *Genes Chromosomes Cancer* 2016;55:821–33.
47. Jesinghaus M, Konukiewitz B, Keller G, et al. Colorectal mixed adenoneuroendocrine carcinomas and neuroendocrine carcinomas are genetically closely related to colorectal adenocarcinomas. *Mod Pathol* 2017;30:610–9.
48. Cooks T, Pateras IS, Tarcic O, et al. Mutant p53 prolongs NF-kappaB activation and promotes chronic inflammation and inflammation-associated colorectal cancer. *Cancer Cell* 2013;23:634–46.
49. Vasen HF, Blanco I, Aktan-Collan K, et al. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut* 2013;62:812–23.
50. Mirabelli-Primdahl L, Gryfe R, Kim H, et al. Beta-catenin mutations are specific for colorectal carcinomas with microsatellite instability but occur in endometrial carcinomas irrespective of mutator pathway. *Cancer Res* 1999;59:3346–51.
51. Johnson V, Volikos E, Halford SE, et al. Exon 3 beta-catenin mutations are specifically associated with colorectal carcinomas in hereditary non-polyposis colorectal cancer syndrome. *Gut* 2005;54:264–7.
52. Huels DJ, Ridgway RA, Radulescu S, et al. E-cadherin can limit the transforming properties of activating beta-catenin mutations. *EMBO J* 2015;34:2321–33.
53. Yurgelun MB, Goel A, Hornick JL, et al. Microsatellite instability and DNA mismatch repair protein deficiency in Lynch syndrome colorectal polyps. *Cancer Prevent Res* 2012;5:574–82.

Chapter 8. References

Aaltonen, L., Peltomäki, P., Leach, F., Sistonen, P., Pylkkänen, L., Mecklin, J., Järvinen, H., Powell, S., Jen, J., Hamilton, S. and al., e. (1993) 'Clues to the pathogenesis of familial colorectal cancer.', *Science*, 260(5109), pp. 812-6.

Aaltonen, L., Salovaara, R., Kristo, P., Canzian, F., Hemminki, A., Peltomäki, P., Chadwick, R., Kääriäinen, H., Eskelinen, M., Järvinen, H., Mecklin, J. and Chapelle, A.d.l. (1998) 'Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease.', *New England Journal of Medicine*, 338(21), pp. 1481-7.

Aarnio, M., Mecklin, J., Aaltonen, L., Nyström-Lahti, M. and Järvinen, H. (1995) 'Life-time risk of different cancers in hereditary non-polyposis colorectal cancer (HNPCC) syndrome.', *International Journal of Cancer*, 64(6), pp. 430-3.

Ahadova, A., Gallon, R., Gebert, J., Ballhausen, A., Endris, V., Kirchner, M., Stenzinger, A., Burn, J., von Knebel Doeberitz, M., Blaker, H. and Kloor, M. (2018) 'Three molecular pathways model colorectal carcinogenesis in Lynch syndrome', *International Journal of Cancer*, 143(1), pp. 139-150.

Ahadova, A., von Knebel Doeberitz, M., Bläker, H. and Kloor, M. (2016) 'CTNNB1-mutant colorectal carcinomas with immediate invasive growth: a model of interval cancers in Lynch syndrome.', *Familial Cancer*, 15(4), pp. 579-86.

Alazzouzi, H., Domingo, E., González, S., Blanco, I., Armengol, M., Espín, E., Plaja, A., Schwartz, S., Capella, G. and Schwartz, S.J. (2005) 'Low levels of microsatellite instability characterize MLH1 and MSH2 HNPCC carriers before tumor diagnosis.', *Human Molecular Genetics*, 14(2), pp. 235-9.

Alexandrov, L., Nik-Zainal, S., Wedge, D., Aparicio, S., Behjati, S., Biankin, A., Bignell, G., Bolli, N., Borg, A., Børresen-Dale, A., Boyault, S., Burkhardt, B., Butler, A., Caldas, C., Davies, H., Desmedt, C., Eils, R., Eyfjörd, J., Foekens, J., Greaves, M., Hosoda, F., Hutter, B., Ilicic, T., Imbeaud, S., Imielinski, M., Jäger, N., Jones, D., Jones, D., Knappskog, S., Kool, M., Lakhani, S., López-Otín, C., Martin, S., Munshi, N., Nakamura, H., Northcott, P., Pajic, M., Papaemmanuil, E., Paradiso, A., Pearson, J., Puente, X., Raine, K., Ramakrishna, M., Richardson, A., Richter, J.,

Rosenstiel, P., Schlesner, M., Schumacher, T., Span, P., Teague, J., Totoki, Y., Tutt, A., Valdés-Mas, R., van Buuren, M., van 't Veer, L., Vincent-Salomon, A., Waddell, N., Yates, L., Initiative, A.P.C.G., Consortium, I.B.C., Consortium, I.M.-S., PedBrain, I., Zucman-Rossi, J., Futreal, P., McDermott, U., Lichter, P., Meyerson, M., Grimmond, S., Siebert, R., Campo, E., Shibata, T., Pfister, S., Campbell, P. and Stratton, M. (2013) 'Signatures of mutational processes in human cancer.', *Nature*, 500(7463), pp. 415-21.

Alhilal, G. (2016) 'Development and validation of a next generation sequencing based microsatellite instability assay for routine clinical use', *Doctoral Thesis, Newcastle University*.

Allegra, C., Jessup, J., Somerfield, M., Hamilton, S., Hammond, E., Hayes, D., McAllister, P., Morton, R. and Schilsky, R. (2009) 'American Society of Clinical Oncology provisional clinical opinion: testing for KRAS gene mutations in patients with metastatic colorectal carcinoma to predict response to anti-epidermal growth factor receptor monoclonal antibody therapy', *Journal of Clinical Oncology*, 27(12), pp. 2091-6.

Alotaibi, H., Ricciardone, M. and Ozturk, M. (2008) 'Homozygosity at variant MLH1 can lead to secondary mutation in NF1, neurofibromatosis type I and early onset leukemia', *Mutat Res*, 637(1-2), pp. 209-14.

Amsen, D., van Gisbergen, K., Hombrink, P. and van Lier, R. (2018) 'Tissue-resident memory T cells at the center of immunity to solid tumors', *Nature Immunology*, 19(6), pp. 538-546.

Ananda, G., Walsh, E., Jacob, K., Krasilnikova, M., Eckert, K., Chiaromonte, F. and Makova, K. (2013) 'Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome.', *Genome Biology and Evolution*, 5(3), pp. 606-20.

Anderson, K. and LaBaer, J. (2005) 'The sentinel within: exploiting the immune system for cancer biomarkers', *Journal of Proteome Research*, 4(4), pp. 1123-33.

Andre, T., de Gramont, A., Vernerey, D., Chibaudel, B., Bonnetain, F., Tijeras-Raballand, A., Scriver, A., Hickish, T., Tabernero, J., Van Laethem, J., Banzi, M., Maertense, E., Shmueli, E., Carlsson, G., Scheithauer, W., Papamichael, D., Moehler, M., Landolfi, S., Demetter, P., Colote, S., Tournigand, C., Louvet, C., Duval, A., Flejou, J. and de Gramont, A. (2015) 'Adjuvant Fluorouracil, Leucovorin, and Oxaliplatin in Stage II to III Colon Cancer: Updated

10-Year Survival and Outcomes According to BRAF Mutation and Mismatch Repair Status of the MOSAIC Study', *Journal of Clinical Oncology*, 33(35), pp. 4176-87.

Angelopoulou, K., Stratis, M. and Diamandis, E. (1997) 'Humoral immune response against p53 protein in patients with colorectal carcinoma', *International Journal of Cancer*, 70(1), pp. 46-51.

Angelopoulou, K., Yu, H., Bharaj, B., Giali, M. and Diamandis, E. (2000) 'p53 gene mutation, tumor p53 protein overexpression, and serum p53 autoantibody generation in patients with breast cancer', *Clin Biochem*, 33(1), pp. 53-62.

Aquilina, G., Crescenzi, M. and Bignami, M. (1999) 'Mismatch repair, G(2)/M cell cycle arrest and lethality after DNA damage', *Carcinogenesis*, 20(12), pp. 2317-26.

Aravanis, A., Lee, M. and Klausner, R. (2017) 'Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection', *Cell*, 168(4), pp. 571-574.

Armbruster, D. and Pry, T. (2008) 'Limit of Blank, Limit of Detection and Limit of Quantitation', *The Clinical Biochemist Reviews*, 29(Suppl 1), pp. S49-S52.

Aronson, M., Gallinger, S., Cohen, Z., Cohen, S., Dvir, R., Elhasid, R., Baris, H., Kariv, R., Druker, H., Chan, H., Ling, S., Kortan, P., Holter, S., Semotiuk, K., Malkin, D., Farah, R., Sayad, A., Heald, B., Kalady, M., Penney, L., Rideout, A., Rashid, M., Hasadsri, L., Pichurin, P., Riegert-Johnson, D., Campbell, B., Bakry, D., Al-Rimawi, H., Alharbi, Q., Alharbi, M., Shamvil, A., Tabori, U. and Durno, C. (2016) 'Gastrointestinal Findings in the Largest Series of Patients With Hereditary Biallelic Mismatch Repair Deficiency Syndrome: Report from the International Consortium', *American Journal of Gastroenterology*, 111(2), pp. 275-84.

Baas, A., Gabbett, M., Rimac, M., Kansikas, M., Raphael, M., Nievelstein, R., Nicholls, W., Offerhaus, J., Bodmer, D., Wernstedt, A., Krabichler, B., Strasser, U., Nystrom, M., Zschocke, J., Robertson, S., van Haelst, M. and Wimmer, K. (2013) 'Agenesis of the corpus callosum and gray matter heterotopia in three patients with constitutional mismatch repair deficiency syndrome', *European Journal of Human Genetics*, 21(1), pp. 55-61.

Bacher, J., Flanagan, L., Smalley, R., Nassif, N., Burgart, L., Halberg, R., Megid, W. and Thibodeau, S. (2004) 'Development of a fluorescent multiplex assay for detection of MSI-High tumors.', *Disease Markers*, 20(4-5), pp. 237-50.

Bacolla, A., Larson, J., Collins, J., Li, J., Milosavljevic, A., Stenson, P., Cooper, D. and RD, W. (2008) 'Abundance and length of simple repeats in vertebrate genomes are determined by their structural properties.', *Genome Research*, 18(10), pp. 1545-53.

Bakry, D., Aronson, M., Durno, C., Rimawi, H., Farah, R., Alharbi, Q., Alharbi, M., Shamvil, A., Ben-Shachar, S., Mistry, M., Constantini, S., Dvir, R., Qaddoumi, I., Gallinger, S., Lerner-Ellis, J., Pollett, A., Stephens, D., Kelies, S., Chao, E., Malkin, D., Bouffet, E., Hawkins, C. and Tabori, U. (2014) 'Genetic and clinical determinants of constitutional mismatch repair deficiency syndrome: report from the constitutional mismatch repair deficiency consortium', *Eur J Cancer*, 50(5), pp. 987-96.

Balmana, J., Balaguer, F., Cervantes, A. and Arnold, D. (2013) 'Familial risk-colorectal cancer: ESMO Clinical Practice Guidelines', *Annals of Oncology*, 24 Suppl 6, pp. vi73-80.

Bartley, A., Thompson, P., Buckmeier, J., Kepler, C., Hsu, C., Snyder, M., Lance, P., Bhattacharyya, A. and Hamilton, S. (2010) 'Expression of gastric pyloric mucin, MUC6, in colorectal serrated polyps.', *Modern Pathology*, 23(2), pp. 169-76.

Beck, T., Mullikin, J., Program, N.C.S. and Biesecker, L. (2016) 'Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants.', *Clinical Chemistry*, 62(4), pp. 647-54.

Beckman, R. and Loeb, L. (2006) 'Efficiency of carcinogenesis with and without a mutator mutation.', *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), pp. 14140-5.

Berg, K., Glaser, C., Thompson, R., Hamilton, S., Griffin, C. and Eshleman, J. (2000) 'Detection of microsatellite instability by fluorescence multiplex polymerase chain reaction.', *Journal of Molecular Diagnostics*, 2(1), pp. 20-8.

Bertagnolli, M., Niedzwiecki, D., Compton, C., Hahn, H., Hall, M., Damas, B., Jewell, S., Mayer, R., Goldberg, R., Saltz, L., Warren, R. and Redston, M. (2009) 'Microsatellite instability

predicts improved response to adjuvant therapy with irinotecan, fluorouracil, and leucovorin in stage III colon cancer: Cancer and Leukemia Group B Protocol 89803', *Journal of Clinical Oncology*, 27(11), pp. 1814-21.

Bettegowda, C., Sausen, M., Leary, R., Kinde, I., Wang, Y., Agrawal, N., Bartlett, B., Wang, H., Luber, B., Alani, R., Antonarakis, E., Azad, N., Bardelli, A., Brem, H., Cameron, J., Lee, C., Fecher, L., Gallia, G., Gibbs, P., Le, D., Giuntoli, R., Goggins, M., Hogarty, M., Holdhoff, M., Hong, S., Jiao, Y., Juhl, H., Kim, J., Siravegna, G., Laheru, D., Lauricella, C., Lim, M., Lipson, E., Marie, S., Netto, G., Oliner, K., Olivi, A., Olsson, L., Riggins, G., Sartore-Bianchi, A., Schmidt, K., Shih, I., Oba-Shinjo, S., Siena, S., Theodorescu, D., Tie, J., Harkins, T., Veronese, S., Wang, T., Weingart, J., Wolfgang, C., Wood, L., Xing, D., Hruban, R., Wu, J., Allen, P., Schmidt, C., Choti, M., Velculescu, V., Kinzler, K., Vogelstein, B., Papadopoulos, N. and Diaz, L. (2014) 'Detection of Circulating Tumor DNA in Early- and Late-Stage Human Malignancies', *Science Translational Medicine*, 6(224), pp. 224ra24-224ra24.

Bhardwaj, M., Gies, A., Werner, S., Schrotz-King, P. and Brenner, H. (2017) 'Blood-Based Protein Signatures for Early Detection of Colorectal Cancer: A Systematic Review', *Clinical and Translational Gastroenterology*, 8(11), p. e128.

Bhattacharyya, N., Skandalis, A., Ganesh, A., Groden, J. and Meuth, M. (1994) 'Mutator phenotypes in human colorectal carcinoma cell lines.', *Proceedings of the National Academy of Sciences of the United States of America*, 91(14), pp. 6319-23.

Bindea, G., Mlecnik, B., Tosolini, M., Kirilovsky, A., Waldner, M., Obenauf, A., Angell, H., Fredriksen, T., Lafontaine, L., Berger, A., Bruneval, P., Fridman, W., Becker, C., Pagès, F., Speicher, M., Trajanoski, Z. and Galon, J. (2013) 'Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer.', *Immunity*, 39(4), pp. 782-95.

Blay, J., Lacombe, D., Meunier, F. and Stupp, R. (2012) 'Personalised medicine in oncology: questions for the next 20 years', *Lancet Oncology*, 13(5), pp. 448-9.

Blount, J. and Prakash, A. (2017) 'The Changing Landscape of Lynch Syndrome due to PMS2 Mutations.', *Clinical Genetics*, [Epub ahead of print].

Bocker, T., Diermann, J., Friedl, W., Gebert, J., Holinski-Feder, E., Karner-Hanusch, J., von-Knebel-Doeberitz, M., Koelble, K., Moeslein, G., Schackert, H., Wirtz, H., Fishel, R. and

Rüschoff, J. (1997) 'Microsatellite instability analysis: a multicenter study for reliability and quality control.', *Cancer Research*, 57(21), pp. 4739-43.

Bodo, S., Colas, C., Buhard, O., Collura, A., Tinat, J., Lavoine, N., Guilloux, A., Chalastanis, A., Lafitte, P., Coulet, F., Buisine, M., Ilencikova, D., Ruiz-Ponte, C., Kinzel, M., Grandjouan, S., Brems, H., Lejeune, S., Blanché, H., Wang, Q., Caron, O., Cabaret, O., Svrcek, M., Vidaud, D., Parfait, B., Verloes, A., Knappe, U., Soubrier, F., Mortemousque, I., Leis, A., Auclair-Perrossier, J., Frébourg, T., Fléjou, J., Entz-Werle, N., Leclerc, J., Malka, D., Cohen-Haguénauer, O., Goldberg, Y., Gerdes, A., Fedhila, F., Mathieu-Dramard, M., Hamelin, R., Wafaa, B., Gauthier-Villars, M., Bourdeaut, F., Sheridan, E., Vasen, H., Brugières, L., Wimmer, K., Muleris, M., Duval, A. and CMMRD", E.C.C.f. (2015) 'Diagnosis of Constitutional Mismatch Repair-Deficiency Syndrome Based on Microsatellite Instability and Lymphocyte Tolerance to Methylating Agents.', *Gastroenterology*, 149(4), pp. 1017-29.

Boks, D., Trujillo, A., Voogd, A., Morreau, H., Kenter, G. and Vasen, H. (2002) 'Survival analysis of endometrial carcinoma associated with hereditary nonpolyposis colorectal cancer.', *International Journal of Cancer*, 102(2), pp. 198-200.

Boland, C., Thibodeau, S., Hamilton, S., Sidransky, D., Eshleman, J., Burt, R., Meltzer, S., Rodriguez-Bigas, M., Fodde, R., Ranzani, G. and Srivastava, S. (1998) 'A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer.', *Cancer Research*, 58(22), pp. 5248-57.

Borràs, E., Pineda, M., Cadiñanos, J., Del Valle, J., Brieger, A., Hinrichsen, I., Cabanillas, R., Navarro, M., Brunet, J., Sanjuan, X., Musulen, E., van der Klift, H., Lázaro, C., Plotz, G., Blanco, I. and Capellá, G. (2013) 'Refining the role of PMS2 in Lynch syndrome: germline mutational analysis improved by comprehensive assessment of variants.', *Journal of Medical Genetics*, 50(8), pp. 552-63.

Borrebaeck, C. (2017) 'Precision diagnostics: moving towards protein biomarker signatures of clinical utility in cancer', *Nature Reviews Cancer*, 17(3), pp. 199-204.

Bossuyt, P., Reitsma, J., Bruns, D., Gatsonis, C., Glasziou, P., Irwig, L., Moher, D., Rennie, D., Vet, H.d., Lijmer, J. and Accuracy, S.f.R.o.D. (2003) 'The STARD statement for reporting

studies of diagnostic accuracy: explanation and elaboration.', *Annals of Internal Medicine*, 138(1), pp. W1-12.

