Development and validation of a next generation sequencing based microsatellite instability assay for routine clinical use

Mohammed Ghanim M. Alhilal

Thesis submitted in partial fulfilment of the requirements

for the degree of Doctor of Philosophy

Newcastle University

Faculty of Medical Sciences

Institute of Genetic Medicine

December 2016

Abstract

Colorectal cancer (CRC) is the second most common cancer in both men and women. Approximately 3-5% of CRCs show microsatellite instability (MSI) caused by germline defects in mismatch repair genes. In addition, 12% of sporadic CRCs show MSI. Currently, MSI is tested using a fragment analysis based assay not suitable for high throughput testing. Knowledge of microsatellite instability affects prognosis, surveillance and treatment of CRCs and MSI testing is now recommended for all newly diagnosed CRCs. As a result, development of high throughput approaches is desirable. The focus of my work was to develop and validate a high throughput sequence based MSI assay.

Initially, I tested 25 (7-9bp) mononucleotide markers, previously identified from *in silico* analyses, using a cohort of 55 CRCs, and selected 8 markers which collectively could discriminate between MSI-high (MSI-H) and microsatellite stable (MSS) cases. To define the optimal parameters to discriminate between MSI-H and MSS samples, I tested these 8 markers and 9 long (8-12bp) mononucleotide markers identified in a parallel study, across a panel of 141 CRC samples. This allowed development of a scoring scheme for the 17 markers, which achieved 96% sensitivity and 100% specificity. I validated this scheme using an independent cohort of 70 CRCs without knowing their MSI status. The assay achieved a 100% sensitivity and specificity.

Finally, I assessed the ability of short repeats to allow inference of the clonal variation within both FFPE (7) and fresh (4) MSI-H CRCs by analysing multiple samples from each cancer. I was able to infer the lineage relationship between primary tumour and lymph node metastasis in three cases and to construct phylogenetic trees for all cancers for which multiple samples were available illustrating the utility of these markers for understanding of CRC clonal variation.

Certificate of approval

I confirm that, to the best of my knowledge, this thesis is from the student's own work and effort, and all other sources of information used have been acknowledged. This thesis has been submitted with my approval.

Supervisor

Professor Sir John Burn

Acknowledgements

I would like to thank my supervisors Professor Sir John Burn, Dr Michael Jackson and Dr Mauro Santibanez-Koref for their help and support throughout this PhD. In particular John Burn for the help and support he provided to keep the project progressing on time, Michael Jackson for his close follow up of the work on day by day basis and because he worked diligently to keep me on track and Mauro Santibaez-Koref for his supervision and guidance in performing the bioinformatic part of the study.

I would like to thank the Higher Committee for Education Development in Iraq (HCED Iraq) for funding this PhD.

I would like to thank Ángel Alonso Sánchez and Sira Moreno Laguna for their help in providing samples from their lab in Spain. I also would like to thank Mark Arends and Anca Oniscu for their help in providing samples from their lab in Edinburgh.

I would also like to thank Dr Helen Turner and Dr Stephanie Needham for their help in retrieval of tumour samples and examination of histological slides of samples analysed in this thesis.

I would like to thank Dr Lisa Redford for her help since the start of my PhD and for her active participation in my study. I also would like to thank the students Leonardo Amorim, Paloma Andrade and Iona V. Middleton for their assistance and contributions.

Finally, I am very grateful to my family (my parents, brothers, sister, wife and my lovely son) for their support over the past 3 years I have spent working on my PhD project.

Table of Contents

Abstract i

Acknowledgements v

Table of Contents vii

List of Figures xiii

List of Tables xvii

List of Abbreviations xix

Chapter 1. Introduction 1

1.1. Colorectal Cancer and microsatellite instability 1

1.2. *MMR* Genes 2

1.2.1. MMR Function 2

1.2.2. MMR Defects 4

1.3. Molecular testing of *MMR* defects 5

1.3.1. Immunohistochemistry Analysis 5

1.3.2. Sequencing of MMR genes to detect point mutations 6

1.3.3. Detection of large genomic rearrangements (deletion and duplication) 6

1.3.4. Other tests 7

1.4. Microsatellite Instability 7

1.4.1. Repetitive DNA 7

1.4.2. Microsatellites 8

1.4.3. Multistep Evolution of Microsatellites 10

1.4.4. Coding Microsatellites 10

1.4.5. Microsatellite instability 10

1.4.6. Mechanisms of microsatellite instability 11

1.4.7. Factors affecting the microsatellite mutation rate 12

1.4.8. MSI testing: 13

1.5. MSI and its impact on prognosis in CRCs 13
1.6. MSI and its impact on treatment of CRCs 15
1.7. Clonality and Microsatellite Instability 18
1.7.1. The role of microsatellites in assessment of clonality 20
1.8. Aims of study and chapters outlines 23
Chapter 2. Methods 25
2.1. Ethical approvals 25
2.2. Clinical samples 25
2.2.1. Tumour samples for MSI assay 25
2.2.2. Tumour samples for clonality assay 26
2.3. DNA Extraction 29
2.3.1. DNA extraction from FFPE tissue samples using the BiOstic FFPE Tissue DNA Isolation Kit 29
2.3.2. DNA extraction from Fresh tissue samples using the ReliaPrep™ gDNA Tissue Miniprep System kit 29
2.4. DNA Quantity and Quality Assessment 29
2.4.1. Quantitative Assessment 29
2.4.2. Qualitative Assessment 30
2.5. Primer Design 30
2.6. PCR 31
2.6.1. Amplicon generation by PCR for MiSeq analysis 31
2.6.2. PCR amplification for fragment analysis 31
2.7. Post-PCR detection 32
2.7.1. Gel Electrophoresis: 32
2.7.2. QIAxcel Electrophoresis 32
2.8. Fragment Analysis 33
2.9. Library preparation and High Throughput MiSeq Sequencing 33

2.9.1. Amplicon Pooling....33
2.9.2. Clean-Up of pooled amplicons....33
2.9.3. Size determination....34
2.9.4. Barcoding of pooled amplicons....34
2.9.5. Clean Up of the barcoded amplicons and QIAxcel electrophoresis....35
2.9.6. DNA Quantitaficaton and Dilution....35
2.9.7. Library preparation and MiSeq sequencing....35
2.10. Data analysis....36
2.10.1. Sequencing data....36
2.10.2. Data visualization....36
2.10.3. Variant calling....36
2.10.4. Deletion frequency....37
2.10.5. Fisher's Exact test....37
2.10.6. Threshold setting....37
2.10.7. Determination of MSI status....38
2.10.8. Constructing the phylogenetic tree for clonality assay....38
Chapter 3. Establishing a consensus short mononucleotide repeat panel for NGS-based MSI testing....41
3.1. Introduction and aims....41
3.1.1. Introduction....41
3.1.2. Aims....46
3.2. Results....46
3.2.1. Assessment of 29 mononucleotide repeats to identify the most variable repeats....46
3.2.2. Analysis of 25 variable short (7-9bp) mononucleotide repeats to assess the criteria for calling instability in CRCs....53

3.3. Discussion 77

3.4. Conclusions 80

Chapter 4. Assessment of arbitrary threshold sets and determination of an optimal MSI scoring system 81

4.1. Introduction and aims 81

4.1.1. Introduction 81

4.1.2. Aims 83

4.2. Results 84

4.2.1. Amplification and Sequencing of the new 17 marker panel 84

4.2.2. Analysis of MiSeq data and calculation of deletion frequency 90

4.2.3. Assessment of different threshold sets to conclude the most informative cutoff values 91

4.2.4. Assignment of a new scoring system for calling instability 95

4.2.5. Stratification of the new MSI scoring system against MMR IHC status of the 141 CRC samples 96

4.2.6. Comparison of deletion curves for all markers with those of the previous cohorts 101

4.2.7. Assessment of DNA fragmentation of a selected subset of Spanish samples 110

4.3. Discussion 111

4.4. Conclusions 114

Chapter 5. Analytical validation of the weighted MSI score using an independent cohort of CRCs 115

5.1. Introduction and aims 115

5.1.1. Introduction 115

5.1.2. Aims 117

5.2. Results 118

5.2.1. Amplification and sequencing of the short mononucleotide panel using a cohort of 100 CRC samples....118

5.2.2. Analysis of the sequencing data and assessment of the weighted MSI scoring system (Threshold 7)....120

5.2.3. Comparison of threshold curves for all markers in the 3 independent cohorts....123

5.2.4. Assessment of the inter-cohort inconsistency of marker GM14-11 deletion....131

5.3. Discussion....135

5.3.1. Validation of the MSI scoring system....135

5.3.2. Fulfilment of the recommended requirements for validation of targeted NGS assay....135

5.3.3. Assessment of the run performance....136

5.4. Conclusions....138

Chapter 6. Analysis of clonal characteristics of MSI-H CRC using short mononucleotide markers....139

6.1. Introduction and aim....139

6.1.1. Introduction....139

6.1.2. Aim....143

6.2. Results....143

6.2.1. Collection of MSI-H CRC fresh tissue samples....143

6.2.2. Collection of FFPE MSI-H CRC samples....145

6.2.3. PCR amplification and MiSeq sequencing of the collected samples to assess the clonal composition....149

6.2.4. Clonal composition of the FFPE tumours....151

6.2.5. Clonal composition of the Fresh MSI-H CRC samples....162

6.3. Discussion....169

6.4. Conclusions....172

Chapter 7. General Discussion and Future work 173

7.1. General discussion 173

7.1.1. Comparison of the current assay with other methods 174

7.1.2. Comparison of deletion curves suggests further assessment of robustness and reproducibility is required 177

7.1.3. Suspicion of polymorphism in 2 independent markers in 2 different samples 178

7.1.4. Determination of the optimal quality metrics for an NGS based MSI approach 179

7.1.5. Cost analysis and turnaround time 180

7.1.6. Future improvements in assay design 182

7.1.7. Using short mononucleotide markers to assess intratumour heterogeneity 183

Chapter 8. Appendix 185

Chapter 9. References 199

List of Figures

Figure 1-1: The sequential action of *MMR* genes in DNA repair 4

Figure 3-1: Illustration of the overall number of markers and the study workflow. 47

Figure 3-2: The frequencies of variant reads of 6 short (7bp) homopolymers. 48

Figure 3-3: The frequencies of variant reads of 5 short (8bp) homopolymers. 49

Figure 3-4: The frequencies of variant reads of 9 short (9bp) homopolymers 50

Figure 3-5: The frequencies of variant reads of 4 long (10bp) homopolymers 51

Figure 3-6: The frequencies of variant reads of 4 long (11bp) homopolymers 52

Figure 3-7: The frequencies of variant reads of the 12bp long homopolymers (GM18). 52

Figure 3-8: Initial check of DNA amplifiability using a 300bp amplicon. 55

Figure 3-9: Amplifications of DNA samples with 100bp amplicons. 56

Figure 3-10: The number of amplifiable colorectal cancer samples using the 300bp and 100bp amplicons. 56

Figure 3-11: Comparative assessment of DNA quality between 2 groups of MSI-H samples. 57

Figure 3-12: The distribution of sequencing reads for all tested samples. 59

Figure 3-13: Q30 of the sequencing reads generated in MiSeq run. 59

Figure 3-14: The organisation of sequencing reads. 60

Figure 3-15: The COPReC data format for a single amplicon in a single sample. 61

Figure 3-16: Threshold curves of the marker LR24. 62

Figure 3-17: Sensitivity and specificity curves for the 7bp markers. 63

Figure 3-18: Sensitivity and specificity of the 8bp markers. 64

Figure 3-19: Sensitivity and specificity of the 9bp markers. 65

Figure 3-20: Deletion frequencies (Y axis) of all the 25 markers in 2 samples 66

Figure 3-21: The deletion frequencies for the marker LR24 across all samples. 67

Figure 3-22: Number of 7-9bp markers called as unstable in each sample using different threshold sets..... 69

Figure 3-23: Deletion frequencies (Y-axis) of all markers (X-axis) in 2 different samples 71

Figure 3-24: Sequencing reads of the marker LR24 linked to a specific SNP in one of the MSI-H samples. 72

Figure 3-25: The deletion frequency and allelic bias of the 25 markers. 73

Figure 3-26: The instability (shown in percentage) of the 15 markers in MSI-H samples. 74

Figure 3-27: The instability (shown in percentage) of the final panel 75

Figure 3-28: The 8 informative markers selected from the 25 tested markers. 76

Figure 3-29: A: Base composition of the homopolymers tested in this study. 77

Figure 4-1: Illustration of the overall workflow in this study. 84

Figure 4-2: The informative mononucleotide markers from 2 independent studies... 85

Figure 4-3: Schematic representation of the primer designing..... 87

Figure 4-4: Q score distribution of MiSeq run 89

Figure 4-5: The distribution of samples with and without deletion and allelic bias according to cutoff values in the 6 threshold sets 94

Figure 4-6: The MSI score for all samples that were called as unstable by the weighted scoring system..... 97

Figure 4-7: The MSI score for all samples that were called as stable by the new scoring system..... 98

Figure 4-8: MSI results from fragment analysis of the equivocal cases. 99

Figure 4-9: The final sensitivity and specificity of the weighted MSI scoring system 100

Figure 4-10: Deletion frequencies of all the 17 markers in the 6 equivocal samples in 2 MiSeq runs..... 101

Figure 4-11: Threshold curves of the 7bp, 8bp and Poly G/C markers in both Newcastle and Spanish cohorts..... 105

Figure 4-12: Threshold curves of the 9bp markers in both Newcastle and Spanish cohorts. 106

Figure 4-13: Threshold curves of the 11bp and 12bp markers in both Newcastle and Spanish cohorts.....109

Figure 4-14: The DNA integrity test for a subset of Spanish samples. 110

Figure 5-1: Q score (on the left) and cluster density (on the right) of the MiSeq run. 119

Figure 5-2: The distribution of samples, according to cutoff values in T7 121

Figure 5-3: The overall MSI score for all tested Edinburgh samples (69 samples)..122

Figure 5-4: The deletion curve of the 7bp, 8bp and poly G/C markers in both MSS and MSI-H samples in the 3 independent cohorts.....126

Figure 5-5: The deletion curve of the 9bp markers in both MSS and MSI-H samples in 3 independent cohorts.....128

Figure 5-6: The deletion curve of the 11 and 12bp markers in both MSS and MSI-H samples in 3 independent cohorts.....130

Figure 5-7: The alignment of the sequencing reads of the marker GM14-11 in 4 MSI-H samples. 132

Figure 5-8: The alignment of the sequencing reads of the marker GM14-11 in 4 MSS samples.....133

Figure 5-9: Deletion frequencies of the marker GM14-11 in 24 samples from 2 different cohorts.....134

Figure 6-1: The orientation of specimens retrieved from fresh CRC tumours.....144

Figure 6-2: Electropherogram showing the fragment analysis of 2 fresh CRC tumours. 145

Figure 6-3: The clonal characteristics of the tumour PR53139/13.....154

Figure 6-4: The clonal characteristics of the tumour PR32079/14.....156

Figure 6-5: The clonal characteristics of the tumour PR7146/13.....157

Figure 6-6: The clonal characteristics of the tumour PR34630/03.....158

Figure 6-7: The clonal characteristics of the tumour PR049276/12.....159

Figure 6-8: The clonal characteristics of the tumour PR53996/14. 161

Figure 6-9: The clonal characteristics of the tumour PR45703/14. 162

Figure 6-10: The clonal characteristics of the tumour PR10654/14. 164

Figure 6-11: The clonal characteristics of the tumour PR17848/14. 165

Figure 6-12: The clonal characteristics of the tumour PR51896/13. 167

Figure 6-13: The clonal characteristics of the tumour PR32516/14. 169

Figure 8-1: Poster shows the initial selection of the markers. 191

Figure 8- 2: Poster shows the overall workflow..... 192

Figure 8-3: Deletion frequencies of a subset of 15 markers in the tumour PR32079/14..... 193

Figure 8-4: Deletion frequencies of a subset of 16 markers in the tumour PR53139/13..... 194

Figure 8-5: Deletion frequencies of a subset of 14 markers in the tumour PR10654/14..... 195

Figure 8-6: Deletion frequencies of a subset of 18 markers in the tumour PR17848/14..... 196

Figure 8-7: Deletion frequencies of a subset of 14 markers in the tumour PR51896/13..... 197

Figure 8-8: Deletion frequencies of a subset of 13 markers in the tumour PR32516/14..... 198

List of Tables

Table 2-1: Colorectal cancer tumours that were analysed in this study and their provider. 28

Table 2-2: The PCR program that used to generate products for fragment analysis. 32

Table 3-1: DNA samples that were used in the initial assessment. 47

Table 3-2: The 25 primers that were used to analyse the short repeats....54

Table 3-3: Cutoff values and false calling rates in the 4 threshold sets. 68

Table 3-4: Outputs and performance metrics of the Final (0.05, 0.1) threshold set...71

Table 4-1: List of the 17 primers used in the MSI analysis. 88

Table 4-2: The number of samples that were successfully amplified, sequenced and called to adequate depth (≥100 reads). 90

Table 4-3: Cutoff values that used for threshold setting. 93

Table 4-4: Analytical parameters of the 6 threshold sets. 95

Table 4-5: Sensitivity and specificity of T7. 96

Table 4-6: Sensitivity and specificity of all markers in the 2 tested cohorts (N= Newcastle and S= Spanish) at the cutoff values specified in T7. 103

Table 5-1: Number of amplicons that were amplified, sequenced and called to adequate depth (i.e. ≥100 reads) in Edinburgh cohort. 120

Table 5-2: Threshold sets and their cutoff values. 122

Table 5-3: The sensitivity, specificity for all threshold sets in both DF and AB subgroups....123

Table 5-4: Sensitivity and specificity of all markers in the 3 cohorts....124

Table 6-1: The FFPE CRC tumours that were used to assess ITH. 147

Table 6-2: The list of primers that were used in the clonal assessment 150

Table 6-3: Cutoff values for calling instability in the clonality assay. 151

Table 6-4: Number of unstable markers and sequencing coverage for the 4 specimens of the tumour PR53139/13. 152

Table 6-5: Number of unstable markers in the 6 specimens of tumour PR32079/14. 155

Table 7-1: Specifications of the marker GM14-11 with discordant reads in 2 different samples (S74 and S79). 178

Table 7-2: Quality metrics for the 3 MiSeq runs using 3 different library concentrations. 180

Table 8-1: Samples that were included in the Newcastle cohort..... 186

Table 8-2: The 17 primer sets that were used in the analysis of both Spanish and Edinburgh cohorts..... 188

Table 8-3: Cost analysis and comparison between MiSeq based assay and the conventional fragment analysis based assay..... 189

Table 8-4: Total turnaround time (TAT) expected from the MiSeq based assay. 190

List of Abbreviations

ACMG: American college of medical genetics

ACGS: Association of clinical genetic sciences

BAM: Binary Alignment/Map

BLAST: Basic Local Alignment Search Tool

BWA: Burrows–Wheeler Aligner

CAPP: Cancer Prevention Programme

CIMP: CpG island methylator phenotype

COPReC: Concordant overlapping paired reads caller

CRC: Colorectal cancer

dbSNP: Single Nucleotide Polymorphism Database

DFS: Disease free survival

dH2O: deionised water

dsDNA: double stranded DNA

EMAST: Elevated microsatellite alterations at selected tetranucleotide repeats

EpCAM: Epithelial Cell Adhesion Molecule

EXO1: Exonuclease I

FFPE: formalin fixed paraffin embedded

FNR: False negative rate

FPR: False positive rate

FSP: Frameshift peptides

HNPCC: Hereditary non polyposis colorectal cancer

IGV: Integrated genome viewer

IHC: Immunohistochemistry

ITH: Intratumour Heterogeneity

IGV: Integrative Genomics Viewer

Indel: insertion and deletion

K/mm^2: 1000 clusters per square millimetre (for cluster density)

LINES: Long Interspersed Nuclear Elements

MAF: Minor allele frequency

MIP: Molecular inversion probe

MMR: mismatch repair

MSI: microsatellite instability

MSI-H: high levels of microsatellite instability

MSI-L: low levels of microsatellite instability

MSS: microsatellite stable

MTX: Methotrexate

NHS: National Health Service

OS: Overall survival

PD-1: Programmed death receptor-1

PDL1: Programmed death ligand -1

Per: paired end reads

pM: picomolar

RVI: Royal Victoria Infirmary

SelTarBase: Selective Targets in Human MSI-H Tumorigenesis Database

SINES: Short Interspersed Nuclear Elements

SNP: Single Nucleotide Polymorphism

TAE: tris-acetate-EDTA

TCGA: The Cancer Genome Atlas

TE: Tris-EDTA

TIL: Tumour infiltrating lymphocytes

TN: True negative

TP: True positive

UCSC: University of California Santa Cruz

5-FU: 5-fluorouracil

Chapter 1. Introduction

1.1. Colorectal Cancer and microsatellite instability

Colorectal cancer (CRC) is the third most common cancer both in men and women. Every year, approximately 40,000 new colorectal cancer cases are diagnosed in the UK and one person out of 20 is estimated to develop colorectal cancer at some point in the life (Cancer-Research-UK, 2015).

Like other types of cancer, CRC is accompanied by the emergence of many genetic and epigenetic alterations. In about 15% of early stages CRCs, a specific form of genetic alteration, called microsatellite instability (MSI), is observed (Ward et al., 2001, Kim et al., 1994). Microsatellites are short, repetitive DNA sequences scattered across Eukaryotic genomes. Microsatellites are more liable to length changes as a result of suboptimal fidelity of DNA polymerase in replicating these sequences.

Under normal circumstances, errors that happen during DNA replication of microsatellites are corrected by a special family of proteins known as mismatch repair proteins encoded by mismatch repair genes (*MMR*). Germline mutations of *MMR* genes are the hallmark of the most common hereditary type of CRC, which is known as Lynch syndrome. It is expected, therefore, to find a certain level of microsatellite instability in the context of mutated *MMR* genes (Lynch and de la Chapelle, 2003, De la Chapelle, 2004).

Microsatellite instability (MSI) has a biological and clinical impact in CRC. It has been found that MSI positive CRCs (MSI-H) have better outcomes than microsatellite stable (MSS) CRC in terms of survival, response to certain chemotherapies and low recurrence rate (Sinicrope et al., 2011, Saridaki et al., 2014). Furthermore, MSI positive CRCs are less likely to distantly metastasize and more prone to reside in the right side of the colon (Caecum and Ascending colon) compared to MSS CRCs (Kim et al., 1994).

As a result of the aforementioned importance, finding a reliable laboratory test to detect microsatellite instability is desirable. This need has been addressed using a variety of methods over the last 2 decades and a consensus guidelines were published in 1997 (Rodriguez-Bigas et al., 1997) and amended in 2004 (Umar et al., 2004) to guide diagnosis of microsatellite instability. In practice, MSI testing has been

applied in the form of a widely used commercial kit sold by Promega (Promega, Madison, WI, USA). This MSI test is based on a multiplex amplification of 5 quasimonomorphic long (more than 20 nucleotides in length) mononucleotide markers followed by a fragment analysis. Although it is the most commonly used one, this test has several drawbacks such as the reliance on long repetitive tracts, convoluted interpretation, low throughput and a suboptimal sensitivity and specificity in cancers with *MSH6* mutations and tumours other than CRCs (Lynch et al., 2009, Berg et al., 2009).

As a result of the clinical benefits of MSI detection and its impact on the subsequent choice of treatment, seminal guidelines have recently recommended offering the MSI test for all newly diagnosed CRCs (Berg et al., 2009, de la Chapelle and Hampel, 2010, Vasen et al., 2013, Loughrey et al., 2014). It is desirable, therefore, to develop a new MSI testing approach accurate and robust enough to satisfy these evolving needs on a large-scale basis.

The advent of massive parallel sequencing techniques (also known as Next Generation Sequencing (NGS)) has opened the way for new research and diagnostic developments. This has made it feasible to use the multiplexing potential of NGS to analyse thousands of targets in the same sequencing run with a relatively short turnaround time. It may be possible, therefore, to develop an NGS based MSI assay robust and sensitive enough to be used in clinical laboratories.

1.2. *MMR* Genes

1.2.1. MMR Function

Every day, millions of cells in a human body are lost and replaced by new cells from skin, hair follicles and other organs. This replacement (which is achieved by compensatory cell division) is accompanied by millions of DNA replication events that are essential to ensure the newly synthesized cells get the exact copy of the parental DNA and thus maintain viability of the organism. Due to the huge number of cell divisions, nucleotide errors during DNA replications are inevitable events, and under normal circumstances, these occur at a low rate of approximately 1.3×10^{-8} base pairs per generation in humans (Scally and Durbin, 2012, Nachman and Crowell, 2000). In most eukaryotes, the rate of mutation is controlled by the characteristic 3'-5' exonuclease activity of DNA polymerase which is responsible for proofreading of the

newly synthesized DNA strand to ensure an accurate DNA replication. However, in repetitive sequences, *in vivo* studies showed that this proofreading function is efficient with short sequences only (Tran et al., 1997).

A further level of controlling DNA errors is achieved by a group of specialized proteins, which function within the Mismatch Repair system encoded by the Mismatch Repair genes. Mismatch Repair (*MMR*) genes that are involved in predisposition to human cancers include *MLH1* (*mut*L Homolog1), *MSH2* (*mut*S Homolog 2), *MSH6* (*mut*S Homolog 6) and *PMS2* (Postmieotic Segregation increased 2).

Roles of *MMR* genes in DNA repair have a sequential pattern (as shown in the Figure 1-1 comprising:

1) **Recognition of the DNA error**: This is primarily achieved by the human homologue of the MSH family (*hMSH*). *hMSHα* heterodimer (composed of one molecule of *MSH2* and one molecule of *MSH6*) which recognizes small insertion/deletion loops (IDLs) and base mismatches, while hMSHβ (*MSH2* and *MSH3*) are capable of recognizing larger IDLs (Jascur and Boland, 2006). There is evidence indicating that recognition of the mismatch in the newly synthesized strand requires the finding of a nick (e.g. that created by Okazaki fragments) which in turn can be used as an access point for subsequent enzymes involved in repair (e.g. Exonuclease I) (Jascur and Boland, 2006).
2) **Recruitment of the repair enzyme machinery**: This involves the hMutL complex (which is composed of *hMLH1* and *hPMS2*) which modifies the enzymatic activity of DNA polymerase, displacing the proliferating cell nuclear antigen (PCNA) and recruiting Exonuclease I (Exo 1), which in turn is required for excision of mismatched sequences.
3) **Removal (excision) of the incorrectly matched bases**: This step is mediated by Exo 1 which excises a single strand (may reach up to 1000 nucleotides) from a nick, to the position of the mismatch in 5'-3' or 3'-5' direction (Genschel et al., 2002).
4) **Resynthesis of the required bases by DNA polymerase**: This step is achieved by DNA polymerase δ (Wilson et al., 2005, Jascur and Boland, 2006).

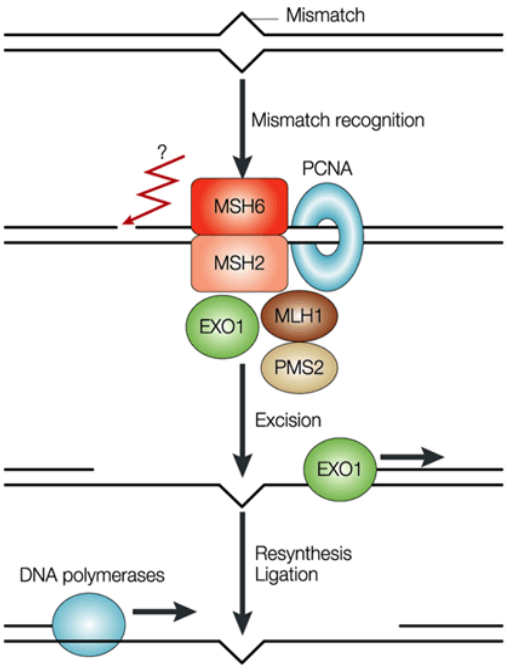


Figure 1-1: The sequential action of *MMR* genes in DNA repair (Martin and Scharff, 2002).

MMR genes are also involved in several other cellular functions (other than mismatch repair) which include repairing of double strand breaks, induction of apoptosis, anti-recombination effect, and destabilzation of DNA (Hegde et al., 2014).

1.2.2. MMR Defects

Based on the *MMR* functions outlined above, it is clear how defective *MMR* genes (whether complete absence of a protein or just a defective protein) could initiate events leading to the accumulation of genomic mutations. Lynch syndrome (LS, also called hereditary non polyposis colorectal cancer), is a hereditary disorder where inherited mutations in *MMR* genes increase the susceptibility to certain cancers, particularly CRCs. In LS, germline mutations in *MMR* genes are most often detected in *MLH1* and *MSH2* (about 90%) and less frequently in *MSH6* and *PMS2* (10%) (Lynch and de la Chapelle, 2003).

The most common mutations in the *MMR* genes are substitutions followed by deletions (Li et al., 2013a). However, *MMR* gene defects can also be found in other forms than point mutation, including the following:

- **Genomic rearrangements** of *MLH1* and *MSH2:* This kind of mutation represents a proportion of *MMR* gene mutations. Large genomic rearrangement extends from a single exon to a full gene deletion. In a study

conducted on 365 suspected LS patients, prevalence of large genomic rearrangements was reported to be 17.9% and 45.3% in *MLH1* and *MSH2*, respectively in mutation positive cases (=153 cases) (Baudhuin et al., 2005).

- **Epigenetic changes**: 60%-90% of the CpG islands in the human genome are methylated in the 5^{th} carbon atom of cytosine bases. In the majority of expressed genes, these CpG islands are located in the promoter region, thus methylation leads to alterations in the function of these genes (Gazzoli et al., 2002). *MLH1* promoter hypermethylation has been reported in sporadic colorectal cancer cases with microsatellite instability (Kane et al., 1997). Li et al. (2013b) observed *MLH1* promoter hypermethylation in 20.3% of unselected colorectal cancer cases. Interestingly, the prevalence of *MLH1* hypermethylation was reported to be as high as 80% of all MSI-H carriers (Lynch et al., 2009). Simultaneous methylation of *MLH1* promoter and *p16* in CRCs have recently been analysed and found to be associated with prognostic and clinicopathological features like right sided CRC, poorly differentiated, *BRAF* mutation positive and MSI-H phenotype (Veganzones et al., 2015). This kind of combined methylation has an impact on the disease free survival and might help in determination of new therapeutic approaches in MSI-H CRCs.
- **Promoter abnormalities**: *MSH2* promoter mutations have been investigated and a specific SNP (- 80 A insertion) has been found to change the binding activity of *MSH2* and form a novel binding complex. However, these mutations seem to have a limited role in the initiation of carcinogenesis in both Lynch syndrome and early CRCs (Shin et al., 2002).

1.3. Molecular testing of *MMR* defects

1.3.1. Immunohistochemistry Analysis

Routinely, *MMR* gene defects are often tested first by immunohistochemistry to check the functionality of these genes by targeting their protein products. This is usually done using commercially available antibodies against MMR proteins to test for expression in tissue samples. MMR immunohistochemistry (IHC) is a sensitive testing strategy and it detects *MMR* gene dysfunctions in colorectal cancer cases with sensitivity up to 100%. Its low cost and simplicity have made it a widely used testing option for *MMR* function. IHC, however, requires experienced personnel to

perform and to interpret the results, and can show a relatively low specificity (as low as 82%) (Stormorken et al., 2005). Furthermore, the existence of two independent somatic *MMR* gene mutations had been observed in up to 70% of *MMR* defective cases where no germline mutation is present. This is important to be detected as it will affect the subsequent surveillance options (Haraldsdottir et al., 2014). As a result of these limitations, IHC cannot be considered as the gold standard test to identify patients with defective *MMR* genes.

1.3.2. Sequencing of MMR genes to detect point mutations

PCR amplification of all coding exons and flanking sequences followed by direct gene scanning represents the gold standard testing strategy to detect point mutations of *MMR* genes. However, direct *PMS2* gene sequencing can be complicated by the fact that this gene is located on chromosome 7 where several homologous copies also exist (i.e. *PMS2* psuedogenes) (Nakagawa et al., 2004).