Bouffet, E., Larouche, V., Campbell, B., Merico, D., de Borja, R., Aronson, M., Durno, C., Krueger, J., Cabric, V., Ramaswamy, V., Zhukova, N., Mason, G., Farah, R., Afzal, S., Yalon, M., Rechavi, G., Magimairajan, V., Walsh, M., Constantini, S., Dvir, R., Elhasid, R., Reddy, A., Osborn, M., Sullivan, M., Hansford, J., Dodgshun, A., Klauber-Demore, N., Peterson, L., Patel, S., Lindhorst, S., Atkinson, J., Cohen, Z., Laframboise, R., Dirks, P., Taylor, M., Malkin, D., Albrecht, S., Dudley, R., Jabado, N., Hawkins, C., Shlien, A. and Tabori, U. (2016) 'Immune Checkpoint Inhibition for Hypermutant Glioblastoma Multiforme Resulting From Germline Biallelic Mismatch Repair Deficiency', *Journal of Clinical Oncology*, 34(19), pp. 2206-11.

Bougeard, G., Charbonnier, F., Moerman, A., Martin, C., Ruchoux, M., Drouot, N. and Frebourg, T. (2003) 'Early onset brain tumor and lymphoma in MSH2-deficient children', *American Journal of Human Genetics*, 72(1), pp. 213-6.

Bougeard, G., Renaux-Petel, M., Flaman, J., Charbonnier, C., Fermey, P., Belotti, M., Gauthier-Villars, M., Stoppa-Lyonnet, D., Consolino, E., Brugières, L., Caron, O., Benusiglio, P., Bressac-de Paillerets, B., Bonadona, V., Bonaïti-Pellié, C., Tinat, J., Baert-Desurmont, S. and Frebourg, T. (2015) 'Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers.', *Journal of Clinical Oncology*, 33(21), pp. 2345-52.

Boyer, J., Umar, A., Risinger, J., Lipford, J., Kane, M., Yin, S., Barrett, J., Kolodner, R. and Kunkel, T. (1995) 'Microsatellite instability, mismatch repair deficiency, and genetic defects in human cancer cell lines', *Cancer Research*, 55(24), pp. 6063-70.

Boyle, E., O'Roak, B., Martin, B., Kumar, A. and Shendure, J. (2014a) 'MIPgen: optimized modeling and design of molecular inversion probes for targeted resequencing', *Bioinformatics*, 30(18), pp. 2670-2.

Boyle, T., Bridge, J., Sabatini, L., Nowak, J., Vasalos, P., Jennings, L., Halling, K. and Committee, C.o.A.P.M.O. (2014b) 'Summary of microsatellite instability test results from laboratories participating in proficiency surveys: proficiency survey results from 2005 to 2012.', *Archives of Pathology and Laboratory Medicine*, 138(3), pp. 363-70.

- Broggio, J. and Bannister, N. (2016) *Cancer survival by stage at diagnosis for England (experimental statistics): Adults diagnosed 2012, 2013 and 2014 and followed up to 2015*.
- Bronner, C., Baker, S., Morrison, P., Warren, G., Smith, L., Lescoe, M., Kane, M., Earabino, C., Lipford, J., Lindblom, A., Tannergard, P., Bollag, R., Godwin, A., Ward, D., Nordenskjold, M., Fishel, R., Kolodner, R. and Liskay, R. (1994) 'Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer.', *Nature*, 368(6468), pp. 258-61.
- Buchanan, D., Roberts, A., Walsh, M., Parry, S. and Young, J. (2011) 'Lessons from Lynch syndrome: a tumor biology-based approach to familial colorectal cancer.', *Future Oncology*, 6(4), pp. 539-49.
- Buckowitz, A., Knaebel, H., Benner, A., Bläker, H., Gebert, J., Kienle, P., Doeberitz, M.v.K. and Kloor, M. (2005) 'Microsatellite instability in colorectal cancer is associated with local lymphocyte infiltration and low frequency of distant metastases', *British Journal of Cancer*, 92(9), pp. 1746-53.
- Buhard, O., Suraweera, N., Lectard, A., Duval, A. and Hamelin, R. (2004) 'Quasimonomorphic mononucleotide repeats for high-level microsatellite instability analysis.', *Disease Markers*, 20(4-5), pp. 251-7.
- Burn, J., Gerdes, A., Macrae, F., Mecklin, J., Moeslein, G., Olschwang, S., Eccles, D., Evans, D., Maher, E., LBertario, Bisgaard, M., Dunlop, M., Ho, J., Hodgson, S., Lindblom, A., Lubinski, J., Morrison, P., Murday, V., Ramesar, R., Side, L., Scott, R., Thomas, H., Vasen, H., Barker, G., Crawford, G., Elliott, F., Movahedi, M., Pylvanainen, K., Wijnen, J., Fodde, R., Lynch, H., Mathers, J. and Bishop, D. (2011) 'Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial.', *The Lancet*, 378(9809), pp. 2081-7.
- Calonge, N., Petitti, D., DeWitt, T., Gordis, L., Gregory, K., Harris, R., Isham, G., LeFevre, M., Loveland-Cherry, C., Marion, L., Moyer, V., Ockene, J., Sawaya, G., Siu, A., Teutsch, S. and Yawn, B. (2009) 'Aspirin for the prevention of cardiovascular disease: U.S. Preventive Services Task Force recommendation statement', *Annals of Internal Medicine*, 150(6), pp. 396-404.

Canard, G., Lefevre, J., Colas, C., Coulet, F., Svrcek, M., Lascols, O., Hamelin, R., Shields, C., Duval, A., Fléjou, J., Soubrier, F., Tiret, E. and Parc, Y. (2012) 'Screening for Lynch syndrome in colorectal cancer: are we doing enough?', *Annals of Surgical Oncology*, 19(3), pp. 809-16.

Carethers, J. and Stoffel, E. (2015) 'Lynch syndrome and Lynch syndrome mimics: The growing complex landscape of hereditary colon cancer.', *World Journal of Gastroenterology*, 21(31), pp. 9253-61.

Carlson, K., Sudmant, P., Press, M., Eichler, E., Shendure, J. and Queitsch, C. (2015) 'MIPSTR: a method for multiplex genotyping of germline and somatic STR variation across many individuals', *Genome Research*, 25(5), pp. 750-61.

Casbon, J., Osborne, R., Brenner, S. and Lichtenstein, C. (2011) 'A method for counting PCR template molecules with application to next-generation sequencing.', *Nucleic Acids Research*, 39(12), p. e81.

Castillejo, A., Vargas, G., Castillejo, M., Navarro, M., Barberá, V., González, S., Hernández-Illán, E., Brunet, J., Ramón y Cajal, T., Balmaña, J., Oltra, S., Iglesias, S., Velasco, A., Solanes, A., Campos, O., Sánchez Heras, A., Gallego, J., Carrasco, E., González Juan, D., Segura, A., Chirivella, I., Juan, M., Tena, I., Lázaro, C., Blanco, I., Pineda, M., Capellá, G. and Soto, J. (2014) 'Prevalence of germline MUTYH mutations among Lynch-like syndrome patients.', *European Journal of Cancer*, 50(13), pp. 2241-50.

Chalmers, Z., Connelly, C., Fabrizio, D., Gay, L., Ali, S., Ennis, R., Schrock, A., Campbell, B., Shlien, A., Chmielecki, J., Huang, F., He, Y., Sun, J., Tabori, U., Kennedy, M., Lieber, D., Roels, S., White, J., Otto, G., Ross, J., Garraway, L., Miller, V., Stephens, P. and Frampton, G. (2017) 'Analysis of 100,000 human cancer genomes reveals the landscape of tumor mutational burden', *Genome Medicine*, 9(1), p. 34.

Chapusot, C., Martin, L., Bouvier, A., Bonithon-Kopp, C., Ecartot-Laubriet, A., Rageot, D., Ponnelle, T., Laurent Puig, P., Faivre, J. and Piard, F. (2002) 'Microsatellite instability and intratumoural heterogeneity in 100 right-sided sporadic colon carcinomas', *British Journal of Cancer*, 87(4), pp. 400-4.

Chapusot, C., Martin, L., Puig, P., Ponnelle, T., Cheynel, N., Bouvier, A., Rageot, D., Roignot, P., Rat, P., Faivre, J. and Piard, F. (2004) 'What is the best way to assess microsatellite

instability status in colorectal cancer? Study on a population base of 462 colorectal cancers', *American Journal of Surgical Pathology*, 28(12), pp. 1553-9.

Clendenning, M., Buchanan, D., Walsh, M., Nagler, B., Rosty, C., Thompson, B., Spurdle, A., Hopper, J., Jenkins, M. and Young, J. (2011) 'Mutation deep within an intron of MSH2 causes Lynch syndrome.', *Familial Cancer*, 10(2), pp. 297-301.

Coelho, H., Jones-Hughes, T., Snowsill, T., Briscoe, S., Huxley, N., Frayling, I. and Hyde, C. (2017) 'A systematic review of test accuracy studies evaluating molecular micro-satellite instability testing for the detection of individuals with lynch syndrome', *BMC Cancer*, 17(1), p. 836.

Coolbaugh-Murphy, M., Xu, J., Ramagli, L., Brown, B. and Siciliano, M. (2005) 'Microsatellite instability (MSI) increases with age in normal somatic cells', *Mechanisms of Ageing and Development*, 126(10), pp. 1051-9.

Coolbaugh-Murphy, M., Xu, J., Ramagli, L., Ramagli, B., Brown, B., Lynch, P., Hamilton, S., Frazier, M. and Siciliano, M. (2010) 'Microsatellite Instability in the Peripheral Blood Leukocytes of HNPCC Patients', *Human mutation*, 31(3), pp. 317-324.

Cooper, S., Hall, C., Palmer, T., Sidhu, R., Hayre, J. and Powell, J. (2017) 'Cetuximab and panitumumab for previously untreated metastatic colorectal cancer', *National Institute for Health and Care Excellence Technology Appraisal 439*, <http://nice.org.uk/guidance/ta439>.

Coulie, P., Lehmann, F., Lethe, B., Herman, J., Lurquin, C., Andrawiss, M. and Boon, T. (1995) 'A mutated intron sequence codes for an antigenic peptide recognized by cytolytic T lymphocytes on a human melanoma', *Proceedings of the National Academy of Sciences of the United States of America*, 92(17), pp. 7976-80.

Cross, W., Kovac, M., Mustonen, V., Temko, D., Davis, H., Baker, A., Biswas, S., Arnold, R., Chegwiddden, L., Gatenbee, C., Anderson, A., Koelzer, V., Martinez, P., Jiang, X., Domingo, E., Woodcock, D., Feng, Y., Kovacova, M., Maughan, T., Jansen, M., Rodriguez-Justo, M., Ashraf, S., Guy, R., Cunningham, C., East, J., Wedge, D., Wang, L., Palles, C., Heinemann, K., Sottoriva, A., Leedham, S., Graham, T. and Tomlinson, I. (2018) 'The evolutionary landscape of colorectal tumorigenesis', *Nature Ecology and Evolution*.

Cummings, A. and Garon, E. (2017) 'The ascent of immune checkpoint inhibitors: is the understudy ready for a leading role?', *Cancer Biology & Medicine*, 14(4), pp. 341-347.

Dashti, S., Win, A., Hardikar, S., Glombicki, S., Mallenahalli, S., Thirumurthi, S., Peterson, S., You, Y., Buchanan, D., Figueiredo, J., Campbell, P., Gallinger, S., Newcomb, P., Potter, J., Lindor, N., Le Marchand, L., Haile, R., Hopper, J., Jenkins, M., Basen-Engquist, K., Lynch, P. and Pande, M. (2018) 'Physical activity and the risk of colorectal cancer in Lynch syndrome', *International Journal of Cancer*.

de Jong, A., Hendriks, Y., Kleibeuker, J., Boer, S.d., Cats, A., Griffioen, G., Nagengast, F., Nelis, F., Rookus, M. and Vasen, H. (2006) 'Decrease in mortality in Lynch syndrome families because of surveillance.', *Gastroenterology*, 130(3), pp. 665-71.

de Jong, A., Morreau, H., van Puijnenbroek, M., Eilers, P., Wijnen, J., Nagengast, F., Griffioen, G., Cats, A., Menko, F., Kleibeuker, J. and Vasen, H. (2004a) 'The role of mismatch repair gene defects in the development of adenomas in patients with HNPCC.', *Gastroenterology*, 126(1), pp. 42-8.

de Jong, A., Puijnenbroek, M.v., Hendriks, Y., Tops, C., Wijnen, J., Ausems, M., Meijers-Heijboer, H., Wagner, A., Os, T.v., Bröcker-Vriends, A., Vasen, H. and Morreau, H. (2004b) 'Microsatellite instability, immunohistochemistry, and additional PMS2 staining in suspected hereditary nonpolyposis colorectal cancer.', *Clinical Cancer Research*, 10(3), pp. 972-80.

De Roock, W., Claes, B., Bernasconi, D., De Schutter, J., Biesmans, B., Fountzilas, G., Kalogeras, K., Kotoula, V., Papamichael, D., Laurent-Puig, P., Penault-Llorca, F., Rougier, P., Vincenzi, B., Santini, D., Tonini, G., Cappuzzo, F., Frattini, M., Molinari, F., Saletti, P., De Dosso, S., Martini, M., Bardelli, A., Siena, S., Sartore-Bianchi, A., Tabernero, J., Macarulla, T., Di Fiore, F., Gangloff, A., Ciardiello, F., Pfeiffer, P., Qvortrup, C., Hansen, T., Van Cutsem, E., Piessevaux, H., Lambrechts, D., Delorenzi, M. and Tejpar, S. (2010a) 'Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis.', *The Lancet Oncology*, 11(8), pp. 753-62.

De Roock, W., Jonker, D., Di Nicolantonio, F., Sartore-Bianchi, A., Tu, D., Siena, S., Lamba, S., Arena, S., Frattini, M., Piessevaux, H., Van Cutsem, E., O'Callaghan, C., Khambata-Ford, S.,

Zalcborg, J., Simes, J., Karapetis, C., Bardelli, A. and Tejpar, S. (2010b) 'Association of KRAS p.G13D mutation with outcome in patients with chemotherapy-refractory metastatic colorectal cancer treated with cetuximab.', *Journal of the American Medical Association*, 304(16), pp. 1812-20.

De Rosa, M., Fasano, C., Panariello, L., Scarano, M., Belli, G., Iannelli, A., Ciciliano, F. and Izzo, P. (2000) 'Evidence for a recessive inheritance of Turcot's syndrome caused by compound heterozygous mutations within the PMS2 gene', *Oncogene*, 19(13), pp. 1719-23.

De Sousa E Melo, F., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L., de Jong, J., de Boer, O., van Leersum, R., Bijlsma, M., Rodermond, H., van der Heijden, M., van Noesel, C., Tuynman, J., Dekker, E., Markowitz, F., Medema, J. and Vermeulen, L. (2013) 'Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions', *Nature Medicine*, 19(5), pp. 614-8.

De Vos, M., Hayward, B., Charlton, R., Taylor, G., Glaser, A., Picton, S., Cole, T., Maher, E., McKeown, C., Mann, J., Yates, J., Baralle, D., Rankin, J., Bonthron, D. and Sheridan, E. (2006) 'PMS2 mutations in childhood cancer', *Journal of the National Cancer Institute*, 98(5), pp. 358-61.

De Vos, M., Hayward, B., Picton, S., Sheridan, E. and Bonthron, D. (2004) 'Novel PMS2 pseudogenes can conceal recessive mutations causing a distinctive childhood cancer syndrome', *American Journal of Human Genetics*, 74(5), pp. 954-64.

Deacu, E., Mori, Y., Sato, F., Yin, J., Olaru, A., Sterian, A., Xu, Y., Wang, S., Schulmann, K., Berki, A., Kan, T., Abraham, J. and Meltzer, S. (2004) 'Activin type II receptor restoration in ACVR2-deficient colon cancer cells induces transforming growth factor-beta response pathway genes.', *Cancer Research*, 64(21), pp. 7690-6.

Deng, G., Chen, A., Hong, J., Chae, H. and Kim, Y. (1999) 'Methylation of CpG in a small region of the hMLH1 promoter invariably correlates with the absence of gene expression.', *Cancer Research*, 59(9), pp. 2029-33.

Desmetz, C., Mange, A., Maudelonde, T. and Solassol, J. (2011) 'Autoantibody signatures: progress and perspectives for early cancer detection', *Journal of Cellular and Molecular Medicine*, 15(10), pp. 2013-24.

Diaz, L., Williams, R., Wu, J., Kinde, I., Hecht, J., Berlin, J., Allen, B., Bozic, I., Reiter, J., Nowak, M., Kinzler, K., Oliner, K. and Vogelstein, B. (2012) 'The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers', *Nature*, 486(7404), pp. 537-540.

Dienstmann, R., Mason, M., Sinicrope, F., Phipps, A., Tejpar, S., Nesbakken, A., Danielsen, S., Sveen, A., Buchanan, D., Clendenning, M., Rosty, C., Bot, B., Alberts, S., Milburn Jessup, J., Lothe, R., Delorenzi, M., Newcomb, P., Sargent, D. and Guinney, J. (2017) 'Prediction of overall survival in stage II and III colon cancer beyond TNM system: a retrospective, pooled biomarker study.', *Annals of Oncology*, 28(5), pp. 1023-31.

Dieringer, D. and Schlötterer, C. (2003) 'Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species.', *Genome Research*, 13(10), pp. 2242-51.

Dietmaier, W., Wallinger, S., Bocker, T., Kullmann, F., Fishel, R. and Rüschoff, J. (1997) 'Diagnostic microsatellite instability: definition and correlation with mismatch repair protein expression.', *Cancer Research*, 57(21), pp. 4749-56.

Dinh, T., Rosner, B., Atwood, J., Boland, C., Syngal, S., Vasen, H., Gruber, S. and Burt, R. (2011) 'Health benefits and cost-effectiveness of primary genetic screening for Lynch syndrome in the general population', *Cancer Prevention Research*, 4(1), pp. 9-22.

Domingo, E., Laiho, P., Ollikainen, M., Pinto, M., Wang, L., French, A., Westra, J., Frebourg, T., Espín, E., Armengol, M., Hamelin, R., Yamamoto, H., Hofstra, R., Seruca, R., Lindblom, A., Peltomäki, P., Thibodeau, S., Aaltonen, L. and Schwartz, S.J. (2004) 'BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing', *Journal of Medical Genetics*, 41(9), pp. 664-8.

Dorner, T. and Radbruch, A. (2007) 'Antibodies and B cell memory in viral immunity', *Immunity*, 27(3), pp. 384-92.

Douglas, J., Gruber, S., Meister, K., Bonner, J., Watson, P., Krush, A. and Lynch, H. (2005) 'History and molecular genetics of Lynch syndrome in family G: a century later', *JAMA*, 294(17), pp. 2195-202.

- Drummond, J., Li, G., Longley, M. and Modrich, P. (1995) 'Isolation of an hMSH2-p160 heterodimer that restores DNA mismatch repair to tumor cells.', *Science*, 268(5219), pp. 1909-12.
- Dukes, C. (1952) 'Familial intestinal polyposis', *Annals of Eugenics*, 17(Part 1), pp. 1-29.
- Dunn, G., Bruce, A., Ikeda, H., Old, L. and Schreiber, R. (2002) 'Cancer immunoediting: from immunosurveillance to tumor escape.', *Nature Immunology*, 3(11), pp. 991-8.
- Dunstone, G. and Knaggs, T. (1972) 'Familial cancer of the colon and rectum', *Journal of Medical Genetics*, 9(4), pp. 451-6.
- Durno, C., Aronson, M., Tabori, U., Malkin, D., Gallinger, S. and Chan, H. (2012) 'Oncologic surveillance for subjects with biallelic mismatch repair gene mutations: 10 year follow-up of a kindred', *Pediatric Blood Cancer*, 59(4), pp. 652-6.
- Durno, C., Boland, C., Cohen, S., Dominitz, J., Giardiello, F., Johnson, D., Kaltenbach, T., Levin, T., Lieberman, D., Robertson, D. and Rex, D. (2017) 'Recommendations on surveillance and management of biallelic mismatch repair deficiency (BMMRD) syndrome: a consensus statement by the US Multi-Society Task Force on Colorectal Cancer', *Gastrointestinal Endoscopy*, 85(5), pp. 873-882.
- Duval, A. and Hamelin, R. (2002) 'Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability.', *Cancer research*, 62(9), pp. 2447-54.
- Edelbrock, M., Kaliyaperumal, S. and Williams, K. (2009) 'DNA mismatch repair efficiency and fidelity are elevated during DNA synthesis in human cells', *Mutation Research*, 662(1-2), pp. 59-66.
- Ellegren, H. (2004) 'Microsatellites: simple sequences with complex evolution.', *Nature Reviews. Genetics*, 5(6), pp. 435-45.
- Etzioni, R., Urban, N., Ramsey, S., McIntosh, M., Schwartz, S., Reid, B., Radich, J., Anderson, G. and Hartwell, L. (2003) 'The case for early detection', *Nature Reviews Cancer*, 3(4), pp. 243-52.

- Fan, H. and Chu, J. (2007) 'A brief review of short tandem repeat mutation', *Genomics Proteomics Bioinformatics*, 5(1), pp. 7-14.
- Fazekas, A., Steeves, R. and Newmaster, S. (2010) 'Improving sequencing quality from PCR products containing long mononucleotide repeats.', *Biotechniques*, 48(4), pp. 277-85.
- Fearon, E. and Vogelstein, B. (1990) 'A genetic model for colorectal tumorigenesis.', *Cell*, 61(5), pp. 759-67.
- Findeisen, P., Kloor, M., Merx, S., Sutter, C., Woerner, S., Dostmann, N., Benner, A., Dondog, B., Pawlita, M., Dippold, W., Wagner, R., Gebert, J. and von Knebel Doeberitz, M. (2005) 'T25 repeat in the 3' untranslated region of the CASP2 gene: a sensitive and specific marker for microsatellite instability in colorectal cancer.', *Cancer Research*, 65(18), pp. 8072-8.
- Fishel, R., Lescoe, M., Rao, M., Copeland, N., Jenkins, N., Garber, J., Kane, M. and Kolodner, R. (1993) 'The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer.', *Cell*, 75(5), pp. 1027-38.
- Fisher, D., Maple, J., Ben-Menachem, T., Cash, B., Decker, G., Early, D., Evans, J., Fanelli, R., Fukami, N., Hwang, J., Jain, R., Jue, T., Khan, K., Malpas, P., Sharaf, R., Shergill, A. and Dornitz, J. (2011) 'Complications of colonoscopy', *Gastrointestinal Endoscopy*, 74(4), pp. 745-52.
- Fong, S., Zhang, J., Lim, C., Eu, K. and Liu, Y. (2009) 'Comparison of 7 methods for extracting cell-free DNA from serum samples of colorectal cancer patients', *Clinical Chemistry*, 55(3), pp. 587-9.
- Frampton, G., Fichtenholtz, A., Otto, G., Wang, K., Downing, S., He, J., Schnall-Levin, M., White, J., Sanford, E., An, P., Sun, J., Juhn, F., Brennan, K., Iwanik, K., Maillet, A., Buell, J., White, E., Zhao, M., Balasubramanian, S., Terzic, S., Richards, T., Banning, V., Garcia, L., Mahoney, K., Zwick, Z., Donahue, A., Beltran, H., Mosquera, J., Rubin, M., Dogan, S., Hedvat, C., Berger, M., Puztai, L., Lechner, M., Boshoff, C., Jarosz, M., Vietz, C., Parker, A., Miller, V., Ross, J., Curran, J., Cronin, M., Stephens, P., Lipson, D. and Yelensky, R. (2013) 'Development and validation of a clinical cancer genomic profiling test based on massively parallel DNA sequencing', *Nature Biotechnology*, 31(11), pp. 1023-31.

Fu, D., Calvo, J. and Samson, L. (2012) 'Balancing repair and tolerance of DNA damage caused by alkylating agents', *Nature Reviews Cancer*, 12(2), pp. 104-20.

Furutani, E. and Shimamura, A. (2017) 'Germline Genetic Predisposition to Hematologic Malignancy', *Journal of Clinical Oncology*, 35(9), pp. 1018-1028.

Gallinger, S., Aronson, M., Shayan, K., Ratcliffe, E., Gerstle, J., Parkin, P., Rothenmund, H., Croitoru, M., Baumann, E., Durie, P., Weksberg, R., Pollett, A., Riddell, R., Ngan, B., Cutz, E., Lagarde, A. and Chan, H. (2004) 'Gastrointestinal cancers and neurofibromatosis type 1 features in children with a germline homozygous MLH1 mutation', *Gastroenterology*, 126(2), pp. 576-85.

Gardner, E. (1951) 'A genetic and clinical study of intestinal polyposis, a predisposing factor for carcinoma of the colon and rectum', *American Journal of Human Genetics*, 3(2), pp. 167-176.

Genovese, G., Kahler, A., Handsaker, R., Lindberg, J., Rose, S., Bakhoum, S., Chambert, K., Mick, E., Neale, B., Fromer, M., Purcell, S., Svantesson, O., Landen, M., Hoglund, M., Lehmann, S., Gabriel, S., Moran, J., Lander, E., Sullivan, P., Sklar, P., Gronberg, H., Hultman, C. and McCarroll, S. (2014) 'Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence', *New England Journal of Medicine*, 371(26), pp. 2477-87.

Genschel, J., Littman, S., Drummond, J. and Modrich, P. (1998) 'Isolation of MutSbeta from human cells and comparison of the mismatch repair specificities of MutSbeta and MutSalpha.', *The Journal of Biological Chemistry*, 273(31), pp. 19895-901.

Geurts-Giele, W., Leenen, C., Dubbink, H., Meijssen, I., Post, E., Sleddens, H., Kuipers, E., Goverde, A., van den Ouweland, A., van Lier, M., Steyerberg, E., van Leerdam, M., Wagner, A. and Dinjens, W. (2014) 'Somatic aberrations of mismatch repair genes as a cause of microsatellite-unstable cancers.', *The Journal of Pathology*, 234(4), pp. 548-59.

Goodenberger, M., Thomas, B., Riegert-Johnson, D., Boland, C., Plon, S., Clendenning, M., Win, A., Senter, L., Lipkin, S., Stadler, Z., Macrae, F., Lynch, H., Weitzel, J., de la Chapelle, A., Syngal, S., Lynch, P., Parry, S., Jenkins, M., Gallinger, S., Holter, S., Aronson, M., Newcomb, P., Burnett, T., Le Marchand, L., Pichurin, P., Hampel, H., Terdiman, J., Lu, K., Thibodeau, S. and

Lindor, N. (2016) 'PMS2 monoallelic mutation carriers: the known unknown', *Genetic Medicine*, 18(1), pp. 13-9.

Goverde, A., Spaander, M., van Doorn, H., Dubbink, H., van den Ouweland, A., Tops, C., Kooi, S., de Waard, J., Hoedemaeker, R., Bruno, M., Hofstra, R., de Bekker-Grob, E., Dinjens, W., Steyerberg, E. and Wagner, A. (2016) 'Cost-effectiveness of routine screening for Lynch syndrome in endometrial cancer patients up to 70years of age', *Gynecologic Oncology*, 143(3), pp. 453-459.

Gray, P., Tsai, P., Chen, D., Wu, S., Hoo, J., Mu, W., Li, B., Vuong, H., Lu, H., Batth, N., Willett, S., Uyeda, L., Shah, S., Gau, C., Umali, M., Espenschied, C., Janicek, M., Brown, S., Margileth, D., Dobra, L., Wagman, L., Rana, H., Hall, M., Ross, T., Terdiman, J., Cullinane, C., Ries, S., Totten, E. and Elliott, A. (2018) 'TumorNext-Lynch-MMR: a comprehensive next generation sequencing assay for the detection of germline and somatic mutations in genes associated with mismatch repair deficiency and Lynch syndrome.', *Oncotarget*, 9(29), pp. 20304-22.

Greaves, M. and Maley, C. (2012) 'Clonal evolution in cancer', *Nature*, 481(7381), pp. 306-13.

Gregory, M., Bertout, J., Ericson, N., Taylor, S., Mukherjee, R., Robins, H., Drescher, C. and Bielas, J. (2016) 'Targeted single molecule mutation detection with massively parallel sequencing', *Nucleic Acids Research*, 44(3), p. e22.

Groden, J., Thliveris, A., Samowitz, W., Carlson, M., Gelbert, L., Albertsen, H., Joslyn, G., Stevens, J., Spirio, L., Robertson, M., et al. (1991) 'Identification and characterization of the familial adenomatous polyposis coli gene', *Cell*, 66(3), pp. 589-600.

Biomarkers Definitions Working Group (2001) 'Biomarkers and surrogate endpoints: preferred definitions and conceptual framework', *Clinical Pharmacol and Therapeutics*, 69(3), pp. 89-95.

Gu, J., Ricker, C. and Barzi, A. (2018) 'Value of germline multi-gene panel next generation sequencing (NGS) in identification of hereditary cancer syndromes (HCS) in colorectal cancer population (CRC)', *Journal of Clinical Oncology*, 36(15_suppl), pp. 1587-1587.

Gubin, M., Zhang, X., Schuster, H., Caron, E., Ward, J., Noguchi, T., Ivanova, Y., Hundal, J., Arthur, C., Kriebber, W., Mulder, G., Toebes, M., Vesely, M., Lam, S., Korman, A., Allison, J.,

Freeman, G., Sharpe, A., Pearce, E., Schumacher, T., Aebbersold, R., Rammensee, H., Melief, C., Mardis, E., Gillanders, W., Artyomov, M. and Schreiber, R. (2014) 'Checkpoint blockade cancer immunotherapy targets tumour-specific mutant antigens', *Nature*, 515(7528), pp. 577-81.

Guinney, J., Dienstmann, R., Wang, X., de Reyniès, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B., Morris, J., Simon, I., Gerster, S., Fessler, E., De Sousa E Melo, F., Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J., Hanahan, D., Tabernero, J., Bernards, R., Friend, S., Laurent-Puig, P., Medema, J., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L. and Tejpar, S. (2015) 'The consensus molecular subtypes of colorectal cancer.', *Nature Medicine*, 21(11), pp. 1350-6.

Gur, C., Ibrahim, Y., Isaacson, B., Yamin, R., Abed, J., Gamliel, M., Enk, J., Bar-On, Y., Stanietsky-Kaynan, N., Copenhagen-Glazer, S., Shussman, N., Almogy, G., Cuapio, A., Hofer, E., Mevorach, D., Tabib, A., Ortenberg, R., Markel, G., Miklic, K., Jonjic, S., Brennan, C., Garrett, W., Bachrach, G. and Mandelboim, O. (2015) 'Binding of the Fap2 protein of *Fusobacterium nucleatum* to human inhibitory receptor TIGIT protects tumors from immune cell attack', *Immunity*, 42(2), pp. 344-355.

Gurin, C., Federici, M., Kang, L. and Boyd, J. (1999) 'Causes and consequences of microsatellite instability in endometrial carcinoma.', *Cancer Research*, 59(2), pp. 462-6.

Gylling, A., Abdel-Rahman, W., Juhola, M., Nuorva, K., Hautala, E., Järvinen, H., Mecklin, J., Aarnio, M. and Peltomäki, P. (2007) 'Is gastric cancer part of the tumour spectrum of hereditary non-polyposis colorectal cancer? A molecular genetic study', *Gut*, 56(7), pp. 926-933.

Haanstra, J., Vasen, H., Sanduleanu, S., van der Wouden, E., Koornstra, J., Kleibeuker, J. and de Vos Tot Nederveen Cappel, W. (2013) 'Quality colonoscopy and risk of interval cancer in Lynch syndrome', *International Journal of Colorectal Disease*, 28(12), pp. 1643-9.

Halford, S., Sasieni, P., Rowan, A., Wasan, H., Bodmer, W., Talbot, I., Hawkins, N., Ward, R. and Tomlinson, I. (2002) 'Low-level microsatellite instability occurs in most colorectal

cancers and is a nonrandomly distributed quantitative trait.', *Cancer Research*, 62(1), pp. 53-7.

Hamilton, S. (2018) 'Status of Testing for High-Level Microsatellite Instability/Deficient Mismatch Repair in Colorectal Carcinoma.', *JAMA Oncology*, 4(2), p. e173574.

Hamilton, S., Liu, B., Parsons, R., Papadopoulos, N., Jen, J., Powell, S., Krush, A., Berk, T., Cohen, Z., Tetu, B. and et al. (1995) 'The molecular basis of Turcot's syndrome', *N Engl J Med*, 332(13), pp. 839-47.

Hammel, P., Boissier, B., Chaumette, M., Piedbois, P., Rotman, N., Kouyoumdjian, J., Lubin, R., Delchier, J. and Soussi, T. (1997) 'Detection and monitoring of serum p53 antibodies in patients with colorectal cancer', *Gut*, 40(3), pp. 356-61.

Hampel, H. and de la Chapelle, A. (2011) 'The search for unaffected individuals with Lynch syndrome: do the ends justify the means?', *Cancer Prevention Research*, 4(1), pp. 1-5.

Hampel, H., Frankel, W., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., Clendenning, M., Sotamaa, K., Prior, T., Westman, J., Panescu, J., Fix, D., Lockman, J., LaJeunesse, J., Comeras, I. and Chapelle, A.d.I. (2008) 'Feasibility of screening for Lynch syndrome among patients with colorectal cancer.', *Journal of Clinical Oncology*, 26(35), pp. 5783-8.

Hampel, H., Frankel, W., Martin, E., Arnold, M., Khanduja, K., Kuebler, P., Nakagawa, H., Sotamaa, K., Prior, T., Westman, J., Panescu, J., Fix, D., Lockman, J., Comeras, I. and Chapelle, A.d.I. (2005a) 'Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer).', *New England Journal of Medicine*, 352(18), pp. 1851-60.

Hampel, H., Frankel, W., Panescu, J., Lockman, J., Sotamaa, K., Fix, D., Comeras, I., La Jeunesse, J., Nakagawa, H., Westman, J., Prior, T., Clendenning, M., Penzone, P., Lombardi, J., Dunn, P., Cohn, D., Copeland, L., Eaton, L., Fowler, J., Lewandowski, G., Vaccarello, L., Bell, J., Reid, G. and A, d.I.C. (2006) 'Screening for Lynch syndrome (hereditary nonpolyposis colorectal cancer) among endometrial cancer patients.', *Cancer Research*, 66(15), pp. 7810-7.

Hampel, H., Pearlman, R., Beightol, M., Zhao, W., Jones, D., Frankel, W., Goodfellow, P., Yilmaz, A., Miller, K., Bacher, J., Jacobson, A., Paskett, E., Shields, P., Goldberg, R., de la

Chapelle, A., Shirts, B., Pritchard, C. and Group, O.C.C.P.I.S. (2018) 'Assessment of Tumor Sequencing as a Replacement for Lynch Syndrome Screening and Current Molecular Tests for Patients With Colorectal Cancer.', *JAMA Oncology*.

Hampel, H., Stephens, J., Pukkala, E., Sankila, R., Aaltonen, L., Mecklin, J. and Chapelle, A.d.l. (2005b) 'Cancer risk in hereditary nonpolyposis colorectal cancer syndrome: later age of onset.', *Gastroenterology*, 129(2), pp. 415-21.

Hanahan, D. and Weinberg, R. (2011) 'Hallmarks of Cancer: The Next Generation', *Cell*, 144(5), pp. 646-74.

Haraldsdottir, S., Hampel, H., Tomsic, J., Frankel, W., Pearlman, R., de la Chapelle, A. and Pritchard, C. (2014) 'Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations.', *Gastroenterology*, 147(6), pp. 1308-16.

Haraldsdottir, S., Rafnar, T., Frankel, W., Einarsdottir, S., Sigurdsson, A., Hampel, H., Snaebjornsson, P., Masson, G., Weng, D., Arngrimsson, R., Kehr, B., Yilmaz, A., Haraldsson, S., Sulem, P., Stefansson, T., Shields, P., Sigurdsson, F., Bekaii-Saab, T., Moller, P., Steinarsdottir, M., Alexiusdottir, K., Hitchins, M., Pritchard, C., de la Chapelle, A., Jonasson, J., Goldberg, R. and Stefansson, K. (2017) 'Comprehensive population-wide analysis of Lynch syndrome in Iceland reveals founder mutations in MSH6 and PMS2.', *Nature Communications*, 8, p. 14755.

Hartwell, L., Mankoff, D., Paulovich, A., Ramsey, S. and Swisher, E. (2006) 'Cancer biomarkers: a systems approach', *Nature Biotechnology*, 24(8), pp. 905-8.