The results of the preceding MMR IHC test can guide the decision of which gene needs to be tested, thus cutting the time and cost. Furthermore, with the advent of next generation sequencing, it has become possible to sequence all coding sequences of *MMR* genes in a single run. Nevertheless, both sequencing approaches (i.e. Sanger sequencing and next generation sequencing) would not be able to detect full exon or multi- exon deletions/ duplications.

1.3.3. Detection of large genomic rearrangements (deletion and duplication)

Multiplex Ligation-dependent Probe Amplification (MLPA) is the method of choice to detect large genomic rearrangements which are too small to be detected by standard cytogenetic methods. The existence of 2 or more sequentially deleted exons could be considered as a dependable result. However, when there is a deletion in a single exon (or a single probe deletion when multiple probes are designed for a single exon), results need to be confirmed by other techniques (e.g. sanger sequencing to check for a SNP at site of probe binding or Southern Blotting) (Hegde et al., 2013).

1.3.4. Other tests

Other tests can be selectively performed in targeted cases. *MLH1* promoter hypermethylation can be tested using bisulfite conversion followed by real time PCR to compare both wildtype and methylated plot graphs.

In 10-40% of tumours with defective *MSH2/MSH6* protein in IHC with no *MMR* germline mutations, the underlying reason is a deletion in the epithelial cell adhesion molecule (*EpCAM) gene* (Ligtenberg et al., 2009, Kovacs et al., 2009, Niessen et al., 2009, Guarinos et al., 2010). *EpCAM* (also known as *TACSTD1* gene) deletions usually lead to loss of the most 3' exons (2 or 4 exons) in addition to the polyadenlation signal of the gene. This ,consequently, impairs the proper termination of transcription and leads to promoter methylation of the downstream gene, which is *MSH2* (in tissues that express *EpCAM*) (Ligtenberg et al., 2009). *EPCAM* mutation status is tested for either by MLPA or Southern blot.

Other less common genetic alterations might be involved in loss of *MMR* function, such as:

1) Inversion of *MSH2* (exons 1-7): In a study conducted in 2014, this inversion was observed in 6 out of 10 CRC patients with unexplained defective *MSH2* expression (Rhees et al., 2014).
2) *MMR* germline mutations are usually associated with MSI, but in some cases where no germline mutation in *MMR* is existed, other genes might be responsible. Polymerase E (*POLE)* mutations have been found to be associated with MSI-H with the absence of *MMR* germline mutations (but was associated with *MSH2* and *MSH6* somatic mutations) (Elsayed et al., 2014).

1.4. Microsatellite Instability

1.4.1. Repetitive DNA

Repetitive DNA refers to DNA sequences that are repeated many times (multiple copies) throughout the genome. These repetitive DNA sequences can be subdivided into:

- **Interspersed repeats:** These are sequences that are repeated in different genomic positions, due to replication by transposition. Interspersed sequences can be further divided into long interspersed nuclear elements (LINEs) (more

than 300bp in length) such as L1 repeats or short interspersed nuclear elements (SINEs) (100-300bp) such as *Alu* elements.

- **Short tandem repeats** (also known as microsatellites): DNA repeats of 2 bases or more repeated 2- million times adjacent to each other. Some chromosomal regions have more abundance of these tandem repeats than others, e.g. centromere and telomere.
- **Segmental duplications**: These are arbitrarily defined as any complex sequence tract of 1kb or more, which shares >90% identity to another region of the genome. These constitute 1-14% of different human chromosomes and highly enriched in pericentromeric and subtelomric regions (Zhang et al., 2005, Treangen and Salzberg, 2012).

Satellite is a “catch all” term for all non-mobile, highly repetitive sequences with a clear repeat periodicity within complex genomes. These repeats can be classified based on repeat periodicity into Satellites (>20 bases), Minisatellites (7-20 bases) and Microsatellites (1-6 bases) (Ellegren, 2004).

1.4.2. Microsatellites

Microsatellites constitute about 3% of the human genome and can be classified in different ways. Based on repeat length, microsatellites can be subclassified into those with a single nucleotide (mononucleotide repeat), two (dinucleotides), three (trinucleotides), four (tetranucleotides), five (pentanucleoide) or 6 (hexanucleotide). Alternatively, microsatellites can be classified based on the tract structure into perfect (or simple) microsatellites (in which the tract is continuous and made up of a single repetitive unit), imperfect (or complex) microsatellites (where the tract is interrupted by another sequences) or compound, where 2 or more microsatellite units coexist together (Urquhart et al., 1994, Sharma et al., 2007). However, microsatellites might undergo mutations, e.g. transition/transversion, which lead to a single event causing an interruption in the microsatellite tract. This interruption is usually increases the stability of the microsatellite.

The vast majority of microsatellites (MSs) are located in the non-coding regions, however, about 8% of them can also exist within coding sequences (Ellegren, 2000) where the majority of them are tri- and hexanucleotides MSs (Subramanian et al., 2003). On the chromosomal level, microsatellites are unevenly

distributed in different human chromosomes, but chromosome 19 has the greatest population of these sequences (Subramanian et al., 2003).

Mononucleotides are an abundant form of microsatellites across all chromosomes but comparatively more abundant in coding regions of chromosome 7 and 16 while the least abundance is in the exonic regions of the Y chromosome. Poly A and poly T tracts are about 300- fold more abundant than Poly G and poly C across the genome (Subramanian et al., 2003).

The exact function of microsatellites is not clearly understood, and they have historically been considered as junk DNA for this reason. However, some functions have been ascribed to specific microsatellites in a number of analyses:

1- A highly conserved feature of many proteins involved in transcription regulation is that they contain glutamine- rich domains (Glutamine is encoded by a triplet sequence) (Escher et al., 2000). These glutamine rich activation domains were found to behave very similarly in both yeast and mammals in regulating gene expression. They are activated by the binding of acidic activators to remote enhancers.
2- It has been suggested that microsatellites are involved in regulation of transcription of several genes such as early growth response 1 (*EGR1*), where increase in the length of polymorphic CA microsatellites in the first intron was found to be associated with reduced expression (Gebhardt et al., 1999), tyrosine hydroxylase, where a tetranucleotide microsatellite in the first intron was found to impact upon the gene expression and be associated with predisposition to schizophrenia (Meloni et al., 1995, Meloni et al., 1998) and P53 inducible gene 3 (*PIG3*), where a pentanucleotide microsatellite in the *PIG3* promoter is necessary to facilitate the binding of the wildtype *P53* and thus induce apoptosis (Contente et al., 2002).
3- Microsatellites may also have an effect on recombination (Wahls et al., 1990), nucleosome positioning (Wang and Griffith, 1995) and chromatin structure (Heale and Petes, 1995).
4- It has recently found that repetitive DNA sequences play a role in determination of the 3 dimensional structure of the genome in different organisms including human (Axel Cournac, 2015).

1.4.3. Multistep Evolution of Microsatellites

The mechanisms of microsatellite development have been extensively studied. It has been suggested that microsatellites are developed initially from a specific sequences called 'proto-STR'. These sequences have the ability to mutate and acquire instability. A threshold number of repeats need to be crossed prior to developing into a full-blown microsatellite. It has been proposed that 4 for di- and 2 for tetranucleotides are the minimum number of repeats that are required for pro-STR to develop into microsatellites (Messier et al., 1996).

1.4.4. Coding Microsatellites

Microsatellites (MS) are distributed randomly in the genome, thus they can be located in the coding, regulatory, intronic or intergenic sequences. When they are found in coding or regulatory sequences, microsatellite alterations are most likely to have functional consequences. Mononucleotide microsatellite instability in some target genes (e.g. *TGFBRII*, *BAX*, *MSH3* and *MSH6*) has been found to be associated with the transition from early to late cancer stage (Yashiro et al., 2010). Recently, alterations of dinucleotide MSs in the 3'UTR of the microsomal prostaglandin E synthase-1 (*mPGES-1*) gene were shown to alter the expression of this gene in CRC (Cherukuri et al., 2015).

1.4.5. Microsatellite instability

Under normal circumstances, the mutation rate of the human genome is very low (down to 1.3×10^{-8} per nucleotide per generation). The mutation rate can be defined as the number of mutations per a specific genomic region with a specified length (e.g. megabase) or per time (e.g. cell generation) (Roberts and Gordenin, 2014). The mutation rate in microsatellites is higher than normal sequences (up to 1.2×10^{-3} per locus per gamete per generation) due to the high frequency of DNA polymerase slippage (Weber and Wong, 1993, Fan and Chu, 2007).

Initially, Ionov and colleagues observed that colorectal cancer patients who have somatic mutations in particular poly (dA, dT) sequences have additional mutations in other simple DNA repeats. They initially named this condition as "mutator mutation" or "mutator phenotype" (Ionov et al., 1993). These terms correspond to what is currently known as "Microsatellite instability".

Microsatellite instability first became implicated in hereditary colorectal cancer when an anonymous repetitive marker on chromosome 2p was genetically linked to the occurrence of early onset colorectal cancer (Peltomaki et al., 1993). At the same time, another related locus on chromosome 3p was mapped by another group of researchers (Lindblom et al., 1993). These findings were contributory to the discovery of *MSH2* and *MLH1* genes, respectively.

Microsatellite instability (MSI) has been observed in up to 96-100% (Moslein et al., 1996, Mueller et al., 2009) of hereditary non polyposis colorectal cancer (HNPCC) where one or more of the *MMR* genes (*MLH1*, *MSH2*, *MSH6* and *PMS2*) is mutated. It has also been detected in up to 15% of sporadic colorectal cancers, but in these cases, contrary to HNPCC, MSI is almost always caused by promoter hypermethylaion of the *MLH1* gene (Boland, 2000). However, MSI was reported to be a rare event in rectal cancers and when found, most likely refers to hereditary background (i.e. Lynch syndrome) (Nilbert et al., 1999, de Rosa et al., 2016).

1.4.6. Mechanisms of microsatellite instability

It is not clearly understood what is the exact mechanism that lies behind the development of microsatellites, however 4 possible mechanisms have been proposed:

1- **Unequal crossing over**: This results in transfer of large tracts of satellite DNA between homologous chromosomes during recombination (Schug et al., 1998).

2- **Poly A extension of Retrotranscripts**: Some microsatellites are likely to be products of the extension of the 3' poly A tail retrotranscripts after retroposition. This suggestion has been supported by the association of poly A microsatellites and transposable elements. However, this is not always true and at best, it only explains a single type of microsatellite (Nadir et al., 1996).

3- Repetitive DNA usually assumes a non B DNA conformation; which is more mutation-labile (Wang et al., 2008).

4- **Replication slippage**: It was postulated that slippage of DNA polymerase over a repetitive sequence is more likely to happen. This slippage leads to

generation of extra bases that form a loop and, if not eliminated, will be incorporated in the subsequent replication (Kornberg et al., 1964).

Usually, mismatches occur during DNA replication by inserting a different (mismatched) base into the new DNA strand and, driven by the proofreading function of DNA polymerase and MMR proteins, these mismatches are corrected and the proper base inserted into the DNA strand. In the presence of intact *MMR* genes, these trivial changes are eliminated. If, for any reason, these erroneous bases are not corrected, the accumulation of these errors in repetitive sequences results in **microsatellite instability**. It has been found that *MMR* gene mutations (specifically, *MLH1*, *MSH2* and *PMS1*) increase instability 100-700 fold. Moreover, loss of the proofreading function of DNA polymerase has been found to have less impact on the instability than *MMR* mutations (Strand et al., 1993).

1.4.7. Factors affecting the microsatellite mutation rate

Several factors have been suggested to influence the mutation rate of microsatellites. These are:

1- **Repeat number**: In human, the slippage rate is exponentially increased with the repeat number (Lai and Sun, 2003).

2- **Repeat unit**: Dinucleotides have been found to mutate more frequently than other kinds of microsatellites. Dinucleotides have also been found to undergo expansion more than contraction compared to tetranucleotides. This may, partly, explain the higher abundance of dinucleotide repeats (Ellegren, 2000). Non-disease causing dinucleotide repeats have higher mutation rates than trinucleotide microsatellites, while disease relevant trinucleotide repeats show mutation rates of up to 7 times higher than tetranucleotides (Chakraborty et al., 1997).

3- **Base composition** of repeat unit: It has been shown that Poly G/C homopolymers have a higher mutation rate, even in the context of intact *MMR* genes (Boyer et al., 2002).

4- **Flanking sequences**: The mutation rate of microsatellites is also affected by the sequence context within which the microsatellite is located e.g. high GC in the surrounding sequences reduces the mutation rate (Glenn et al., 1996).

5- **Age**: microsatellite instabilities can accumulate with time resulting in higher incidence in older age groups (Brinkmann et al., 1998).
6- **Sex**: microsatellite mutations were found to behave in a sex-dependant manner with a higher mutation rate in male than female (this might be due to the higher number of replication in spermatogenesis than oogenesis) (Brinkmann et al., 1998, Ellegren, 2000).

1.4.8. MSI testing:

Direct MSI testing is performed using a PCR based technique to compare the fragment lengths of target microsatellites in cancer tissues with those of matched normal tissues. From 1997, MSI testing in diagnostic settings depended on a consensus panel, called the Bethesda panel, which was composed of two mononucleotide markers (BAT-25 and BAT-26) and three dinucleotide markers (D2S123, D5S346 and D17S250). Using that panel, tumours with 2 or more unstable markers were classified as MSI-High (MSI-H), those with one unstable marker are classified as MSI-Low (MSI-L) and those with no unstable marker are classified as MS- Stable (MSS) (Rodriguez-Bigas et al., 1997). To improve sensitivity of the Bethesda panel, revised guidelines have suggested the expansion of that panel to include more markers that are known to be less polymorphic (quasimonomorphic markers) (Umar et al., 2004). Currently, in diagnostic practice, most centers rely on a commercially available panel of quasimonomorphic mononucleotide markers sold by Promega (Promega, Madison, WI, USA). The fact that most people are homozygous at these markers means that tumours can be tested without the need for normal control tissue, thereby cutting costs.

1.5. MSI and its impact on prognosis in CRCs

In addition to the role of MSI status in diagnosis, it has a role in prognosis. MSI-H CRCs are found to be poorly differentiated, less aggressive and have a better prognosis (Saridaki et al., 2014). Moreover, MSI-H CRCs have been found to have a high risk of loco-regional recurrence (Søreide et al., 2009) and stage I MSI-H CRCs are unlikely to show lymph node metastasis (Kang et al., 2015). Therefore, stratification of patients for MSI status could affect the post-resection surveillance (endoscopy vs radioimaging) and prediction of lymph node involvement.

In a meta-analysis, MSI- H colorectal cancers were found to have a 15% better prognosis compared to MSS tumours (Popat et al., 2005). It has been reported that MSI-H status is an independent positive prognostic factor in stage II CRCs following surgical removal (Merok et al., 2013). In 2011, a large study concluded that a previously treated patients with stage II and III MSI-H CRCs have a better prognosis with improved disease free and overall survival (DFS and OS) (Sinicrope et al., 2011). The exact aetiology underpinning the better prognosis in MSI-H patients is not well understood, however, *DCC*, *TP53* and *KRAS* gene mutations (which are all known to be associated with poorer prognosis in CRCs) are less common in MSI-H cancers (Klump et al., 2004). Moreover, MSI-H associated CRCs have been shown to be heavily infiltrated by lymphocytes that might reflect immune-mediated mechanisms which may contribute to the favourable prognosis (Linnebacher et al., 2001).

The prognostic importance of MSI status appears to depend upon cancer stage. MSI-H is observed in 14-21% of stage II CRC and is associated with a good prognosis (Merok et al., 2013, Klingbiel et al., 2015), 7-12% of stage III CRCs with less prognostic benefit (Klingbiel et al., 2015) and only 4% of stage IV CRCs where it is associated with a poor prognosis (Hoffmeister et al., 2013, de Cuba et al., 2015). Based on these findings, it seems that there is an inverse relationship between the prognosis and stage of disease in MSI-H CRCs. The co-existence of other genetic mutations with MSI-H CRCs might be an underlying reason for this difference. *BRAF* mutations occur in about 40% of MSI-H CRC and it infers a poor prognosis. The coexistence of *BRAF* mutations might be the reason for the poor survival prediction in advanced MSI-H CRCs (Tran et al., 2011). It has been recommended therefore to test both *KRAS* and *BRAF* in MSI-H CRCs when a prognostic stratification is required (de Cuba et al., 2015).

Elevated microsatellite alterations at selected tetranucleotide repeats (EMAST) was found to be elevated in MSI-H metastatic CRCs, and its existence associated with poor prognosis (worse overall survival) (Birgisson et al., 2015). Moreover, the existence of other genetic alterations (e.g. 1p36 deletion) was found to be associated with higher risk of dissemination in MSS, rather than MSI-H, CRCs (Mayrhofer et al., 2014).

Finally, as *MMR* gene mutations have been suggested to be an early event during colorectal carcinogenesis, it has been suggested that testing the colorectal adenomas obtained from routine follow up colonoscopy for MSI might be a useful tool for early detection and determination of surveillance option (Loukola et al., 1999).

1.6. MSI and its impact on treatment of CRCs

MSI status, to some extent, plays a role in choosing effective and safe chemotherapies in specific cancers. From the molecular point of view, it has been suggested that an intact *MMR* system is required to effectively induce apoptosis on exposure to 5-fluorouracil (5-FU) (Carethers et al., 1999). The MSI status has long been used to be a limiting factor of responsiveness to 5-FU (no response, or it may worsen the condition). Ribic et al (2003) concluded that Flourouracil (5-FU) based chemotherapy was effective in stage II or stage III MSS or MSI-Low, but not MSI-H colorectal cancers. In a meta-analysis conducted in 2009, stage I and II MSI-H CRC patients did not show a significant difference in both recurrence free and overall survival, whether or not they received chemotherapy (Des Guetz et al., 2009). In practice, all patients with stage III CRC, and some high risk stage II CRCs, receive FU-based adjuvant therapy, but in view of the above data, it has been proposed that MSI-H stage II CRCs should be excluded from this therapeutic scheme (de la Chapelle and Hampel, 2010).

Some drugs seem to be selectively active in MSI-H CRCs. Methotrexate (MTX), a dihydrofolate reductase (DHFR) inhibitor, is a chemotherapeutic agent that is used to treat many human cancers including CRCs. MTX was found to be an effective therapeutic choice, particularly in those with mutated *MSH2* (Martin et al., 2009). Moreover, the addition of bevacizumab (antiangiogenic therapy) to the standard oxaliplatin-based therapy has significantly improved survival in stage II & III *MMR* deficient CRCs (Pogue-Geile et al., 2013).

The finding that MSI-H CRCs have a high density of lymphocyte infiltration has been suggested to be a basis for a promising immune–mediated, MSI-targeted therapy for HNPCC patients and their at risk relatives (Linnebacher et al, 2001). A recent study suggested that the accumulation of coding microsatellite instabilities results in synthesis of neoantigens which attract more CD8 tumour infiltrating lymphocytes (TILs), opening the window for immune-modulating therapies, both in

Lynch syndrome and sporadic MSI-H CRCs (Maby et al., 2015). Despite the fact that MSI-H CRCs usually show heavy lymphocytic infiltrations, they are not naturally eradicated, perhaps due to the associated immune checkpoint upregulations e.g. (Programmed death receptor-1) *PD-1*, (Programmed death ligand -1) *PDL-1*. Thus, developing new immune modulator therapies was suggested to be useful for MSI-H CRCs (Xiao and Freeman, 2015, Llosa et al., 2015).

A further potential impact of MSI testing is that knowing the *MMR* status prior to surgery might change surgical intervention because of the associated high possibility of metachronous cancer in d*MMR* CRCs. Extended resection (rather than segmental resection) in such patients has been found to reduce the risk of recurrence (Aronson et al., 2015).

In approximately 80% of MSI-H CRCs, the underlying reason is hypermethylation of the *MLH1* promoter (Lynch et al., 2009), giving rise to a distinct type of CRCs called CIMP (Cytosine Islands Methylation Phenotype) positive CRCs. The development of metachronous CRC after right hemicolectomy is uncommon in CIMP+ MSI-H CRC, while Lynch patients (who are CIMP- MSI-H) are still at increased risk of metachronous cancer after segmental resection (Messick et al., 2014). Thus, classifying patients according to MSI status could have a significant impact on subsequent surgical decision.

Quite recently, a landmark study concluded that the use of pembrolizumab (PD-1 blocking agent) showed observable clinical benefit in *MMR* deficient tumours (Le et al., 2015). This will likely open the window to development of a more target therapy in this particular group of tumours in the near future. Furthermore, the ongoing trials to develop a frameshift peptide (FSP) vaccines in MSI-H tumours represent a unique preventive scenario in this well-defined group of patients (Kloor et al., 2015).

Although the current testing strategy performs well in classifying CRCs into MSI-H, MSI-L and MSS, it has several notable limitations:

1) It is done now using a low throughput approach and usually complicated by a convoluted subjective interpretation based mainly on the visual inspection of fragment profiles.

2) It shows suboptimal sensitivity and specificity in detecting patients with *MLH1* and *MSH2* gene mutations (91% and 90%, respectively) *(Berg et al., 2009).*
3) MSI-H is not detected in a number of colorectal cancer cases with *MSH6* (or *PMS2*) germline mutations (55-77% sensitivity and 90% specificity) (Lynch et al., 2009, Berg et al., 2009).
4) The low sensitivity to detect MSI in *MMR* deficient tumours other than CRC (85.7%) (Kuismanen et al., 2002).

The majority of colorectal cancers are MSS, but a substantial body of evidence suggests that all colorectal cancers have some degree of instability which cannot be detected unless a satisfactory number of markers is used at least to differentiate between the MSI-L and MSS groups of CRCs (Laiho et al., 2002). A study conducted in 2013 showed that using a hexapanel has a superior sensitivity compared to the currently used pentaplex panel in terms of both detecting *MSH6* mutated tumours (96.7% vs 84%) and *MMR* deficient non-colonic tumours (92.9% vs 85.7%) (Pagin et al., 2013).

As a result of the clinical benefits of MSI detection and its impact on the subsequent choice of treatment, several studies and guidelines have recently recommended offering the MSI test for all newly diagnosed CRCs (Berg et al., 2009, de la Chapelle and Hampel, 2010, Vasen et al., 2013, Loughrey et al., 2014, Kloor et al., 2015). It is desirable, therefore, to develop a new MSI testing strategy accurate enough to satisfy these evolving needs on a large-scale basis.

The advent of next generation sequencing (NGS) has opened new research and diagnostic avenues to adopt genetic tests on a large scale. In the last few years, researchers have started to look at the possibility of using high throughput techniques to test for MSI. In 2012, a comprehensive analysis conducted by the cancer genome atlas network (TCGA), showed, in part, that microsatellite instability can be detected using the NGS approach (Cancer Genome Atlas, 2012). One year later, a large-scale genome and transcriptome-wide study (Yoon et al., 2013) demonstrated the impact of MSI on the expression profile (it was associated with downregulation of 139 genes) in human gastric cancers, despite the majority of these instabilities (90.5%) being localized to the noncoding regions (UTRs). This study also showed that MSI could be detected at a huge number of repeats of varying length,

suggesting that sequence typing of short repeats may be a viable alternative option to current MSI testing techniques.

The main aim of this study is to develop and validate an NGS based MSI test that can be used efficiently as a screening tool to fulfil the increasing demand for MSI screening. The development of such a test is promising because:

1) It could be significantly cheaper due to the high number of cases being analysed, simplification of methodology, and removal of the need for manual visual analysis of raw data.
2) It could potentially enable MSI testing to be offered routinely as a screening tool.
3) It could help to improve the management of MSI-H tumours, as it would enable efficient identification of Lynch syndrome families (as the current method has suboptimal sensitivity and specificity in diagnosing Lynch patients).
4) It could also have important implications in terms of surveillance of at risk individuals.
5) Being offered in an NGS- based version, it could provide more options in the future to be collated with mutation testing of relevant genes in the same test (e.g. *BRAF*, *MMR* or *KRAS*)

1.7. Clonality and Microsatellite Instability

Cancer cells are usually known to share common features collectively known as “the hallmarks of cancer” like uncontrolled cell division, angiogenesis, evasion of the growth suppression, antiapoptotic features and others (Hanahan and Weinberg, 2011). However, individual cancers are genetically heterogeneous and this heterogeneity reflects its variable biological features. Cancers are usually composed of more than one clone, cells within each clone share the same phenotypic and genotypic characteristics which might be different from those of other clones, creating a state of intratumour heterogeneity (ITH). This intrinsic intratumour variability imposes burdens in terms of both proper diagnosis and management (Michor and Polyak, 2010).

It has been postulated that tumours are initiated from a vulnerable cell after exposure to a carcinogen. Genetic instability of that progenitor cell produces cells with a growth advantage and thus allows the clonal propagation. Most of the

produced cells die due to metabolic or environmental pressure, but occasionally, one cell survives that pressure (because it has additional selective advantage) and, with further division, results in a clone that has the same biological characteristics of the original progenitor cell in addition the acquired selective advantage mutations (Nowell, 1976). According to Nowell's assumption, the tumour mass is derived solely from a single progenitor cell (unicellular origin). However, some tumours might be considered as exceptions from Nowell's model like viral induced tumours e.g. condylomata acuminata (where viral infection might affect the surrounding cells) and hereditary cancers e.g. Neurofibromatosis where a familial gene mutation affects all cells and increase susceptibility to cancer. This concept has been extended later into linear and branched evolution. In linear evolution, it has been proposed that cancer cells will acquire genetic mutations with time and the fittest cell will create the dominant clone. According to the linear approach, most cancer cells belong to a single clone (the fittest clone) and that clone will carry all mutations that happened during the evolution of tumour. The branched model, on the other hand, states that cancer cells acquire different mutations resulting in different individual clones that develop concomitantly, but independently, in the tumour (Polyak, 2008, de Bruin et al., 2013). Consistent with the branched model, tumour genomic analyses revealed the coexistence of many independent clones and at the time of diagnosis, one becomes the dominant subclone. These dominant cell populations were estimated to constitute more than 50% of the tumour mass while other subdominant subclones equally or unequally share the rest of tumour cell populations (Nik-Zainal et al., 2012).

To assess the genetic heterogeneity, a clonal marker needs to be used to trace a cell or a group of cells and compare them with other cells within the same tumour mass or with different tumour lesions. X- Chromosome inactivation was used in preliminary trials. One of the X chromosomes in female mammals would be randomly inactivated during embryonic life in a phenomenon called Lyonization (Lyon, 1961). On propagation, daughter cells will inherit the same pattern of X chromosome inactivation from the parental cells. This concept was used to trace clonal characteristics in tumours. However, this approach is limited to females and would not give further information about the genetic characteristics of tumour cells (Wang et al., 2009).

The clonal development of cancer is associated with mutations in different genes. Mutations that confer phenotypic effect on cancer cells and are positively selected during cancer development are defined as driver mutations, while those that have no clear effect (or neutral) are named as passenger mutations (Stratton et al., 2009). The vast majority of cancer associated mutations are passenger while the minority belonging to the driver class (Vogelstein et al., 2013). A genome wide screening of drivers and passengers was used to assess ITH in many human tumours (Futreal et al., 2004, McFarland et al., 2013, Carreira et al., 2014). However, the distribution of driver mutations does not reflect the proper phylogenetic relationship as these mutations impose selective advantage to cells bearing them (Naxerova et al., 2014). Passenger mutations, on the other hand, represent useful markers to detect early lesions that develop prior to tumour development, but a notable drawback of screening the passengers is the need to screen hundreds of thousands of genomic loci. However, the advent of massive parallel sequencing might help to make such an approach doable in the future (Salk and Horwitz, 2010).

Analysis of specific genomic regions was used as an alternative approach to construct phylogenetic trees. A study conducted in 2014 utilised deep sequencing of X chromosome to assess the phylogenetic relationship in *MMR*-deficient colorectal adenomas (De Grassi et al., 2014). In that assay, all protein-coding genes in addition to selected intergenic segments in X chromosome were analysed in four male patients with colorectal adenomas with matched normal tissues to compare the mutational profile. It was possible, with this approach, to construct lineage relationship in each adenoma, however, such an approach is limited to markers that are located in chromosome X only.

The rapid proliferation of cancer cells is associated with increased DNA replication. In repetitive sequences (microsatellites), the replication fidelity decreases due to the high possibility of polymerase slippage during replication (Lai and Sun, 2003). Microsatellites can, therefore, be used as a genome wide approach to assess the intratumour heterogeneity.

1.7.1. The role of microsatellites in assessment of clonality

The successive accumulation of genetic and epigenetic alterations is crucial for the development of cancer and it has been suggested that the normal mutation

rate together with clonal expansion is enough to allow emergence of these alterations (Tomlinson et al., 1996). However, additional genomic instabilities (whether at the chromosomal or nucleotide levels) were found to be existed in most tumours (Lengauer et al., 1998).

Microsatellites can be used as a biological clock to count the genetic events that occur during tumour development. Replication slippage is the most likely reason for the development of microsatellite instability. Therefore, instability (whether a deletion or insertion) develops during DNA replication, which happens with each cell division. During each slippage, a single base either deleted or inserted, but with intact repair genes, these slippage events are corrected and errors are eradicated from the genome. Based on that one division-one event concept, microsatellite instability can be used to document the time since *MMR* genes mutations have happened.

It is believed that one of the first kind of microsatellites to mutate following *MMR* mutations is the mononucleotide (poly A and poly T) repeats (Ionov et al., 1993, Blake et al., 2001). In yeast, the (A) homopolymers of > 8 bases length are more prone to mutations than shorter tracts. It has been suggested that both polymerase proofreading and *MMR* are able to deal with the alteration of short repeats, while only *MMR* proteins are able to repair mutation in the longer repeats. Therefore, genes containing coding homopolymers are at risk of inactivation when *MMR* genes are mutated (Tran et al., 1997).

Most tumours appear thousands of replications after the loss of *MMR* function. Assuming a rate of 1 division per day, an MSI-H tumour requires a decade to develop after the loss of *MMR* genes compared to sporadic CRCs which requires more than a decade to evolve (Blake et al., 2001). It has been suggested that tumour clones with earlier *MMR* mutations acquire more MSI events and the longer the duration since *MMR* mutations have happened, the higher the frequency of microsatellite mutations. However, different cells, and hence different clones, will have different timing of *MMR* gene mutations. Therefore, these cells manifest variable microsatellites profiles. Moreover, both deletions and insertions could result in an allelic variation for each individual cell and this allelic variation can be used to discriminate between different cells (and hence, different clones) within the same tumour. Shibata *et al* (1996) have suggested that recently developed and adjacent clones are more likely to have a similar (or nearly similar) allelic profile, while the older and spatially distant clones are

likely to show more allelic diversity owing to the accumulations of different microsatellite mutational events.

Microsatellite instability has been used to test the intratumour heterogeneity (ITH) in different human cancers. Dinucleotide microsatellite instability was successfully utilised to investigate the multileneage development in benign and tumour samples from Cutaneous T-cell lymphoma (CTCL) (Rübben et al., 2004). In intestinal metaplasia (IM), which is a premalignant condition of the intestinal type gastric cancer, microsatellite instability was used to investigate the genetic heterogeneity in different lesions from the same patient. Different allelic profiles of the tested microsatellites were observed in the spatially different lesions (Guo et al., 2015). Beggs et al (2013), has extensively investigated the heterogeneity in MSI status in sporadic polyps using a unique approach (gland by gland approach). The study found a heterogenous MSI profile within the same polyp. The MSI in a heterogeneous polyp (MSI +/-) are thought to develop later during progression, while MSI+ polyps were thought to develop MSI earlier in the original adenomatous crypt. The existence of MSI clones within MSS polyps might accelerate carcinogenesis (as the MSI clone will be positively selected) (Beggs et al., 2013).

Similarly, microsatellite instability can be used to compare the clonal characteristics between primary tumours and their metastases. One study tested the MS alterations both in the primary human tumours (lung and bladder cancers) and their distant foci, found that the alterations observed in the cytopathological samples (urine and sputum) and tumour edges, were identical to those in the primary site (Mao et al., 1994).

The existence of genetic heterogeneity in terms of microsatellite instability can be indirectly used to distinguish between benign tumours and those which are potentially malignant lesions. PolyG monnucleoides, for instance, have been used to detect clonal lineages in preneoplastic conditions like ulcerative colitis (Salk et al., 2009).

A secondary aim of this study is to assess the utility of short mononucleotide microsatellites in determining the clonal changes between different specimens of MSI-H CRCs. This will help to determine the lineage relationship between the tested samples and assessing the tumour age since the event of *MMR* loss.

1.8. Aims of study and chapters outlines

This study aims to develop a high throughput MSI testing approach convenient to be used in routine clinical diagnostics. This panel is ultimately aiming to efficiently discriminate between MSI-H and MSS CRC cases. As a secondary aim, part of the study aimed to assess the utility of a panel of short mononucleotides to assess the clonal characteristics in MSI-H CRCs.

In the first result chapter (Chapter 3), a panel of 25 short (7-9bp) mononucleotide markers was assessed across a cohort of 55 CRCs, of them 25 are MSI-H. The aim of this chapter is to assess the most informative markers in terms of differentiation between MSS and MSI-H cases. This chapter also includes my contribution in an initial assessment of a panel of 120 short repeats to find out the informative markers in assessing the MSI status.

In the second result chapter (Chapter 4), the most informative markers from Chapter 3 in addition to informative mononucloetide markers from a parallel study (17 markers in total, 7-12bp in length) are assessed using a large cohort composed of 141 CRC samples (referred from Spain). The work of this Chapter aims to assess the degree of discrimination of that panel of markers, and to establish an informative calling system for classifying samples into MSI-H and MSS.

In the third result chapter (Chapter 5), the calling system that was developed in chapter 4 is validated across an independent cohort composed of 70 CRC samples referred from Edinburgh. The utility of neighbouring SNPs to establish allele specificity is assessed as an additional informative criterion.

In the fourth result chapter (Chapter 6), a panel of 23 short mononucleotide markers and the calling system developed and validated in previous chapters are used to assess the clonal characteristics of both fresh and FFPE MSI-H CRC tumours. For that purpose, different specimens from the same tumour were used.

Chapter 2. Methods

2.1. Ethical approvals

Samples for MSI analysis have been collected under the ethical approval referenced IRAS project ID: 99148 (REC reference: 13/LO/1514) entitled “The use of rapid DNA extraction and genetic testing on silicone nanowires to screen for microsatellite instability in tumour tissue as a matter of routine”. The clinical and pathological data of the collected cases were retrospectively checked, aided by the NHS passport under the reference number REF: LOA/CP.

Amendment of the original approval was applied in order to cover the collection and processing of colorectal cancer fresh tissue (CRC) samples from the Royal Victoria Infirmary (RVI). The amendment was approved on the 10th of March 2015.

The samples that have been referred from abroad were covered by ethical approvals from the original referring laboratories.

2.2. Clinical samples

2.2.1. Tumour samples for MSI assay

1) Eleven tumour DNA samples (5 MSI-H, 6 MSS and one matched normal sample) were used to assess the variability of 120 short repeats. These samples were processed by Dr Lisa Redford (Newcastle University, UK). The 5 MSI-H samples were obtained from the Cancer Prevention Program 2 (CAPP2) Biobank. Microsatellite instability was assessed using MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA).
2) A group of 248 formalin fixed- paraffin embedded (FFPE) tissue curls from different tumours were obtained from the Northern Genetics Service (Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK) in 2014. These tumours have previously been MSI tested using MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA), and their MSI status, clinical and pathological data were obtained from the National Health Service (NHS) database. These samples were classified into microsatellite instability-High (MSI-H) and microsatellite stable (MSS) accordingly. To increase the number of MSI-H tumours, 9 additional DNA samples from MSI-H CRC tumours were

provided from the Northern Genetics Service (Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK).

3) A batch of 201 CRC DNA samples (labelled as S1-201) was provided by the Genetics Service, Complejo Hospitalario de Navarra and the Oncogenetics and Hereditary Cancer Group, IDISNA (Biomedical Research Institute of Navarre, 31008 ESPANA). MSI status was determined for these samples using MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA) and MMR IHC. From them, 141 CRC samples were chosen to be analysed and I was kept blind for their MSI status during analysis.
4) A batch of 100 CRC DNA samples was provided by collaborators in Edinburgh (Dr Mark Arends, Department of Molecular Pathology, University of Edinburgh, UK). These samples were extracted and quantified in the original laboratory and, therefore, sent to us in the form of extracted DNA. The MSI status was determined based on MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA) and MMR IHC. Our research group remained blind for the clinical and pathological data of these samples, but the cohort had approximately equal number of MSI-H and MSS samples.

2.2.2. Tumour samples for clonality assay

1) Thirteen groups of fresh CRC tissue samples were obtained from the Department of Cellular Pathology (RVI, Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK). Each group was composed of 8 fresh tissue samples retrieved from different locations in clockwise orientations within the same CRC tumour. The samples were taken from locations corresponding to 3, 6, 9 and 12 o'clock both by fine needle aspiration using BD Microlance 21-gage needles (BD, New Jersey, United States of America) (to retrieve deep samples from within the tumour mass) and scalpel (to retrieve more superficial tissue samples). The closest margin to the antimesentric border represents the 12 o'clock position. Each tumour sample was accompanied by matched normal tissue sample from the same patient for the purpose of downstream analysis. The normal tissue samples were retrieved from normal mucosa 7-10cm away from the tumour mass to avoid tumour cell contamination. The retrieval of these samples was carried out by Dr Stephanie Needham (Department of Cellular Pathology, RVI, Newcastle Upon Tyne Hospitals NHS Foundation

Trust, UK). In addition, 3 fresh MSI-H CRC tumours were obtained from Dr Lisa Redford (Newcastle university, UK) to increase the number of MSI-H tumours. MSI status of these 3 additional MSI-H tumours was determined based on MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA).

2) The NHS database was mined to identify MSI-H CRC tumours (using MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA)) with lymph node involvement (or those with a multiple tumour samples) that had been referred from the RVI in the time from 2000-2015. The search was limited to the RVI referred cases due to technical and ethical issues. The criteria I was looking for, were to get at least 2 different tumour samples for each patient and/or an involved lymph node (when possible). Eighteen MSI-H CRC's tumours were found to have adequate pathological information (their histopathological reports were examined) and all were referred from the RVI. Overall, a total of 7 FFPE CRC tumours were collected and for each of them, there was at least one additional sample from an involved lymph node or from a different colonic location (caecal, ascending, transverse, descending, sigmoid or rectal). The corresponding slides were requested from the department of cellular pathology and reviewed by a candidate pathologist (Dr Helen Turner, Department of Cellular Pathology, RVI, Newcastle Hospitals NHS Foundation Trust, UK). The pathologist reviewed the slides and selected the appropriate slides fulfilling the requested criteria. The corresponding blocks (from which the nominated slides were retrieved) were then requested and collected from the Department of Cellular Pathology. Four slices of 10μM thickness were cut from each block by our in house microtome (Leica Microsystems GmbH, Wetzlar, Germany) for DNA extraction.

The clinical samples included in this study and their details are summarized in Table 2-1.

	Year	Group	No. of cases	Description	Provider	Form
Samples used in the MSI assay						
1	2014	CRC tumours	11	5 MSI-H tumour sample obtained from CAPP2 and 6 MSS tumours in addition to a single normal sample	CAPP2	DNA
2	2014	Tumours	248	Composed of different cancers and all have been MSI tested previously	NGS/ NUTH	FFPE curls
3	2014	MSI-H CRCs	9	These are 9 additional samples to increase the number of MSI-H tumours.	NGS/ NUTH	DNA
4	2015	CRC tumours	201	141 samples were chosen for analysis. Almost half of them are MSI-H and the rest are MSS samples	Spain	DNA
5	2015	CRC tumours	100	50 MSI-H and 50 MSS samples	Edinburgh	DNA
Samples used in the clonality assay						
1	2015	CRC tumours	13	Each group represents a single CRC tumour. For each CRC tumour, 8 specimens in addition to 1 matching normal sample were provided.	RVI/ NUTH	Fresh tissue
2	2015	MSI-H CRC tumours	3	These are 3 MSI-H tumours prepared and extracted previously	NU	DNA
3	2016	MSI-H CRC tumours	7	At least 2 samples from 2 different locations were provided for each tumour	RVI/ NUTH	FFPE blocks

Table 2-1: Colorectal cancer tumours that were analysed in this study and their provider. **CAPP2**= Cancer Prevention Program, **NGS**= Northern Genetic Service, **NUTH**= Newcastle Upon Tyne Hospitals, **RVI**= Royal Victoria Infirmary hospital, **NU**= Newcastle University.

2.3. DNA Extraction

2.3.1. DNA extraction from FFPE tissue samples using the BiOstic FFPE Tissue DNA Isolation Kit

FFPE samples were provided either as sliced curls or as paraffin blocks. Tissue blocks were sliced using microtome (Leica Microsystems GmbH, Wetzlar, Germany) to prepare the required slices. For both groups, DNA was extracted using the Bi*O*stic FFPE Tissue DNA Isolation Kit (MO BIO Laboratories, CA, USA) following the manufacturer's instructions. Briefly, samples were incubated with an optimised wax melting solution and proteinase K at 55°C for 2 hours, then the lysate was incubated in 90°C for an hour to remove the cross links and allow for successful PCR. Finally, DNA was eluted in 100µl of elution buffer (10 mM Tris pH 8.0) and the extracted DNA was stored at -20°C until it was used in subsequent analysis.

2.3.2. DNA extraction from Fresh tissue samples using the ReliaPrep™ gDNA Tissue Miniprep System kit

Immediately upon collection, fresh tissue samples were kept at -20°C until they were processed for DNA extraction. The time from collection to DNA extraction was kept as short as possible (ranging from 1 hour-1 week). DNA was extracted from the fresh tissue samples using the ReliaPrep™ gDNA Tissue Miniprep System kit (Promega, Madison, WI, USA) according to the manufacturer's instructions. Briefly, the frozen tissue sections were placed in a 1.5ml tube containing Phosphate Buffer Solution (PBS) and lysed by 20µl Proteinase K. The lysate was then incubated in 56°C for 2 hours and finally, eluted in 100µl of Nuclease-free water. The extracted DNA was kept at -20°C until it was used in subsequent analysis.

2.4. DNA Quantity and Quality Assessment

2.4.1. Quantitative Assessment

2.4.1.1. Qubit Fluorometer

DNA was quantified using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) using the high sensitivity dsDNA assay kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA) according to the manufacturer's instructions. Briefly, 2µl of DNA was mixed with 198µl of a dsDNA HS reagent and Qubit dsDNA HS buffer to determine the DNA concentration.

2.4.1.2. QIAxcel Automated Electrophoresis

The PCR products were quantified by a QIAxcel® automated capillary electrophoresis system (Qiagen, Hilden, Germany). The quantification was performed using a QIAxcel DNA screening kit (2400) according to manufacturer's instructions.

2.4.2. Qualitative Assessment

DNA quality was assessed using the Agilent® Bioanalyser (Agilent Technologies, CA, USA). The quality assessment was carried out using the Agilent High Sensitivity DNA Kit (Agilent Technologies, CA, USA) following the manufacturers' instructions. To check DNA integrity, DV100 and DV200 (which refer to the percentage of DNA with a size ≥100 and 200 respectively in the tested sample) were used.

2.5. Primer Design

Primer design was performed using a combination of Primer-BLAST: http://www.ncbi.nlm.nih.gov/tools/primer-blast/ and Primer 3 (Rozen and Skaletsky, 1999). All primers were checked for sequence specificity using UCSC Genome Browser's *in silico* PCR online tool (Kent et al., 2002).

Primers were designed to be compatible with one of 2 library protocols:

1) Tagmentation based library preparation: The first set of primers was specifically designed to amplify a genomic segment of about 300bp that is compatible with the tagmentation based library preparation protocol. All primers were tested using a normal control DNA sample to assess their function.
2) Tagmentation free library preparation or direct amplicon sequencing: The second set of primers was optimized to amplify a genomic sequence of about 150bp and was used in the Tagmentation- free library preparation protocol. For these primers, overhanging adapter sequences were directly incorporated into the 3' end of primers. The adapters sequences were designed based on the technical data listed by Illumina (https://support.illumina.com/content/dam/illuminal support/documents/ documentation/ chemistry_documentation/16s/16s-metagenomic-library-prep-

guide-15044223-b.pdf) (Illumina, California, USA). Both the primers and their incorporated adapter sequences were then ordered from Metabion (Metabion international AG, Steinkirchen, Germany).

2.6. PCR

2.6.1. Amplicon generation by PCR for MiSeq analysis

Monoplex PCR was carried out to generate all amplicons from all samples in this work. The high fidelity Herculase II Fusion DNA polymerase (Agilent Technologies, CA, USA) was used for that purpose because it can replicate repetitive sequences with a very low error rate (Fazekas et al., 2010). For each reaction, a total reaction volume of 25µl was used which included 17.25µl of dH2O, 5µl 5X reaction buffer, 0.25µl of dNTP (100 mM), 0.25µl Herculase II polymerase, 0.63µl of both forward and reverse primers (10µM) and 1µl of genomic DNA in 10 mM Tris pH 8.0 elution buffer (10-50 ng/µl). PCR was performed using a SensoQuest® thermal cycler (Sensoquest, Goettingen, Germany) with an initial denaturation at 95^{o}C for 2 minutes, followed by 35 cycles of 95^{o}C for 20 seconds, 58^{o}C for 20 seconds and 72^{o}C for 30 seconds and a final extension at 72^{o}C for 3 minutes. The annealing temperature of 58°C was used for all primer pairs except two primer pairs; IM16-9 and GM14-11, in which 57°C was used as an optimal annealing temperature.

2.6.2. PCR amplification for fragment analysis

Genomic DNA from fresh tissue samples was amplified using the MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, USA). For each PCR reaction, 5.85µl of dH2O was mixed with 1µl of Gold STAR 10X Buffer, 1µl of MSI 10X Primer Pair Mix, 0.15µl of AmpliTaq Gold® DNA polymerase (5u/ µl) and 2 µl of genomic DNA (1-2 ng/µl). The PCR amplification was run on a SensoQuest thermal cycler (Sensoquest, Goettingen, Germany) following the PCR program shown in Table 2-2.

PCR program	Temperature	Time	Cycles
Preinitiation denaturation	95°C	11 min.	1
Preinitiation denaturation	96°C	1 min	1
Stage 1 (denaturation)	94°C	30 sec.	10
Stage 2 (Annealing)	68°C- 58°C (0.53°C/sec)	hold for 30 sec.	
Stage 3 (Extension)	ramp to 70°C in 50 seconds (0.24°C/sec)	hold for 1 min.	
Stage 1 (denaturation)	90°C	30 sec.	22
Stage 2 (Annealing)	Ramp to 58°C in 60 seconds (0.53°C/sec)	Hold for 30 Sec.	
Stage 3 (Extension)	Ramp to 70°C in 50 seconds (0.24°C/sec)	Hold for 1 min.	
Delay (Post extension)	60°C	30 min.	1
Hold	4°C		

Table 2-2: The PCR program that used to generate products for fragment analysis.

2.7. Post-PCR detection

2.7.1. Gel Electrophoresis:

Agarose gel electrophoresis was used to check PCR products when the number of products is low. PCR products were run in 1.5% agarose gel, which was prepared by dissolving 1.5 g of agarose (NBS Biologicals, Cambridgeshire, UK) in 100ml of 1X Tris Acetate buffer (which is prepared from 0.04M TRIS acetate and 0.001M EDTA in water) and melted in the microwave for 2 minutes. 10µl of GelRed Nucleic Acid Gel Stain 10,000x (Biotium, California, USA) was mixed with the gel to visualise the products. 7µl of 2X loading dye (Promega, Madison, USA) was mixed with 7µl of PCR product prior to loading into the gel wells. DNA ladder was prepared by mixing 10µl of 1Kb DNA ladder (Promega, Madison, WI, USA) with 15µl of dH2O and 25µl of 2X Blue/Orange loading dye (Promega, Madison, WI, USA). The gel was run at 90 volts for 1 hour using electrophoresis system (BIO-RAD, CA, USA). PCR products were visualised using a GelDoc-It™ (UVP, CA, USA) documentation system.

2.7.2. QIAxcel Electrophoresis

QIAxcel automated electrophoresis system (Qiagen, Hilden, Germany) was used according to the manufacturer's instructions to check PCR products when the

number of products is high. All PCR amplicons in this study were visualised on the QIAxcel system, which adds further advantage of product quantification.

2.8. Fragment Analysis

For MSI testing by fragment analysis, 2µl of PCR products from section 2.6.2 were mixed with 11µl of Hi-Di Formamide (Thermo Fisher Scientific, Waltham, Massachusetts, USA) and 1µl of Internal Lane System 600 (ILS 600). The whole mix was denatured at 95°C for 3 minutes, followed by an immediate chilling for another 3 minutes. The products were then analysed by the Genetic Analyzer 3130xl (Thermo Fisher Scientific, Waltham, Massachusetts, USA). When run is finished, the data were exported and analysed by the Gene Mapper (Thermo Fisher Scientific, Waltham, Massachusetts, USA) or GeneMarker (Softgenetics, State College, PA, USA) software.

2.9. Library preparation and High Throughput MiSeq Sequencing

2.9.1. Amplicon Pooling

All amplicons for each sample were quantified by QIAxcel electrophoresis system and pooled at approximately equal concentrations. For some amplicons with very low concentrations, the entire product volume was used for pooling. Amplicons that failed to be generated by PCR were not pooled.

2.9.2. Clean-Up of pooled amplicons

The pooled PCR products from each sample were purified using AMPure clean up kit (Beckman Coulter, CA, USA) following the manufacturer's instructions. Briefly, a specific volume of AMPure magnetic beads (=pooled amplicon volume x 1.8) was used for each sample. This was followed by incubation at room temperature for 5 minutes and wash with 70% ethanol. Finally, the purified products were eluted in 50µl of dH2O. AMPure clean-up is the recommended clean up method by both the Nextera XT DNA and 16S metagenomic sample preparation protocols which are the library preparation protocols that used in this study.

2.9.3. Size determination

The lengths of purified products were measured using QIAxcel automated electrophoresis (Qiagen, Hilden, Germany). This initial size determination was used for comparison with the product length in the subsequent steps.

2.9.4. Barcoding of pooled amplicons

Each sample was tagged with a unique set of 8 base indexes (index i7 and index i5) by a reduced cycle PCR following the Nextera XT DNA library preparation protocol (Illumina, San Diego, CA, USA) for the longer amplicons (~300bp) and following the 16S metagenomic sample preparation protocol for the shorter amplicons (~150bp). For both protocols, library was prepared as recommended by manufacturer with minor modifications such as using QIAxcel instead of the Agilent Bioanalyzer to compare the change in band sizes of pooled amplicons before and after barcoding.

Two sets of indexes were used; the first set is Nextera XT index kit **FC-131-1001** (96 index, 384 samples) (Illumina, San Diego, CA, USA) which consisted of the i5 indexes (S502- S508 and S517) and the i7 indexes (N701-N712). 5µl of each index (i5 and i7) were mixed with 15µl of the Nextera proprietary master mix and 5µl of the cleaned amplicons. These indexes were distributed so that to give a 96 unique combinations of barcodes and incorporated with the amplicons by a 12 cycle PCR program with a pre-heating at 72°C for 3 minutes, followed by an initial denaturation at 95°C for 30 seconds, 12 cycles of 95°C for 10 seconds, 55°C for 30 seconds and 72°C for 30 seconds, and a final extension at 72°C for 5 minutes.

The second index set was the Nextera XT Index Kit v2 **Set D FC-131-2004** (96 indexes, 384 samples) (Illumina, San Diego, CA, USA). This kit was composed of i5 indexes (S513, 515-518,520-522 and 508) and i7 indexes (N701-712). In the PCR reaction, Herculase II Fusion DNA polymerase (Agilent Technologies, CA, USA) was used instead of the 2x HiFi Kappa enzyme (which is the recommended enzyme both in the Nextera and 16S Metagenomic protocols) for its superior fidelity and its relatively low cost. For each individual sample, the Herculase PCR mix was prepared by combining 0.5µl of the Herculase II Fusion enzyme solution, 0.5µl dNTPs (100mM), 10µl of the Herculase PCR buffer and 24µl of dH2O. 5µl of both i5 and i7 indexes were mixed with 35µl of the Herculase PCR mix and 5µl of the cleaned

amplicon. The whole mix was then amplified using a reduced cycle PCR with an initial denaturation at 95°C for 1 minute, then a 10 cycles of 95°C for 30 seconds, 55°C for 30 seconds and 72°C for 30 seconds with a final extension at 72°C for 5 minutes.

2.9.5. Clean Up of the barcoded amplicons and QIAxcel electrophoresis

The PCR products from 2.9.4 were then washed and cleaned up using an AMPure clean up kit (Beckman Coulter, CA, USA) and the success of the barcoding procedure was checked by QIAxcel automated electrophoresis (Qiagen, Hilden, Germany) to visually check for shift in bands sizes compared to the band profiles observed before adding the indexes from section 2.9.3.

2.9.6. DNA Quantitaficaton and Dilution

The DNA concentrations of the pooled products were quantified using the Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) as recommended by the library preparation protocols using 2µl from each pooled sample. The concentrations of the barcoded amplicons were measured in ng/µl, while it is recommended that it be calculated in nM. The conversion from ng/µl to nM was done by the following equation:

Concentration (ng/µl) X 10^6 / 660 (g/mol) X (average Library size) = [nM]

The average library size was calculated by combining the average product sizes for all amplicons after adding the overhang (50bp), adapters (60) and indexes (8) = 118bp). The average library size was estimated to be 430bp (in the tagmentation based protocol) and 260bp (in the tagmentation free protocol). All samples were then diluted to a DNA concentration of about 0.7ng/µl using elution buffer to achieve the required 4nM concentration as recommended in protocols.

2.9.7. Library preparation and MiSeq sequencing

After normalisation of all indexed samples at the required concentration, 5µl of each sample (with an optimised concentration of 4 nM of the barcoded amplicons) was added to a 1.5ml eppendorf tube to create the pooled library, which contained all the amplicons of the all pooled samples. The pooled amplicon library was then chemically denatured (using a freshly prepared 0.2N NaOH) and diluted to an optimal

concentration of 4pM in the first and second experiments and 10pM in the third experiment.

The control PhiX library was prepared in a similar way by dilution to the same amplicon library final concentration (i.e. 4pM and 10pM). Both libraries (pooled amplicon library and PhiX library) were then pooled together and prior to be loaded into the sequencing machine, they were denatured at 96°C for 3 minutes and chilled on ice until analysed by the MiSeq.

Sequencing of pooled amplicons was done on the Illumina MiSeq platform (Illumina, San Diego, CA, USA) using the MiSeq Reagent kit V3 (600 cycles) (Illumina, San Diego, CA, USA) for targeted resequencing with paired end read sequencing (251 cycles for both read 1 and read 2). The Basespace® web based cloud on-site informatics tool (http://www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace.html) was used to monitor NGS data in real time and to store and retrieve the data later on.

2.10. Data analysis

2.10.1. Sequencing data

Initial analysis was done by the MiSeq reporter system. This includes adapter trimming (the removal of the adapter sequences) and demultiplexing. Sequences were then aligned using BWA aligner against GRCh37/hg19 assembly as reference. Sequencing data were retrieved from the Illumine MiSeq machine in the form of FASTQ files and analysed primarily by Dr Mauro Santibanez Koref (Institute of Genetic Medicine, Newcastle University). The data were then processed using the R studio environment (R Core Team) to generate the text file format.

2.10.2. Data visualization

Alignment files (as BAM files) were visualized using the Integrated Genomics Viewer software (IGV) (Robinson et al., 2011).

2.10.3. Variant calling

For variant calling, an in-house caller called Concordant Overlapping Paired Reads Caller (COPReC), was designed and run by Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University). This software was used to retrieve insertions and deletions in the concordant overlapping paired end reads. The

COPReC generates an output data in the form of a table for each homopolymer and its adjacent SNP, which shows sequencing reads for each homopolymer length (variant and wildtype) and the called base of the SNP for each sequencing read that span both the homopolymers and SNP.

2.10.4. Deletion frequency

Deletion frequency for each amplicon was calculated by comparing the sequencing reads of each variant genotype to the total reads from each SNP allele. This was done for all amplicons that have at least 100 paired end sequencing reads from all SNP alleles.

2.10.5. Fisher's Exact test

A 2x2 contingency table was constructed for each amplicon to assess the allelic distribution of the deletion. This was called as deletion with allelic bias. Allelic bias was used as an additional criterion to assess instability. Calculations were made manually by using an online tool to calculate the Fisher's Exact (http://scistatcalc.blogspot.co.uk/2013/11/fishers-exact-test-calculator). Values below 0.0001 were considered as extremely significant, between 0.0001-0.001 were very significant, 0.001-0.05 are significant and those more than 0.05 were considered as not significant.

2.10.6. Threshold setting

Deletion frequencies for each marker were plotted across all MSS and MSI-H samples to generate deletion curves. For each marker, deletion curves in the MSI-H and MSS samples represent sensitivity and specificity curves respectively. Threshold sets were arbitrarily selected, starting from low deletion frequency (0.01) and gradually increased. For each threshold set, the performance of results of our MSI assay was compared to the reported phenotype done by the gold standard assay (MSI Analysis System, Version1.2: Promega, Madison, USA). This yielded in recognition of true positive, true negative, false positive and false negative samples. True positives (TP) are defined as samples that were reported as MSI-H and predicted as MSI-H as well, true negatives (TN) which are samples that have been reported and predicted as MSS, false negatives (FN) which are samples that were

reported as MSI-H and predicted as MSS and false positives (FP) which are samples that were reported as MSS and predicted as MSI-H.

These values were used to generate sensitivity and specificity for each threshold set as follows:

Sensitivity= True Positives / (True Positives + False Negatives)

Specificity= True Negatives / (True Negatives + False Positives)

The false positive and false negative rates were calculated as explained below:

FPR= False positives (FP) / False positives (FP) + True negatives (TN)

FNR= False negatives (FN)/ False negatives (FN) + True positives (TP)

2.10.7. Determination of MSI status

For purposes of classifications, markers that showed deletion frequencies above the threshold value and having an allelic bias were considered unstable.

In the subsequent validation assays, in addition to the aforementioned classification, a new scoring system was applied by giving a score of 1 for the marker that showed a deletion above the length-specific threshold and a score of 2 for those with both a deletion frequency above the threshold and an allelic bias. Samples that have an overall score equal or more than 3 were called as MSI-H and those with an overall score below 3 were considered as MSS.

2.10.8. Constructing the phylogenetic tree for clonality assay

The phylogenetic tree for samples that were assessed in the clonality assay was constructed using the Mesquite software (http://mesquiteproject.wikispaces.com/). The following steps were used to construct a tree:

1) **Data input:** the system requires inputs of Taxa and characters. Tumour specimens were considered as Taxa and the markers were considered as characters in the character matrix. The number of both taxa and characters were specified accordingly. In the project window, a value of 0 was entered for ancestral character and 1 for a derived one. In accordance with this scheme, 0

value was representing stable and 1 represents unstable markers in that particular specimen. The file needs to be saved in order to be used in the subsequent tree construction.

2) **Constructing phylogenetic trees:** after entering the inputs, the software is now ready to construct a phylogenetic tree. This was done by the option (Taxa & Trees> Make New Trees Block from> Tree search> Mesquite Heuristic Search (Add & Rearrange) > Tree length> SPR rearrange> OK> the MAX number of trees was set to 100. Then the tree was constructed based on the data entered in the character matrix. The first tree is usually unrooted, therefore it needs to be manually re-rooted to a specific outgroup (ancestral taxon). The outgroup was selected to be the normal specimen or tumour specimen that has the least number of unstable markers. The rooted tree then was saved by: Tree> store tree.
3) **Construction of a consensus tree:** The high number of constructed trees needs to be reduced to a single representative tree, this is done by constructing a consensus tree. The consensus tree was constructed as follows: Taxa & Trees> Make New Trees Block from> Consensus Tree> store trees> OK> Majority Rule Consensus> Ok> tick the options of consider tree weights and write group frequency list and treat trees as rooted as specified in the first tree> OK. This will bring the consensus tree window, which is again unrooted, so it should be rooted manually by choosing an outgroup as described earlier.

Chapter 3. Establishing a consensus short mononucleotide repeat panel for NGS-based MSI testing

3.1. Introduction and aims

3.1.1. Introduction

3.1.1.1. Microsatellite instability of mononucleotide repeats

Microsatellites (also called simple repeat sequences) are short repetitive sequences scattered throughout the genomes of most complex eukaryotes including human. They can be classified into mono, di, tri, tetra, penta or hexanucloetides depending on the length of the repeat unit. In *MMR* deficient cells, these sequences are susceptible to length alteration leading to a phenotype called Microsatellite Instability (MSI), which represents the hallmark of the mismatch repair gene mutations (*MMR*). Di and Tetranucleotide microsatellites are more polymorphic than other kinds of microsatellites, therefore, they are less suitable for use in the detection of the microsatellite instability (Umar et al., 2004, Sutter et al., 1999). Mononucleotides, on the other hand, show the lowest degree of variability and thus represent a sensitive target to be used in the assessment of microsatellite instability (Sutter et al., 1999, Zhou et al., 1998, Cicek et al., 2011). The earliest finding of mononucleotide repeat mutations in colorectal cancers was reported in 1993 (Ionov et al.), where they found that polyA monotonous repeats mutations were present in up to 12% of colorectal cancers (CRC) and, interestingly, they were associated with a distinct pathological and molecular phenotype. These mutations were found to be inversely correlated to mutations in other genes like *KRAS* and *P53* and metastasis at time of diagnosis, while they were positively correlated with poorly differentiated histology, right sided localisation and tumours in blacks. They concluded that this special kind of alterations in mononucleotide repeats is mediated by mutations that compromise DNA replication fidelity.

Subsequent analyses have established that the frequency of mononucleotide mutations was dependent upon both the length and sequence content of the repeat. (A) homopolymers were found to be more prone to mutation when the number of bases are more than 20bp and their susceptibility to mutation decreases gradually with the shortening of repeat length to become less likely when a repeat is less than 10bp (Parsons et al., 1995). Moreover, *In vivo* studies showed that the polymerase

proofreading function is length sensitive, as this proofreading is highly effective in dealing with errors in short homopolymers, while having a marginal effect on long runs. Mismatch repair genes (*MMR*) seem to be the only proofreaders that can effectively deal with errors in long homopolymers (Tran et al., 1997).

Microsatellite instability has several clinical advantages, colorectal cancers with microsatellite instability was found to have a better prognosis compared to MSS CRCs (Popat et al., 2005). Furthermore, the detection of microsatellite instability carries additional advantage of determining the convenient kind of therapy. MSI-H CRCs are less sensitive to 5 Fluoro-Uracil (5-FU) based therapy (Ribic et al., 2003), while there is evidence of a relatively high sensitivity to methotrexate in those with *MSH2* mutations (Martin et al., 2009). Recently, *MMR* mutated colorectal cancers showed a clinical benefit of immunotherapy pembrolizumab (Le et al., 2015). Additional advantages of assessing the microsatellite instability is the stratification of cancer risk and then screening those with high risk for hereditary cancers (Lynch syndrome).

For all the above mentioned advantages of MSI testing, it is imperative to develop a laboratory test sensitive and specific enough to assess the microsatellite instability.

3.1.1.2. Development and evolution of the MSI testing

MSI was initially tested by using the original panel recommended by the National Cancer Institute (NCI) workshop held at 1998 and it was composed of 2 mononucleotide and 3 dinucleotide markers (Boland et al., 1998). However, additional alternative markers were proposed in the same workshop including long mononucleotide and dinucleotide repeats. The adoption of this panel was based on a validation assay conducted by Ruschoff and colleagues (1997) that assessed the utility of 31 markers. They suggested that whatever the number of the markers used in the assay, instability in ≥ 40% of the markers used is enough to call the case as microsatellite instability- High (MSI-H). However, when a single marker exhibits instability, it was recommended to test an extra five markers to confirm the diagnosis.