Haug, U., Kuntz, K., Knudsen, A., Hundt, S. and Brenner, H. (2011) 'Sensitivity of immunochemical faecal occult blood testing for detecting left- vs right-sided colorectal neoplasia', *British Journal of Cancer*, 104(11), pp. 1779-85.

Hayes, D. (2018) 'Precision Medicine and Testing for Tumor Biomarkers-Are All Tests Born Equal?', *JAMA Oncology*, 4(6), pp. 773-774.

- Hempelmann, J., Scroggins, S., Pritchard, C. and Salipante, S. (2015) 'MSIplus for Integrated Colorectal Cancer Molecular Testing by Next-Generation Sequencing.', *Journal of Molecular Diagnostics*, 17(6), pp. 705-14.
- Henry, N. and Hayes, D. (2012) 'Cancer biomarkers', *Molecular Oncology*, 6(2), pp. 140-6.
- Herman, J., Umar, A., Polyak, K., Graff, J., Ahuja, N., Issa, J., Markowitz, S., Willson, J., Hamilton, S., Kinzler, K., Kane, M., Kolodner, R., Vogelstein, B., Kunkel, T. and Baylin, S. (1998) 'Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma.', *Proceedings of the National Academy of Sciences of the United States of America*, 95(12), pp. 6870-5.
- Hiatt, J., Pritchard, C., Salipante, S., O'Roak, B. and Shendure, J. (2013) 'Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation.', *Genome Research*, 23(5), pp. 843-54.
- Hickman, M. and Samson, L. (1999) 'Role of DNA mismatch repair and p53 in signaling induction of apoptosis by alkylating agents', *Proceedings of the National Academy of Sciences USA*, 96(19), pp. 10764-9.
- Horak, P., Frohling, S. and Glimm, H. (2016) 'Integrating next-generation sequencing into clinical oncology: strategies, promises and pitfalls', *ESMO Open*, 1(5), p. e000094.
- Hu, P., Lee, C., Xu, J., Simien, C., Fan, C., Tam, M., Ramagli, L., Brown, B., Lynch, P., Frazier, M., Siciliano, M. and Coolbaugh-Murphy, M. (2011) 'Microsatellite instability in saliva from patients with hereditary non-polyposis colon cancer and siblings carrying germline mismatch repair gene mutations', *Annals of Clinical and Laboratory Science*, 41(4), pp. 321-30.
- Huang, J., Papadopoulos, N., McKinley, A., Farrington, S., Curtis, L., Wyllie, A., Zheng, S., Willson, J., Markowitz, S., Morin, P., Kinzler, K., Vogelstein, B. and Dunlop, M. (1996) 'APC mutations in colorectal tumors with mismatch repair deficiency.', *Proceedings of the National Academy of Sciences of the United States of America*, 93(17), pp. 9049-54.
- Huang, Q., Xu, F., Shen, H., Deng, H., Liu, Y., Liu, Y., Li, J., Recker, R. and Deng, H. (2002) 'Mutation patterns at dinucleotide microsatellite loci in humans', *American Journal of Human Genetics*, 70(3), pp. 625-34.

Huang, Z., Huang, D., Ni, S., Peng, Z., Sheng, W. and Du, X. (2010) 'Plasma microRNAs are promising novel biomarkers for early detection of colorectal cancer', *International Journal of Cancer*, 127(1), pp. 118-26.

Iino, H., Simms, L., Young, J., Arnold, J., Winship, I., Webb, S., Furlong, K., Leggett, B. and Jass, J. (2000) 'DNA microsatellite instability and mismatch repair protein loss in adenomas presenting in hereditary non-polyposis colorectal cancer.', *Gut*, 47(1), pp. 37-42.

Ilyas, S. and Yang, J. (2015) 'Landscape of Tumor Antigens in T Cell Immunotherapy', *Journal of Immunology*, 195(11), pp. 5117-22.

Illumina Inc. (2017) 'Effects of Index Misassignment on Multiplexing and Downstream Analysis'.

Ingham, D., Diggle, C., Berry, I., Bristow, C., Hayward, B., Rahman, N., Markham, A., Sheridan, E., Bonthron, D. and Carr, I. (2013) 'Simple detection of germline microsatellite instability for diagnosis of constitutional mismatch repair cancer syndrome.', *Human Mutation*, 34(6), pp. 847-52.

Ionov, Y., Peinado, M., Malkhosyan, S., Shibata, D. and Perucho, M. (1993) 'Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis.', *Nature*, 363(6429), pp. 558-61.

Ionov, Y., Yamamoto, H., Krajewski, S., Reed, J. and Perucho, M. (2000) 'Mutational inactivation of the proapoptotic gene BAX confers selective advantage during tumor clonal evolution.', *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), pp. 10872-7.

Ishikawa, T., Fujita, T., Suzuki, Y., Okabe, S., Yuasa, Y., Iwai, T. and Kawakami, Y. (2003) 'Tumor-specific immunological recognition of frameshift-mutated peptides in colon cancer with microsatellite instability.', *Cancer research*, 63(17), pp. 5564-72.

Jackson, S. and Bartek, J. (2009) 'The DNA-damage response in human biology and disease.', *Nature*, 461(7267), pp. 1071-8.

Jahr, S., Hentze, H., Englisch, S., Hardt, D., Fackelmayer, F., Hesch, R. and Knippers, R. (2001) 'DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells.', *Cancer Research*, 61(4), pp. 1659-65.

Järvinen, H., Aarnio, M., Mustonen, H., Aktan-Collan, K., Aaltonen, L., Peltomäki, P., De La Chapelle, A. and Mecklin, J. (2000) 'Controlled 15-year trial on screening for colorectal cancer in families with hereditary nonpolyposis colorectal cancer.', *Gastroenterology*, 118(5), pp. 829-34.

Jass, J. and Stewart, S. (1992) 'Evolution of hereditary non-polyposis colorectal cancer.', *Gut*, 33(6), pp. 783-6.

Jass, J., Stewart, S., Stewart, J. and Lane, M. (1994) 'Hereditary non-polyposis colorectal cancer--morphologies, genes and mutations.', *Mutation Research*, 310(1), pp. 125-33.

Jennings, L., Arcila, M., Corless, C., Kamel-Reid, S., Lubin, I., Pfeifer, J., Temple-Smolkin, R., Voelkerding, K. and Nikiforova, M. (2017) 'Guidelines for Validation of Next-Generation Sequencing-Based Oncology Panels: A Joint Consensus Recommendation of the Association for Molecular Pathology and College of American Pathologists.', *Journal of Molecular Diagnostics*, 19(3), pp. 341-65.

Jiricny, J. (2006) 'The multifaceted mismatch-repair system.', *Nature Reviews. Molecular Cell Biology*, 7(5), pp. 335-46.

Jover, R., Zapater, P., Castells, A., Llor, X., Andreu, M., Cubiella, J., Piñol, V., Xicola, R., Bujanda, L., Reñé, J., Clofent, J., Bessa, X., Morillas, J., Nicolás-Pérez, D., Payá, A., Alenda, C. and Association, G.O.G.o.t.S.G. (2006) 'Mismatch repair status in the prediction of benefit from adjuvant fluorouracil chemotherapy in colorectal cancer.', *Gut*, 55(6), pp. 848-55.

Juliusson, G. and Liliemark, J. (1993) 'High complete remission rate from 2-chloro-2'-deoxyadenosine in previously treated patients with B-cell chronic lymphocytic leukemia: response predicted by rapid decrease of blood lymphocyte count', *Journal of Clinical Oncology*, 11(4), pp. 679-89.

Kambara, T., Simms, L., Whitehall, V., Spring, K., Wynter, C., Walsh, M., Barker, M., Arnold, S., McGivern, A., Matsubara, N., Tanaka, N., Higuchi, T., Young, J., Jass, J. and Leggett, B.

(2004) 'BRAF mutation is associated with DNA methylation in serrated polyps and cancers of the colorectum', *Gut*, 53(8), pp. 1137-44.

Kariola, R., Hampel, H., Frankel, W., Rævaara, T., Chapelle, A.d.l. and Nyström-Lahti, M. (2004) 'MSH6 missense mutations are often associated with no or low cancer susceptibility.', *British Journal of Cancer*, 91(7), pp. 1287-92.

Karran, P. and Attard, N. (2008) 'Thiopurines in current medical practice: molecular mechanisms and contributions to therapy-related cancer', *Nature Reviews Cancer*, 8(1), pp. 24-36.

Kaz, A., Kim, Y., Dzieciatkowski, S., Lynch, H., Watson, P., Kay Washington, M., Lin, L. and Grady, W. (2007) 'Evidence for the role of aberrant DNA methylation in the pathogenesis of Lynch syndrome adenomas.', *International Journal of Cancer*, 120(9), pp. 1922-9.

Kehl, K., Shen, C., Litton, J., Arun, B. and Giordano, S. (2016) 'Rates of BRCA1/2 mutation testing among young survivors of breast cancer.', *Breast Cancer Research and Treatment*, 155(1), pp. 165-73.

Kelkar, Y., Tyekucheva, S., Chiaromonte, F. and Makova, K. (2008) 'The genome-wide determinants of human and chimpanzee microsatellite evolution.', *Genome Research*, 18(1), pp. 30-8.

Kennedy, S., Schmitt, M., Fox, E., Kohn, B., Salk, J., Ahn, E., Prindle, M., Kuong, K., Shen, J., Risques, R. and Loeb, L. (2014) 'Detecting ultralow-frequency mutations by Duplex Sequencing', *Nature Protocols*, 9(11), pp. 2586-606.

Kim, S., Klempner, S., Park, S., Park, J., Park, Y., Lim, H., Kang, W., Kim, K. and Lee, J. (2017) 'Correlating programmed death ligand 1 (PD-L1) expression, mismatch repair deficiency, and outcomes across tumor types: implications for immunotherapy', *Oncotarget*, 8(44), pp. 77415-77423.

Kim, T., Laird, P. and Park, P. (2013) 'The landscape of microsatellite instability in colorectal and endometrial cancer genomes.', *Cell*, 155(4), pp. 858-68.

- Kinde, I., Wu, J., Papadopoulos, N., Kinzler, K. and Vogelstein, B. (2011) 'Detection and quantification of rare mutations with massively parallel sequencing.', *Proceedings of the National Academy of Sciences of the United States of America*, 108(23), pp. 9530-5.
- Klein, E., Ben-Bassat, H., Neumann, H., Ralph, P., Zeuthen, J., Polliack, A. and Vánky, F. (1976) 'Properties of the K562 cell line, derived from a patient with chronic myeloid leukemia.', *International Journal of Cancer*, 18(4), pp. 421-31.
- Kloor, M., Huth, C., Voigt, A., Benner, A., Schirmacher, P., Doeberitz, M.v.K. and Bläker, H. (2012) 'Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study.', *The Lancet Oncology*, 13(6), pp. 598-606.
- Kloor, M., Michel, S., Buckowitz, B., Rüschoff, J., Büttner, R., Holinski-Feder, E., Dippold, W., Wagner, R., Tariverdian, M., Benner, A., Schwitalle, Y., Kuchenbuch, B. and Doeberitz, M.v.K. (2007) 'Beta2-microglobulin mutations in microsatellite unstable colorectal tumors', *International Journal of Cancer*, 121(2), pp. 454-8.
- Kloor, M. and von Knebel Doeberitz, M. (2016) 'The Immune Biology of Microsatellite-Unstable Cancer', *Trends in Cancer*, 2(3), pp. 121-133.
- Knudson, A. (2001) 'Two genetic hits (more or less) to cancer.', *Nature Reviews. Cancer*, 1(2), pp. 157-62.
- Koi, M., Umar, A., Chauhan, D., Cherian, S., Carethers, J., Kunkel, T. and Boland, C. (1994) 'Human chromosome 3 corrects mismatch repair deficiency and microsatellite instability and reduces N-methyl-N'-nitro-N-nitrosoguanidine tolerance in colon tumor cells with homozygous hMLH1 mutation.', *Cancer Research*, 54(16), pp. 4308-12.
- Kornberg, A., Bertsch, L., Jackson, J. and Khorana, H. (1964) 'Enzymatic synthesis of deoxyribonucleic acid, XVI. Oligonucleotides as templates and the mechanism of their replication', *Proceedings of the National Academy of Sciences USA*, 51, pp. 315-23.
- Kruger, S., Kinzel, M., Walldorf, C., Gottschling, S., Bier, A., Tinschert, S., von Stackelberg, A., Henn, W., Gorgens, H., Boue, S., Kolble, K., Buttner, R. and Schackert, H. (2008) 'Homozygous PMS2 germline mutations in two families with early-onset haematological malignancy, brain

tumours, HNPCC-associated tumours, and signs of neurofibromatosis type 1', *European Journal of Human Genetics*, 16(1), pp. 62-72.

Ku, C., Cooper, D., Wu, M., Roukos, D., Pawitan, Y., Soong, R. and Iacopetta, B. (2012) 'Gene discovery in familial cancer syndromes by exome sequencing: prospects for the elucidation of familial colorectal cancer type X.', *Modern Pathology*, 25(8), pp. 1055-68.

Kunkel, T. and Erie, D. (2005) 'DNA mismatch repair', *Annual Review of Biochemistry*, 74, pp. 681-710.

Labianca, R., Nordlinger, B., Beretta, G., Brouquet, A. and Cervantes, A. (2010) 'Primary colon cancer: ESMO Clinical Practice Guidelines for diagnosis, adjuvant treatment and follow-up', *Annals of Oncology*, 21 Suppl 5, pp. v70-7.

Ladabaum, U., Wang, G., Terdiman, J., Blanco, A., Kuppermann, M., Boland, C., Ford, J., Elkin, E. and Phillips, K. (2011) 'Strategies to identify the Lynch syndrome among patients with colorectal cancer: a cost-effectiveness analysis.', *Annals of Internal Medicine*, 155(2), pp. 69-79.

Laiho, P., Launonen, V., Lahermo, P., Esteller, M., Guo, M., Herman, J., Mecklin, J., Järvinen, H., Sistonen, P., Kim, K., Shibata, D., Houlston, R. and Aaltonen, L. (2002) 'Low-level microsatellite instability in most colorectal carcinomas.', *Cancer Research*, 62(4), pp. 1166-70.

Le, D., Durham, J., Smith, K., Wang, H., Bartlett, B., Aulakh, L., Lu, S., Kemberling, H., Wilt, C., Luber, B., Wong, F., Azad, N., Rucki, A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G., Crocenzi, T., Lee, J., Greten, T., Duffy, A., Ciombor, K., Eyring, A., Lam, B., Joe, A., Kang, S., Holdhoff, M., Danilova, L., Cope, L., Meyer, C., Zhou, S., Goldberg, R., Armstrong, D., Bever, K., Fader, A., Taube, J., Housseau, F., Spetzler, D., Xiao, N., Pardoll, D., Papadopoulos, N., Kinzler, K., Eshleman, J., Vogelstein, B., Anders, R. and Diaz, L.J. (2017) 'Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade.', *Science*, 357(6349), pp. 409-13.

Le, D., Uram, J., Wang, H., Bartlett, B., Kemberling, H., Eyring, A., Skora, A., Luber, B., Azad, N., Laheru, D., Biedrzycki, B., Donehower, R., Zaheer, A., Fisher, G., Crocenzi, T., Lee, J., Duffy, S., Goldberg, R., Chapelle, A.d.l., Koshiji, M., Bhajjee, F., Huebner, T., Hruban, R.,

Wood, L., Cuka, N., Pardoll, D., Papadopoulos, N., Kinzler, K., Zhou, S., Cornish, T., Taube, J., Anders, R., Eshleman, J., Vogelstein, B. and Diaz, L. (2015) 'PD-1 Blockade in Tumors with Mismatch-Repair Deficiency.', *New England Journal of Medicine*, 372(26), pp. 2509-20.

Leach, F., Polyak, K., Burrell, M., Johnson, K., Hill, D., Dunlop, M., Wyllie, A., Peltomaki, P., Chapelle, A.d.l., Hamilton, S., Kinzler, K. and Vogelstein, B. (1996) 'Expression of the human mismatch repair gene hMSH2 in normal and neoplastic tissues.', *Cancer Research*, 56(2), pp. 235-40.

Lee, B., Lee, E., Jung, E., Chun, H., Chang, D., Song, S., Park, J. and Kim, D. (2009) 'Aberrant methylation of APC, MGMT, RASSF2A, and Wif-1 genes in plasma as a biomarker for early detection of colorectal cancer', *Clinical Cancer Research*, 15(19), pp. 6185-91.

Lee, J., Li, L., Gretz, N., Gebert, J. and Dihlmann, S. (2012) 'Absent in Melanoma 2 (AIM2) is an important mediator of interferon-dependent and -independent HLA-DRA and HLA-DRB gene expression in colorectal cancers.', *Oncogene*, 31(10), pp. 1242-53.

Lee, J., Liles, E., Bent, S., Levin, T. and Corley, D. (2014) 'Accuracy of Fecal Immunochemical Tests for Colorectal Cancer: Systematic Review and Meta-analysis', *Annals of Internal Medicine*, 160(3), pp. 171-171.

Leenders, E., Westdorp, H., Bruggemann, R., Loeffen, J., Kratz, C., Burn, J., Hoogerbrugge, N. and Jongmans, M. (2018) 'Cancer prevention by aspirin in children with Constitutional Mismatch Repair Deficiency (CMMRD)', *European Journal of Human Genetics*.

Leenen, C., Geurts-Giele, W., Dubbink, H., Reddingius, R., van den Ouweland, A., Tops, C., van de Klift, H., Kuipers, E., van Leerdam, M., Dinjens, W. and Wagner, A. (2011) 'Pitfalls in molecular analysis for mismatch repair deficiency in a family with biallelic pms2 germline mutations', *Clin Genet*, 80(6), pp. 558-65.

Leenen, C., van Lier, M., van Doorn, H., van Leerdam, M., Kooi, S., de Waard, J., Hoedemaeker, R., van den Ouweland, A., Hulspas, S., Dubbink, H., Kuipers, E., Wagner, A., Dinjens, W. and Steyerberg, E. (2012) 'Prospective evaluation of molecular screening for Lynch syndrome in patients with endometrial cancer ≤ 70 years.', *Gynaecologic Oncology*, 125(2), pp. 414-20.