Over the next decade, three main caveats were observed in the performance of the original NCI panel, these are:

1) Dinucleotides are highly polymorphic and less sensitive than mononucleotides in detecting MSI-H, thus it was mandatory to test matched normal samples for correct MSI phenotyping.
2) The need to test extra-markers in those with a single unstable marker.
3) The existence of instability in two dinucleotides with absence of mononucleotide instability may result in misclassification of MSI-H cases.

In 2002, a new panel composed of 5 long (>20 bp) mononucleotide markers (BAT26, BAT25, NR-21, NR-22 and NR-24) was tested and found to have a high sensitivity and specificity (Suraweera et al., 2002). Two years later, the NCI held another workshop and recommended a 5 mononucleotide panel as the most sensitive panel compared to the previous NCI panel (Umar et al., 2004).

The current trend of MSI testing relies mainly on PCR amplification of 5 mononucleotide microsatellite (BAT26, BAT25, NR-21, MONO27 and NR-24) panel, followed by a fragment analysis to examine the electropherogram profiles in tumours compared to its matched normal tissue samples. These 5 mononucleotide markers were validated by Bacher et al (2004) and based on this panel, three distinct phenotypes can be concluded, microsatellite instability- High (MSI-H) when 2 or more of the markers shows instability, microsatellite –Low (MSI-L) when a single marker shows instability and microsatellite stable (MSS) when none of the 5 markers show instability. Commercially, those 5 mononucleotide markers were gathered in a single multiplex PCR kit sold by Promega (MSI Analysis System, Version 1.2 kit, Promega, Madison, WI, USA) and this has led to a wide adoption of that panel (Boyle et al., 2014). The modified NCI panel showed a very high sensitivity (95.6-100%) (Goel et al., 2010, Suraweera et al., 2002, Bacher et al., 2004) in detecting microsatellite instability. However, the ultimate gold standard for MSI assay should ideally be the detection of a pathogenic mutation in the *MMR* genes.

Despite the fact that this mononucleotide panel has a high sensitivity and specificity and has improved the practice of MSI testing, it has its own limitations. The use of these longer homopolymers has long been associated with generation of PCR- induced errors that likely complicate the downstream phenotype interpretation. These errors occur due the inefficiency of commercial polymerases to faithfully replicate these longer repeats and manifested in the electropherogram in the form of stutter peaks (Shinde et al., 2003). Another limitation of the current MSI test is the

convoluted and subjective interpretation as it is solely based on the visual inspection of the fragment profile in the electropherogram rather than exploration of the sequence contents within the fragment.

Given the importance of microsatellite instability in diagnosis, prognosis and determination of therapeutic options especially in colorectal cancer, several studies started to recommend MSI testing for all newly diagnosed CRC cases (Julié et al., 2008, de la Chapelle and Hampel, 2010, Vasen et al., 2013, Medscape, 2015). This means testing around 40,000 new CRC cases each year in the UK alone (and more than 1,300,000 new CRC cases worldwide) (Cancer Research UK, 2015). It would not be possible to fulfil these recommendations with the current test with its inherent limitations of low throughput and the generation of the polymerase induced errors in the long homopolymers. Therefore, it is becoming increasingly important to develop an MSI assay robust and accurate enough to be able to cover the huge number of cases that need to be tested.

3.1.1.3. The development of a sequence based MSI test

The advent of next generation sequencing (NGS) has opened new research and diagnostic avenues to adopt genetic tests on a large scale. Repetitive sequences (like microsatellites) potentially represent a caveat in NGS applications as they are more prone to generate sequencing errors and their low sequence diversity might compromise the subsequent analysis (Clarke et al., 2001). However, PCR induced errors have been shown to be dramatically reduced or obliterated by using the commercial polymerases with non- specific dsDNA binding domain (Herculase II Fusion polymerase). This kind of enzymes was shown to have the best proofreading function in homopolymers, approaching error free replication for mononucleotide repeats ≤13bp in length after 35 PCR cycles (Fazekas et al., 2010)

In the last few years, researchers have started to look at the possibility of using high throughput techniques to test for MSI. In 2012, a comprehensive analysis conducted by the cancer genome atlas network (TCGA) to explore the somatic mutations and classification of CRC. That study, showed, in part, that microsatellite instability can be detected using the NGS approach (Cancer Genome Atlas, 2012). Since that time, an emerging body of studies attempted to apply this technique to MSI testing (Salipante et al., 2014, Zhao et al., 2014, Gan et al., 2015). In 2014, our

group has analysed the sequencing data form the low depth whole genome sequencing conducted by TCGA (unpublished data) to investigating the difference in the deletion of short homopolymers only (7-12bp) both in MSI-H, MSS and matched normal samples. This analysis showed that it was possible to infer the microsatellite alterations by NGS platforms and there was a clear difference in the instability between MSI-H tumours and their matched normal counterparts and MSS samples as well.

One of the major obstacles associated with the sequencing of microsatellites is the generation of sequence errors (Clarke et al., 2001). The existence of high frequency single nucleotide polymorphisms (SNP) in the nearby sequences (about 30bp on either side) of the target homopolymers can be used to assess instability. In heterozygous cases, the variant reads that span both the homopolymers and the adjacent high frequency SNP were used to assess the distribution of instability with SNP alleles. The instability that is significantly biased toward one of the SNP alleles is, therefore, more likely to be a real instability compared to sequencing error that is not. This feature is called allelic bias and was considered by our team as an additional classifier for subsequent selection of informative polymorphic repeats.

Our lab extensively mined the data generated from that whole genome analysis undertaken by TCGA to look for the most unstable short repetitive sequences in a selected subset of MSI-H samples compared to matched normal and MSS samples. The initial huge list of the variable repeats was then narrowed down to create a short list containing those markers that were not polymorphic, most variable in the MSI-H cases, have an adjacent high frequency SNP and were sequenced to adequate depth (≥20 reads) both in MSS and MSI-H groups. About 200,000 short (7-12bp) repeats were found to fulfil the above mentioned criteria, of them, 120 markers which primers could be easily generated were used in the subsequent analysis. These 120 markers were initially assessed to examine their variability both in MSS and MSI-H samples and, ultimately, determine the most informative markers among them. These 120 markers were randomly divided between 3 researchers. I have been involved in designing and amplification of 29 markers using 5 MSI-H and 6 MSS tumour samples. The rest of the 120 markers were analysed by Dr Lisa Redford (Newcastle University, UK) and Iona Middleton (Newcastle University). That initial analysis revealed that short homopolymers, sequence based- MSI test is feasible

and, subsequently, a panel of 66 (out of the 120 tested repeats) highly variable repeats was identified for further analysis against a larger group of CRC samples to consolidate the initial finding and find out the most unstable markers amongst them.

3.1.2. Aims

The whole genome analysis revealed a huge number of polymorphic markers, of them, 120 (7-12bp) for which primers could be generated were assessed in the MSI-H samples compared to both MSS and normal samples. This assay established that instability can be assessed in homopolymers and 66 markers were found to be the most variable among them. However, longer (10-12bp) homopolymers were more variable, but associated with more sequencing errors compared to short (7-9bp) repeats. To further investigate the utility of both short and long repeats, the 66 markers were split into two groups with the plan that I would analyse the short ones (7-9bp) using a cohort composed of 55 CRC samples, of which 25 samples are MSI-H, 25 MSS and 5 MSI-L. The overall aims of this chapter are:

- Design and amplify primers for 29 markers (derived from the 120 markers) to be analysed against a small cohort of CRC samples.
- To test a cohort of CRC samples composed of 25 MSI-H, 25 MSS and 5 MI-L samples using the 25 short (7-9bp) repeats identified from the whole genome analysis.
- Establish and assess convenient length- specific thresholds for calling instability based on analysis of sensitivity and specificity for each individual marker.
- Assess the utility of the allelic bias as an additional parameter could be utilised to call instability.
- Determine the minimal number of the most informative markers that can efficiently discriminate between MSI-H and MSS samples.

3.2. Results

3.2.1. Assessment of 29 mononucleotide repeats to identify the most variable repeats

As a part of 120 variable markers retrieved from the whole genome analysis, I initially designed and amplified 29 of these markers using a cohort composed of 5

MSI-H lynch syndrome tumours, 6 MSS and one normal tissue samples as explained above and shown in Table 3-1 and Figure 3-1.

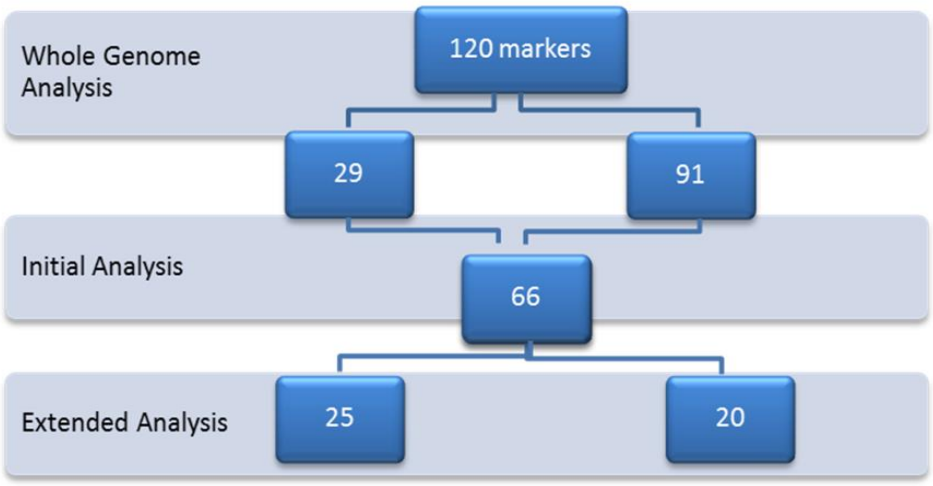


Figure 3-1: Illustration of the overall number of markers and the study workflow. In the initial 120 markers, 29 markers were designed and tested across a cohort of 11 tumour and normal samples.

MSI-H samples	Number	MSS samples	Number	Normal sample	Number
T1	U029	**N1**	169259	**NR**	U096T
T3	U179H03	**N2**	169736		
T4	U179 H12	**N3**	170146		
T5	U303	**N4**	170402		
T6	U312	**N5**	171223		
		N6	169836		

Table 3-1: DNA samples that were used in the initial assessment. 5 MSI-H, 6 MSS and one normal samples were used to be assessed by the 120 markers.

All primers were designed to generate amplicons of around 300bp in size as this is the recommended amplicon length by the Nextera XT DNA library preparation protocol (Illumina, San Diego, CA, USA) which is the kit that was used for library preparation in this experiment. The 300bp amplicons were designed to span both the homopolymer of interest (the marker) and an adjacent high frequency SNP. The minor allele frequency (MAF) of the included SNPs was ranging between 0.05-0.9, and a MAF near to 0.5 was preferred when possible. Approximately 15ng of DNA from each sample was used to generate PCR amplicons using the Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA, USA). The amplified products were then visualized and quantified by QIAxcel (Qiagen, Limburg, Netherlands).

Amplicons for all 120 markers were subsequently pooled, sequenced and analysed. Out of 120 repeats, 66 showed evidence of instability in 1 or more tumours. The results were then analysed by Dr Lisa Redford (Newcastle University, UK). The results of the 29 markers that were designed and amplified by myself showed that for the 7bp markers, there was no clear instability both in MSI-H and MSS tumours as shown in Figure 3-2.

Figure 3-2: The frequencies of variant reads of 6 short (7bp) homopolymers. MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T). Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base) and 1 (insertion of a single base). There is no clear deletion in all markers. (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

For the 8bp group, five markers were designed and tested. Only a single marker (GM09) showed a clear degree of instability in a single MSI-H tumour (U029T) compared to MSS tumours as shown in Figure 3-3.

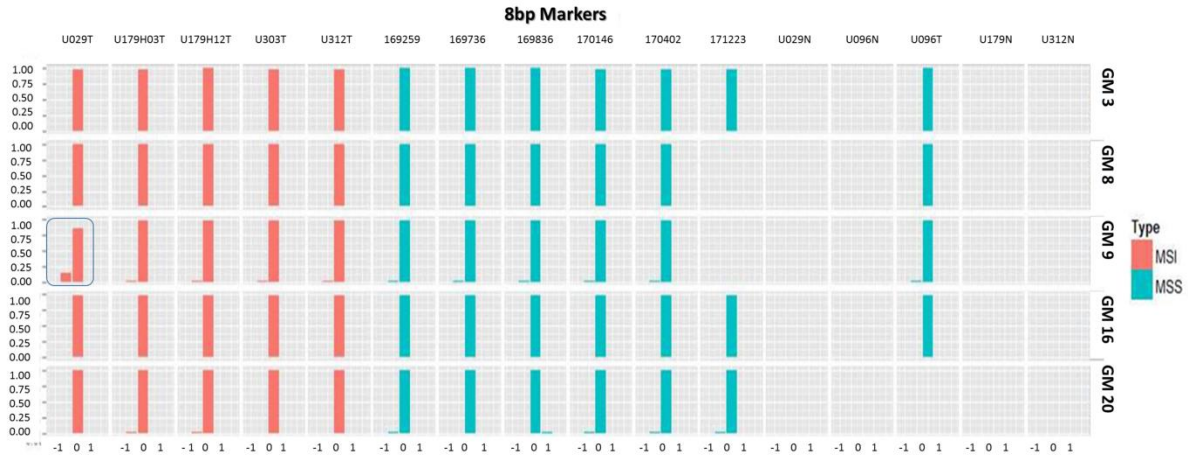


Figure 3-3: The frequencies of variant reads of 5 short (8bp) homopolymers. MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T). Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base) and 1 (insertion of a single base). The marker GM09 shows 1bp deletion in sample U029T (blue square). (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

For the 9bp group, among the 9 markers that were designed and tested, 2 markers (GM11 and GM17) showed a deletion frequency of up to 25% in 2 MSI-H tumours while another 4 markers (GM06, GM10, GM15 and GM23) showed deletion up to 20% in a single MSI-H tumour as shown in Figure 3-4. The remaining 3 markers (GM05, GM21 and M28) did not show deletion in any of the tested samples. None of the markers showed deletion frequency greater than 10% in both MSS tumours and the normal tissue sample.

Figure 3-4: The frequencies of variant reads of 9 short (9bp) homopolymers in MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T).Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base) and 1 (insertion of a single base). The markers (GM11 and GM17) show clear (up to 25%) 1bp deletion in 2 MSI-H samples and another 4 markers (GM06, GM10, GM15 and GM23) showed 1bp deletion frequency (up to 20%) in a single MSI-H sample (blue squares). None of the markers show a clear instability in MSS tumours or in normal sample. (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

For the 10bp group of markers, all the tested markers (4 markers) exhibited instability (up to 20% deletion frequency) in at least 2 MSI-H samples, but on the other hand, all markers showed a low level of instability in the MSS tumours and even in the normal tissue sample as shown in Figure 3-5.

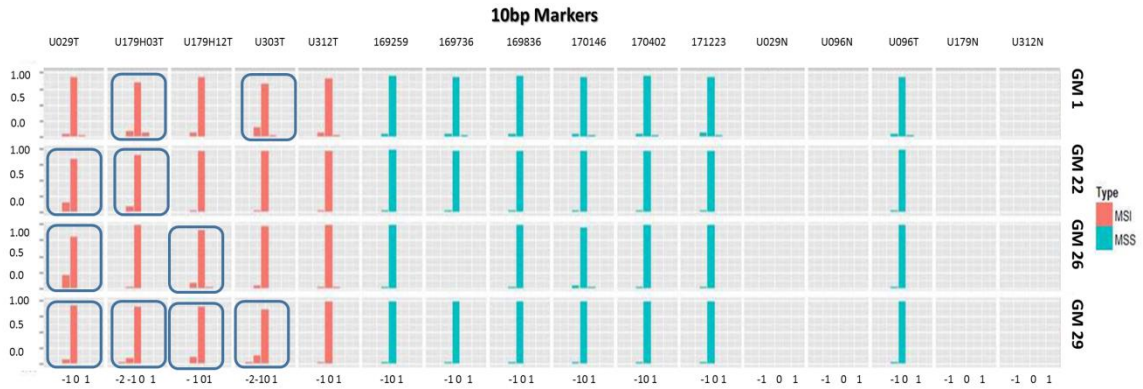


Figure 3-5: The frequencies of variant reads of 4 long (10bp) homopolymers in MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T). Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base) and 1 (insertion of a single base). The marker GM29 show instability up to 15% in 4 MSI-H samples and the other 3 markers (GM01, GM22, GM26) showed deletion frequency (up to 20%) in 2 MSI-H samples (blue squares). Al markers show a low level of instability in MSS tumours and the normal tissue. (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

For the 11bp group of markers, 4 markers were tested. One of the markers (GM07), showed a high degree of instability (up to 70% deletion frequency) in MSI-H samples. Although GM07 showed instability in the MSS tumours as well (up to 15%), instability was higher in the MSI-H samples (up to 70%). The other 2 markers (GM13 and GM14) showed instability in the MSI-H samples (up to 40% deletion frequency) while they didn't show such a high degree of deletions in the MSS group as shown in Figure 3-6. The last marker (GM02) showed instability in both MSI-H group (up to 40% deletion frequency) and MSS and normal group as well (up to 20% deletion frequency). Although the deletion frequency in the MSI-H group is higher than that in the MSS group, the existence of such behaviour indicates that this marker is unlikely to be useful in discriminating between MSI- H and MSS samples.

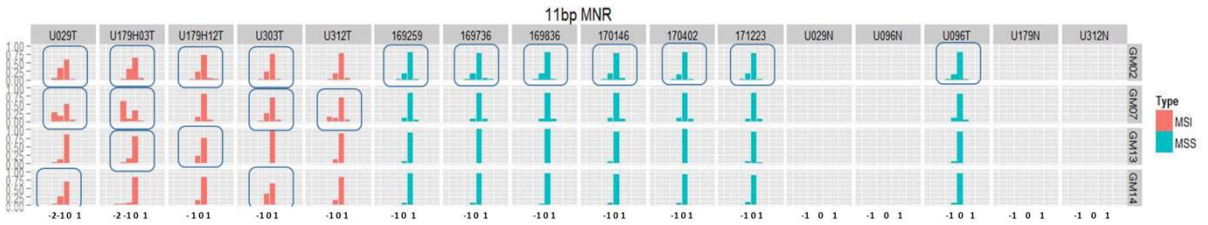


Figure 3-6: The frequencies of variant reads of 4 long (11bp) homopolymers MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T).Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base), 1 (insertion of a single base), -2= deletion of 2bp, 1: insertion of 1bp and 2: insertion of 2bp. The markers (GM07, GM13 and GM14) show instability up to 80% in at least 2 MSI-H samples (blue squares) and they were relatively stable in the MSS samples. The marker (GM02) showed instability both in MSI-H and MSS groups (blue squares). (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

A single 12bp marker (GM18) has been designed and tested. It showed a clear instability (more than 50% deletion frequency) in all MSI-H samples and up to 20% in both MSS tumours and normal tissue as shown in Figure 3-7 indicating that this marker would unlikely to be useful to specifically detect MSI-H samples.

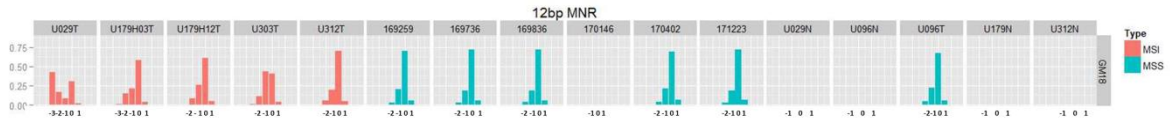


Figure 3-7: The frequencies of variant reads of the 12bp long homopolymers (GM18). MSI-H (red bars), MSS tissue samples (green bars) and normal tissue sample (U096T).Sequencing reads were plotted for the wildtype (0), -1 (deletion of a single base), 1 (insertion of a single base), -2= deletion of 2bp, 1: insertion of 1bp and 2: insertion of 2bp. The markers show instability in all samples (MSI-H, MSS and even the normal tissue) but the instability was higher in the MSI-H samples (more than 50%). (**Y-axis**: The relative frequency of the variant reads, **X-axis**: Allele length, sample IDs are shown in the upper row, the marker IDs are shown in the rightmost column).

From the above results, it was obvious that the variability was length dependant as the short repeats were stable both in MSI-H and MSS groups. The longer repeats (10-12bp), on the other hand, were more variable in the MSI-H samples, but at the same time, were associated with noticeable instability in MSS samples and even in normal tissue. This variability (in MSS and normal samples) is most likely to be PCR induced errors and it is clearly associated with longer repeats compared to the short ones.

By the end of this part of the study, it has been found that 39% of shorter repeats (7-9bp) showed evidence of microsatellite instability in MSI-H samples, compared to 80% of longer repeats (10-12bp). However, longer repeats also

generated more variant reads in normal and stable tumour samples, indicative of sequence error.

The results obtained from this part of the study established that:

1) A short homopolymer sequence-based MSI test is potentially feasible, and a subset of 66 variable repeats was identified, which could be used as a basis for such a test.
2) There is a length-related variation in both instability and sequence error rates.
3) This analysis confirmed that both MSS and normal samples were tested stable, suggesting that sequence-based test may not require matched normal on a routine basis.
4) Extending the analysis using a larger number of tumours would be required to define the optimal combination of repeats, and the criteria for tumour classification.

As the optimal size of repeat for use is unclear, and analysis would require multiple next generation sequencing runs, the subsequent work was split into the extended analysis of longer repeats (8-12bp, undertaken by Dr Lisa Redford) and analysis of 25 short repeats (7-9bp, undertaken by myself).

3.2.2. Analysis of 25 variable short (7-9bp) mononucleotide repeats to assess the criteria for calling instability in CRCs

Having established that it is feasible to utilise short homopolymers in MSI testing, it is essential to define the parameters that can be used to assess instability using a larger cohort of CRC samples. From the 66 most variable markers that were nominated from the previous part of the study (explained above), a batch of 25 of the most unstable short repeats (7-9bp) has been chosen for further assessment using a relatively larger number of samples than the previously tested cohort (which was composed of 5 MSI-H tumours samples, 6 MSS tumours and a single normal sample). The panel of 25 markers was composed of ten (7bp) markers, five (8bp) markers and ten (9bp) markers as shown in Table 3-2. This panel was selected based on the accumulated result from the previous initial analysis done by Dr Lisa Redford (Newcastle University, UK). All these markers showed a high discriminatory power between MSI-H and MSS samples in that initial assay. Out of the 29 markers designed and analysed initially by myself (mentioned in 3.2.1), 6 markers ended up in these 25 markers (1 is 8bp and 5 are 9bp in length).

	Amplicon	Repeat size	SNP1	SNP2	SNP3	Repeat Position
1	LR49-7	7	rs80323298	rs201097746	rs12903384	Chr15: 93619048
2	LR51-7	7	rs8474			Chr10: 51026725
3	IM14-7	7	rs11760281			Chr7: 80104531
4	IM19-7	7	rs72736428	rs186539440	rs4877153	Chr9: 82475001
5	IM43-7	7	rs9981507			Chr21: 32873761
6	IM55-7	7	rs13099818			Chr3: 143253845
7	IM66-7	7	rs147847688	rs141474571	rs4794136	Chr17: 48433967
8	IM67-7	7	rs67082587	rs57484333		Chr7: 22290895
9	LR08-7	7	rs181578273	rs7117269		Chr11: 56546206
10	LR15-7	7	rs56084507			Chr8: 92077210
11	IM59-8	8	rs10156232			Chr8: 108359001
12	LR20-8	8	rs146973215	rs191572633	rs217474	Chr1: 64029634
13	LR46-8	8	rs143884078	rs182346625	rs6040079	Chr20: 10660085
14	IM41-8	8	rs1944640	rs112075239		Chr6: 147948941
15	GM09-8	8	rs6038623			Chr20: 6836977
16	IM16-9	9	rs114923415	rs73367791	rs59912715	Chr18: 1108767
17	LR10-9	9	rs111814302	rs1768398	rs1768397	Chr1: 81591388
18	LR24-9	9	rs192329538	rs1127091		Chr1: 153779429
19	LR21-9	9	rs182900605	rs80237898	rs2413976	Chr15: 50189465
20	LR40-9	9	rs6432372			Chr2: 13447470
21	GM17-9	9	rs666398			Chr11: 95551111
22	GM21-9	9	rs185182			Chr3: 142695339
23	GM23-9	9	rs184237728	rs32123		Chr5: 11345921
24	GM28-9	9	rs4130799			Chr5: 29209381
25	GM11-9	9	rs347435			Chr5: 166099891

Table 3-2: The 25 primers that were used to analyse the short repeats with their repeat length, associated SNPs and genomic positions.

To develop a panel of DNAs from tumours of known MSI status, I first examined DNA from **248** FFPE tissues previously analysed for MSI status by the Northern Genetics Service (as explained in Chapter 2 section 2.2.1). Clinical, pathological and molecular data of the samples were checked retrospectively from the NHS database.

The cases that were diagnosed with primary cancers other than colorectal were excluded from the cohort because I was mainly interested in colorectal cancer

cases. The only exceptions were 2 endometrial cancer samples which were allowed to be included in the study because they were MSI-H.

After this initial filtration, **87** samples (**55** MSS, **27** MSI-H and **5** MSI-L) were selected for further molecular work. DNA has been extracted from all tumours by B*IO*stick DNA extraction Kit (MO BIO Laboratories, CA, USA) following the manufacturer's instructions. The extracted DNA was then quantified by Qubit 2.0 Flourometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA).

As an initial assessment of DNA quality, approximately 15 ng of DNA from each sample was used to generate PCR amplicons using the Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA, USA) as shown in Figure 3-8.

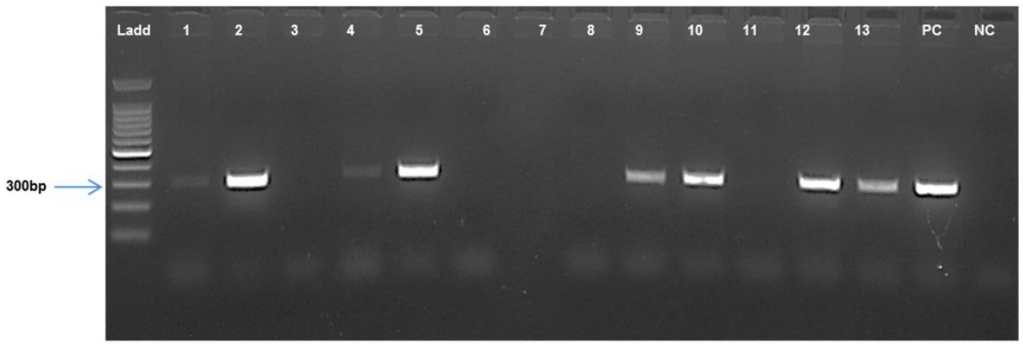


Figure 3-8: Initial check of DNA amplifiability using a 300bp amplicon. Samples in lanes 3, 6, 7, 8 & 11 failed to be amplified. Samples in lanes 1 & 4 showed faint bands while the remaining samples were successfully amplified. (**Ladd**: 100bp-1kb Ladder, **PC**: Positive Control, **NC**: Negative Control)

58.6% (51 out of 87) of the samples were successfully amplified using the 300bp primer set. To investigate the possible reasons for PCR failures, I then amplified the failed samples (those which failed to be amplified with the 300bp primer set) using a 100bp amplicon. Overall, 42% (15/36) of samples which failed to amplify the 300bp amplicon were successfully amplified using the small primer set (100bp) as shown in Figure 3-9. The results of successive amplifications are summarised in Figure 3-10.

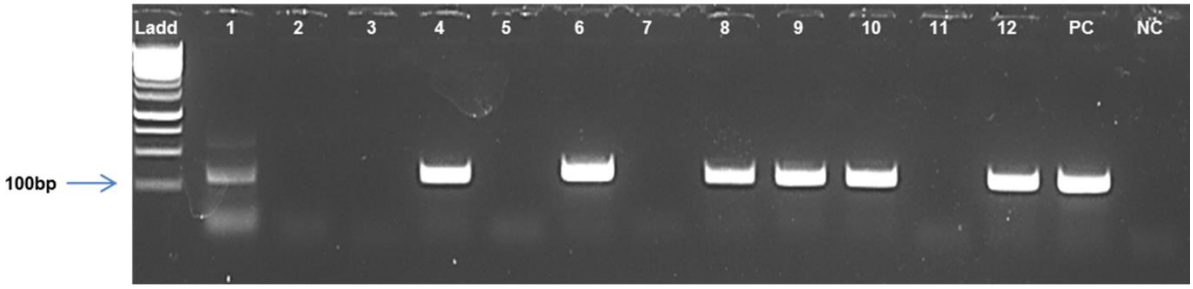


Figure 3-9: Amplifications of DNA samples with 100bp amplicons. Samples in lanes 2,3,5,7 &11 failed to be amplified. (**Ladd**: Ladder, **PC**: Positive Control, **NC**: Negative Control).

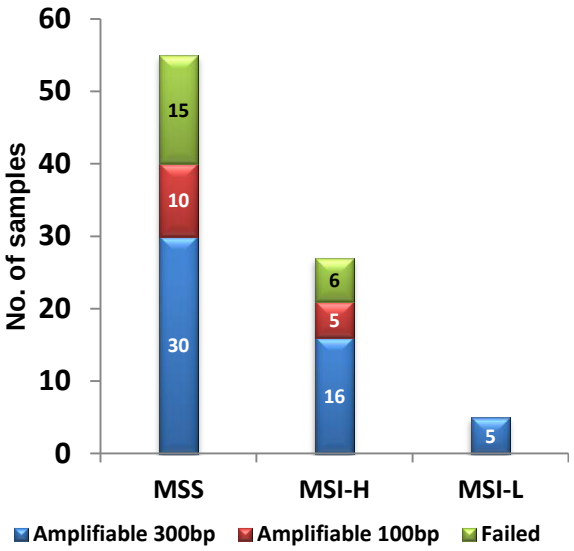


Figure 3-10: The number of amplifiable colorectal cancer samples using the 300bp and 100bp amplicons. 300bp amplifiable samples are those that successfully amplified 300bp amplicons (which should amplify 100bp amplicons as well) (represented as blue bars), 100bp amplifiable samples refer to those that amplify the 100bp amplicon only (represented as red bars). Most of the samples have successfully amplified the short amplicon. However, 6 MSI-H and 15 MSS samples have failed to amplify even the 100bp amplicons.

To further investigate the possible reasons behind the variable PCR results, a subset of MSI-H samples were tested for DNA quality using the Agilent Bioanalyser (Agilent, Santa Clara, CA, USA). DV200 (percentage of DNA with a size equal and above 200bp in the tested sample) was used as a parameter of the DNA quality (and hence, the DNA integrity) in the samples. Two groups of MSI-H samples were subjected to quality investigation, 300bp amplifiable (samples G1, G10, and G3) and non-amplifiable (samples 18, 47 and 58) samples. Results showed that the first 3 samples (the amplifiable) had a DV200 of at least 15%, which is consistent with the successful amplifications in these cases. In 2 out of the 3 (non-amplifiable) samples, DV200 was very low (undetected) while the third one had 10% DV200 as shown in

Figure 3-11. These results are broadly consistent with the assumption that successful amplification of long amplicons (~300bp) is dependent upon DNA quality.