Levi, Z., Kariv, R., Barnes-Kedar, I., Goldberg, Y., Half, E., Morgentern, S., Eli, B., Baris, H., Vilkin, A., Belfer, R., Niv, Y., Elhasid, R., Dvir, R., Abu-Freha, N. and Cohen, S. (2015) 'The gastrointestinal manifestation of constitutional mismatch repair deficiency syndrome: from a single adenoma to polyposis-like phenotype and early onset cancer', *Clinical Genetics*, 88(5), pp. 474-8.

Levy, S. and Myers, R. (2016) 'Advancements in Next-Generation Sequencing', *Annual Review of Genomics and Human Genetics*, 17, pp. 95-115.

Li, L., Morales, J., Veigl, M., Sedwick, D., Greer, S., Meyers, M., Wagner, M., Fishel, R. and Boothman, D. (2009) 'DNA mismatch repair (MMR)-dependent 5-fluorouracil cytotoxicity and the potential for new therapeutic targets.', *British Journal of Pharmacology*, 158(3), pp. 679-92.

Li, H. and Durbin, R. (2010) 'Fast and accurate long-read alignment with Burrows-Wheeler transform.', *Bioinformatics*, 26(5), pp. 589-95.

Li, L., Hamel, N., Baker, K., McGuffin, M., Couillard, M., Gologan, A., Marcus, V., Chodirker, B., Chudley, A., Stefanovici, C., Durandy, A., Hegele, R., Feng, B., Goldgar, D., Zhu, J., De Rosa, M., Gruber, S., Wimmer, K., Young, B., Chong, G., Tischkowitz, M. and Foulkes, W. (2015) 'A homozygous PMS2 founder mutation with an attenuated constitutional mismatch repair deficiency phenotype', *Journal of Medical Genetics*, 52(5), pp. 348-52.

Li, Y., Karjalainen, A., Koskinen, H., Hemminki, K., Vainio, H., Shnaidman, M., Ying, Z., Pukkala, E. and Brandt-Rauf, P. (2005) 'p53 autoantibodies predict subsequent development of cancer.', *International Journal of Cancer*, 114(1), pp. 157-60.

Lièvre, A., Bachet, J., Le Corre, D., Boige, V., Landi, B., Emile, J., Côté, J., Tomasic, G., Penna, C., Ducreux, M., Rougier, P., Penault-Llorca, F. and Laurent-Puig, P. (2006) 'KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer.', *Cancer Research*, 66(8), pp. 3992-5.

Ligtenberg, M., Kuiper, R., Chan, T., Goossens, M., Hebeda, K., Voorendt, M., Lee, T., Bodmer, D., Hoenselaar, E., Hendriks-Cornelissen, S., Tsui, W., Kong, C., Brunner, H., van Kessel, A., Yuen, S., van Krieken, J., Leung, S. and Hoogerbrugge, N. (2009) 'Heritable somatic

methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3' exons of TACSTD1', *Nature Genetics*, 41(1), pp. 112-7.

Lindor, N., Rabe, K., Petersen, G., Haile, R., Casey, G., Baron, J., Gallinger, S., Bapat, B., Aronson, M., Hopper, J., Jass, J., LeMarchand, L., Grove, J., Potter, J., Newcomb, P., Terdiman, J., Conrad, P., Moslein, G., Goldberg, R., Ziogas, A., Anton-Culver, H., Andrade, M.d., Siegmund, K., Thibodeau, S., Boardman, L. and Seminara, D. (2005) 'Lower cancer incidence in Amsterdam-I criteria families without mismatch repair deficiency: familial colorectal cancer type X.', *293*, 16(1979-85).

Lindor, N., Smalley, R., Barker, M., Bigler, J., Krumroy, L., Lum-Jones, A., Plummer, S., Selander, T., Thomas, S., Youash, M., Seminara, D., Casey, G., Bapat, B. and Thibodeau, S. (2006) 'Ascending the learning curve--MSI testing experience of a six-laboratory consortium.', *Cancer Biomarkers*, 2(1-2), pp. 5-9.

Linnebacher, M., Gebert, J., Rudy, W., Woerner, S., Yuan, Y., Bork, P. and Doeberitz, M.v.K. (2001) 'Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens.', *International Journal of Cancer*, 93(1), pp. 6-11.

Lipkin, S., Moens, P., Wang, V., Lenzi, M., Shanmugarajah, D., Gilgeous, A., Thomas, J., Cheng, J., Touchman, J., Green, E., Schwartzberg, P., Collins, F. and Cohen, P. (2002) 'Meiotic arrest and aneuploidy in MLH3-deficient mice.', *Nature Genetics*, 31(4), pp. 385-90.

Liu, B., Parsons, R., Papadopoulos, N., Nicolaides, N., Lynch, H., Watson, P., Jass, J., Dunlop, M., Wyllie, A., Peltomäki, P., Chapelle, A.d.I., Hamilton, S., Vogelstein, B. and Kinzler, K. (1996) 'Analysis of mismatch repair genes in hereditary non-polyposis colorectal cancer patients.', *Nature Medicine*, 2(2), pp. 169-74.

Liu, Q., Hesson, L., Nunez, A., Packham, D., Williams, R., Ward, R. and Sloane, M. (2016) 'A cryptic paracentric inversion of MSH2 exons 2-6 causes Lynch syndrome.', *Carcinogenesis*, 37(1), pp. 10-7.

Locker, G., Hamilton, S., Harris, J., Jessup, J., Kemeny, N., Macdonald, J., Somerfield, M., Hayes, D. and Bast Jr., R. (2006) 'ASCO 2006 update of recommendations for the use of tumor markers in gastrointestinal cancer', *Journal of Clinical Oncology*, 24(33), pp. 5313-27.

Lorans, M., Dow, E., Macrae, F., Winship, I. and Buchanan, D. (2018) 'Update on Hereditary Colorectal Cancer: Improving the Clinical Utility of Multigene Panel Testing.' *Clinical Colorectal Cancer*, 17(2), pp. e293-e305

Lothe, R., Peltomäki, P., Meling, G., Aaltonen, L., Nyström-Lahti, M., Pylkkänen, L., Heimdal, K., Andersen, T., Møller, P., Rognum, T. and al, e. (1993) 'Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history.', *Cancer Research*, 53(24), pp. 5849-52.

Loughrey, M., Quirke, P. and Shepherd, N. (2014) 'Dataset for colorectal cancer histopathology reports', *Standards and datasets for reporting cancers, The Royal College of Pathologists*.

Lu, Y., Soong, T. and Elemento, O. (2013) 'A novel approach for characterizing microsatellite instability in cancer cells.', *PLoS One*, 8(5), p. e63056.

Lynch, H., Kimberling, W., Albano, W., Lynch, J., Elston, R., Biscione, K., Schuelke, G., Sandberg, A., Lipkin, M., Deschner, E., Mikol, Y., Bailey-Wilson, J. and Danes, B. (1985) 'Hereditary nonpolyposis colorectal cancer (Lynch syndromes I and II). I. Clinical description of resource.', *Cancer*, 56(4), pp. 934-8.

Lynch, H. and Krush, A. (1971) 'Cancer family "G" revisited: 1895-1970', *Cancer*, 27(6), pp. 1505-11.

Lynch, H., Lynch, P., Lanspa, S., Snyder, C., Lynch, J. and Boland, C. (2009) 'Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications', *Clinical Genetics*, 76(1), pp. 1-18.

Lynch, H., Smyrk, T. and Lynch, J. (1998) 'Molecular genetics and clinical-pathology features of hereditary nonpolyposis colorectal carcinoma (Lynch syndrome): historical journey from pedigree anecdote to molecular genetic confirmation', *Oncology*, 55(2), pp. 103-8.

Maby, P., Tougeron, D., Hamieh, M., Mlecnik, B., Kora, H., Bindea, G., Angell, H., Fredriksen, T., Elie, N., Fauquemberg, E., Drouet, A., Leprince, J., Benichou, J., Mauillon, J., Pessot, F.L., Sesboué, R., Tuech, J., Sabourin, J., Michel, P., Frébourg, T., Galon, J. and Latouche, J. (2015) 'Correlation between Density of CD8+ T-cell Infiltrate in Microsatellite Unstable Colorectal

Cancers and Frameshift Mutations: A Rationale for Personalized Immunotherapy', *Cancer Research*, 75(17), pp. 3446-55.

Mandelker, D., Schmidt, R., Ankala, A., McDonald Gibson, K., Bowser, M., Sharma, H., Duffy, E., Hegde, M., Santani, A., Lebo, M. and Funke, B. (2016) 'Navigating highly homologous genes in a molecular diagnostic setting: a resource for clinical next-generation sequencing.', *Genetics in Medicine*, 18(12), pp. 1282-89.

Mangold, E., Pagenstecher, C., Friedl, W., Fischer, H., Merkelbach-Bruse, S., Ohlendorf, M., Friedrichs, N., Aretz, S., Buettner, R., Propping, P. and Mathiak, M. (2005) 'Tumours from MSH2 mutation carriers show loss of MSH2 expression but many tumours from MLH1 mutation carriers exhibit weak positive MLH1 staining.', *The Journal of Pathology*, 207(4), pp. 385-95.

Marino, P., Touzani, R., Perrier, L., Rouleau, E., Kossi, D., Zhaomin, Z., Charrier, N., Goardon, N., Preudhomme, C., Durand-Zaleski, I., Borget, I., Baffert, S. and Group, N. (2018) 'Cost of cancer diagnosis using next-generation sequencing targeted gene panels in routine practice: a nationwide French study.', *European Journal of Human Genetics*, 26(3), pp. 314-23.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R., Zborowska, E., Kinzler, K., Vogelstein, B. and al., e. (1995) 'Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability.', *Science*, 268(5215), pp. 1336-8.

Marx, V. (2016) 'Cancer: hunting rare somatic mutations', *Nature Methods*, 13(4), pp. 295-9.

Matsumura, M., Fremont, D., Peterson, P. and Wilson, I. (1992) 'Emerging principles for the recognition of peptide antigens by MHC class I molecules', *Science*, 257(5072), pp. 927-34.

McGranahan, N., Rosenthal, R., Hiley, C., Rowan, A., Watkins, T., Wilson, G., Birkbak, N., Veeriah, S., Van Loo, P., Herrero, J. and Swanton, C. (2017) 'Allele-Specific HLA Loss and Immune Escape in Lung Cancer Evolution', *Cell*, 171(6), pp. 1259-1271.e11.

Menko, F., Kaspers, G., Meijer, G., Claes, K., van Hagen, J. and Gille, J. (2004) 'A homozygous MSH6 mutation in a child with cafe-au-lait spots, oligodendroglioma and rectal cancer', *Familial Cancer*, 3(2), pp. 123-7.

Mensenkamp, A., Vogelaar, I., van Zelst-Stams, W., Goossens, M., Ouchene, H., Hendriks-Cornelissen, S., Kwint, M., Hoogerbrugge, N., Nagtegaal, I. and Ligtenberg, M. (2014) 'Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors.', *Gastroenterology*, 146(3), pp. 643-6.

MERCK&Co.Inc. (2017) 'KEYTRUDA® (pembrolizumab) for injection, for intravenous use', *FDA Reference ID: 4101813*.

Merlo, A., Rochlitz, C. and Scott, R. (1996) 'Survival of patients with Turcot's syndrome and glioblastoma', *N Engl J Med*, 334(11), pp. 736-7.

Mills, A., Liou, S., Ford, J., Berek, J., Pai, R. and Longacre, T. (2014) 'Lynch Syndrome Screening Should Be Considered for All Patients With Newly Diagnosed Endometrial Cancer', *The American journal of surgical pathology*, 38(11), pp. 1501-1509.

Mittal, D., Gubin, M., Schreiber, R. and Smyth, M. (2014) 'New insights into cancer immunoediting and its three component phases--elimination, equilibrium and escape', *Current Opinion in Immunology*, 27, pp. 16-25.

Miyaki, M., Konishi, M., Tanaka, K., Kikuchi-Yanoshita, R., Muraoka, M., Yasuno, M., Igari, T., Koike, M., Chiba, M. and Mori, T. (1997) 'Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer.', *Nature Genetics*, 17(3), pp. 271-2.

Mlecnik, B., Bindea, G., Angell, H., Maby, P., Angelova, M., Tougeron, D., Church, S., Lafontaine, L., Fischer, M., Fredriksen, T., Sasso, M., Bilocq, A., Kirilovsky, A., Obenauf, A., Hamieh, M., Berger, A., Bruneval, P., Tuech, J., Sabourin, J., Le Pessot, F., Mauillon, J., Rafii, A., Laurent-Puig, P., Speicher, M., Trajanoski, Z., Michel, P., Sesboue, R., Frebourg, T., Pages, F., Valge-Archer, V., Latouche, J. and Galon, J. (2016) 'Integrative Analyses of Colorectal Cancer Show Immunoscore Is a Stronger Predictor of Patient Survival Than Microsatellite Instability', *Immunity*, 44(3), pp. 698-711.

Møller, P., Seppälä, T., Bernstein, I., Holinski-Feder, E., Sala, P., Evans, D., Lindblom, A., Macrae, F., Blanco, I., Sijmons, R., Jeffries, J., Vasen, H., Burn, J., Nakken, S., Hovig, E., Rødland, E., Tharmaratnam, K., de Vos Tot Nederveen Cappel, W., Hill, J., Wijnen, J., Green, K., Lalloo, F., Sunde, L., Mints, M., Bertario, L., Pineda, M., Navarro, M., Morak, M., Renkonen-Sinisalo, L., Frayling, I., Plazzer, J., Pylvanainen, K., Sampson, J., Capella, G.,

Mecklin, J., Möslein, G. and Group, M. (2017a) 'Cancer incidence and survival in Lynch syndrome patients receiving colonoscopic and gynaecological surveillance: first report from the prospective Lynch syndrome database.', *Gut*, 66(3), pp. 464-72.

Møller, P., Seppälä, T., Bernstein, I., Holinski-Feder, E., Sala, P., Evans, D., Lindblom, A., Macrae, F., Blanco, I., Sijmons, R., Jeffries, J., Vasen, H., Burn, J., Nakken, S., Hovig, E., Rødland, E., Tharmaratnam, K., de Vos Tot Nederveen Cappel, W., Hill, J., Wijnen, J., Jenkins, M., Green, K., Laloo, F., Sunde, L., Mints, M., Bertario, L., Pineda, M., Navarro, M., Morak, M., Renkonen-Sinisalo, L., Valentin, M., Frayling, I., Plazzer, J., Pylvanainen, K., Genuardi, M., Mecklin, J., Moeslein, G., Sampson, J., Capella, G. and Group, M. (2017b) 'Cancer risk and survival in path_MMR carriers by gene and gender up to 75 years of age: a report from the Prospective Lynch Syndrome Database.', *Gut*.

Morak, M., Heidenreich, B., Keller, G., Hampel, H., Laner, A., de la Chapelle, A. and Holinski-Feder, E. (2014) 'Biallelic MUTYH mutations can mimic Lynch syndrome.', *European Journal of Human Genetics*, 22(11), pp. 1334-7.

Moreira, L., Balaguer, F., Lindor, N., Chapelle, A.d.l., Hampel, H., Aaltonen, L., Hopper, J., Marchand, L.L., Gallinger, S., Newcomb, P., Haile, R., Thibodeau, S., Gunawardena, S., Jenkins, M., Buchanan, D., Potter, J., Baron, J., Ahnen, D., Moreno, V., Andreu, M., Leon, M.P.d., Rustgi, A., Castells, A. and Consortium, E. (2012) 'Identification of Lynch syndrome among patients with colorectal cancer.', *JAMA*, 308(15), pp. 1555-65.

Moreira, L., Muñoz, J., Cuatrecasas, M., Quintanilla, I., Leoz, M., Carballal, S., Ocaña, T., López-Cerón, M., Pellise, M., Castellví-Bel, S., Jover, R., Andreu, M., Carracedo, A., Xicola, R., Llor, X., Boland, C., Goel, A., Castells, A., Balaguer, F. and Association, G.O.G.o.t.S.G. (2015) 'Prevalence of somatic mutl homolog 1 promoter hypermethylation in Lynch syndrome colorectal cancer.', *Cancer*, 121(9), pp. 1395-404.

Mork, M., Borrás, E., Taggart, M., Cuddy, A., Bannon, S., You, Y., Lynch, P., Ramirez, P., Rodriguez-Bigas, M. and Vilar, E. (2016) 'Identification of a novel PMS2 alteration c.505C>G (R169G) in trans with a PMS2 pathogenic mutation in a patient with constitutional mismatch repair deficiency', *Familial Cancer*, 15(4), pp. 587-91.

Movahedi, M., Bishop, D., Macrae, F., Mecklin, J., Moeslein, G., Olschwang, S., Eccles, D., Evans, D., Maher, E., Bertario, L., Bisgaard, M., Dunlop, M., Ho, J., Hodgson, S., Lindblom, A., Lubinski, J., Morrison, P., Murday, V., Ramesar, R., Side, L., Scott, R., Thomas, H., Vasen, H., Burn, J. and Mathers, J. (2015) 'Obesity, Aspirin, and Risk of Colorectal Cancer in Carriers of Hereditary Colorectal Cancer: A Prospective Investigation in the CAPP2 Study.', *Journal of Clinical Oncology*, 33(31), pp. 3591-7.

Murphy, K., Zhang, S., Geiger, T., Hafez, M., Bacher, J., Berg, K. and Eshleman, J. (2006) 'Comparison of the microsatellite instability analysis system and the Bethesda panel for the determination of microsatellite instability in colorectal cancers.', *Journal of Molecular Diagnostics*, 8(3), pp. 305-11.

Muzny, D., Bainbridge, M., Chang, K., Dinh, H., Drummond, J., Fowler, G., Kovar, C., Lewis, L., al, e. and Network, C.G.A. (2012) 'Comprehensive molecular characterization of human colon and rectal cancer.', *Nature*, 487(7407), pp. 330-7.

Mvundura, M., Grosse, S., Hampel, H. and Palomaki, G. (2010) 'The cost-effectiveness of genetic testing strategies for Lynch syndrome among newly diagnosed patients with colorectal cancer.', *Genetic Medicine*, 12(2), pp. 93-104.

Nadir, E., Margalit, H., Gallily, T. and Ben-Sasson, S. (1996) 'Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications', *Proceedings of the National Academy of Sciences USA*, 93(13), pp. 6470-5.

Nakagawa, H., Lockman, J., Frankel, W., Hampel, H., Steenblock, K., Burgart, L., Thibodeau, S. and de la Chapelle, A. (2004) 'Mismatch Repair Gene PMS2', *Cancer Research*, 64(14), pp. 4721-27.