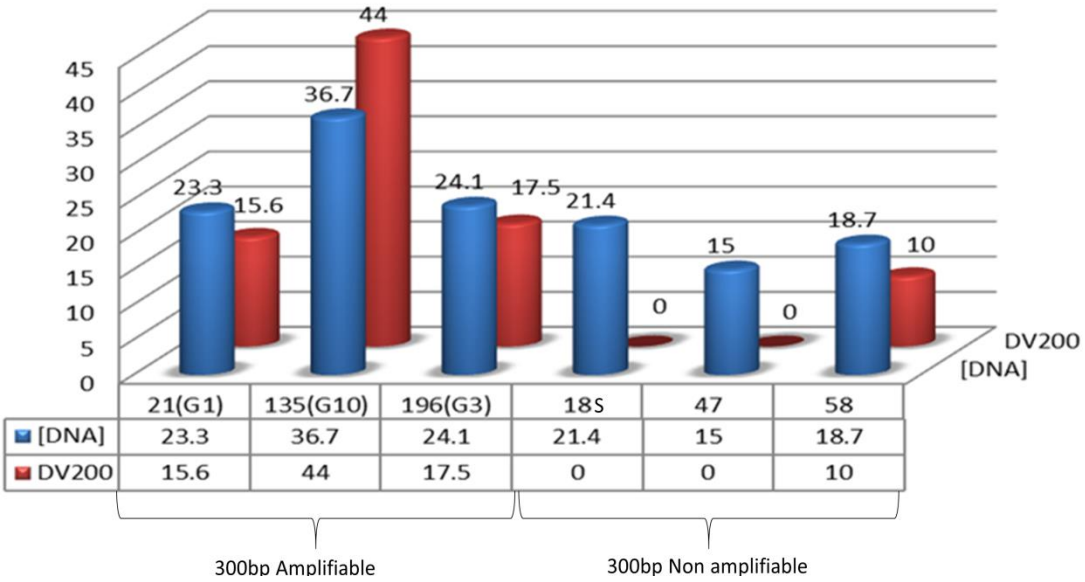


Figure 3-11: Comparative assessment of DNA quality between 2 groups of MSI-H samples. Amplifiable (G1, G10 and G3) and non-amplifiable (18, 47 and 58) using the 300bp primers. **Y-axis** refers to the value of DNA concentration (for blue bars) or the value of DV200 (for red bars).The corresponding DNA concentrations [DNA] in ng/µl were shown for each sample. **DV200**: the percentage of DNA present in fragments equal or greater than 200bp in size. Amplifiable samples clearly showed higher DV200 than the non-amplifiable ones.

Only those samples which were successfully amplified with the 300bp amplicons were used for the subsequent analysis because this is the amplicon length that is needed for library preparation using the Nextera XT protocol.

Because of the low number of amplifiable MSI-H samples within this cohort, DNA from additional 9 MSI-H samples (from outside the 248 pool) were obtained from the northern genetics service (Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK) as explained in Chapter 2 section 2.2.1.

Following these stages of sample filtration and additions, the final panel of amplifiable CRC samples became comprised of **25** MSS, **25** MSI-H and **5** MSI-L samples as shown in Appendix Table 8-1. All these samples were amplified using the primers listed in Table 3-1. PCR amplification was done using the Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA, USA) and approximately 15 ng/µl of the genomic DNA was used for amplification. However, if initial PCR failed, this was increased up to 50ng.

The vast majority of the selected samples were successfully amplified using all 25 primer pairs. Those samples which failed to be amplified in the first round, were then re-amplified (with more template concentration) to fill in the gaps in the amplification profile of the tested cohort. A total number of 1406 amplicons were generated. All amplicons were then quantified using the QIAxcel automated electrophoresis (Qiagen, Hilden, Germany) and pooled. The pooled amplicons were then cleaned up using the Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, USA) to get rid of the extra primers, dNTPs and dsDNA fragments other than the target PCR product. The cleaned amplicons were then quantified by Qubit 2.0 fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), tagmented, barcoded and normalized by diluting the amplicon concentration down to about 0.2ng/µl to achieve a library concentration of 10 pM of pooled amplicons with 5% PhiX library. Pooled libraries were, then sequenced by MiSeq system using MiSeq Reagent kit v3 (600 cycles) (Illumina, San Diego, CA, USA).

The raw data obtained from the MiSeq run were real time monitored, extracted and stored using the Basespace web based cloud on-site informatics tool (Illumina, San Diego, CA, USA). The run yielded 41,383,444 total reads, of which 92.7% passed filter. The sequencing reads were successfully scored for all samples. The lowest percentage of raw reads were observed in sample 29 (0.17% of the total reads) while the highest percentage of raw reads identified in sample 2 (4.6%) of the total reads as shown in Figure 3-12. The coverage depth achieved by this run was approximately 10,000 paired end reads per amplicon.

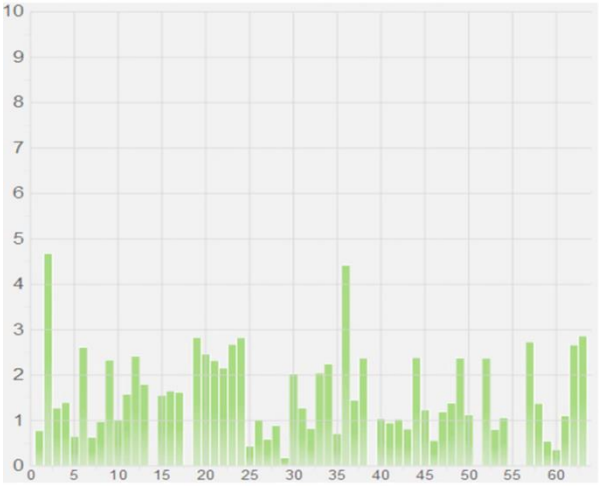


Figure 3-12: The distribution of sequencing reads for all tested samples.X-axis refers to sample number and **Y-axis** refers to the percentage of reads scored for each sample compared to overall sequencing reads.

Q30 is a fundamental quality metric for the MiSeq run as it reflects the probability that one base is called erroneously in each 1000 bases called. The MiSeq run achieved a Q30 value of 66.7% as shown in the Figure 3-13

Figure 3-13: Q30 of the sequencing reads generated in MiSeq run.Green bars= number of bases with Q-score >30, **Blue bars**= number of bases with Q-score<30. **Y-axis** refers to the number of bases (in millions). **X-axis** refers to Q-score.

Variant calling was performed using an in-house variant caller developed and run by Dr Mauro Santibanez Koref (Institute of Genetic Medicine, Newcastle University) called Concordant Overlapping Paired Reads Caller (COPReC) and the sequencing reads were quantified for each allele of a unique length in each marker individually. The data were exported in the form of text files that were processed using the R studio (R Core Team) to generate the spreadsheet format of the sequencing reads.

In the previous analysis conducted by Dr Lisa Redford (Newcastle University, UK), deletions of homopolymers were found far more abundant than insertions in MSI-H samples and data, therefore, were analysed based on the calculation of deletion frequency (Redford, 2016). Similarly, the data analysis of the current study was based on deletion frequency in addition to allelic bias.

3.2.2.1. Calculation of deletion frequency for all the 25 short repeats across all samples

Sequencing reads were called and reported for each marker across all samples separately. The sequencing reads were sorted in a way to recognise all alleles and genotypes. The genotypes were categorized as wildtype (the reference sequence in which no insertion or deletion was observed, which is, commonly, the most prevalent genotype), -1 (deletion of 1bp), +1 (insertion of 1bp) as shown in Figure 3-14.

	G	T	<NA>
-1	29	40	15
0	1681	1827	646
1	12	13	5
<NA>	335	254	0

Figure 3-14: The organisation of sequencing reads. The leftmost column represents genotypes, the most prevalent is 0 genotype (Wildtype genotype), **-1** and 1 refers to 1bp deletion and 1bp insertion genotypes respectively. The uppermost raw represents the called alleles for the adjacent SNP. **NA** refers to the reads that have skipped the homopolymer of interest (Horizontal) or the SNP (Vertical).

All amplicons with an overall number of sequencing reads exceeding 100 were included in the subsequent analysis. If the total number (for all observed alleles and variants) didn't reach that threshold, then the marker was excluded from the baseline calculations. For purposes of normalisation, the number of reads for the deletion allele was compared to those of the most abundant genotype (wildtype) to determine the ratio (in the form of a percentage) between them as shown in Figure 3-15. This percentage was called "**deletion frequency**" and used as a main classifier in the subsequent analyses.

	A	G	<NA>
-2	0	3	0
-1	2	2254	28
0	2	7790	59
1	0	32	0
<NA>	0	431	0

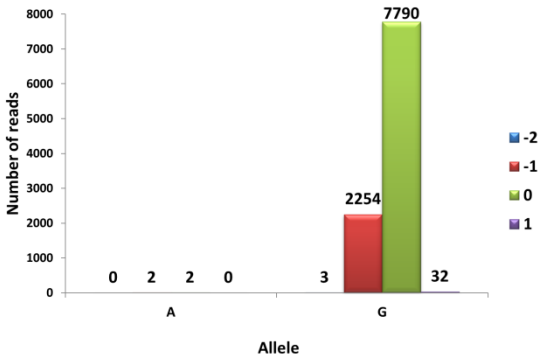


Figure 3-15: The COPReC data format for a single amplicon in a single sample. Numbers corresponding to (-2, -1, 0 and 1) refer to the sequencing reads registered for each variant length. The most abundant genotype is the 0 (wildtype). The next frequent genotype is the -1 (deletion of 1 bp). A and G represents both alleles for the adjacent SNP. **NA** refers to the sequencing reads that skip either the SNP (vertical NA) or the homopolymer (Horizontal NA). The reads are represented in the form of bar charts shown on the right side.

Deletion frequencies were then calculated for all markers both in MSI-H and MSS samples to generate deletion curves as shown in Figure 3-16. These deletion curves can provide estimates of sensitivity and specificity respectively. For purposes of comparisons, the reported phenotype was based on the MSI status provided by the original lab, which based on testing by the MSI Analysis System, Version1.2 (Promega, Madison, WI, USA). According to this test, samples that show instability in ≥ 2 of the 5 tested markers were called as MSI-H and those which didn't show instability in any of the tested markers were called as MSS.

Sensitivity and specificity curves were generated for each marker by plotting the deletion frequencies observed across all MSI-H and MSS samples respectively. The sensitivity curves were constructed in a way to show the proportion of MSI-H samples (in Y-axis) that have a deletion frequency equal to or above the value shown in the X-axis. The specificity curves, on the other hand, show the proportion of MSS samples (on the Y-axis) that have a deletion frequency equal to or above the value indicated in the X-axis (displayed in the form of 1-specificity). Based on these curves, it was possible to define threshold values of deletion frequency for each marker which most effectively separated MSI-H from MSS samples (maximising specificity relative to sensitivity) as potentially suitable cutoff values to be considered in potential assays (shown in Figure 3-16).

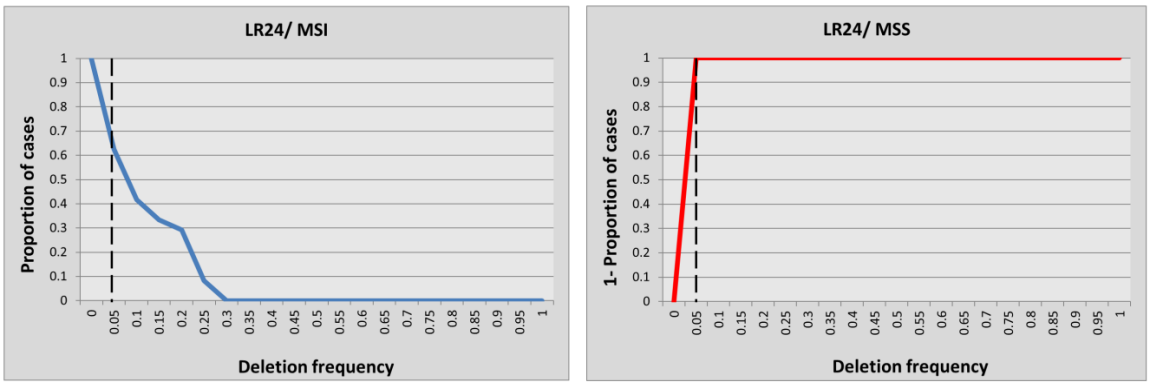


Figure 3-16: Threshold curves of the marker LR24. The sensitivity curve represents the deletion frequencies in MSI-H samples and specificity curve represent deletion frequencies in MSS samples. The specificity was designed in the form of 1-specificity. The X-axis represents deletion frequencies observed of that marker across all samples. Y-axis refers to the proportion of samples. In this particular example, 62% of MSI-H samples show a deletion of ≥0.05, while none of the MSS samples have that value of deletion frequency. A deletion frequency of 0.05 (5%) therefore can be considered as a reasonable cutoff value for this marker.

In the initial analysis, short repeats (7-9bp) showed a very low degree of variability, while longer repeats (10-12bp) were more variable both in the MSI-H and MSS samples. It was expected, therefore, that a relatively low threshold of deletion frequency could be used to assess the variability of the repeats in the current assay.

The deletion curves of 7bp markers are shown in Figure 3-17. Three markers out of the 10 tested markers (LR51-7, IM43-7 and IM55-7) didn't show any noticeable deletion both in MSS and MSI-H samples and a very low sensitivity was observed for another 2 markers (LR15-7 and LR8-7) with a sensitivity of 4.3% and 4.1% respectively at a deletion frequency of ≥5% (1/23 MSI-H samples for LR15-7 and 1/24 for LR48-7). Four out of the above mentioned 5 markers (LR51-7, IM43-7, IM55-7 and LR15-7) exhibited 100% specificity (no false positives) at deletion frequency of 5%, while one (LR8-7) had a single false positive sample. However, the specificity of the marker LR8-7 only reached 100% at a deletion frequency of 10%, where the sensitivity was 4.1%.

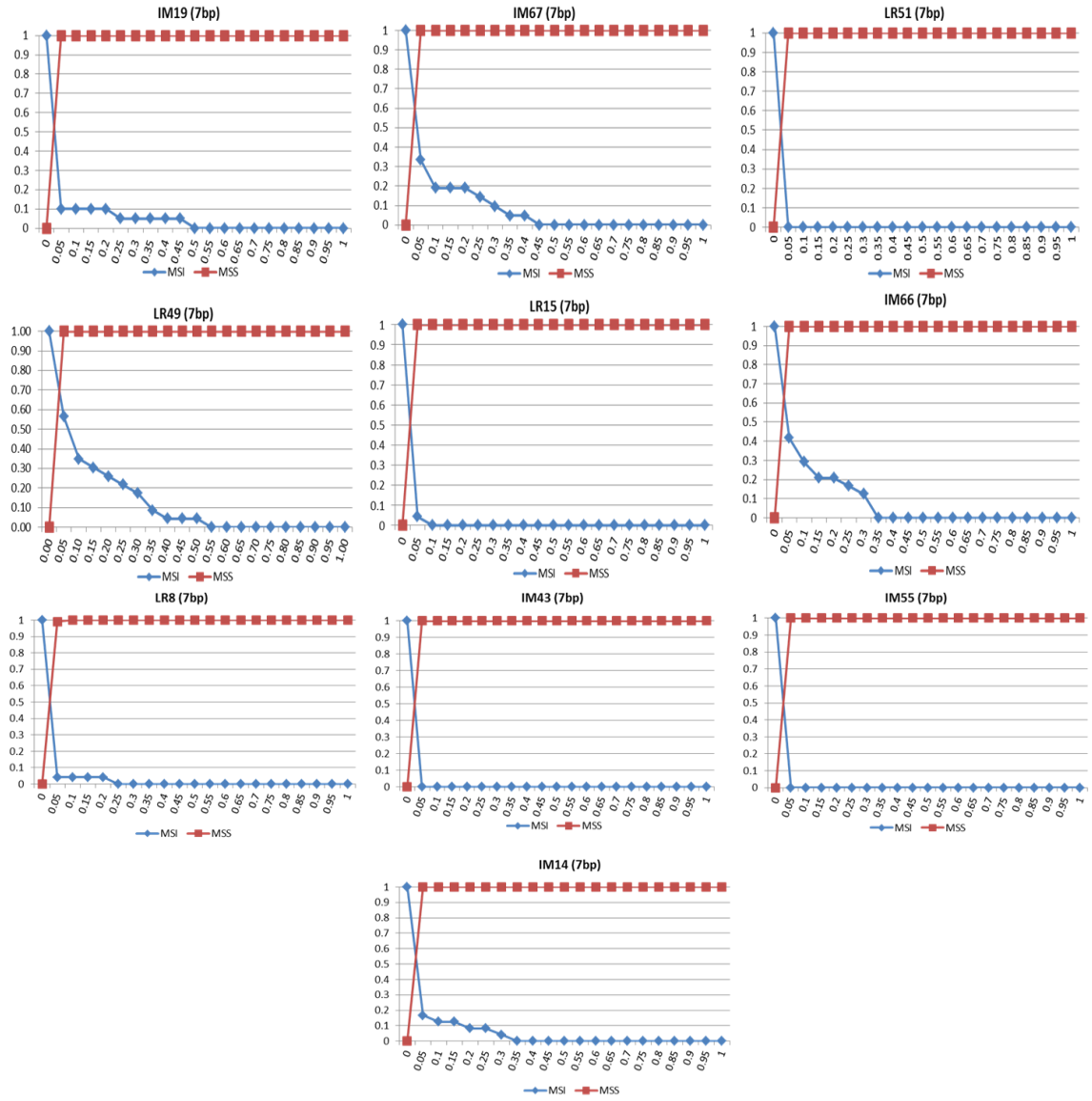


Figure 3-17: Sensitivity and specificity curves for the 7bp markers. X-axis represents deletion frequency observed in proportion of MSI-H cases (blue curves) and MSS cases (red curves). The specificity curve was constructed in the form of 1-Specifcity. At a deletion frequency of ≥5%, the marker LR49-7 shows the highest sensitivity and specificity.

The remaining five 7bp markers exhibited 100% specificity (no false positives) at a deletion frequency of 5%, with sensitivity ranging between 10%-57%. The highest sensitivity at that deletion frequency (i.e. 5%) was achieved with the marker LR49-7, where 57% of MSI-H samples showed a ≥ 5% deletion in that marker, while none of the MSS cases show any degree of instability at or above 5%. For that marker (i.e. LR49-7), up to 17% of MSI-H samples, have had deletion frequencies up to 30% and a single sample (out of the 23 tested MSI-H cases) had a 51% deletion frequency.

In conclusion, at a deletion frequency of 5%, all the 7bp markers show a perfect specificity (= 100%) except a single marker, while sensitivity was relatively low. Thus, in order to have an informative panel, more than one marker would need to be included to improve the overall sensitivity.

For the 8bp repeats, all of the 5 markers showed 100% specificity at and above 5% deletion frequency. At that deletion frequency, the marker IM41-8 had the lowest sensitivity (4.1%), while the marker LR20-8 had the highest sensitivity (47.8%). At the same deletion frequency (i.e. 5%), the other 3 markers (LR46-8, GM9-8 and IM59-8) have had a sensitivity ranging between 29.1%- 37.5% as shown in Figure 3-18. This gives a conclusion that these markers are very specific but less sensitive, thus again, many repeats would need to be included in a panel to improve sensitivity.

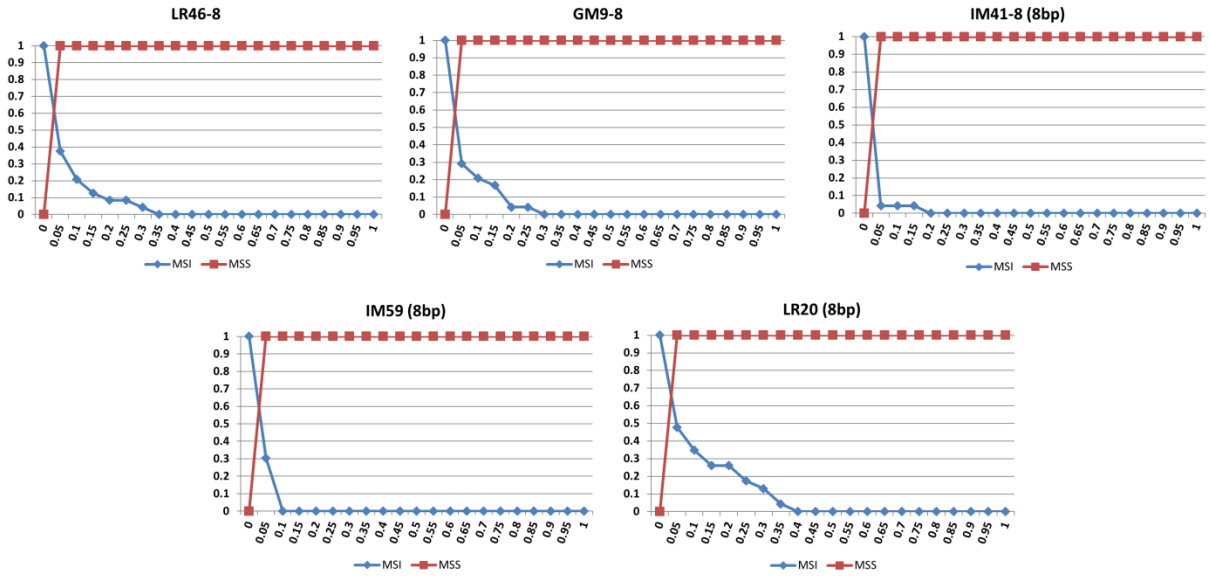


Figure 3-18: Sensitivity and specificity of the 8bp markers. The X-axis represents deletion frequency observed in proportion of MSI-H cases (blue curves) and MSS cases (red curves). The specificity curve was constructed in the form of 1-Specifcity. The highest sensitivity and specificity at ≥5% deletion frequency was observed in the marker LR20-8.

For the 9bp repeats, 6 markers (LR24-9, GM21-9, GM28-9, IM16-9, LR21-9 and LR40-9) exhibited 100% specificity at a deletion frequency of 5% and a sensitivity ranging from 1%-71%, with the highest sensitivity (17 out of the 24 tested MSI-H samples) being observed in the marker IM16-9, and the lowest sensitivity in the marker GM21-9 (1%). At a deletion frequency of 10%, an additional 2 markers (GM11-9 and GM17-9) achieved 100% specificity while sensitivity was 58.3% and 25% respectively. The remaining 2 markers (GM23-9 and LR10-9) have achieved

100% specificity at a deletion frequency of 15% and 20% respectively as shown in Figure 3-19.

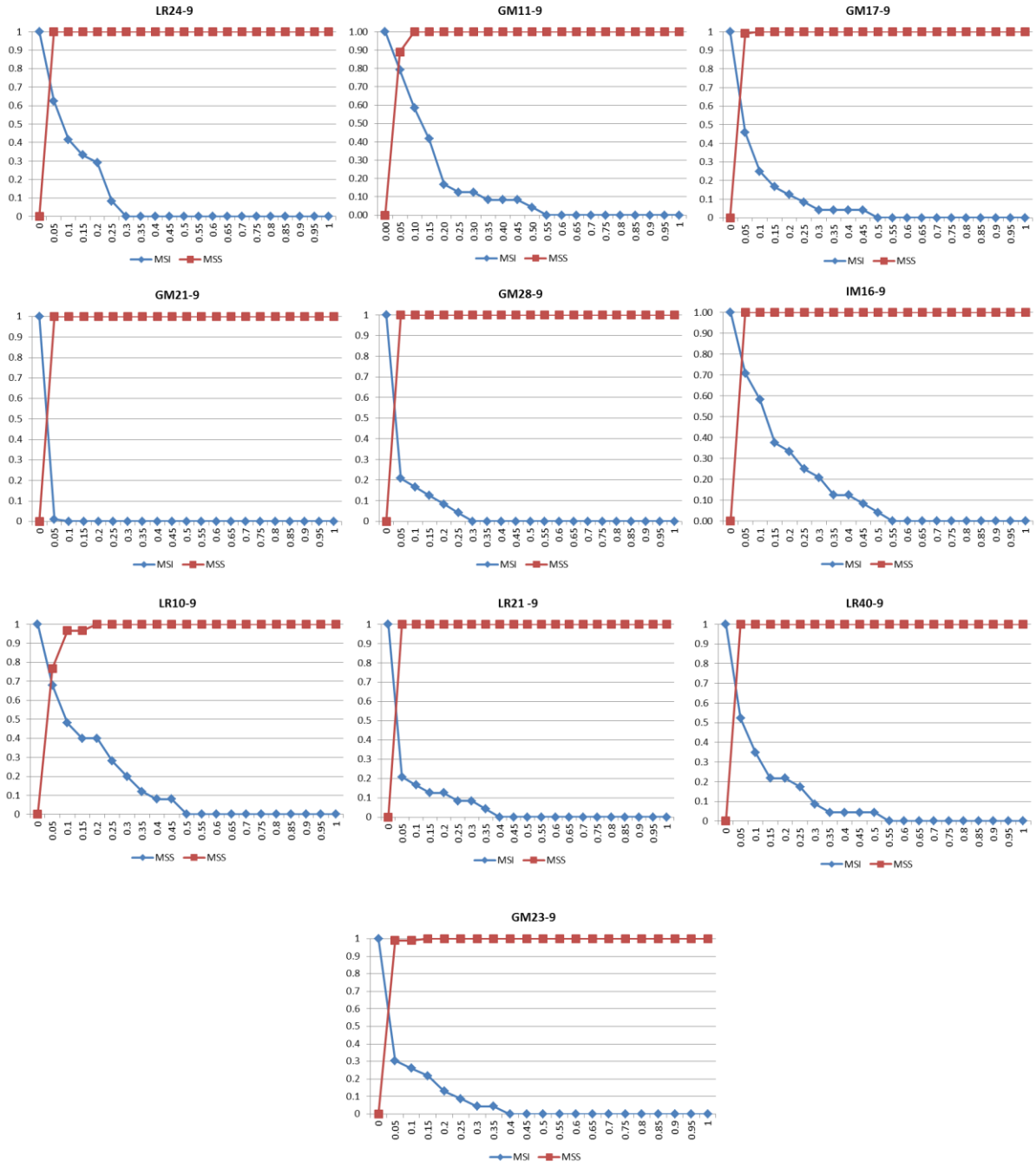


Figure 3-19: Sensitivity and specificity of the 9bp markers. The X-axis represents deletion frequency observed in proportion of MSI-H cases (blue curves) and MSS cases (red curves). The specificity curve is constructed in the form of 1-Specifcity. The highest sensitivity and specificity at ≥5%, was observed in the marker IM16-9.

Compared to other markers (i.e. 7 and 8bp markers), relatively higher deletion frequency is required to achieve a 100% specificity in the 9bp markers. However, 6 of the 9bp markers achieved 100% specificity at 5% deletion frequency.

All markers in all the 3 groups (7, 8 and 9bp) exhibited a high specificity but relatively low sensitivity at the same deletion frequency (e.g. 5%). This means, to improve the sensitivity, more than one marker from each group need to be included in a collated panel.

The performance of all markers, then, analysed on a sample by sample basis. The above calculations were performed for all markers across all samples to compare the deletion profile for each marker in the MSS and MSI-H groups. Below in Figure 3-20, is a representation of the variant repeat frequencies for all markers in 2 samples, one is MSI-H (S42= G42) and one is MSS (S19= G19). Nine markers (out of the 25 tested markers) showed a high proportion of variant reads (in the form of 1bp deletions) in the MSI-H sample, while none of the 25 tested markers show any observable deletion in the MSS sample (see the legend of Figure 3-20 for details).

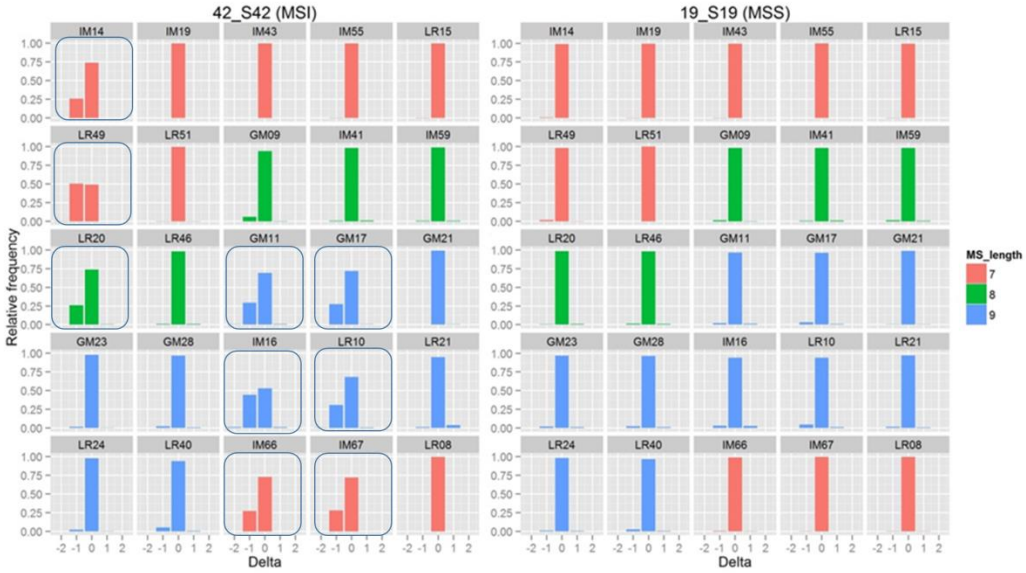


Figure 3-20: Deletion frequencies (Y axis) of all the 25 markers in 2 samples, one is MSI-H (S42) and the other is MSS (S19).**X-axis** represents the nature of the observed genotype (0= Wildtype, -1= 1bp deletion, -2= 2bp deletion, 1= 1bp insertion). A clear degree of instability (1bp deletion) can be observed in 9 markers in the MSI-H sample (IM14-7, LR49-7, LR20-8, GM11-9, GM17-9, IM16-9, LR10-9, IM66-7 and IM67-7) (blue squares). None of the markers exhibit such a frequency of instability in the MSS sample.

In the Figure 3-20, it was clear that 4 markers from the 7bp group and another 4 from the 9bp group exhibited deletion frequency in the MSI-H sample far more that those observed in the MSS sample. An additional 8bp marker showed a 25% deletion frequency in the MSI-H sample compared to 1% in the MSS sample. This indicates that the inclusion of more than one marker from each group would add more to the informativeness of the panel.

Almost all markers showed variable frequencies of deletion, but deletion frequencies were higher in MSI-H than MSS samples as shown in Figure 3-21 as an example.

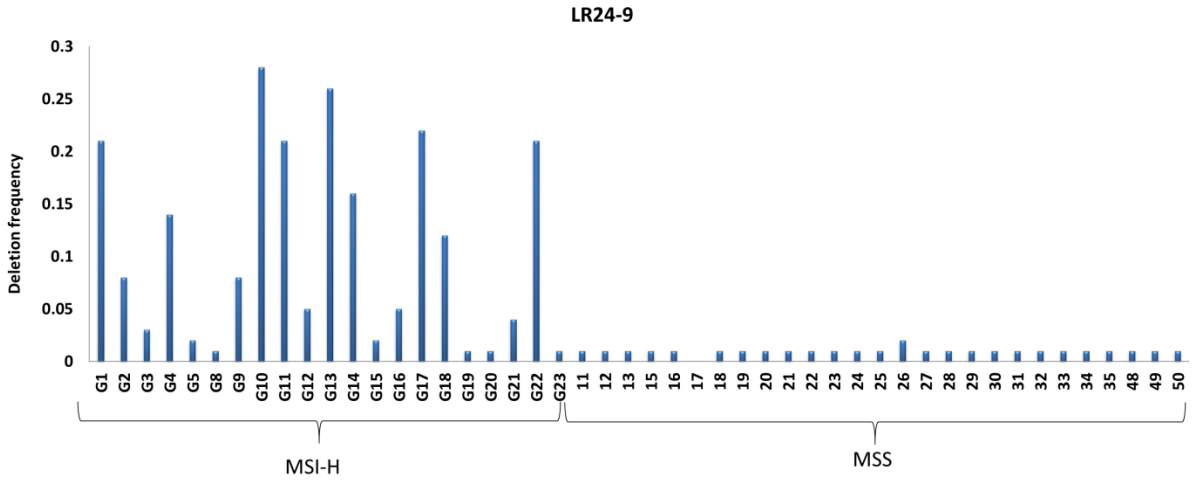


Figure 3-21: The deletion frequencies for the marker LR24 across all samples. Deletion is more abundant in MSI-H than MSS samples.