Negrini, S., Gorgoulis, V. and Halazonetis, T. (2010) 'Genomic instability--an evolving hallmark of cancer', *Nature Reviews Molecular Cell Biology*, 11(3), pp. 220-8.

Neveling, K., Mensenkamp, A., Derks, R., Kwint, M., Ouchene, H., Steehouwer, M., van Lier, B., Bosgoed, E., Rikken, A., Tychon, M., Zafeiropoulou, D., Castelein, S., Hehir-Kwa, J., Tjwan Thung, D., Hofste, T., Lelieveld, S., Bertens, S., Adan, I., Eijkelenboom, A., Tops, B., Yntema, H., Stokowy, T., Knappskog, P., Hoberg-Vetti, H., Steen, V., Boyle, E., Martin, B., Ligtenberg,

M., Shendure, J., Nelen, M. and Hoischen, A. (2017) 'BRCA Testing by Single-Molecule Molecular Inversion Probes', *Clinical Chemistry*, 63(2), pp. 503-512.

Newland, A., Kroese, M., Akehurst, R., Bagshaw, J., Chambers, P., Crawford, S., Denton, E., Edwards, S., Fleming, S., Gray, J., Hitchman, J., McGinley, P., Messenger, M., Moseley, A., Naylor, P., Neely, D., Richards, S., Ryan, D., Sculpher, M., Thomas, S., Wierzbicki, A., Latchford, A., Georgiou, D., Laloo, F., Monahan, K., Ilyas, M., Skarrott, P., Glynne-Jones, R., Wallis, Y., Mullaney, B., Walker, T., Byron, S., Albrow, R. and Fernley, R. (2017) 'Molecular testing strategies for Lynch syndrome in people with colorectal cancer (DG27)', *NICE Diagnostics Guidance 27*.

Nicolaides, N., Carter, K., Shell, B., Papadopoulos, N., Vogelstein, B. and Kinzler, K. (1995) 'Genomic organization of the human PMS2 gene family', *Genomics*, 30(2), pp. 195-206.

Nicolaides, N., Papadopoulos, N., Liu, B., Wei, Y., Carter, K., Ruben, S., Rosen, C., Haseltine, W., Fleischmann, R., Fraser, C., Adams, M., Venter, J., Dunlop, M., Hamilton, S., Petersen, G., Chapelle, A.d.l., Vogelstein, B. and Kinzler, K. (1994) 'Mutations of two PMS homologues in hereditary nonpolyposis colon cancer.', *Nature*, 371(6492), pp. 75-80.

Niedzicka, M., Fijarczyk, A., Dudek, K., Stuglik, M. and Babik, W. (2016) 'Molecular Inversion Probes for targeted resequencing in non-model organisms.', *Scientific Reports*, 6, p. 24051.

Nielsen, J., Sahota, R., Milne, K., Kost, S., Nesslinger, N., Watson, P. and Nelson, B. (2012) 'CD20+ tumor-infiltrating lymphocytes have an atypical CD27- memory phenotype and together with CD8+ T cells promote favorable prognosis in ovarian cancer', *Clinical Cancer Research*, 18(12), pp. 3281-92.

Nikiforov, Y., Steward, D., Robinson-Smith, T., Haugen, B., Klopper, J., Zhu, Z., Fagin, J., Falciglia, M., Weber, K. and Nikiforova, M. (2009) 'Molecular testing for mutations in improving the fine-needle aspiration diagnosis of thyroid nodules.', *The Journal of Clinical Endocrinology and Metabolism*, 94(6), pp. 2092-8.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M., Wendl, M. and Ding, L. (2014) 'MSIsensor: microsatellite instability detection using paired tumor-normal sequence data', *Bioinformatics*, 30(7), pp. 1015-6.

Norton, S., Lechner, J., Williams, T. and Fernando, M. (2013) 'A stabilizing reagent prevents cell-free DNA contamination by cellular DNA in plasma during blood sample storage and shipping as determined by digital PCR', *Clinical Biochemistry*, 46(15), pp. 1561-5.

O'Leary, N., Wright, M., Brister, J., Ciufu, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V., Kodali, V., Li, W., Maglott, D., Masterson, P., McGarvey, K., Murphy, M., O'Neill, K., Pujar, S., Rangwala, S., Rausch, D., Riddick, L., Schoch, C., Shkeda, A., Storz, S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R., Vatsan, A., Wallin, C., Webb, D., Wu, W., Landrum, M., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. and Pruitt, K. (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic Acids Research*, 44(D1), pp. D733-45.

O'Roak, B., Vives, L., Fu, W., Egertson, J., Stanaway, I., Phelps, I., Carvill, G., Kumar, A., Lee, C., Ankenman, K., Munson, J., Hiatt, J., Turner, E., Levy, R., O'Day, D., Krumm, N., Coe, B., Martin, B., Borenstein, E., Nickerson, D., Mefford, H., Doherty, D., Akey, J., Bernier, R., Eichler, E. and Shendure, J. (2012) 'Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders.', *Science*, 338(6114), pp. 1619-22.

Oliveira, C., Westra, J., Arango, D., Ollikainen, M., Domingo, E., Ferreira, A., Velho, S., Niessen, R., Lagerstedt, K., Alhopuro, P., Laiho, P., Veiga, I., Teixeira, M., Ligtenberg, M., Kleibeuker, J., Sijmons, R., Plukker, J., Imai, K., Lage, P., Hamelin, R., Albuquerque, C., Schwartz, S.J., Lindblom, A., Peltomaki, P., Yamamoto, H., Aaltonen, L., Seruca, R. and Hofstra, R. (2004) 'Distinct patterns of KRAS mutations in colorectal carcinomas according to germline mismatch repair defects and hMLH1 methylation status.', *Human Molecular Genetics*, 13(19), pp. 2303-11.

Ostergaard, J., Sunde, L. and Okkels, H. (2005) 'Neurofibromatosis von Recklinghausen type I phenotype and early onset of cancers in siblings compound heterozygous for mutations in MSH6', *American Journal of Medical Genetics*, 139a(2), pp. 96-105; discussion 96.

Padua, D. and Massagué, J. (2009) 'Roles of TGFbeta in metastasis.', *Cell Research*, 19(1), pp. 89-102.

Pal, T., Permeth-Wey, J. and Sellers, T. (2008) 'A review of the clinical relevance of mismatch-repair deficiency in ovarian cancer.', *Cancer*, 113(4), pp. 733-42.

Papadopoulos, N., Nicolaidis, N., Wei, Y., Ruben, S., Carter, K., Rosen, C., Haseltine, W., Fleischmann, R., Fraser, C., Adams, M., Venter, J., Hamilton, S., Petersen, G., Watson, P., Lynch, H., Peltomaki, P., Mecklin, J., de la Chapelle, A., Kinzler, K. and Vogelstein, B. (1994) 'Mutation of a mutL homolog in hereditary colon cancer.', *Science*, 263(5153), pp. 1625-9.

Pardoll, D. (2012) 'The blockade of immune checkpoints in cancer immunotherapy.', *Nature Reviews Cancer*, 12(4), pp. 252-64.

Parry, S., Win, A., Parry, B., Macrae, F., Gurrin, L., Church, J., Baron, J., Giles, G., Leggett, B., Winship, I., Lipton, L., Young, G., Young, J., Lodge, C., Southey, M., Newcomb, P., Marchand, L.L., Haile, R., Lindor, N., Gallinger, S., Hopper, J. and Jenkins, M. (2011) 'Metachronous colorectal cancer risk for mismatch repair gene mutation carriers: the advantage of more extensive colon surgery.', *Gut*, 60(7), pp. 950-7.

Parsons, M., Buchanan, D., Thompson, B., Young, J. and Spurdle, A. (2012) 'Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification.', *Journal of Medical Genetics*, 49(3), pp. 151-7.

Parsons, R., Li, G., Longley, M., Modrich, P., Liu, B., Berk, T., Hamilton, S., Kinzler, K. and Vogelstein, B. (1995) 'Mismatch repair deficiency in phenotypically normal human cells.', *Science*, 268(5211), pp. 738-40.

Paschen, A., Mendez, R., Jimenez, P., Sucker, A., Ruiz-Cabello, F., Song, M., Garrido, F. and Schadendorf, D. (2003) 'Complete loss of HLA class I antigen expression on melanoma cells: a result of successive mutational events', *International Journal of Cancer*, 103(6), pp. 759-67.

Patel, S., Sanjana, N., Kishton, R., Eidizadeh, A., Vodnala, S., Cam, M., Gartner, J., Jia, L., Steinberg, S., Yamamoto, T., Merchant, A., Mehta, G., Chichura, A., Shalem, O., Tran, E., Eil, R., Sukumar, M., Guijarro, E., Day, C., Robbins, P., Feldman, S., Merlino, G., Zhang, F. and Restifo, N. (2017) 'Identification of essential genes for cancer immunotherapy', *Nature*, 548(7669), pp. 537-542.

Pereira-Faca, S., Kuick, R., Puravs, E., Zhang, Q., Krasnoselsky, A., Phanstiel, D., Qiu, J., Misek, D., Hinderer, R., Tammemagi, M., Landi, M., Caporaso, N., Pfeiffer, R., Edelstein, C., Goodman, G., Barnett, M., Thornquist, M., Brenner, D. and Hanash, S. (2007) 'Identification of 14-3-3 theta as an antigen that induces a humoral response in lung cancer', *Cancer Research*, 67(24), pp. 12000-6.

Pérez-Carbonell, L., Alenda, C., Payá, A., Castillejo, A., Barberá, V., Guillén, C., Rojas, E., Acame, N., Gutiérrez-Aviñó, F., Castells, A., Llor, X., Andreu, M., Soto, J. and Jover, R. (2010) 'Methylation analysis of MLH1 improves the selection of patients for genetic testing in Lynch syndrome.', *Journal of Molecular Diagnostics*, 12(4), pp. 498-504.

Pérez-Carbonell, L., Ruiz-Ponte, C., Guarinos, C., Alenda, C., Payá, A., Brea, A., Egoavil, C., Castillejo, A., Barberá, V., Bessa, X., Xicola, R., Rodríguez-Soler, M., Sánchez-Fortún, C., Acame, N., Castellví-Bel, S., Piñol, V., Balaguer, F., Bujanda, L., De-Castro, M., Llor, X., Andreu, M., Carracedo, A., Soto, J., Castells, A. and Jover, R. (2012) 'Comparison between universal molecular screening for Lynch syndrome and revised Bethesda guidelines in a large population-based cohort of patients with colorectal cancer.', *Gut*, 61(6), pp. 865-72.

Perkins, S., Rubin, J., Leonard, J., Smyth, M., El Naqa, I., Michalski, J., Simpson, J., Limbrick, D., Park, T. and Mansur, D. (2011) 'Glioblastoma in children: a single-institution experience', *International Journal of Radiation Oncology, Biology, Physics*, 80(4), pp. 1117-21.

Peron, S., Metin, A., Gardes, P., Alyanakian, M., Sheridan, E., Kratz, C., Fischer, A. and Durandy, A. (2008) 'Human PMS2 deficiency is associated with impaired immunoglobulin class switch recombination', *Journal of Experimental Medicine*, 205(11), pp. 2465-72.

Phipps, A., Buchanan, D., Makar, K., Win, A., Baron, J., Lindor, N., Potter, J. and Newcomb, P. (2013) 'KRAS-mutation status in relation to colorectal cancer survival: the joint impact of correlated tumour markers.', *British Journal of Cancer*, 108(8), pp. 1757-64.

Piñol, V., Castells, A., Andreu, M., Castellví-Bel, S., Alenda, C., Llor, X., Xicola, R., Rodríguez-Moranta, F., Payá, A., Jover, R., Bessa, X. and Association., G.O.G.o.t.S.G. (2005) 'Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer.', *JAMA*, 293(16), pp. 1986-94.

- Pollard, T. (2010) 'A Guide to Simple and Informative Binding Assays', *Molecular Biology of the Cell*, 21(23), pp. 4061-4067.
- Popat, S., Hubner, R. and Houlston, R. (2005) 'Systematic review of microsatellite instability and colorectal cancer prognosis', *Journal of Clinical Oncology*, 23(3), pp. 609-18.
- Postow, M., Sidlow, R. and Hellmann, M. (2018) 'Immune-Related Adverse Events Associated with Immune Checkpoint Blockade', *New England Journal of Medicine*, 378(2), pp. 158-168.
- Quackenbush, J. (2002) 'Microarray data normalization and transformation.', *Nature Genetics*, 32(Suppl), pp. 496-501.
- Rajagopalan, H., Bardelli, A., Lengauer, C., Kinzler, K., Vogelstein, B. and Velculescu, V. (2002) 'Tumorigenesis: RAF/RAS oncogenes and mismatch-repair status.', *Nature*, 418(6901), p. 934.
- Rampino, N., Yamamoto, H., Ionov, Y., Li, Y., Sawai, H., Reed, J. and Peruchó, M. (1997) 'Somatic frameshift mutations in the BAX gene in colon cancers of the microsatellite mutator phenotype.', *Science*, 275(5302), pp. 967-9.
- Ray, P., Manach, Y.L., Riou, B. and Houle, T. (2010) 'Statistical evaluation of a biomarker.', *Anesthesiology*, 112(4), pp. 1023-40.
- Rebbeck, T., Burns-White, K., Chan, A., Emmons, K., Freedman, M., Hunter, D., Kraft, P., Laden, F., Mucci, L., Parmigiani, G., Schrag, D., Syngal, S., Tamimi, R., Viswanath, K., Yurgelun, M. and Garber, J. (2018) 'Precision Prevention and Early Detection of Cancer: Fundamental Principles', *Cancer Discovery*, 8(7), pp. 803-811.
- Redford, L., Alhilal, G., Needham, S., O'Brien, O., Coaker, J., Tyson, J., Amorim, L., Middleton, I., Izuogu, O., Arends, M., Oniscu, A., Alonso, Á., Laguna, S., Gallon, R., Sheth, H., Santibanez-Koref, M., Jackson, M. and Burn, J. (2018). 'A novel panel of short mononucleotide repeats linked to informative polymorphisms enabling effective high volume low cost discrimination between mismatch repair deficient and proficient tumours.', *PLoS One*, 13(8), e0203052.
- Reese, J., Liu, L. and Gerson, S. (2003) 'Repopulating defect of mismatch repair-deficient hematopoietic stem cells.', *Blood*, 102(5), pp. 1626-33.

Reumkens, A., Rondagh, E., Bakker, C., Winkens, B., Masclee, A. and Sanduleanu, S. (2016) 'Post-Colonoscopy Complications: A Systematic Review, Time Trends, and Meta-Analysis of Population-Based Studies', *American Journal of Gastroenterology*, 111(8), pp. 1092-101.

Reuschenbach, M., Dörre, J., Waterboer, T., Kopitz, J., Schneider, M., Hoogerbrugge, N., Jäger, E., Kloor, M. and Doeberitz, M.v.K. (2014) 'A multiplex method for the detection of serum antibodies against in silico-predicted tumor antigens', *Cancer Immunology, Immunotherapy*, 63(12), pp. 1251-9.

Reuschenbach, M., Kloor, M., Morak, M., Wentzensen, N., Germann, A., Garbe, Y., Tariverdian, M., Findeisen, P., Neumaier, M., Holinski-Feder, E. and Doeberitz, M.v.K. (2010) 'Serum antibodies against frameshift peptides in microsatellite unstable colorectal cancer patients with Lynch syndrome.', *Familial Cancer*, 9(2), pp. 173-9.

Rhees, J., Arnold, M. and Boland, C. (2014) 'Inversion of exons 1-7 of the MSH2 gene is a frequent cause of unexplained Lynch syndrome in one local population.', *Familial Cancer*, 13(2), pp. 219-25.

Ribic, C., Sargent, D., Moore, M., Thibodeau, S., French, A., Goldberg, R., Hamilton, S., Laurent-Puig, P., Gryfe, R., Shepherd, L., Tu, D., Redston, M. and Gallinger, S. (2003) 'Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer.', *New England Journal of Medicine*, 349(3), pp. 247-57.

Ricciardone, M., Ozcelik, T., Cevher, B., Ozdag, H., Tuncer, M., Gurgey, A., Uzunalimoglu, O., Cetinkaya, H., Tanyeli, A., Erken, E. and Ozturk, M. (1999) 'Human MLH1 deficiency predisposes to hematological malignancy and neurofibromatosis type 1', *Cancer Res*, 59(2), pp. 290-3.

Ripperger, T., Beger, C., Rahner, N., Sykora, K., Bockmeyer, C., Lehmann, U., Kreipe, H. and Schlegelberger, B. (2010) 'Constitutional mismatch repair deficiency and childhood leukemia/lymphoma--report on a novel biallelic MSH6 mutation.', *Haematologica*, 95(5), pp. 841-4.

Robinson, W., DiGennaro, C., Hueber, W., Haab, B., Kamachi, M., Dean, E., Fournel, S., Fong, D., Genovese, M., de Vegvar, H., Skriver, K., Hirschberg, D., Morris, R., Muller, S., Pruijn, G., van Venrooij, W., Smolen, J., Brown, P., Steinman, L. and Utz, P. (2002) 'Autoantigen

microarrays for multiplex characterization of autoantibody responses', *Nature Medicine*, 8(3), pp. 295-301.

Rodriguez-Bigas, M., Boland, C., Hamilton, S., Henson, D., Jass, J., Khan, P., Lynch, H., Perucho, M., Smyrk, T., Sobin, L. and Srivastava, S. (1997) 'A National Cancer Institute Workshop on Hereditary Nonpolyposis Colorectal Cancer Syndrome: meeting highlights and Bethesda guidelines.', *Journal of the National Cancer Institute*, 89(23), pp. 1758-62.

Rodríguez-Soler, M., Pérez-Carbonell, L., Guarinos, C., Zapater, P., Castillejo, A., Barberá, V., Juárez, M., Bessa, X., Xicola, R., Clofent, J., Bujanda, L., Balaguer, F., Reñé, J., de-Castro, L., Marín-Gabriel, J., Lanas, A., Cubiella, J., Nicolás-Pérez, D., Brea-Fernández, A., Castellví-Bel, S., Alenda, C., Ruiz-Ponte, C., Carracedo, A., Castells, A., Andreu, M., Llor, X., Soto, J., Payá, A. and Jover, R. (2013) 'Risk of cancer in cases of suspected lynch syndrome without germline mutation.', *Gastroenterology*, 144(5), pp. 926-932.