3.2.2.2. Optimisation of thresholds for calling instability of the short mononucleotide markers

It was clear from deletion curves of individual markers that at certain deletion frequency, certain markers achieved 100% specificity while others are not (even within the same length group). In order to generate a unified scale for calling instability, I have tried to test different cutoff values for each group of markers.

Based on the data extracted from threshold curves, 4 arbitrary threshold sets were tested. Each threshold set has been tested for each marker by examining deletion frequencies both in MSI-H and MSS samples.

For purposes of correct classification of samples, any marker shows a deletion frequency at or above the threshold was classified as “unstable”, while those with lower frequencies (below the threshold value) are classified as “stable”. False positive rate (FPR) and false negative rates (FNR) were calculated as explained in Chapter 2 section 2.10.6

The first threshold set was based on using a very low deletion frequency (0.01) as a cutoff value for all markers (7, 8 and 9bp markers). Applying this low level of deletion frequency as arbitrary threshold resulted in a high false positive and false

negative rates (FPR, FNR) as shown in Table 3-3. Because the cutoff values were very low, all MSS samples have had a high number of unstable markers (ranging from 12 to 19 markers) as shown in Figure 3-22. This clearly indicates a very poor discriminatory power of the panel at those cutoff values.

	Group	Cutoff	FPR	FNR	Sensitivity	Specificity
0.01 Threshold set	7bp	0.01	0.25	0.59	41%	75%
	8bp	0.01	0.76	0.14	86%	24%
	9bp	0.01	0.84	0.11	89%	16%
0.05 Threshold set	7bp	0.05	0.00	0.82	18%	100%
	8bp	0.05	0.01	0.70	30%	99%
	9bp	0.05	0.05	0.52	48%	95%
0.1 Threshold set	7bp	0.1	0.00	0.88	12%	100%
	8bp	0.1	0.00	0.84	16%	100%
	9bp	0.1	0.01	0.65	35%	99%
Final Threshold set	7bp	0.05	0.00	0.82	18%	100%
	8bp	0.05	0.01	0.70	30%	99%
	9bp	0.1	0.01	0.65	35%	99%

Table 3-3: Cutoff values and false calling rates in the 4 threshold sets. The highest FPR and FNR were observed in the 0.01 threshold set while, the lowest were in the final threshold set.

To minimise both FNR and FPR, a second set of thresholds was tested using 5% (0.05) as a cutoff value for all groups, this resulted in a significant decline in both FPR and FNR especially in the 7bp group of markers, but FNR in the 9bp persisted high as shown in Table 3-3.

Applying these cutoff values resulted in a dramatic decrease in the number of unstable markers in the MSS samples and only 12 MSS samples (out of the 30 tested MSS samples) have had ≥ 1 unstable marker as shown in Figure 3-22.

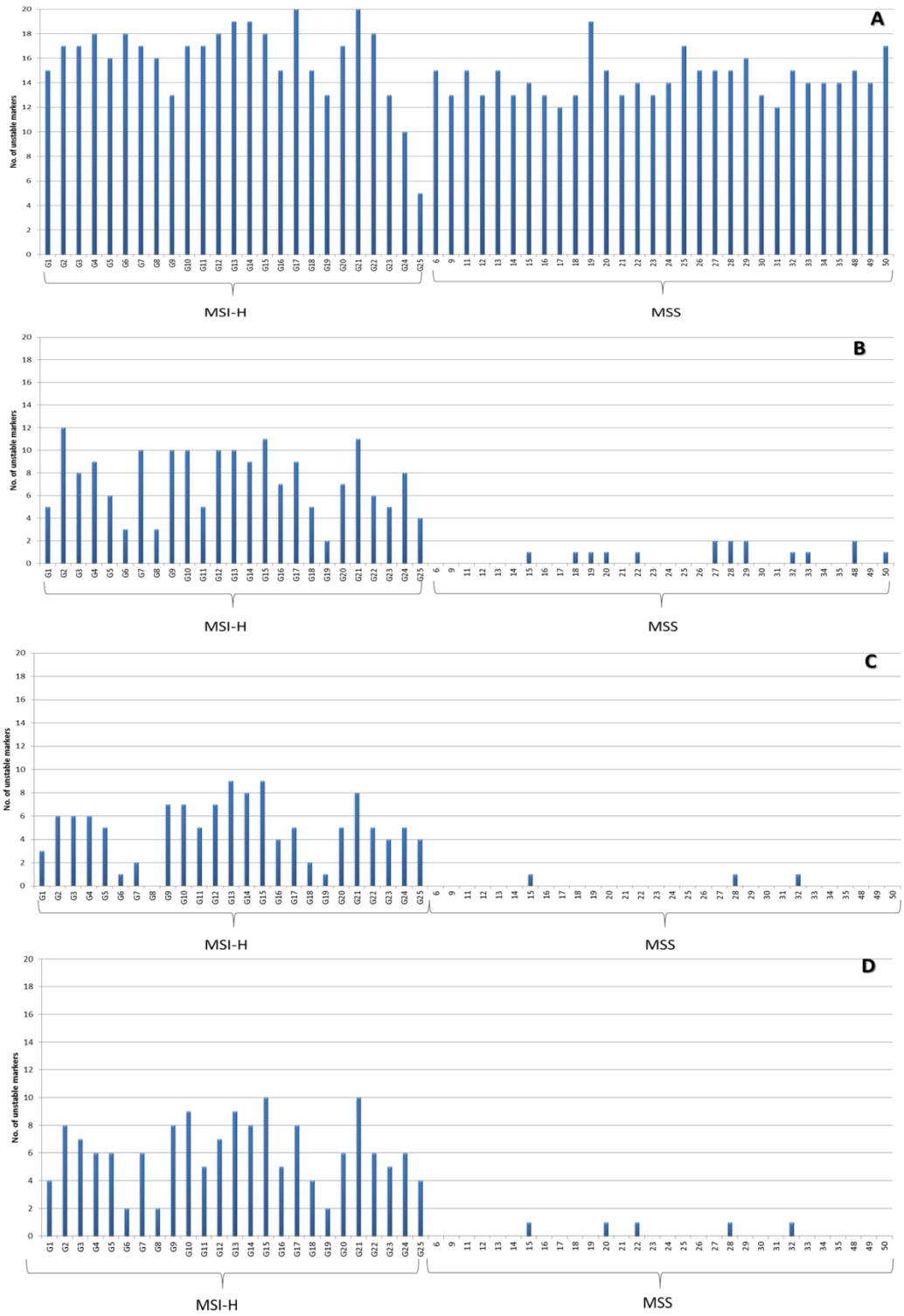


Figure 3-22: Number of 7-9bp markers called as unstable in each sample using different threshold sets. **(A)** 0.01 threshold set, **(B)** 0.05 threshold set, **(C)** 0.1 threshold set and **(D)** the final combined threshold set (0.05 and 0.1). The highest number of markers with deletion above the cutoff values was observed in the 0.01 threshold set, while the lowest in the 0.1 threshold set. In the 0.1 threshold set, one MSI-H sample was miscalled as MSS.

A third set of threshold values was applied using 10% (0.1) as a cutoff value for all groups of markers (7, 8 and 9bp repeats). Using this value as a threshold, FPR

has reduced to 0 in both 7 and 8bp markers and to 0.01 in the 9bp markers as shown in Table 3-3. Using 0.1 as a cutoff value has eliminated more unstable makers in the MSS group, and at that threshold value, only three MSS samples have had a single unstable marker. On the other hand, these relatively high cutoff values resulted in miscalling of a single MSI-H sample as stable as shown in Figure 3-22.

As shown in the above threshold sets, empirical elevation of the cutoff values resulted in an obvious reduction in both FPR and FNR. The false positive rate is more important than false negative rate because of its subsequent clinical and biological impacts in the determination of a wrong MSI status. Therefore, the aim was to establish a cutoff value, high enough to eliminate all (or almost all) false positives.

There is a substantial evidence that instability is proportional to the length of the repeat (Vilkki et al., 2002, Fazekas et al., 2010), suggesting it may be appropriate to generate length- specific threshold values and to investigate such a length specific threshold set using the same parameters. As long as using 0.05 (5%) as a cutoff value resulted in elimination of almost all FPR for the 7 and 8bp markers while the least FPR for the 9bp markers was achieved by the cutoff value of 0.1 (10%), a collated threshold set was designed using 0.05 as the cutoff value for the 7 and 8bp markers and 0.1 as the cutoff value for the 9bp markers. The application of this threshold set resulted in the least FPR and FNR for both groups of markers as shown in Table 3-3.

Then, sensitivity and specificity were calculated for the final threshold set (with the length specific cutoff values) as explained in Chapter 2 section 2.10.6. Applying these values in a single threshold set, and using a criterion of calling instability for those samples that have at least single marker with a deletion frequency above the length-specific threshold value, resulted in 100% sensitivity and 83% specificity as shown in Table 3-4.

TP	TN	FP	FN	Sensitivity	Specificity
25	25	5	0	100%	83%

Table 3-4: Outputs and performance metrics of the Final (0.05, 0.1) threshold set.TP= True positive, TN= True negative, FP= False positive and FN= False negative.

The suboptimal specificity was due to 5 MSS samples were called unstable (false positives). Interestingly, all the 5 false positive samples had a single marker with a deletion frequency equal or more than the cutoff value designed for each group as shown in Figure 3-22.

Microsatellite instability might affect any repetitive sequences in the human genome and there is evidence suggesting that all colorectal cancer cases could exhibit a low level of microsatellite instability when a large enough number of markers are tested (Laiho et al., 2002, Umar et al., 2004). Therefore, it is expected to find a low level of instability even in MSS samples. Based on that assumption and to eliminate the possibility of misclassification, it was possible to use a cutoff value of having at least 2 unstable markers for each sample to be called as MSI-H and those with none or less than 2 unstable markers are called MSS. Applying this rule has raised the specificity to 100%.

All unstable markers showed higher degrees of instabilities in MSI-H than MSS cases, as shown in Figure 3-23.

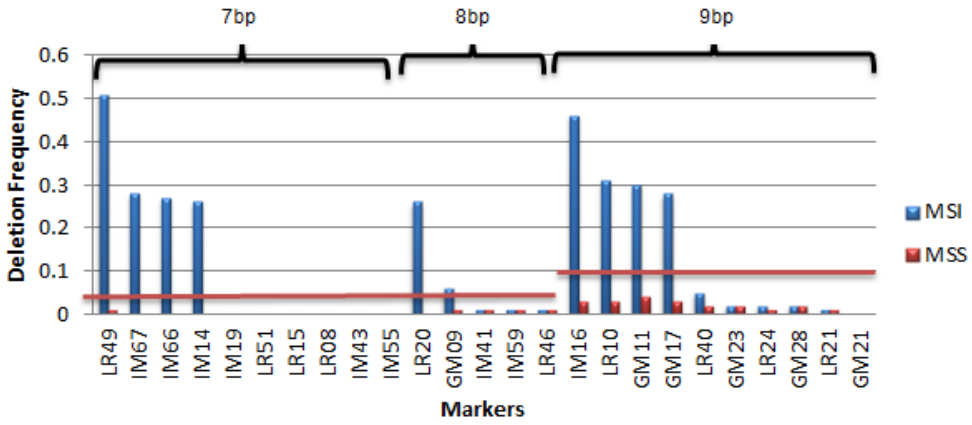


Figure 3-23: Deletion frequencies (Y-axis) of all markers (X-axis) in 2 different samples; one is MSI-H (blue) and the other is MSS (red). Thresholds are 0.05 in 7 and 8bp markers and 0.1 in 9bp markers shown as horizontal red lines. Ten markers have a higher deletion frequency than the specified threshold in the MSI-H sample compared to MSS sample.

3.2.2.3. Assessment of the utility of allelic biased instability as an additional parameter for calling instability

The other parameter against which all markers were stratified, is the allelic bias. All primers were designed to amplify a genomic segment that spans both the target homopolymer and an adjacent high frequency SNP. The amplicon was considered as heterozygous when the overall number of sequencing reads from the minor allele exceeds ≥10% of the overall reads for the other allele. However, in almost all heterozygous amplicons, minor alleles showed ≥40% sequencing reads compared to the other allele. Deletion frequencies were calculated for all alleles of each SNP to observe the difference between them. If that difference is significant, then deletion is said to be allelic biased and it is more likely to be a real instability rather than sequencing errors. To do this, the ratio between sequencing reads of deletion genotypes and wildtype reads for both alleles were calculated as shown and explained in the Figure 3-24 and its legend.

	A	G	T	<NA>
-2	4	2	0	0
-1	1279	250	0	71
0	9571	8181	4	812
1	42	36	0	2
2	1	0	0	0
<NA>	1283	2002	0	0

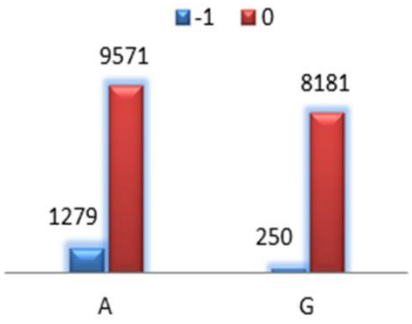


Figure 3-24: Sequencing reads of the marker LR24 linked to a specific SNP in one of the MSI-H samples. The most abundant genotype is the 0 (wildtype). The next frequent genotype is the -1 (deletion of 1 bp). The deletion is observed on both SNP alleles (A and G), but more frequent in allele A where deletion frequency of -1 allele is **13%**. By applying the Fisher's Exact test, deletion in allele A is shown to be significant (*P* value less than 0.05) compared to deletion in allele G. Because the instability is allelic biased and above the allotted cutoff value, this marker is therefore considered unstable in that particular sample.

Fisher's Exact test (of a 2x2 contingency table) was used to test the probability that a deletion has an allelic bias. The significance of instability is expressed as *p* value, with a *p* value less than 0.05 considered significant (as shown in Chapter 2 section 2.10.5).

3.2.2.4. Assigning an informative panel of 8 markers from within the 25 short markers

Allelic bias was used as an extra tool to assess the instability of a marker and its existence increases the likelihood that a registered deletion reflects real instability rather than just a technical artefact. Instability was then called based on the fulfilment of both criteria, deletion frequency above the threshold which shows a significant allelic bias. Those markers that fulfil these 2 criteria were called as unstable markers and each sample was then called unstable if it has at least a single unstable marker (fulfilling both criteria). Deletion frequency and allelic bias were assessed for each marker individually as shown in Figure 3-25.

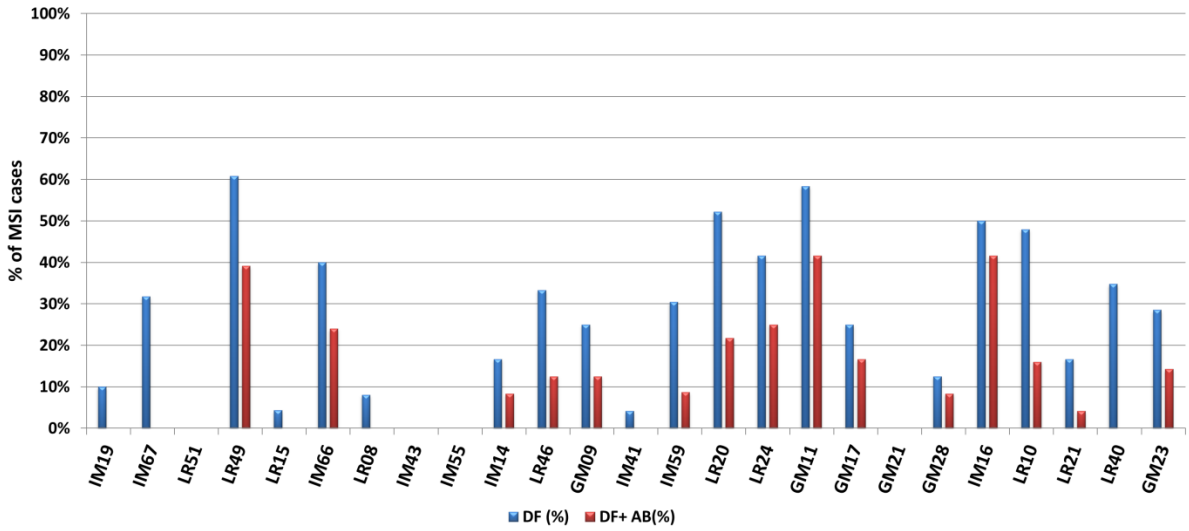


Figure 3-25: The deletion frequency and allelic bias of the 25 markers. Markers are plotted on the X- axis and percentage of MSI-H samples on Y- axis. **DF**= Number of markers that show deletion frequency above threshold value, **DF+AB**= Number of markers that showed deletion frequency above threshold value in addition to allelic bias. Four markers did not show any deletion and an additional 6 markers did not show evidence of allelic bias.

The sensitivity and specificity of the 25 marker panel were calculated by defining instability as having at least a single marker with a significant allelic bias and a deletion frequency above the specified threshold values. This panel has efficiently discriminated MSI-H from MSS samples. At a threshold of at least one marker with a deletion frequency above the length specific cutoff value in addition to a significant allelic bias in each MSI-H sample, the panel achieved 100% sensitivity and 97% specificity. The suboptimal specificity was attributable to the finding that one marker was unstable in one MSS sample (marker GM23). Out of the 25 markers used in the initial panel, 4 markers (LR51, IM43, IM55 and GM21) didn't show observable instability both in MSI-H and MSS samples, indicating that these are less informative

and inclusion of such markers, therefore, is unlikely to add more information about the MSI status, therefore, they were excluded from the subsequent analysis.

An additional 6 markers (IM19, IM67, LR15, LR8, IM41 and LR40) lack evidence of allelic bias throughout the cohort, thus these were also excluded from the panel. The remaining 15 markers (shown in Figure 3-26) were chosen for further analysis. Both sensitivity and specificity remained unaffected by this reduction in the number of markers (100% and 97%, respectively); indicating that within this cohort of samples, omitting these markers had no effect on the overall performance of the panel.

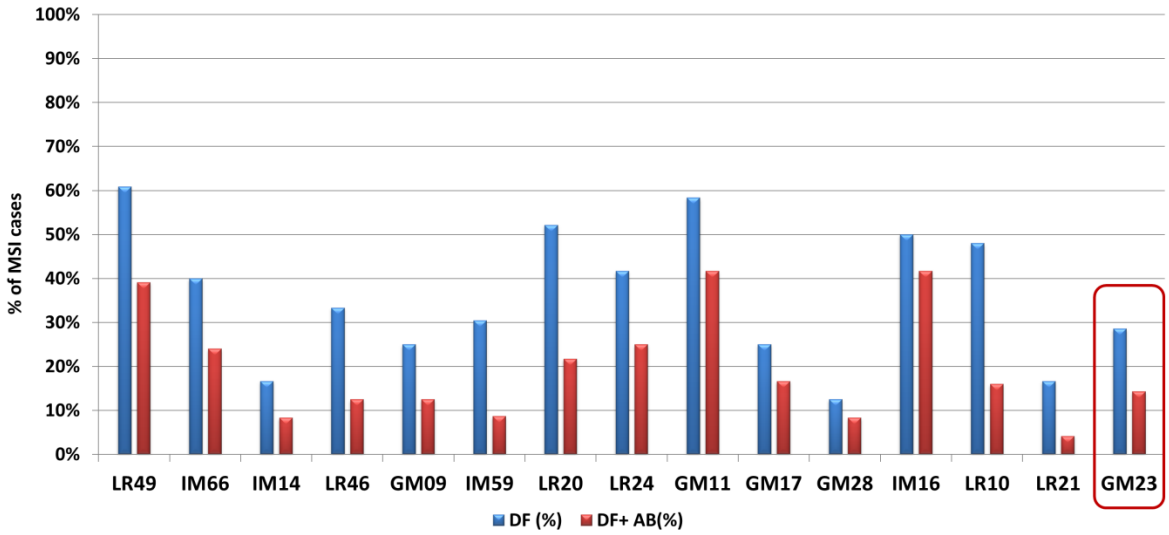


Figure 3-26: The instability (shown in percentage) of the 15 markers in MSI-H samples. DF= Number of markers that show deletion frequency above threshold value, **DF+AB**= Number of markers that showed deletion frequency above threshold value in addition to allelic bias. LR49 is the most unstable marker as it showed allelic bias in 38% of the tested MSI-H samples and stable in all MSS samples. The marker GM23 (red rectangle) was unstable in 14% of MS-H samples and in a single MSS sample.

The overall aim of this part of the study was to find out the most informative markers that can efficiently distinguish the unstable tumours. Therefore, the number of markers has been reduced further to improve specificity and to have the minimal number of informative markers based on results from the tested cohort in this part of the study. The selection of the most informative markers was based on the following criteria:

1) The selected marker should have instability in a high percentage of MSI-H samples with evidence of allelic bias.

2) The selected markers should be stable across all MSS samples (Specificity= 100%).
3) Each MSI-H case should have at least one unstable marker with evidence of allelic bias.

A nominated panel of 8 markers (from the 15markers) was then suggested as shown in Figure 3-27. By applying this 8 markers panel, all false positives were eliminated and the sensitivity and specificity were both 100% in the tested cohort.

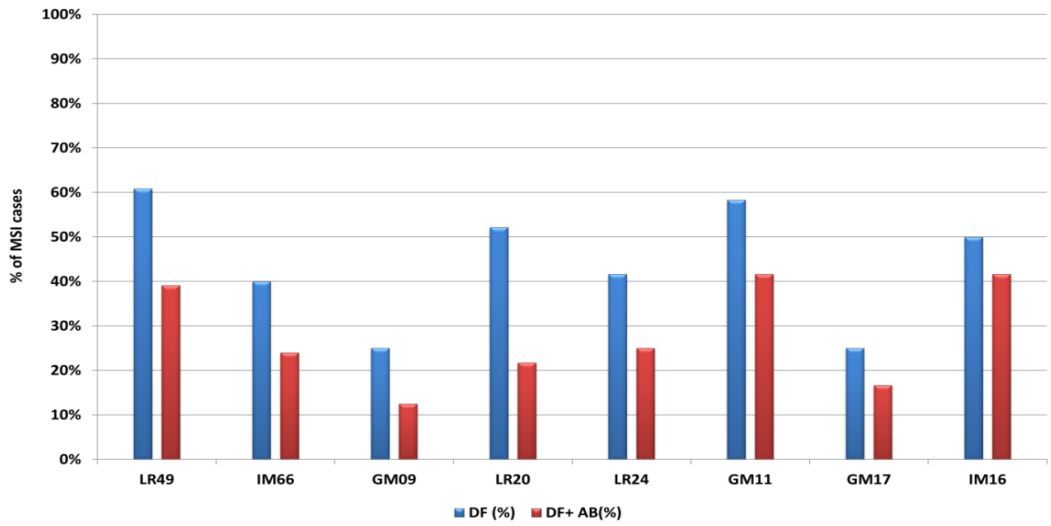


Figure 3-27: The instability (shown in percentage) of the final panel,which composed of 8 markers in MSI-H samples. **DF**= Number of markers that show deletion frequency above threshold value, **DF+AB**= Number of markers that showed deletion frequency above threshold value in addition to allelic bias.

Using this panel, each MSI-H sample within the cohort exhibited at least one unstable marker with additional evidence of allelic bias and none of the MSS samples showed any unstable marker, suggesting this panel may be useful in differentiation between MSS and MSI-H samples in the tested cohort. Interestingly, one 7bp marker (LR49) was unstable in up to 38% of the MSI-H samples (9 out of 24) and additional two 9bp markers (IM16 and GM11) were unstable in up to 42% (10 out of 24) of the MSI-H samples as shown in Figure 3-28.

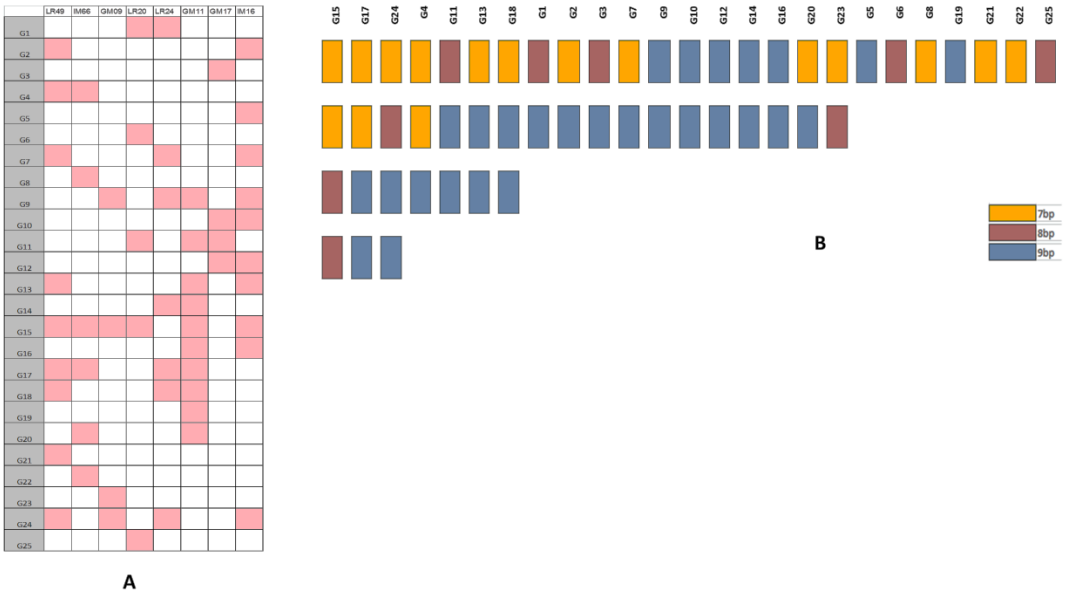


Figure 3-28: The 8 informative markers selected from the 25 tested markers. A: The markers IDs are shown at the top and MSI-H sample numbers are grey highlighted on the leftmost column. Unstable markers in each sample are highlighted in pink. All samples have at least one unstable marker out of the 8 markers in this panel. **B:** The number of unstable markers and their length for each MSI-H sample. The most unstable group of markers is the 9bp group.

The poly A/T homopolymers constituted the majority of the markers, both in the initial 25 marker panel (22 out of 25 markers) and the 8 marker panel (7 out of 8 markers) as shown in Figure 3-29A. The 25 marker panel was composed of exactly the same number of 7bp and 9bp repeats (10 markers for each), while in the 8 marker panel; most of the markers were 9bp repeats as shown in Figure 3-29B.

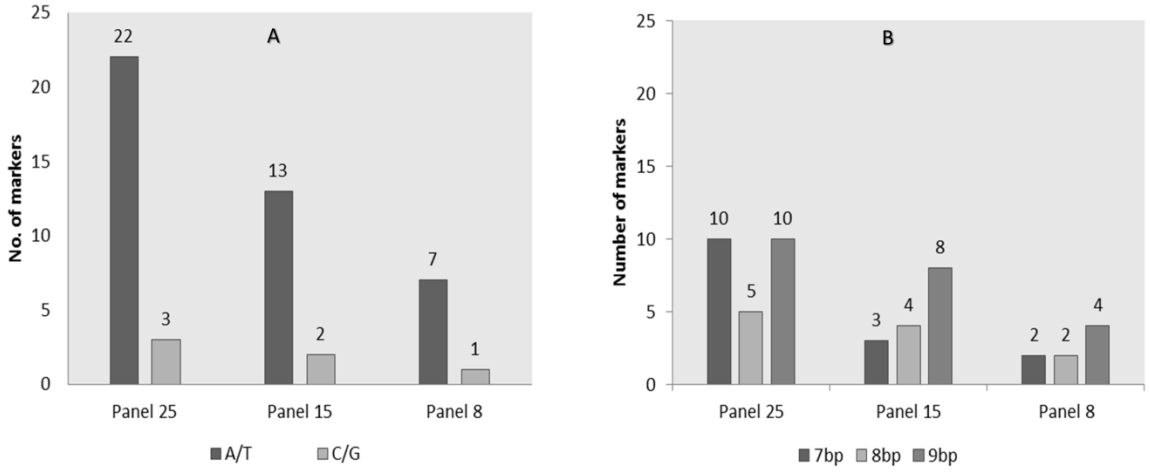


Figure 3-29: A: Base composition of the homopolymers tested in this study. The poly A/T markers represent the main component both in the 25 marker panel (Panel 25) and the 8 marker panel (Panel 8). **B:** The length of the homopolymers used in the 25 marker panel (Panel 25) encompassed an equal number of 7 and 9bp repeats. In the 8 marker panel (Panel 8), half of the markers were 9bp repeats.

3.3. Discussion

The overall aim of this study was to develop a microsatellite panel, convenient enough to supersede the currently used test and able to fulfil the increasing demand for MSI testing. In this part of study, it was possible to assemble a highly informative panel of short homopolymers that can efficiently discriminate between MSI-H and MSS samples.

The work presented in this Chapter showed that deletion of homopolymers is far more frequent than insertion, which is consistent with what was reported by others (Kunkel, 1990, Clarke et al., 2001, Redford, 2016). Therefore, deletion frequency has been used as the main classifier for calling instability.

For all markers across all samples in the cohort, deletion frequency and allelic bias have been calculated and then different threshold sets were tested. A marker was said to be informative if it was unstable in MSI-H and stable in MSS samples, while a marker called as uninformative when it was stable in MSI-H and/or unstable in MSS samples. Different thresholds were tested and in the first instance, very low cutoff values (0.01) were set to examine the deletion profile of all markers. At that cutoff, all MSI-H and MSS samples showed a high number of markers exceeding that cutoff value, and some MSS samples have had up to 19 markers with deletion frequency above that cutoff value. That 0.01 threshold set resulted in a very high false positive rate (up to 0.82) and, therefore, it was inconvenient to be applied. The

raising of cutoff values was mandatory to get rid of the false positive results and therefore, a second set of cutoff values was used. A cutoff value of 0.05 was adopted for all markers (7, 8 and 9bp markers). With that set of threshold, both false positive and false negative rates (FPR and FNR) were dramatically decreased (down to 0.003 in the 7bp markers). However, at that threshold set, 4 MSS samples showed 2 markers with a deletion frequency above that cutoff value (i.e. 0.05). With that threshold set, it was possible to use the criteria of 3 markers or more with deletion above the threshold to differentiate all MSI-H from MSS samples. Another threshold set was tested, in which the cutoff value was raised to 0.1 for all markers (7, 8 and 9bp markers). This resulted in a further decline in the FPR and FNR, but at the same time, resulted in misclassification of one MSI-H sample.

The instability of a homopolymer was known to be increased proportionally with the repeat length (Sammalkorpi et al., 2007, Vilkki et al., 2002, Redford, 2016), therefore, it would be better to design cutoff values so as to be increased with the longer repeat length. The combined threshold set was designed to be 0.05 for both 7bp and 8bp repeats and 0.1 for the 9bp repeats. With that set, all MSI-H samples have got at least 2 unstable markers. For an optimal MSI test, this set of threshold values was able to distinguish all MSI-H samples at a cutoff of having at least 2 unstable markers, within this cohort.

Consistent with other studies (Ward et al., 2001, Laiho et al., 2002, de la Chapelle and Hampel, 2010), there was no observable difference in the deletion frequency of all the tested markers between MSS and MSI-L samples. They, therefore, were considered as a single group (both called as MSS) for purposes of subsequent classification.

Most of the significant allelic bias length alterations in the tested markers were observed in MSI-H samples; therefore, it can be used as an additional criterion for classification of CRC samples. The inclusion of this parameter adds more confidence that the alteration is likely a real instability rather than a technical artefact. However, this parameter is only applicable when the adjacent SNP is heterozygous. At a threshold of having at least one unstable marker which have evidence of allelic bias, all MSI-H samples were called correctly using the final (0.05, 0.1) threshold set, while a single MSS sample was called as unstable (100% sensitivity and 97% specificity). Only 15 markers showed allelic biased instability and out of them, 8 markers (from

those 15 markers) were found to be very informative in calling instability using both criteria (i.e. deletion frequency and allelic bias). At a threshold of having at least a single unstable marker, which shows evidence of allelic bias, the 8 markers panel was able to differentiate all MSI-H samples with a sensitivity and specificity of 100% for both.