Roeder, I., Horn, K., Sieburg, H., Cho, R., Muller-Sieburg, C. and Loeffler, M. (2008) 'Characterization and quantification of clonal heterogeneity among hematopoietic stem cells: a model-based approach', *Blood*, 112(13), pp. 4874-4883.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M., Chresta, C., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L., Bernards, R. and Simon, I. (2014) 'Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition', *International Journal of Cancer*, 134(3), pp. 552-62.

Rutjes, A., Reitsma, J., Vandenbroucke, J., Glas, A. and Bossuyt, P. (2005) 'Case-control and two-gate designs in diagnostic accuracy studies', *Clinical Chemistry*, 51(8), pp. 1335-41.

Sadanandam, A., Lyssiotis, C., Homicsko, K., Collisson, E., Gibb, W., Wullschleger, S., Ostos, L., Lannon, W., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A., Wiedenmann, B., Cantley, L., Gray, J. and Hanahan, D. (2013) 'A colorectal cancer classification system that associates cellular phenotype and responses to therapy', *Nature Medicine*, 19(5), pp. 619-25.

Sade-Feldman, M., Jiao, Y., Chen, J., Rooney, M., Barzily-Rokni, M., Eliane, J., Bjorgaard, S., Hammond, M., Vitzthum, H., Blackmon, S., Frederick, D., Hazar-Rethinam, M., Nadres, B., Van Seventer, E., Shukla, S., Yizhak, K., Ray, J., Rosebrock, D., Livitz, D., Adalsteinsson, V.,

Getz, G., Duncan, L., Li, B., Corcoran, R., Lawrence, D., Stemmer-Rachamimov, A., Boland, G., Landau, D., Flaherty, K., Sullivan, R. and Hacohen, N. (2017) 'Resistance to checkpoint blockade therapy through inactivation of antigen presentation', *Nature Communications*, 8(1), p. 1136.

Saeterdal, I., Bjørheim, J., Lislud, K., Gjertsen, M., Bukholm, I., Olsen, O., Nesland, J., Eriksen, J., Møller, M., Lindblom, A. and Gaudernack, G. (2001) 'Frameshift-mutation-derived peptides as tumor-specific antigens in inherited and spontaneous colorectal cancer.', *Proceedings of the National Academy of Sciences of the United States of America*, 98(23), pp. 13255-60.

Sahin, I., Kazmi, S., Yorio, J., Bhadkamkar, N., Kee, B. and Garrett, C. (2013) 'Rare Though Not Mutually Exclusive: A Report of Three Cases of Concomitant KRAS and BRAF Mutation and a Review of the Literature.', *Journal of Cancer*, 4(4), pp. 320-2.

Salahshor, S., Koelble, K., Rubio, C. and Lindblom, A. (2001) 'Microsatellite Instability and hMLH1 and hMSH2 expression analysis in familial and sporadic colorectal cancer.', *Laboratory Investigation*, 81(4), pp. 535-41.

Salipante, S., Scroggins, S., Hampel, H., Turner, E. and Pritchard, C. (2014) 'Microsatellite instability detection by next generation sequencing.', *Clinical Chemistry*, 60(9), pp. 1192-9.

Sargent, D., Marsoni, S., Monges, G., Thibodeau, S., Labianca, R., Hamilton, S., French, A., Kabat, B., Foster, N., Torri, V., Ribic, C., Grothey, A., Moore, M., Zaniboni, A., Seitz, J., Sinicrope, F. and Gallinger, S. (2010) 'Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer.', *Journal of Clinical Oncology*, 28(20), pp. 3219-26.

Sarvaria, A., Madrigal, J. and Saudemont, A. (2017) 'B cell regulation in cancer and anti-tumor immunity', *Cellular and Molecular Immunology*, 14(8), pp. 662-674.

Schilsky, R. (2010) 'Personalized medicine in oncology: the future is now', *Nature Reviews Drug Discovery*, 9(5), pp. 363-6.

Schlotterer, C. and Tautz, D. (1992) 'Slippage synthesis of simple sequence DNA', *Nucleic Acids Research*, 20(2), pp. 211-5.

Schmitt, M., Kennedy, S., Salk, J., Fox, E., Hiatt, J. and Loeb, L. (2012) 'Detection of ultra-rare mutations by next-generation sequencing.', *Proceedings of the National Academy of Sciences of the United States of America*, 109(36), pp. 14508-13.

Scholer, L., Reinert, T., Orntoft, M., Kassentoft, C., Arnadottir, S., Vang, S., Nordentoft, I., Knudsen, M., Lamy, P., Andreasen, D., Mortensen, F., Knudsen, A., Stribolt, K., Sivesgaard, K., Mouritzen, P., Nielsen, H., Laurberg, S., Orntoft, T. and Andersen, C. (2017) 'Clinical Implications of Monitoring Circulating Tumor DNA in Patients with Colorectal Cancer', *Clinical Cancer Research*, 23(18), pp. 5437-5445.

Schwitalle, Y., Kloor, M., Eiermann, S., Linnebacher, M., Kienle, P., Knaebel, H., Tariverdian, M., Benner, A. and Doeberitz, M.v.K. (2008) 'Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers.', *Gastroenterology*, 134(4), pp. 988-97.

Sekine, S., Mori, T., Ogawa, R., Tanaka, M., Yoshida, H., Taniguchi, H., Nakajima, T., Sugano, K., Yoshida, T., Kato, M., Furukawa, E., Ochiai, A. and Hiraoka, N. (2017) 'Mismatch repair deficiency commonly precedes adenoma formation in Lynch Syndrome-Associated colorectal tumorigenesis.', *Modern Pathology*, 30(8), pp. 1144-51.

Seppala, T., Pylvanainen, K., Evans, D., Jarvinen, H., Renkonen-Sinisalo, L., Bernstein, I., Holinski-Feder, E., Sala, P., Lindblom, A., Macrae, F., Blanco, I., Sijmons, R., Jeffries, J., Vasen, H., Burn, J., Nakken, S., Hovig, E., Rodland, E., Tharmaratnam, K., de Vos Tot Nederveen Cappel, W., Hill, J., Wijnen, J., Jenkins, M., Genuardi, M., Green, K., Laloo, F., Sunde, L., Mints, M., Bertario, L., Pineda, M., Navarro, M., Morak, M., Frayling, I., Plazzer, J., Sampson, J., Capella, G., Moslein, G., Mecklin, J. and Moller, P. (2017) 'Colorectal cancer incidence in path_MLH1 carriers subjected to different follow-up protocols: a Prospective Lynch Syndrome Database report', *Hereditary Cancer in Clinical Practice*, 15, p. 18.

Shaikh, T., Handorf, E., Meyer, J., Hall, M. and Esnaola, N. (2018) 'Mismatch Repair Deficiency Testing in Patients With Colorectal Cancer and Nonadherence to Testing Guidelines in Young Adults.', *JAMA Oncology*, 4(2), p. e173580.

Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature Biotechnology*, 26(10), pp. 1135-45.

Shia, J. (2008) 'Immunohistochemistry versus Microsatellite Instability Testing For Screening Colorectal Cancer Patients at Risk For Hereditary Nonpolyposis Colorectal Cancer Syndrome. Part I. The Utility of Immunohistochemistry.', *Journal of Molecular Diagnostics*, 10(4), pp. 293-300.

Shlien, A., Campbell, B., de Borja, R., Alexandrov, L., Merico, D., Wedge, D., Van Loo, P., Tarpey, P., Coupland, P., Behjati, S., Pollett, A., Lipman, T., Heidari, A., Deshmukh, S., Avitzur, N., Meier, B., Gerstung, M., Hong, Y., Merino, D., Ramakrishna, M., Remke, M., Arnold, R., Panigrahi, G., Thakkar, N., Hodel, K., Henninger, E., Goksenin, A., Bakry, D., Charames, G., Druker, H., Lerner-Ellis, J., Mistry, M., Dvir, R., Grant, R., Elhasid, R., Farah, R., Taylor, G., Nathan, P., Alexander, S., Ben-Shachar, S., Ling, S., Gallinger, S., Constantini, S., Dirks, P., Huang, A., Scherer, S., Grundy, R., Durno, C., Aronson, M., Gartner, A., Meyn, M., Taylor, M., Pursell, Z., Pearson, C., Malkin, D., Futreal, P., Stratton, M., Bouffet, E., Hawkins, C., Campbell, P. and Tabori, U. (2015) 'Combined hereditary and somatic mutations of replication error repair genes result in rapid onset of ultra-hypermuted cancers', *Nature Genetics*, 47(3), pp. 257-62.

Shu, Y., Wu, X., Tong, X., Wang, X., Chang, Z., Mao, Y., Chen, X., Sun, J., Wang, Z., Hong, Z., Zhu, L., Zhu, C., Chen, J., Liang, Y., Shao, H. and Shao, Y. (2017) 'Circulating Tumor DNA Mutation Profiling by Targeted Next Generation Sequencing Provides Guidance for Personalized Treatments in Multiple Cancer Types', *Scientific Reports*, 7(1), p. 583.

Sia, E., Kokoska, R., Dominska, M., Greenwell, P. and Petes, T. (1997) 'Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes', *Molecular and Cellular Biology*, 17(5), pp. 2851-8.

Siegel, R., Miller, K. and Jemal, A. (2017) 'Cancer statistics, 2017.', *CA: a Cancer Journal for Clinicians*, 67, pp. 7-30.

Simpson, A., Caballero, O. and Pena, S. (2001) 'Microsatellite instability as a tool for the classification of gastric cancer.', *Trends in Molecular Medicine*, 7(2), pp. 76-80.

Siniluoto, T., Paivansalo, M. and Alavaikko, M. (1991) 'Ultrasonography of spleen and liver in staging Hodgkin's disease', *European Journal of Radiology*, 13(3), pp. 181-6.

Sint, D., Raso, L. and Traugott, M. (2012) 'Advances in multiplex PCR: balancing primer efficiencies and improving detection success', *Methods in Ecology and Evolution*, 3(5), pp. 898-905.

Siravegna, G., Mussolin, B., Buscarino, M., Corti, G., Cassingena, A., Crisafulli, G., Ponzetti, A., Cremolini, C., Amatu, A., Lauricella, C., Lamba, S., Hobor, S., Avallone, A., Valtorta, E., Rospo, G., Medico, E., Motta, V., Antoniotti, C., Tatangelo, F., Bellosillo, B., Veronese, S., Budillon, A., Montagut, C., Racca, P., Marsoni, S., Falcone, A., Corcoran, R., Di Nicolantonio, F., Loupakis, F., Siena, S., Sartore-Bianchi, A. and Bardelli, A. (2015) 'Monitoring clonal evolution and resistance to EGFR blockade in the blood of metastatic colorectal cancer patients', *Nature Medicine*, 21(7), pp. 795-801.

Sjursen, W., Bjornevoll, I., Engebretsen, L., Fjelland, K., Halvorsen, T. and Myrvold, H. (2009) 'A homozygote splice site PMS2 mutation as cause of Turcot syndrome gives rise to two different abnormal transcripts', *Familial Cancer*, 8(3), pp. 179-86.

Sjursen, W., Haukanes, B., Grindedal, E., Aarset, H., Stormorken, A., Engebretsen, L., Jonsrud, C., Bjørnevoll, I., Andresen, P., Ariansen, S., Lavik, L., Gilde, B., Bowitz-Lothe, I., Maehle, L. and Møller, P. (2010) 'Current clinical criteria for Lynch syndrome are not sensitive enough to identify MSH6 mutation carriers.', *Journal of Medical Genetics*, 47(9), pp. 579-85.

Smits, A., Kummer, J., de Bruin, P., Bol, M., van den Tweel, J., Seldenrijk, K., Willems, S., Offerhaus, G., de Weger, R., van Diest, P. and Vink, A. (2014) 'The estimation of tumor cell percentage for molecular testing by pathologists is not accurate', *Modern Pathology*, 27(2), pp. 168-74.

Snowsill, T., Huxley, N., Hoyle, M., Jones-Hughes, T., Coelho, H., Cooper, C., Frayling, I. and Hyde, C. (2014) 'A systematic review and economic evaluation of diagnostic strategies for Lynch syndrome.', *Health Technology Assessment*, 18(58), pp. 1-406.

Sokol, L., Koelzer, V., Rau, T., Karamitopoulou, E., Zlobec, I. and Lugli, A. (2015) 'Loss of tapasin correlates with diminished CD8(+) T-cell immunity and prognosis in colorectal cancer', *Journal of Translational Medicine*, 13, p. 279.

Soussi, T. (2000) 'p53 Antibodies in the sera of patients with various types of cancer: a review', *Cancer Research*, 60(7), pp. 1777-88.

Southey, M., Jenkins, M., Mead, L., Whitty, J., Trivett, M., Tesoriero, A., Smith, L., Jennings, K., Grubb, G., Royce, S., Walsh, M., Barker, M., Young, J., Jass, J., John, D.S., Macrae, F., Giles, G. and Hopper, J. (2005) 'Use of molecular tumor characteristics to prioritize mismatch repair gene testing in early-onset colorectal cancer.', *Journal of Clinical Oncology*, 23(27), pp. 6524-32.

Spencer, D., Tyagi, M., Vallania, F., Bredemeyer, A., Pfeifer, J., Mitra, R. and Duncavage, E. (2014) 'Performance of common analysis methods for detecting low-frequency single nucleotide variants in targeted next-generation sequence data', *Journal of Molecular Diagnostics*, 16(1), pp. 75-88.

Staffa, L., Echterdiek, F., Nelius, N., Benner, A., Werft, W., Lahrmann, B., Grabe, N., Schneider, M., Tariverdian, M., von Knebel Doeberitz, M., Bläker, H. and Kloor, M. (2015) 'Mismatch repair-deficient crypt foci in Lynch syndrome--molecular alterations and association with clinical parameters.', *PLoS One*, 10(3), p. e0121980.

Ståhlberg, A., Krzyzanowski, P., Jackson, J., Egyud, M., Stein, L. and Godfrey, T. (2016) 'Simple, multiplexed, PCR-based barcoding of DNA enables sensitive mutation detection in liquid biopsies using sequencing.', *Nucleic Acids Research*, 44(11), p. e105.

Cancer Research UK Statistics (2015) 'Bowel Cancer Statistics', <https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer>.

Stoffel, E., Mangu, P., Gruber, S., Hamilton, S., Kalady, M., Lau, M., Lu, K., Roach, N. and Limburg, P. (2015) 'Hereditary colorectal cancer syndromes: American Society of Clinical Oncology Clinical Practice Guideline endorsement of the familial risk-colorectal cancer: European Society for Medical Oncology Clinical Practice Guidelines', *Journal of Clinical Oncology*, 33(2), pp. 209-17.

Strand, M., Prolla, T., Liskay, R. and Petes, T. (1993) 'Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair.', *Nature*, 365(6443), pp. 274-6.

Stuckless, S., Green, J., Morgenstern, M., Kennedy, C., Green, R., Woods, M., Fitzgerald, W., Cox, J. and Parfrey, P. (2012) 'Impact of colonoscopic screening in male and female Lynch syndrome carriers with an MSH2 mutation.', *Clinical Genetics*, 82(5), pp. 439-45.

Su, B., Shi, H. and Wan, J. (2012) 'Role of serum carcinoembryonic antigen in the detection of colorectal cancer before and after surgical resection.', *World Journal of Gastroenterology*, 18(17), pp. 2121-6.

Subramanian, S., Mishra, R. and Singh, L. (2003) 'Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions.', *Genome Biology*, 4(2), p. R13.

Suraweera, N., Duval, A., Reperant, M., Vaury, C., Furlan, D., Leroy, K., Seruca, R., Iacopetta, B. and Hamelin, R. (2002) 'Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR', *Gastroenterology*, 123(6), pp. 1804-11.

Surmann, E., Voigt, A., Michel, S., Bauer, K., Reuschenbach, M., Ferrone, S., von Knebel Doeberitz, M. and 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(3), pp. 357-66.

Susanti, S., Fadhil, W., Ebili, H., Asiri, A., Nestarenkaite, A., Hadjimichael, E., Ham-Karim, H., Field, J., Stafford, K., Matharoo-Ball, B., Hassall, J., Sharif, A., Oniscu, A. and Ilyas, M. (2018) 'N_LyST: a simple and rapid screening test for Lynch syndrome.', *Journal of Clinical Pathology*.

Suter, C., Martin, D. and Ward, R. (2004) 'Germline epimutation of MLH1 in individuals with multiple cancers.', *Nature Genetics*, 36(5), pp. 497-501.

Tafe, L., Riggs, E. and Tsongalis, G. (2014) 'Lynch syndrome presenting as endometrial cancer', *Clinical Chemistry*, 60(1), pp. 111-21.

Taly, V., Pekin, D., Benhaim, L., Kotsopoulos, S., Corre, D.L., Li, X., Atochin, I., Link, D., Griffiths, A., Pallier, K., Blons, H., Bouché, O., Landi, B., Hutchison, J. and Laurent-Puig, P.

(2013) 'Multiplex picodroplet digital PCR to detect KRAS mutations in circulating DNA from the plasma of colorectal cancer patients.', *Clinical Chemistry*, 59(12), pp. 1722-31.

Ten Broeke, S., van der Klift, H., Tops, C., Aretz, S., Bernstein, I., Buchanan, D., de la Chapelle, A., Capella, G., Clendenning, M., Engel, C., Gallinger, S., Gomez Garcia, E., Figueiredo, J., Haile, R., Hampel, H., Hopper, J., Hoogerbrugge, N., von Knebel Doeberitz, M., Le Marchand, L., Letteboer, T., Jenkins, M., Lindblom, A., Lindor, N., Mensenkamp, A., Møller, P., Newcomb, P., van Os, T., Pearlman, R., Pineda, M., Rahner, N., Redeker, E., Olderode-Berends, M., Rosty, C., Schackert, H., Scott, R., Senter, L., Spruijt, L., Steinke-Lange, V., Suerink, M., Thibodeau, S., Vos, Y., Wagner, A., Winship, I., Hes, F., Vasen, H., Wijnen, J., Nielsen, M. and Win, A. (2018) 'Cancer Risks for PMS2-Associated Lynch Syndrome.', *Journal of Clinical Oncology*, 36(29), pp. 2961-8.

Terdiman, J., Gum, J.J., Conrad, P., Miller, G., Weinberg, V., Crawley, S., Levin, T., Reeves, C., Schmitt, A., Hepburn, M., Sleisenger, M. and Kim, Y. (2001) 'Efficient detection of hereditary nonpolyposis colorectal cancer gene carriers by screening for tumor microsatellite instability before germline genetic testing.', *Gastroenterology*, 120(1), pp. 21-30.