The inherent problem of DNA degradation and poor quality in FFPE samples represent one of the most eminent obstacles in the MSI testing. In this study, not all samples have successfully been amplified the 300bp primers and in order to check the possible reasons, the DNA quality has been checked and the parameter DV200 was used for that purpose. DV200 refers to the percentage of DNA sized ≥ 200bp in the tested sample. The DV200 for those samples that successfully amplified the 300bp amplicons was ranging between 15.6-44%, while it was 0-10% in those that failed to amplify 300bp amplicons. This strongly suggests that DNA quality is a fundamental factor in the success of this assay.

Furthermore, a 100bp primer set was tested to assess the impact of DNA degradation on the successful amplification. Out of 36 samples that failed to amplify the 300bp amplicons, 42% (15/36 samples) were successfully amplified the 100bp amplicon. This suggests that a considerable proportion of failures in amplification was due to the FFPE induced- DNA degradation. This could represent a potential caveat in the MSI assay and it would be worthy, therefore, to try to redesign the primers to be smaller than those used in the current assay to increase the amplification success rate.

The application of the final threshold set (0.05, 0.1) has resulted in the lowest FPR and FNR, and when allelic bias used as an additional classifying parameter, it resulted in the highest sensitivity and specificity. However, this threshold set and the allelic bias classification system need to be investigated on a larger independent cohort to find out the informativeness of such a set and conclude the optimal calling system.

A concomitant study conducted by Dr Lisa Redford (Newcastle University, UK) was investigating the instability of longer markers (8-12bp) has concluded 9 markers as to be the most informative in that study. These 9 markers were sensitive and specific for the tested cohort, which was composed of 58 CRC samples (when

thresholds were set arbitrarily) (unpublished data). The informative markers from that study, were then collated with the 8 informative markers from my study to build up a joint informative panel composed of 8 short (7-9bp) and 9 long (8-12bp) markers. This panel will be extensively interrogated in next chapters to assess its accuracy in discrimination between MSI-H and MSS samples.

3.4. Conclusions

In conclusion, it was possible to check the instability with evidence of allelic bias for most markers and 8 markers (from the initially tested 25 markers) were highly sensitive and specific (100% for both) using a cohort of 55 CRC samples. A joint panel can be generated by combining the informative markers from this study (eight 7-9bp markers) and other informative markers from a parallel study (nine 8-12bp markers). It would be important to assess this combined panel (of the 17 markers) using an independent cohort. The threshold set proposed in this chapter also needs to be assessed in order to find out the most informative cutoff values that could be used for calling instability.

Chapter 4. Assessment of arbitrary threshold sets and determination of an optimal MSI scoring system

4.1. Introduction and aims

4.1.1. Introduction

In the previous chapter, 25 short (7-9bp) markers were extensively analysed using a cohort of 55 CRCs and from them, a panel of 8 markers was found to be the most informative in terms of discrimination between MSI-H and MSS samples. In that part of the study, four threshold sets were tested and a final set of length specific cutoff values were chosen as the most informative threshold set based on the values of false positive and false negative rates. At a threshold of having at least a single unstable marker with allelic bias in each MSI-H sample in the tested cohort (i.e. 55 CRCs), the 8 markers panel was able to differentiate between all MSI-H and MSS samples. An investigation of longer markers, performed by Dr Lisa Redford (Newcastle University, UK) identified cutoff values that enabled a panel of 9 longer (8-12bp) markers to discriminate between all MSI-H and MSS samples in the tested cohort which was composed of 58 CRC samples. In this chapter, the aim is to analyse these 17 markers (8 short and 9 long) against a larger cohort (composed of 141 CRC cases) and assess different threshold sets to determine optimal threshold values that can be used to implement the target markers in an MSI assay.

In 2012, TCGA published a seminal paper about the somatic alteration in CRC using the NGS platforms (Cancer Genome Atlas, 2012). In that paper, researchers used the exome data to assess the microsatellite instability. Since that time, several studies have tried to assess the MSI status using an NGS approach. In 2013, a novel method to assess microsatellite instability was proposed using RNA- seq data. In that study, RNA-seq data from 20 different cancer cell lines with known MSI status were analysed. The proportion of insertions in microsatellite loci over all insertions (named as PI) and the proportion of deletions in microsatellite loci over all deletions (named as PD) were calculated. PI/PD was referred to as MSI- seq index and used to assess the instability of each microsatellite locus. RNA- seq data from HapMap lymphoblastoid samples were used as a control as they were known to have no instability. A significant increase in the proportion of indels was observed in the MSI samples compared to those in HapMap samples, while there was no significant difference between MSS and the HapMap samples (Lu et al., 2013).

A new NGS based approach called mSING was proposed and assessed in 2014 (Salipante et al., 2014). In that assay, NGS data from 324 different tumours (colorectal, endometrial, ovarian, breast and others) were curated to assess the instability in 15- 2957 mononucleotide markers. Variant allele frequency was calculated in the MSS group and each allele with reads exceeds 5% of the most abundant allele was tallied. The mean of the number of alleles for each marker was calculated to create a baseline reference value. The marker was said to be unstable if the variation exceeds (mean number of alleles + (3xSD) of the baseline reference). They called that approach as mSING (MSI by NGS) and they found it 96-100% sensitive and 97-100% specific in tested cohorts.

In 2015, another study investigated the possibility of performing an MSI assay together with the inclusion of other relevant target genes panel in colorectal cancer (*KRAS*, *NRAS* and *BRAF*), using an NGS platform (Hempelmann et al., 2015). They assessed the microsatellite instability of 17 mononucleotide markers (15- 28bp in length) by using the mSING approach (mentioned above). The study concluded that the used panel of markers was 97.1% sensitive and 100% specific in detecting the microsatellite instability.

Gan *et al* have used the MiSeq platform to assess the utility of both mononucleotides (3 markers 25-34bp in length) and dinucleotides (2 markers 40bp in length) in microsatellite instability (Gan et al., 2015). They calculated the most prevalent allele in both tumour and a corresponding normal sample; they called the marker as unstable if the deviation compared to normal is ≥2bp for mononucleotides and ≥ 4bp for the dinucleotide markers. Quantitative analysis of the MiSeq data from that study showed that the sensitivity and specificity of the mononucleotides were both 100%, while for dinucleotides, the sensitivity was 47- 59% and specificity was 96- 100%.

It is expected for such studies to evolve more over the next few years, as there is no consensus as to the optimal approach in terms of the number of markers that need to be included in the panel, the threshold for calling instability, use of control (normal) tissue and other quality parameters (e.g. sequencing depth, Quality score and others).

4.1.2. Aims

In the previous chapter, the study has nominated a panel of eight (7- 9bp) markers that were able to discriminate between MSI-H and MSS samples in terms of both deletion frequency and allelic bias. Concomitantly, a parallel study conducted by Dr Lisa Redford (Newcastle University, UK) selected a panel of nine (8-12bp) informative markers (able to discriminate MSI-H from MSS samples). To establish the overall sensitivity and specificity of the collated panel, these markers (8 from my study and 9 from the other study) were analysed using a larger cohort composed of 141 CRCs, as summarized in the Figure 4-1. This chapter will outline the assessment of different threshold sets and determine the optimal MSI scoring system that can be used to implement the 17 markers as an MSI test. The overall aims are to:

- Investigate a subset of 141 CRC samples with known MSI status using the joint panel of markers (17 markers).
- Test several length- specific threshold sets and design a new MSI scoring system.
- Investigate and evaluate the role of the allelic bias in calling instability.
- Compare the threshold curves for all markers with their curves in the previous studies (that were done in chapter 3) to assess reproducibility across the 2 cohorts and detect any anomalies in deletion curves of markers.

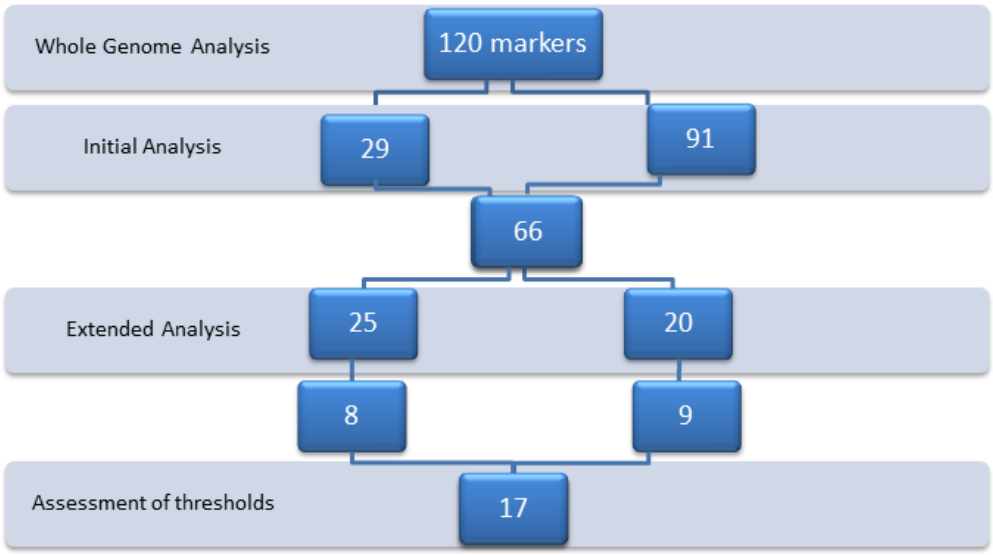


Figure 4-1: Illustration of the overall workflow in this study. The study has started by analysing the 120 variable markers from the whole genome study, and then the highly variable 66 markers were split into short (7-9bp in length) and long (8-12bp in length) markers. Both groups of markers assessed independently and the markers that were able to differentiate MSI-H from MSS samples (17 markers) were gathered from both assays.

4.2. Results

4.2.1. Amplification and Sequencing of the new 17 marker panel

In Chapter 3, 8 out of the 25 tested short (7-9bp) markers were found to be able to discriminate between MSI-H and MSS samples using a cohort composed of 55 CRCs. At a threshold of a single unstable marker (based on the threshold set specified in Chapter 3) for each sample, the panel was able to detect all MSI-H samples in that cohort as clarified in Figure 4-2. Another parallel study (Redford, 2016), concluded a panel of 9 mononucleotide markers (8-12bp in length) as the most informative using an independent cohort of 58 CRCs as shown in Figure 4-2. The results from both independent studies, substantially favours the feasibility of using short homopolymers in MSI testing.

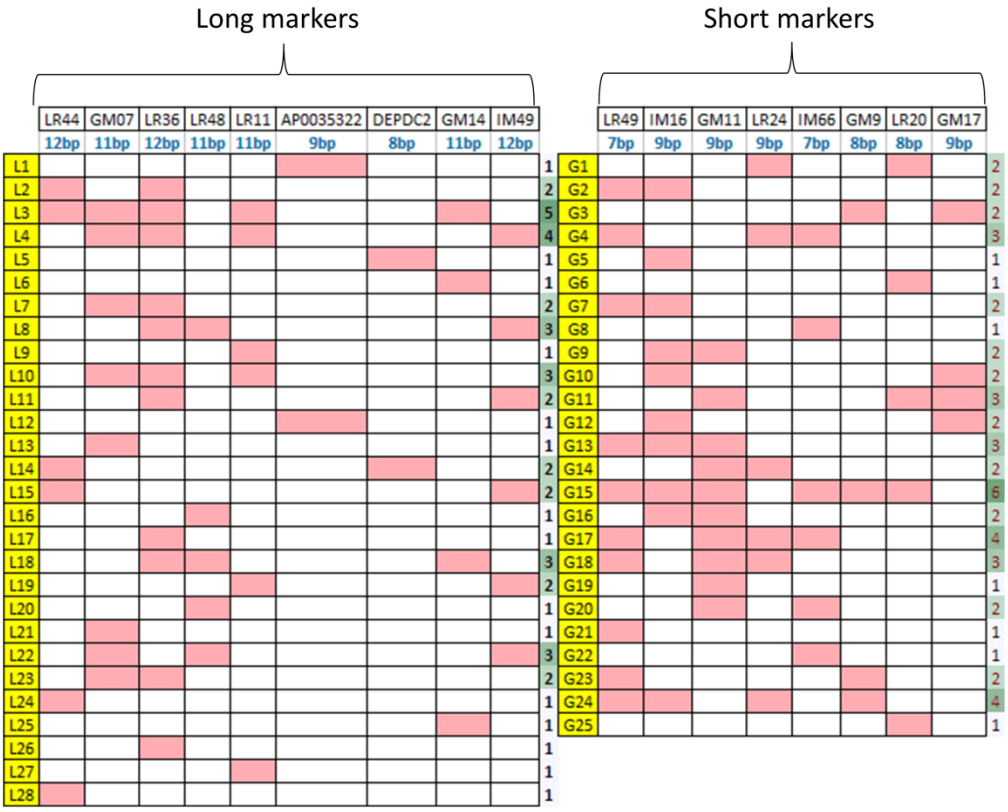


Figure 4-2: The informative mononucleotide markers from 2 independent studies. Long markers= 8-12bp mononucleotide markers, Short markers= 7-9bp mononucleotide markers. The names of the markers are shown at the top and MSI-H sample numbers are highlighted in yellow. Unstable markers (based on threshold sets specified in in each study) in each sample are highlighted in pink. The total number of unstable markers in each case is shown on the right pane (green column). The final panel is composed of 9 long markers on the left side and 8 short markers on the right side. All MSI-H samples from both cohorts have at least one unstable marker out of the 17 markers.

The informative markers from both studies (8 short and 9 long mononucleotide markers= 17 markers) were collated together to be assessed using a large cohort of CRCs with mixed phenotypes. This assessment is to confirm the utility of these markers in detection of MSI and to establish a consensus threshold set can be used for the calling instability.

As the ultimate aim is to develop a panel accurate enough to supersede the currently used one (Promega panel), it is fundamental to standardise the approach and extensively validate that panel prior to it being implemented in clinical laboratories. To assess the test, the collated panel (which is composed of 17 markers), was analysed using a larger cohort of Spanish CRC samples with a mixed MSI phenotype (MSS and MSI-H). The Spanish batch of samples was initially composed of 201 samples and provided by the Genetics Service, Complejo Hospitalario de Navarra and the Oncogenetics and Hereditary Cancer Group, IDISNA (Biomedical Research Institute of Navarre, ESPANA). They were delivered in

the form of extracted DNA and have been quantified and MSI tested in the original lab. I was blinded to their MSI status during initial analysis for purposes of anonymization. Out of the 201 samples, 141 CRC samples (approximately composed of 50% MSI-H and 50% MSS samples) were selected for analysis to establish a standardised calling system for MSI classification.

Primers were designed for the 17 markers of the collated panel. To streamline the assay, 2 main amendments were introduced in the primer designing:

A) Generation of short amplicons

In the previous part of the study, primers were designed to amplify ~300bp amplicons, as this was the recommended amplicon length by the Nextera library preparation protocol (Illumina, California, USA). The amplification of such a relatively long amplicons from FFPE DNA was associated with risk of failure, perhaps due to FFPE- induced DNA degradation, and thus likely to compromise the downstream analysis.

In the current part of the study, the primers were designed to amplify smaller fragments (100-150bp). This delivers 2 main advantages:

1) It will minimise the impact of FFPE associated- DNA degradation on our approach because such a small DNA fragment is more likely to be successfully amplified.
2) In the downstream library preparation, it omits the need for the fragmentation step because PCR products are already relatively small.

Furthermore, the skipping of fragmentation step (the enzymatic fragmentation in case of the Nextera library preparation) brings 2 additional advantages:

1) It reduces the time and effort required for the library preparation and, thus, makes it more easy and straightforward.
2) It obviates the use of Transposase (the enzyme used to fragment DNA) and thus, diminishes sequence errors induced by this step.

B) Primer- Tag incorporation

The second modification is the direct incorporation of the overhang tag sequences to the primers prior to amplification as shown in Figure 4-3. This direct

incorporation will make the process more user friendly, cutting the cost and time required for library preparation.

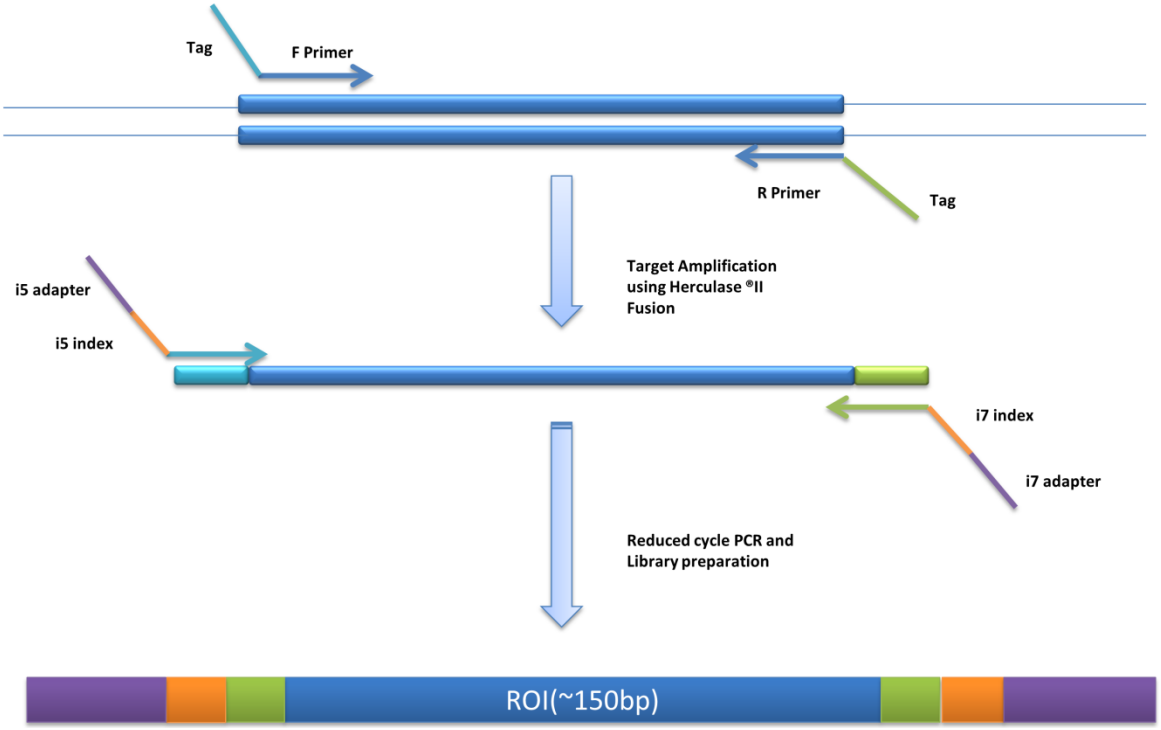


Figure 4-3: Schematic representation of the primer designing and subsequent amplification and library preparation. The primers were designed by direct incorporation of overhang tag sequences to the primer in a single piece. This incorporation allows both sequences (primer and tag) to be incorporated into the target sequences by PCR. The tagged PCR amplicons are then barcoded (i7 and i5) by reduced cycle PCR to prepare the library. **ROI**= region of interest.

As in the earlier analysis, these primers were designed using Primer 3 (Rozen and Skaletsky, 1999) and Primer BLAST, so that to amplify a genomic sequence that contains both a homopolymer and an adjacent high frequency SNP (shown in Table 4-1).

For most primers, the SNPs were chosen to be situated no more than 30bp away from the homopolymer. Then, these primers were *in- silico* checked for theoretical amplifiability of a unique amplicon using the UCSC genome browser (Kent et al., 2002). All markers were derived from the initial 120 variable markers (which were initially retrieved from the whole genome analysis carried out at the beginning of the study as explained in Chapter 3), except 2. These 2 markers were chosen from the literature as they showed a high degree of instability. The 2 markers are DEPDC2-8G (Alhopuro et al., 2012) and AP003532-2-9 (Sammalkorpi et al., 2007) which were tested alongside the longer markers (8-12bp) (Redford, 2016).

	Amplicon	Repeat Size	SNP	SNP alleles	MAF	Repeat Position
1	**LR49-7**	7	rs12903384	A/G	G= 0.4898	Chr15:93619048
2	**IM66-C**	7	rs4794136	C/T	C= 0.3826	Chr17:48433967
			rs141474571	C/G/T	T= 0.0004	
3	**DEPDC2-G**	8	rs4610727	C/T	C= 0.4762	Chr8:68926683
4	**LR20-8**	8	rs217474	C/T	G= 0.3175	Chr1:64029634
			rs146973215	C/T	C= 0.0110	
5	**GM9-8**	8	rs79878287	C/T	C= 0.0006	Chr20:6836977
6	**GM11-9**	9	rs347435	A/G	G= 0.3522	Chr5:166099891
7	**LR24-9**	9	rs192329538	A/G	A= 0.0002	Chr1:153779429
8	**IM16-9**	9	rs73367791	C/T	T= 0.0639	Chr18:1108767
9	**GM17-9**	9	rs666398	C/T	T= 0.2718	Chr11:95551111
10	**AP003532-2-9**	9	rs138081624	A/G	G= 0.0002	Chr11:127625067
11	**GM7-11**	11	rs2283006	A/G	G= 0.3371	Chr7:93085748
12	**LR48-11**	11	rs11105832	C/T	T= 0.2270	Chr12:77988097
13	**LR11-11**	11	rs13011054	A/C	A= 0.4093	Chr2:217217871
14	**GM14-11**	11	rs6804861	C/T	T= 0.3758	Chr3:177328818
15	**IM49-12**	12	rs7642389	C/G	C= 0.2157	Chr3:56682066
16	**LR36-12**	12	rs17550217	A/T	A= 0.2654	Chr4:98999723
17	**LR44-12**	12	rs7905388	C/T	T= 0.3237	Chr10:99898286
			rs7905384	C/T	T= 0.3207	

Table 4-1: List of the 17 primers used in the MSI analysis. For each primer, there is at least one adjacent high frequency SNP. **MAF**= the global minor allele frequency according to dbSNP build 144.

After trials of optimisations, three primer pairs (LR49-7, IM16-9 and GM14-11) failed to give a clear product. Therefore, I redesigned them and the amplification products were successfully obtained. The PCR products (2267 amplicons) were tested and quantified by the QIAxcel automated electrophoresis system (Qiagen, Hilden, Germany), then pooled at approximately equal concentrations and barcoded with unique indexes.

As the amplicons were small in size (~150bp), the 16S metagenomics protocol was used in the library preparation. For barcoding, Herculase II Fusion DNA polymerase (Agilent Technologies, CA, USA) was used instead of the 2x HiFi Kappa enzyme (which is the recommended enzyme in both 16S metagenomics and Nextera protocols). This provide a couple of advantages as it is less costy and brings the

maximal proofreading function (error free replication for mononucleotide repeats ≤13bp in length after 35 PCR cycles) (Fazekas et al., 2010) which, ultimately, is an optimal aim in our assay. Moreover, the number of PCR cycles used to incorporate the barcodes was fewer (10 cycles) compared to that in the Nextera protocol (12 cycles). The barcoded products were then cleaned up by Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, USA), and diluted to achieve a library concentration of 4 pM.

For the sequencing, MiSeq reagent v3 (600 cycles) (Illumina, California, USA) was used and a total reads of 11,862,294 were generated with an average depth of about 2900 paired end reads (per) per amplicon. The average depth was the highest for the marker LR24-9 (5096 per/ amplicon), while was the lowest for the marker LR36-12 (188 per/amplicon). A cluster density of a 510 k/mm^2 was obtained and a Q-score above 30 (99.9% probability of a base being called correctly) was achieved in 69.1% of the sequenced bases as shown in Figure 4-4 A. Towards the last sequencing cycles, the Q30 score started to drop as shown in Figure 4-4 B and that was expected as the sequencing reaches the last bases of amplicons.

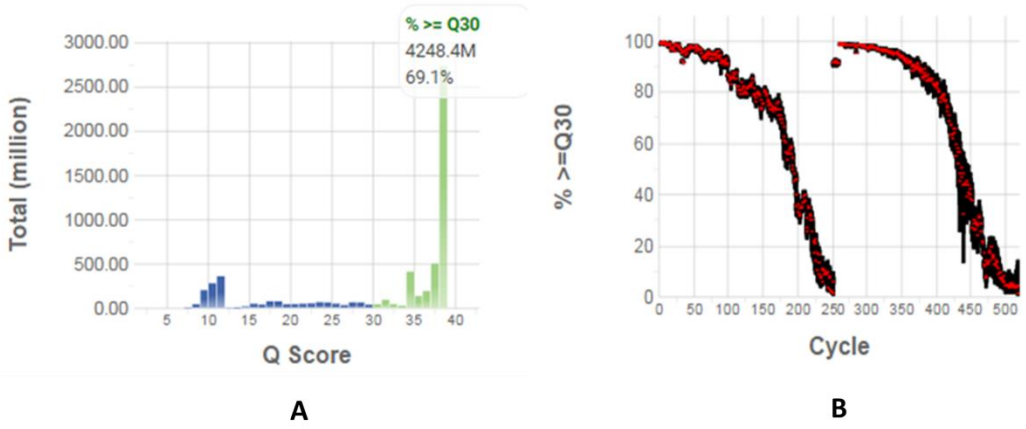


Figure 4-4: Q score distribution of MiSeq run (A) and the distribution of the sequencing reads with Q score >30 across the sequencing cycles (B). It is obvious that reads with a Q score >30 diminish with the progress of sequencing cycles to be extremely low at the last cycles.

The variant caller used in this assay was the same caller that was used in the previous analysis (COPReC variant caller). The amplicons with sequencing reads below 100 were excluded from baseline calculations in order to avoid reads that might be generated erroneously due to PCR duplicates.

Overall, 86% of samples (121/141) were successfully amplified, 85% of samples (120/141) were successfully sequenced and 80% (113/141) were sequenced to adequate depth (i.e. ≥100 per/ amplicon) for 15 markers or more as summarised in Table 4-2.

No. of markers	No. of cases		
	Amplified	Sequenced	Sequenced to ≥100 per/ amplicon
17	78	73	33
16	29	29	52
15	14	18	28
14	12	12	12
13	4	4	10
12	3	4	4
11	1	1	2
10	0	0	0
Total	141	141	141

Table 4-2: The number of samples that were successfully amplified, sequenced and called to adequate depth (≥100 reads). 86% of samples were amplified 15 markers and more, 85% of samples were sequenced 15 markers and more, and 80% of samples were sequenced ≥ 100 reads for 15 markers or more. 100 sequencing reads were used as an arbitrary cutoff value to avoid sequencing reads that originate from sequencing errors. **per**= paired end reads.

4.2.2. Analysis of MiSeq data and calculation of deletion frequency

The COPReC data output were initially retrieved by R studio (R Core Team) to generate the spreadsheet format in order to be analysed. Deletion frequency and allelic bias were used as the main parameters for subsequent analysis. The deletion frequency for each variant genotype was calculated as before.

Sequencing of homopolymers is likely to be accompanied by the generation of non-specific reads due to the inefficiency of PCR and sequencing chemistries to deal with these repetitive sequences. Therefore, it is important to find a tool by which we can differentiate between events which represent a real instability from sequencing errors. Allelic bias is the tool we have used for that purpose. Because each primer pair was designed to amplify a genomic region that contains both a homopolymer and an adjacent SNP, it should be possible to identify allele specific instability, which would be inconsistent with sequence error (as explained in Chapter 3). To increase the confidence that allelic bias reflects a real instability, Fisher's Exact test was used

to calculate the significance of alterations (as explained in Chapter 2 section 2.10.5). With this approach, it is important to differentiate between instability with allele bias and polymorphism. For those markers where the vast majority of reads come from the wildtype for one allele and from a single variant (e.g. -1bp) of the other allele, such a marker is likely to be a polymorphic rather than an allelic biased. However, there is no clear role to be used in order to clearly differentiate between the 2 situations and to be 100% sure, normal tissue needs to be tested.

4.2.3. Assessment of different threshold sets to conclude the most informative cutoff values

Cutoff values can be defined as a specific value of deletion frequency which can be used for classification of samples into MSI-H (have deletion frequency equal or more than that cutoff value) or MSS (have deletion frequency less than that cutoff value). Different thresholds (with length specific cutoff values) were applied to investigate how a change in cutoff values could influence the analysis outputs, with the ultimate aim to define the most informative threshold set. Thresholds were set to be length specific, i.e. markers that belong to each group (e.g. 7, 8, 9, 11 and 12bp) were interrogated using a specific threshold value. These thresholds were chosen based on my previous experiment and another study analysed 9-12bp mononucleotide markers (Redford, 2016). I have started with the threshold values that gave the best results (which showed the lowest false positive rate) from both previous analyses, then; cutoff values were increased gradually in each threshold set. For each threshold set, samples were classified into 3 main subgroups:

1- Samples that have no any marker with a deletion frequency above the threshold value. These samples were deemed microsatellite stable.
2- Samples that have at least one marker with a deletion frequency above the threshold. These samples were deemed unstable by deletion frequency (DF).
3- Samples that have at least one marker with a deletion above the threshold that show allelic bias. These samples were deemed unstable by deletion frequency and allelic bias (AB).

During the setting of thresholds, a single event was individually introduced for each set. Because the allelic bias was suggested to be a reliable evidence of

instability in the previous analysis, so it is likely that samples with both deletion frequency and allelic bias are genuinely unstable. Cutoff values, therefore, were altered in a way to eliminate those samples with deletion frequency only, on the assumption that the optimal threshold set will likely to be in those values. Therefore, prior to unlock the phenotype key, the difference between DF and AB groups was closely observed.

In all threshold sets, a specific threshold value was set to the Poly G/C markers and that value was relatively higher than their mates with corresponding length as this kind of repeats is known to have a higher mutation rate than A/T homopolymers (Boyer et al., 2002).

The first threshold set was designed based on thresholds tested in previous experiments done by myself and Dr Lisa Redford (Newcastle University, UK). In the first 3 threshold sets, cutoff values for short markers (7-9bp) were changed, while those for longer markers (11 and 12bp) were kept constant. In the latter 3 threshold sets, on the other hand, cutoff values for longer markers were arbitrarily raised while those for short markers were kept constant as shown in Table 4-3.

Marker group	7bp	8bp	9bp	PolyG/C	11bp	12bp	
Threshold 1	0.05	0.05	0.05	0.1	0.19	0.19	
Threshold 2	0.05	0.05	0.05	0.1	0.19	0.19	- GM14
Threshold 3	0.05	0.05	0.08	0.1	0.19	0.19	
Threshold 4	0.05	0.05	0.08	0.1	0.19	0.25	
Threshold 5a	0.05	0.05	0.08	0.1	0.19	0.30	-GM14
Threshold 5b	0.05	0.05	0.08	0.1	0.19	0.30	+GM14
Threshold 6a	0.05	0.05	0.08	0.1	0.30	0.30	-GM14
Threshold 6b	0.05	0.05	0.08	0.1	0.30	0.30	+GM14

Table 4-3: Cutoff values that used for threshold setting. For each threshold set, a single event is introduced compared to the preceding set (red coloured). In the first 3 threshold sets, cutoff values of short repeats (7-9bp) were changed, while for the latter 3 sets, cutoff values for long repeats (11 and 12bp) were changed. The cutoff value for the Poly G/C group remained constant for all threshold sets.

When these sets applied on the tested cohort, a relatively high number (101 out of the tested 141 samples) showed instability in at least one marker out of the 17 tested markers in threshold set 1, of them, 70 samples have at least one unstable marker with allelic bias as shown in Figure 4-5. Direct observation of the total number of samples in both groups (i.e. DF and AB subgroups), can clearly show the gradual reduction in the number of cases in the DF subgroup compared to samples in the AB subgroup. Conversely, the number in the AB subgroup showed the least variation with a total difference between T1 and T6 is 7 samples only. This observation is consistent with the assumption that AB is a more reliable event of instability compared to DF alone.