Thibodeau, S., Bren, G. and Schaid, D. (1993) 'Microsatellite instability in cancer of the proximal colon', *Science*, 260(5109), pp. 816-9.

Thibodeau, S., French, A., Cunningham, J., Tester, D., Burgart, L., Roche, P., McDonnell, S., Schaid, D., Vockley, C., Michels, V., Farr, G.J. and O'Connell, M. (1998) 'Microsatellite instability in colorectal cancer: different mutator phenotypes and the principal involvement of hMLH1.', *Cancer Research*, 58(8), pp. 1713-8.

Thibodeau, S., French, A., Roche, P., Cunningham, J., Tester, D., Lindor, N., Moslein, G., Baker, S., Liskay, R., Burgart, L., Honchel, R. and Halling, K. (1996) 'Altered expression of hMSH2 and hMLH1 in tumors with microsatellite instability and genetic alterations in mismatch repair genes.', *Cancer Research*, 56(21), pp. 4836-40.

Thomas, A., Tanaka, M., Trepel, J., Reinhold, W., Rajapakse, V. and Pommier, Y. (2017) 'Temozolomide in the era of precision medicine', *Cancer Research*, 77(4), pp. 823-826.

Thomson, J., Itskovitz-Eldor, J., Shapiro, S., Waknitz, M., Swiergiel, J., Marshall, V. and Jones, J. (1998) 'Embryonic stem cell lines derived from human blastocysts.', *Science*, 282(5391), pp. 1145-7.

Toledano, H., Goldberg, Y., Kedar-Barnes, I., Baris, H., Porat, R., Shochat, C., Bercovich, D., Pikarsky, E., Lerer, I., Yaniv, I., Abeliovich, D. and Peretz, T. (2009) 'Homozygosity of MSH2 c.1906G-->C germline mutation is associated with childhood colon cancer, astrocytoma and signs of Neurofibromatosis type I', *Familial Cancer*, 8(3), pp. 187-94.

Topalian, S., Taube, J., Anders, R. and Pardoll, D. (2016) 'Mechanism-driven biomarkers to guide immune checkpoint blockade in cancer therapy', *Nature Reviews Cancer*, 16(5), pp. 275-87.

Tougeron, D., Fauquembergue, E., Rouquette, A., Le Pessot, F., Sesboüé, R., Laurent, M., Berthet, P., Mauillon, J., Di Fiore, F., Sabourin, J., Michel, P., Tosi, M., Frébourg, T. and Latouche, J. (2009) 'Tumor-infiltrating lymphocytes in colorectal cancers with microsatellite instability are correlated with the number and spectrum of frameshift mutations.', *Modern Pathology*, 22(9), pp. 1186-95.

Turcot, J., Despres, J. and St Pierre, F. (1959) 'Malignant tumors of the central nervous system associated with familial polyposis of the colon: report of two cases', *Diseases of the Colon and Rectum*, 2, pp. 465-8.

Umar, A., Boland, C., Terdiman, J., Syngal, S., Chapelle, A.d.l., Rüschoff, J., Fishel, R., Lindor, N., Burgart, L., Hamelin, R., Hamilton, S., Hiatt, R., Jass, J., Lindblom, A., Lynch, H., Peltomaki, P., Ramsey, S., Rodriguez-Bigas, M., Vasen, H., Hawk, E., Barrett, J., Freedman, A. and Srivastava, S. (2004) 'Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability.', *Journal of the National Cancer Institute*, 96(4), pp. 261-8.

Umar, A., Koi, M., Risinger, J., Glaab, W., Tindall, K., Kolodner, R., Boland, C., Barrett, J. and Kunkel, T. (1997) 'Correction of hypermutability, N-methyl-N'-nitro-N-nitrosoguanidine resistance, and defective DNA mismatch repair by introducing chromosome 2 into human tumor cells with mutations in MSH2 and MSH6.', *Cancer Research*, 57(18), pp. 3949-55.

Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B., Remm, M. and Rozen, S. (2012) 'Primer3—new capabilities and interfaces', *Nucleic Acids Research*, 40(15), pp. e115-e115.

Urganci, N., Genc, D., Kose, G., Onal, Z. and Vidin, O. (2015) 'Colorectal Cancer due to Constitutional Mismatch Repair Deficiency Mimicking Neurofibromatosis I', *Pediatrics*, 136(4), pp. e1047-50.

van der Post, R., Kiemeneij, L., Ligtenberg, M., Witjes, J., Hulsbergen-van de Kaa, C., Bodmer, D., Schaap, L., Kets, C., van Krieken, J. and Hoogerbrugge, N. (2010) 'Risk of urothelial bladder cancer in Lynch syndrome is increased, in particular among MSH2 mutation carriers', *Journal of Medical Genetics*, 47(7), pp. 464-470.

van Lier, M., Leenen, C., Wagner, A., Ramsoekh, D., Dubbink, H., van den Ouweland, A., Westenend, P., de Graaf, E., Wolters, L., Vrijland, W., Kuipers, E., van Leerdam, M., Steyerberg, E., Dinjens, W. and Group, L.S. (2012) 'Yield of routine molecular analyses in colorectal cancer patients ≤ 70 years to detect underlying Lynch syndrome.', *Journal of Pathology*, 226(5), pp. 764-74.

Vasen, H., Blanco, I., Aktan-Collan, K., Gopie, J., Alonso, A., Aretz, S., Bernstein, I., Bertario, L., Burn, J., Capella, G., Colas, C., Engel, C., Frayling, I., Genuardi, M., Heinimann, K., Hes, F., Hodgson, S., Karagiannis, J., Laloo, F., Lindblom, A., Mecklin, J., Møller, P., Myrhoj, T., Nagengast, F., Parc, Y., Leon, M.P.d., Renkonen-Sinisalo, L., Sampson, J., Stormorken, A., Sijmons, R., Tejpar, S., Thomas, H., Rahner, N., Wijnen, J., Järvinen, H., Möslin, G. and group, M. (2013) 'Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts.', *Gut*, 62(6), pp. 812-23.

Vasen, H., Ghorbanoghli, Z., Bourdeaut, F., Cabaret, O., Caron, O., Duval, A., Entz-Werle, N., Goldberg, Y., Ilencikova, D., Kratz, C., Lavoine, N., Loeffen, J., Menko, F., Muleris, M., Sebille, G., Colas, C., Burkhardt, B., Brugieres, L., Wimmer, K. and (C4CMMR-D), E.-C.C.f.C.-D. (2014) 'Guidelines for surveillance of individuals with constitutional mismatch repair-deficiency proposed by the European Consortium "Care for CMMR-D" (C4CMMR-D).', *Journal of Medical Genetics*, 51(5), pp. 283-93.

Vasen, H., Mecklin, J., Khan, P. and Lynch, H. (1991) 'The International Collaborative Group on Hereditary Non-Polyposis Colorectal Cancer (ICG-HNPCC).', *Diseases of the Colon and Rectum*, 34(5), pp. 424-5.

Vasen, H., Watson, P., Mecklin, J. and Lynch, H. (1999) 'New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC.', *Gastroenterology*, 116(6), pp. 1453-6.

Verma, L., Kane, M., Brassett, C., Schmeits, J., Evans, D., Kolodner, R. and Maher, E. (1999) 'Mononucleotide microsatellite instability and germline MSH6 mutation analysis in early onset colorectal cancer.', *Journal of Medical Genetics*, 36(9), pp. 678-82.

Vora, A., Mitchell, C., Lennard, L., Eden, T., Kinsey, S., Lilleyman, J. and Richards, S. (2006) 'Toxicity and efficacy of 6-thioguanine versus 6-mercaptopurine in childhood lymphoblastic leukaemia: a randomised trial', *Lancet*, 368(9544), pp. 1339-48.

Waalkes, A., Smith, N., Penewit, K., Hempelmann, J., Konnick, E., Hause, R., Pritchard, C. and Salipante, S. (2018) 'Accurate Pan-Cancer Molecular Diagnosis of Microsatellite Instability by Single-Molecule Molecular Inversion Probe Capture and High-Throughput Sequencing', *Clinical Chemistry*, 64(6), pp. 950-958.

Wahlberg, S., Schmeits, J., Thomas, G., Loda, M., Garber, J., Syngal, S., Kolodner, R. and Fox, E. (2002) 'Evaluation of microsatellite instability and immunohistochemistry for the prediction of germ-line MSH2 and MLH1 mutations in hereditary nonpolyposis colon cancer families.', *Cancer Research*, 62(12), pp. 3485-92.

Wang, J., Sun, L., Myeroff, L., Wang, X., Gentry, L., Yang, J., Liang, J., Zborowska, E., Markowitz, S., Willson, J. and Brattain, M. (1995) 'Demonstration that mutation of the type II transforming growth factor beta receptor inactivates its tumor suppressor activity in replication error-positive colon carcinoma cells.', *Journal of Biological Chemistry*, 270(37), pp. 22044-9.

Wang, Q., Lasset, C., Desseigne, F., Frappaz, D., Bergeron, C., Navarro, C., Ruano, E. and Puisieux, A. (1999) 'Neurofibromatosis and early onset of cancers in hMLH1-deficient children.', *Cancer Research*, 59(2), pp. 294-7.

Wang, Q., Montmain, G., Ruano, E., Upadhyaya, M., Dudley, S., Liskay, R., Thibodeau, S. and Puisieux, A. (2003) 'Neurofibromatosis type 1 gene as a mutational target in a mismatch repair-deficient cell type', *Human Genetics*, 112(2), pp. 117-23.

Wang, X., Yu, J., Sreekumar, A., Varambally, S., Shen, R., Giacherio, D., Mehra, R., Montie, J., Pienta, K., Sanda, M., Kantoff, P., Rubin, M., Wei, J., Ghosh, D. and Chinnaiyan, A. (2005) 'Autoantibody signatures in prostate cancer', *New England Journal of Medicine*, 353(12), pp. 1224-35.

Weisenberger, D., Siegmund, K., Campan, M., Young, J., Long, T., Faasse, M., Kang, G., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A., Thibodeau, S., Jass, J., Haile, R. and PW, L. (2006) 'CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer.', *Nature Genetics*, 38(7), pp. 787-93.

Westdorp, H., Fennemann, F., Weren, R., Bisseling, T., Ligtenberg, M., Figdor, C., Schreiber, G., Hoogerbrugge, N., Wimmers, F. and Vries, I.d. (2016) 'Opportunities for immunotherapy in microsatellite instable colorectal cancer.', *Cancer Immunology, Immunotherapy*, 65(10), pp. 1249-59.

Westdorp, H., Kolders, S., Hoogerbrugge, N., de Vries, I., Jongmans, M. and Schreiber, G. (2017) 'Immunotherapy holds the key to cancer treatment and prevention in constitutional mismatch repair deficiency (CMMRD) syndrome', *Cancer Letters*, 403, pp. 159-164.

Whiteside, D., McLeod, R., Graham, G., Steckley, J., Booth, K., Somerville, M. and Andrew, S. (2002) 'A homozygous germ-line mutation in the human MSH2 gene predisposes to hematological malignancy and multiple cafe-au-lait spots', *Cancer Research*, 62(2), pp. 359-62.

Wimmer, K. and Etzler, J. (2008) 'Constitutional mismatch repair-deficiency syndrome: have we so far seen only the tip of an iceberg?', *Human Genetics*, 124(2), pp. 105-22.

Wimmer, K., Kratz, C., Vasen, H., Caron, O., Colas, C., Entz-Werle, N., Gerdes, A., Goldberg, Y., Ilencikova, D., Muleris, M., Duval, A., Lavoine, N., Ruiz-Ponte, C., Slavc, I., Burkhardt, B. and Brugieres, L. (2014) 'Diagnostic criteria for constitutional mismatch repair deficiency

syndrome: suggestions of the European consortium 'care for CMMRD' (C4CMMRD)', *J Med Genet*, 51(6), pp. 355-65.

Wimmer, K., Rosenbaum, T. and Messiaen, L. (2017) 'Connections between constitutional mismatch repair deficiency syndrome and neurofibromatosis type 1.', *Clinical Genetics*, 91(4), pp. 507-19.

Win, A., Jenkins, M., Dowty, J., Antoniou, A., Lee, A., Giles, G., Buchanan, D., Clendenning, M., Rosty, C., Ahnen, D., Thibodeau, S., Casey, G., Gallinger, S., Le Marchand, L., Haile, R., Potter, J., Zheng, Y., Lindor, N., Newcomb, P., Hopper, J. and MacInnis, R. (2017) 'Prevalence and Penetrance of Major Genes and Polygenes for Colorectal Cancer', *Cancer Epidemiology, Biomarkers and Prevention*, 26(3), pp. 404-412.

Woerner, S., Benner, A., Sutter, C., Schiller, M., Yuan, Y., Keller, G., Bork, P., von Knebel Doeberitz, M. and Gebert, J. (2003) 'Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes', *Oncogene*, 22(15), pp. 2226-35.

Woerner, S., Kloor, M., Mueller, A., Rueschoff, J., Friedrichs, N., Buettner, R., Buzello, M., Kienle, P., Knaebel, H., Kunstmann, E., Pagenstecher, C., Schackert, H., Moslein, G., Vogelsang, H., von Knebel Doeberitz, M. and Gebert, J. (2005) 'Microsatellite instability of selective target genes in HNPCC-associated colon adenomas', *Oncogene*, 24(15), pp. 2525-35.

Woerner, S., Yuan, Y., Benner, A., Korff, S., Doeberitz, M.v.K. and Bork, P. (2010) 'SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology.', *Nucleic Acids Research*, 38((Database Issue)), pp. D682-9.

Wright, C., Dent, O., Newland, R., Barker, M., Chapuis, P., Bokey, E., Young, J., Leggett, B., Jass, J. and Macdonald, G. (2005) 'Low level microsatellite instability may be associated with reduced cancer specific survival in sporadic stage C colorectal carcinoma.', *Gut*, 54(1), pp. 103-8.

Wu, Y., Berends, M., Mensink, R., Kempinga, C., Sijmons, R., van Der Zee, A., Hollema, H., Kleibeuker, J., Buys, C. and Hofstra, R. (1999) 'Association of hereditary nonpolyposis

colorectal cancer-related tumors displaying low microsatellite instability with MSH6 germline mutations.', *American Journal of Human Genetics*, 65(5), pp. 1291-8.

Yamane, L., Scapulatempo-Neto, C., Reis, R. and Guimarães, D. (2014) 'Serrated pathway in colorectal carcinogenesis', *World Journal of Gastroenterology*, 20(10), pp. 2634-40.

Yang, J., Bhojwani, D., Yang, W., Cai, X., Stocco, G., Crews, K., Wang, J., Morrison, D., Devidas, M., Hunger, S., Willman, C., Raetz, E., Pui, C., Evans, W., Relling, M. and Carroll, W. (2008) 'Genome-wide copy number profiling reveals molecular evolution from diagnosis to relapse in childhood acute lymphoblastic leukemia', *Blood*, 112(10), pp. 4178-83.

Yarchoan, M., Hopkins, A. and Jaffee, E. (2017) 'Tumor Mutational Burden and Response Rate to PD-1 Inhibition.', *New England Journal of Medicine*, 377(25), pp. 2500-1.

You, J., Buhard, O., Ligtenberg, M., Kets, C., Niessen, R., Hofstra, R., Wagner, A., Dinjens, W., Colas, C., Lascols, O., Collura, A., Flejou, J., Duval, A. and Hamelin, R. (2010) 'Tumours with loss of MSH6 expression are MSI-H when screened with a pentaplex of five mononucleotide repeats.', *British Journal of Cancer*, 103(12), pp. 1840-5.

Young, J., Simms, L., Biden, K., Wynter, C., Whitehall, V., Karamatic, R., George, J., Goldblatt, J., Walpole, I., Robin, S., Borten, M., Stitz, R., Searle, J., McKeone, D., Fraser, L., Purdie, D., Podger, K., Price, R., Buttenshaw, R., Walsh, M., Barker, M., Leggett, B. and Jass, J. (2001) 'Features of colorectal cancers with high-level microsatellite instability occurring in familial and sporadic settings: parallel pathways of tumorigenesis', *American Journal of Pathology*, 159(6), pp. 2107-16.

Yun, J., Rago, C., Cheong, I., Pagliarini, R., Angenendt, P., Rajagopalan, H., Schmidt, K., Willson, J., Markowitz, S., Zhou, S., Diaz, L.J., Velculescu, V., Lengauer, C., Kinzler, K., Vogelstein, B. and Papadopoulos, N. (2009) 'Glucose deprivation contributes to the development of KRAS pathway mutations in tumor cells', *Science*, 325(5947), pp. 1555-9.

Yurgelun, M., Allen, B., Kaldate, R., Bowles, K., Judkins, T., Kaushik, P., Roa, B., Wenstrup, R., Hartman, A. and Syngal, S. (2015) 'Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome', *Gastroenterology*, 149(3), pp. 604-13.e20.

Yurgelun, M., Kulke, M., Fuchs, C., Allen, B., Uno, H., Hornick, J., Ukaegbu, C., Brais, L., McNamara, P., Mayer, R., Schrag, D., Meyerhardt, J., Ng, K., Kidd, J., Singh, N., Hartman, A., Wenstrup, R. and Syngal, S. (2017) 'Cancer Susceptibility Gene Mutations in Individuals With Colorectal Cancer.', *Journal of Clinical Oncology*, 35(10), pp. 1086-95.

Zhang, B., Liang, X., Gao, H., Ye, L. and Wang, Y. (2016) 'Models of logistic regression analysis, support vector machine, and back-propagation neural network based on serum tumor markers in colorectal cancer diagnosis', *Genetics and Molecular Research*, 15(2).

Zhang, J., Stevens, M. and Bradshaw, T. (2012) 'Temozolomide: mechanisms of action, repair and resistance', *Current Molecular Pharmacology*, 5(1), pp. 102-14.

Zhang, L. (2008) 'Immunohistochemistry versus microsatellite instability testing for screening colorectal cancer patients at risk for hereditary nonpolyposis colorectal cancer syndrome. Part II. The utility of microsatellite instability testing.', *Journal of Molecular Diagnostics*, 10(4), pp. 301-7.

Zhu, L., Huang, Y., Fang, X., Liu, C., Deng, W., Zhong, C., Xu, J., Xu, D. and Yuan, Y. (2018) 'A Novel and Reliable Method to Detect Microsatellite Instability in Colorectal Cancer by Next-Generation Sequencing.', *Journal of Molecular Diagnostics*, 20(2), pp. 225-31.