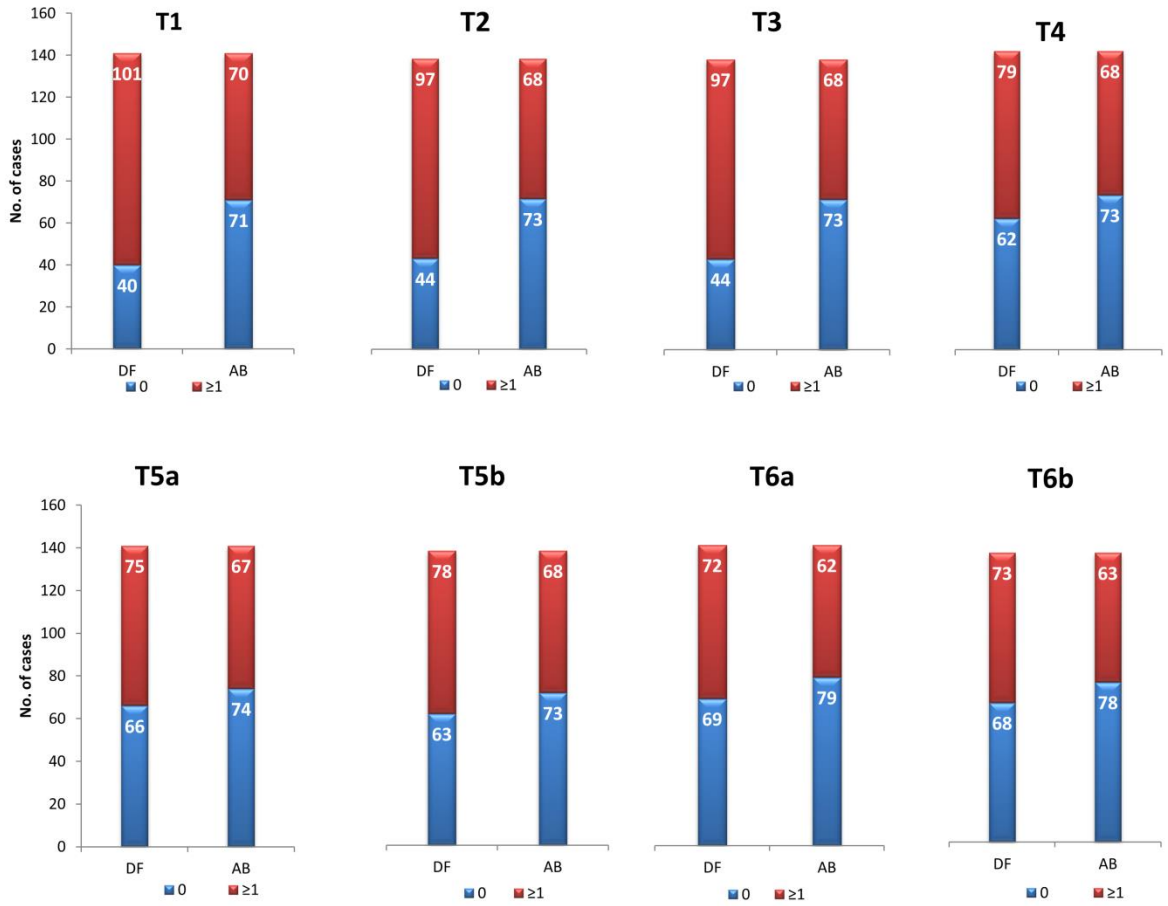


Figure 4-5: The distribution of samples with and without deletion and allelic bias according to cutoff values in the 6 threshold sets. The variation in the number of samples in the DF subgroup is more prominent compared to that of the AB subgroup across threshold sets. 0= samples that have no unstable marker, ≥1= samples that have one or more unstable markers.

Having set these thresholds, the phenotype was predicted for all samples. At that stage, the data became ready to be compared with the reported phenotype in order to elucidate the performance of the different threshold sets compared to a gold standard panel (which is the currently used MSI Analysis System, Version1.2: Promega, Madison, USA). I was then able to calculate Sensitivity and specificity (as explained in Chapter 2 section 2.10.6)

The overall sensitivity and specificity were almost similar within the first three threshold sets. A notable difference was observed in the change of specificity between AB and DF subgroups for each threshold set with a higher specificity in the AB subgroup as shown in Table 4-4. On the other hand, the last 3 threshold sets (i.e. T4, T5 and T6) can clearly show a better overall performance (higher sensitivity and specificity) compared to the initial threshold sets (i.e. T1, T2 and T3), with the highest sensitivity and specificity were observed in T6 as shown in Table 4-4.

Threshold set	>1	0	Sensitivity	Specificity
Threshold 1 DF	101	40	96%	51%
Threshold 1 AB	70	71	88%	82%
Threshold 2 DF	97	44	96%	57%
Threshold 2 AB	68	73	87%	90%
Threshold 3 DF	97	44	96%	57%
Threshold 3 AB	68	73	87%	90%
Threshold 4 DF	79	62	96%	82%
Threshold 4 AB	68	73	87%	90%
Threshold 5a DF	75	66	91%	81%
Threshold 5a AB	67	74	87%	91%
Threshold 5b DF	78	63	93%	81%
Threshold 5b AB	68	73	88%	90%
Threshold 6a DF	72	69	94%	90%
Threshold 6a AB	62	79	86%	96%
Threshold 6b DF	73	68	94%	88%
Threshold 6b AB	63	78	87%	96%

Table 4-4: Analytical parameters of the 6 threshold sets. >1 refers to the number of samples that have a single unstable marker or more, **0** refer to the number of samples that have no unstable marker.

The above results indicate that the increment of the cutoff values in the longer repeats (11 and 12bp repeats) was the main reason behind the improvement in the performance of the last threshold sets. This, also indirectly, indicate that most variant reads were came from the longer repeats and raising cutoff values for these markers was able to get rid of the majority of variant reads in the MSS samples.

4.2.4. Assignment of a new scoring system for calling instability

The presence of AB in markers that have DF above the specified thresholds increase the likelihood that the observed deletion is real instability rather than just sequencing errors. To improve the classification of samples, whether stable or unstable in this panel, an arbitrary scoring system was suggested in which each marker with a deletion frequency above the specified threshold was given a score of 1 and each marker with a deletion frequency above the threshold with evidence of allelic bias was given a score of 2 (because allelic bias represents an additional evidence of instability). According to that scoring system, an overall score of 3 was

set as a cutoff for a sample to be called as unstable. Setting the cutoff value to 3 was to avoid misclassifying MSS samples as MSI-H when there is a single marker with a deletion frequency above the threshold value in addition to allelic bias. Furthermore, it has been reported that all CRC cases have a certain degree of instability (Laiho et al., 2002), so setting the cutoff value to be 3 would be a safe decision to exclude those samples with such a baseline instability. This classification system was termed as the **weighted scoring system** and has been applied to T5 and 6 and the results were as shown in Table 4-5. The weighted MSI scoring system was named as threshold 7 (T7) which uses the same cutoff values specified in T6. When T7 was used, the highest sensitivity and specificity were achieved as shown in Table 4-5.

	Sensitivity	**Specificity**
T7a	93%	99%
T7b	93%	99%

Table 4-5: Sensitivity and specificity of T7. T7a and T7b refer to the inclusion and exclusion of the marker GM14-11 respectively.

The sensitivity and specificity were both below 100% (93% and 99%, respectively) because there were 6 discordant samples. Five out of those 6 discordant samples were miscalled as MSS (i.e. false negative) and a single sample miscalled as MSI-H (i.e. false positive).

4.2.5. Stratification of the new MSI scoring system against MMR IHC status of the 141 CRC samples

The best results, in terms of concordance with reported phenotypes, were achieved by adoption of the weighted scoring system (as discussed above). To confirm the original classification, the Immunohistochemistry of the mismatch repair proteins (IHC MMR) status was requested from the original lab to investigate the consistency of IHC results with the predicted classification.

Almost all samples that have been picked up by our panel as MSI-H (65 samples out of the 141 tested) showed loss of at least one MMR protein by IHC, as shown in Figure 4-6. Only one sample has been classified as MSI-H by our panel while was reported as MSS by the referring lab (i.e. false positive) and, interestingly, this sample was negative for both MLH1 and PMS2 proteins by IHC, raising the possibility that the sample has defective *MMR* genes and has been misclassified.

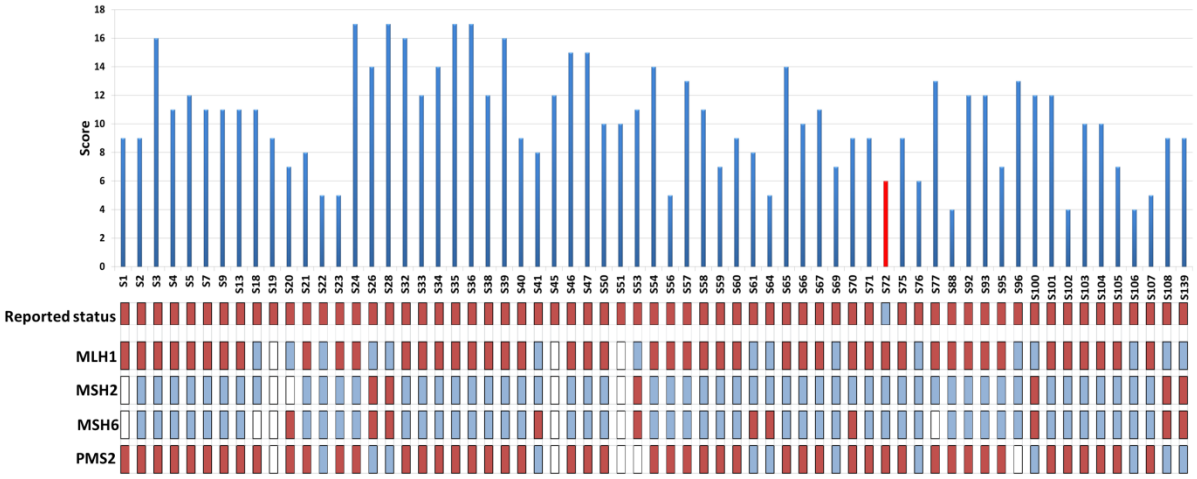


Figure 4-6: The MSI score for all samples that were called as unstable by the weighted scoring system alongside their reported phenotype. **Reported phenotype**: Red colour indicates that a sample was reported as unstable and blue as stable. The corresponding MMR IHC results (MLH1, MSH2, MSH6 and PMS2) are shown for each sample, red colour indicates absent protein, blue indicates the presence of the protein and empty slots indicate no information is available. The single false positive case (S72) (marked as red bar) was negative for MLH1 and PMS2.

Out of the 76 samples which have had an instability score less than 3, 5 samples were classified as unstable by our panel but were originally reported as MSI-H (i.e. false negatives). As shown in Figure 4-7, three out of those 5 false negative samples had normal MMR IHC results, but were reported as MSI-H.

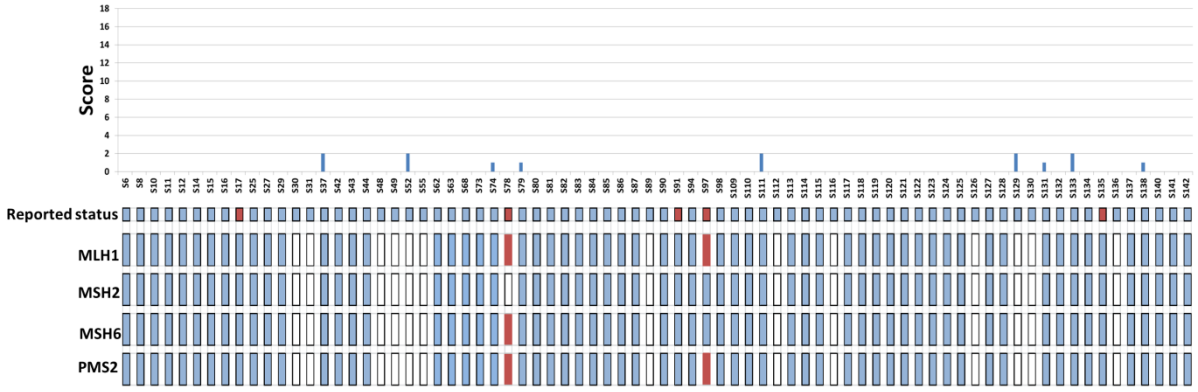


Figure 4-7: The MSI score for all samples that were called as stable by the new scoring system alongside with their reported phenotype. **Reported phenotype**: Red colour indicates that a sample was reported as unstable and blue as stable. The corresponding MMR IHC results (MLH1, MSH2, MSH6 and PMS2) are shown for each case, red colour indicates absent protein, blue indicates the presence of the protein and empty slots indicate no available information. 3 out of the 5 samples that reported as MSI-H show normal MMR IHC staining.

To definitively establish the status of those discordant cases (i.e. one false positive and 5 false negative samples), I re-tested them using the gold standard test (Promega panel). As these samples were referred from abroad, there were no matched normal DNA samples that are ideally required to perform the MSI test by the Promega panel. However, these samples were re-tested without matched normal and blindly interpreted by myself and checked by a qualified clinical scientist (Ottie O'Brien, Northern Genetics Service, Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK). The results presented in Figure 4-8 show that the single false positive sample (S72) is unstable, so was misclassified. One sample out of the 5 false negatives (S17) was found to be stable and another one was found to be MSI-L (S135) rather than MSI-H. The remaining 3 false negatives (S78, S91 and S97) were confirmed to be MSI-H.

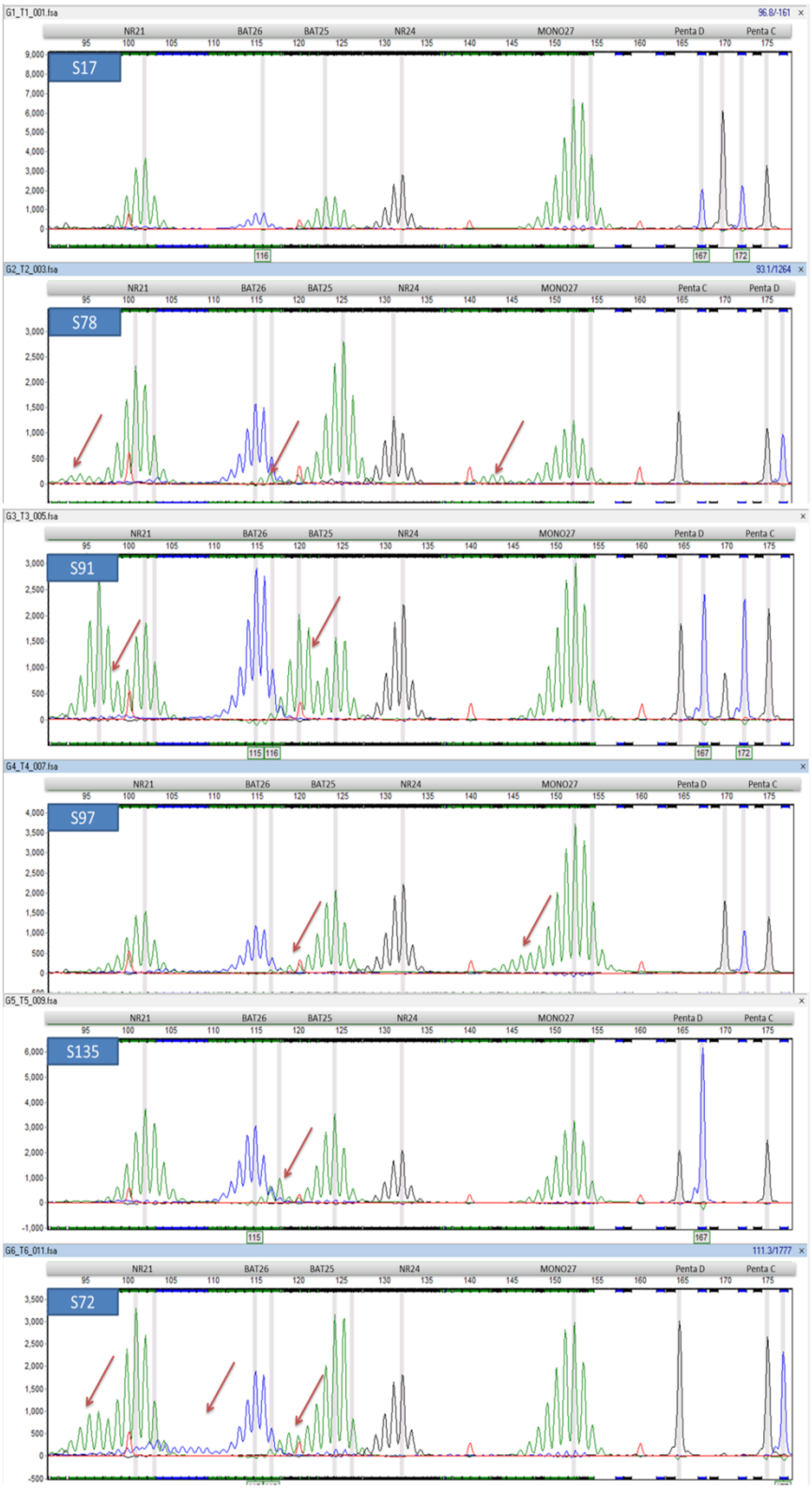


Figure 4-8: MSI results from fragment analysis of the equivocal cases. Sample IDs are labelled in blue boxes and the markers IDs are shown in the top pane. The sample S72 is the false positive, while all the other samples were false negatives. There is no clear instability in the sample S17 (=MSS), while there are 3 unstable markers in samples S78 and S72 (=MSI-H), 2 unstable markers in samples S91 and S97 (=MSI-H), while a single unstable marker in the sample S135 (MSI-L).

With these updates, the final specificity and sensitivity became 100% and 96%, respectively, as shown in Figure 4-9. Because there is no difference between T7a and T7b, this means that the inclusion or exclusion of the marker GM14-11 does not affect the performance when the new scoring system was applied in that particular set.

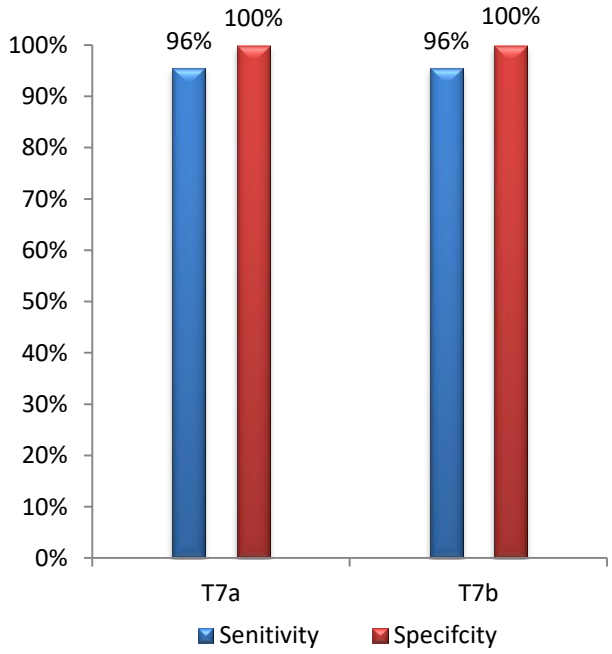


Figure 4-9: The final sensitivity and specificity of the weighted MSI scoring system (T7) after the inclusion of updates. There is no difference between the two T7 values.

The suboptimal sensitivity was because 3 samples were still considered as false negatives, however, one sample (out of the three) has an IHC result that is consistent with our prediction and it would be worthy to test the matched normal tissue samples for these cases (if it became available)

To eliminate the possibility that sample mix up during MiSeq analysis has occurred, the 6 discordant samples were then re-tested using the mononucleotide panel in a new MiSeq run to have a look at their MSI score and compare it in both MiSeq runs. They were re-amplified, re-sequenced and analysed in the same way mentioned earlier. The only difference from the initial sequencing was the library concentration, where 10 pM concentration was used in the re-analysis run. The average depth for the re-analysis was 4499 per/amplicon (higher than the initial analysis which was 2900 per/amplicon). The deletion frequencies were calculated for all markers and compared in both MiSeq runs. The deletions were almost the same

in both runs as shown in Figure 4-10. These results confirm the initial prediction of the discordant cases and give an additional evidence of repeatability of the test.

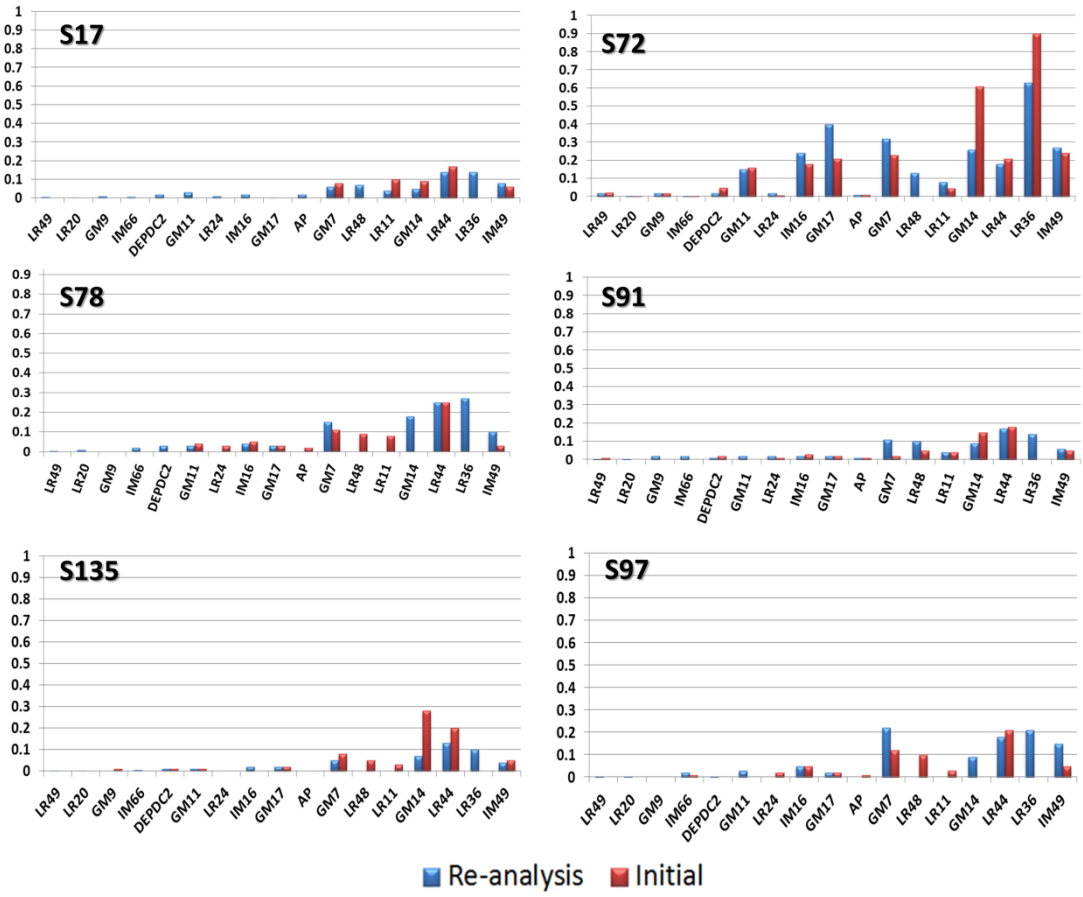


Figure 4-10: Deletion frequencies of all the 17 markers in the 6 equivocal samples in 2 MiSeq runs. X-axis represents the 17 markers and Y-axis represents deletion frequency. The deletion frequencies of the first MiSeq run referred to as **(Initial)** and presented in red bars, the frequencies of the second MiSeq run referred to as **(Re-analysis)** and presented in blue bars. Approximately, deletions are close between the 2 runs, keeping the calling of the sample S72 as MSI-H while all other 5 samples are MSS.

4.2.6. Comparison of deletion curves for all markers with those of the previous cohorts

Deletion curves (or threshold curves) were generated for all markers by plotting the frequency of variant reads on the X- axis and the proportion of samples that have that value of deletion frequency on the Y- axis for both MSI-H and MSS samples. Threshold curves for the MSS samples were constructed in the form of 1-percentage of MSS samples. Threshold curves in MSI-H samples represent sensitivity and those of MSS samples represent specificity for a specific marker.

Deletion frequencies for each marker have been plotted across all MSI-H and MSS samples to examine the distribution of deletions. Furthermore, these values

were compared to those obtained from the previous analysis (done at 2014 using the Newcastle cohort). For purposes of comparison, the 2014 cohort was named as the Newcastle cohort (N) and the 2015 cohort was named as the Spanish cohort (S). To investigate the consistency in the behaviour of markers across different cohorts, *P values* were calculated (as detailed in Materials and Methods) to conclude the significance of the difference in sensitivities and specificities in the different cohorts as shown in Table 4-6

Markers	Cutoff value	Sensitivity			Specificity		
		N	S	*p* value	N	S	*p* value
7bp group							
LR49-7	0.05	57%	44%	0.6760	100%	98%	1.00
8bp group							
LR20-8	0.05	44%	32%	0.3694	99%	98%	0.5439
GM9-8	0.05	29%	24%	0.6343	100%	99%	1.00
G/C group							
IM66-C	0.10	29%	22%	1.00	100%	99%	1.00
DEPDC2-G	0.10	17%	39%	0.2122	100%	100%	1.00
9bp group							
GM11-9	0.10*	54%	39%	0.4057	100%	97%	1.00
LR24-9	0.10*	42%	28%	0.4776	100%	100%	1.00
IM16-9	0.10*	50%	57%	1.00	100%	95%	0.5480
GM17-9	0.10*	27%	20%	0.6141	100%	99%	1.00
AP-9	0.10*	32%	21%	0.4522	100%	97%	1.00
11bp group							
GM7-11	0.30	57%	73%	0.5918	100%	97%	1.00
LR48-11	0.30	35%	51%	0.4107	100%	100%	1.00
LR11-11	0.30	20%	19%	1.00	100%	100%	1.00
GM14-11	0.30	25%	75%	0.0228	100%	89%	0.0930
12bp group							
LR44-12	0.30	64%	77%	0.7265	100%	99%	1.00
LR36-12	0.30	50%	86%	0.1891	100%	92%	0.2199
IM49-12	0.30	39%	57%	0.4339	100%	100%	1.00

Table 4-6: Sensitivity and specificity of all markers in the 2 tested cohorts (N= Newcastle and S= Spanish) at the cutoff values specified in T7. *P values* <0.05 were highlighted in dark red. Cutoff values of the 9bp group markers were shown in 0.10 while it is 0.08 in the T7 set (marked with asterisks). Among the 17 markers, only the marker GM14-11 showed a significant difference between the 2 tested cohorts.

By comparing LR49-7 threshold curve in both Spanish and Newcastle cohorts, approximately similar curves can be observed as shown in Figure 4-11. At a deletion frequency of 5%, the specificity was 100% and 98% in Newcastle and Spanish cohorts, respectively, and no significant difference was observed in both sensitivity and specificity as shown in Table 4-6.

For the 8bp group of markers, a threshold value of 10% performs better in the Newcastle cohort as both markers achieved a specificity of 100% at that threshold and a sensitivity of 34% and 21% for LR20-8 and GM9-8 respectively as shown in Figure 4-11. The comparison between threshold curves for both 8bp markers (LR20-8 and GM9-8) with their curves on the Newcastle cohort show a high degree of consistency between deletion profiles (*p values* for both sensitivity and specificity between the 2 different cohorts were >0.05).

In the Newcastle cohort, specificity for both polyG/C markers was 100% at a deletion frequency of 10% and sensitivity for the marker IM66-C was 22% and 29% in the Spanish and Newcastle cohorts, respectively as shown in Figure 4-11. Sensitivity for DEPDC2-G was higher in the Spanish than that in the Newcastle cohort at a deletion frequency of 10% (39% in the Spanish vs 17% in Newcastle cohort).

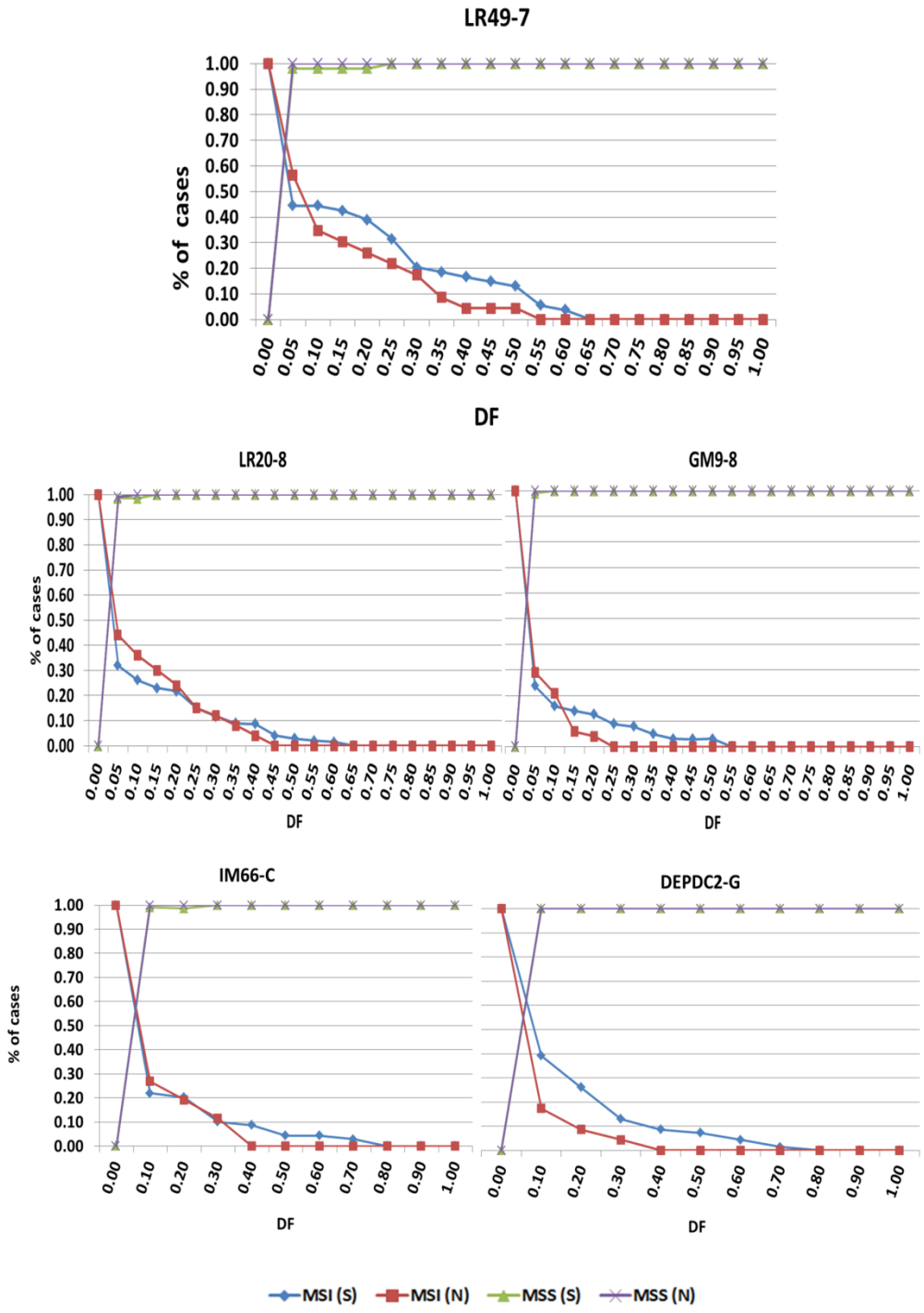


Figure 4-11: Threshold curves of the 7bp, 8bp and Poly G/C markers in both Newcastle and Spanish cohorts. Deletion frequency in the X-axis across the MSI-H samples (blue and red lines) and for MSS samples (purple and green lines). S= Spanish and N= Newcastle cohorts. For MSS samples, 1-proportion of cases was used. There is a clear consistency (no significant difference) in the behaviours of markers across the 2 cohorts.

Comparison of the performance of the 9bp markers in both Spanish and Newcastle cohorts, show that all markers notably have a 100% specificity at a threshold of 10% deletion frequency in the Newcastle cohort, while it was ranging between 95-100% in the Spanish cohort as shown in Figure 4-12 and Table 4-5.

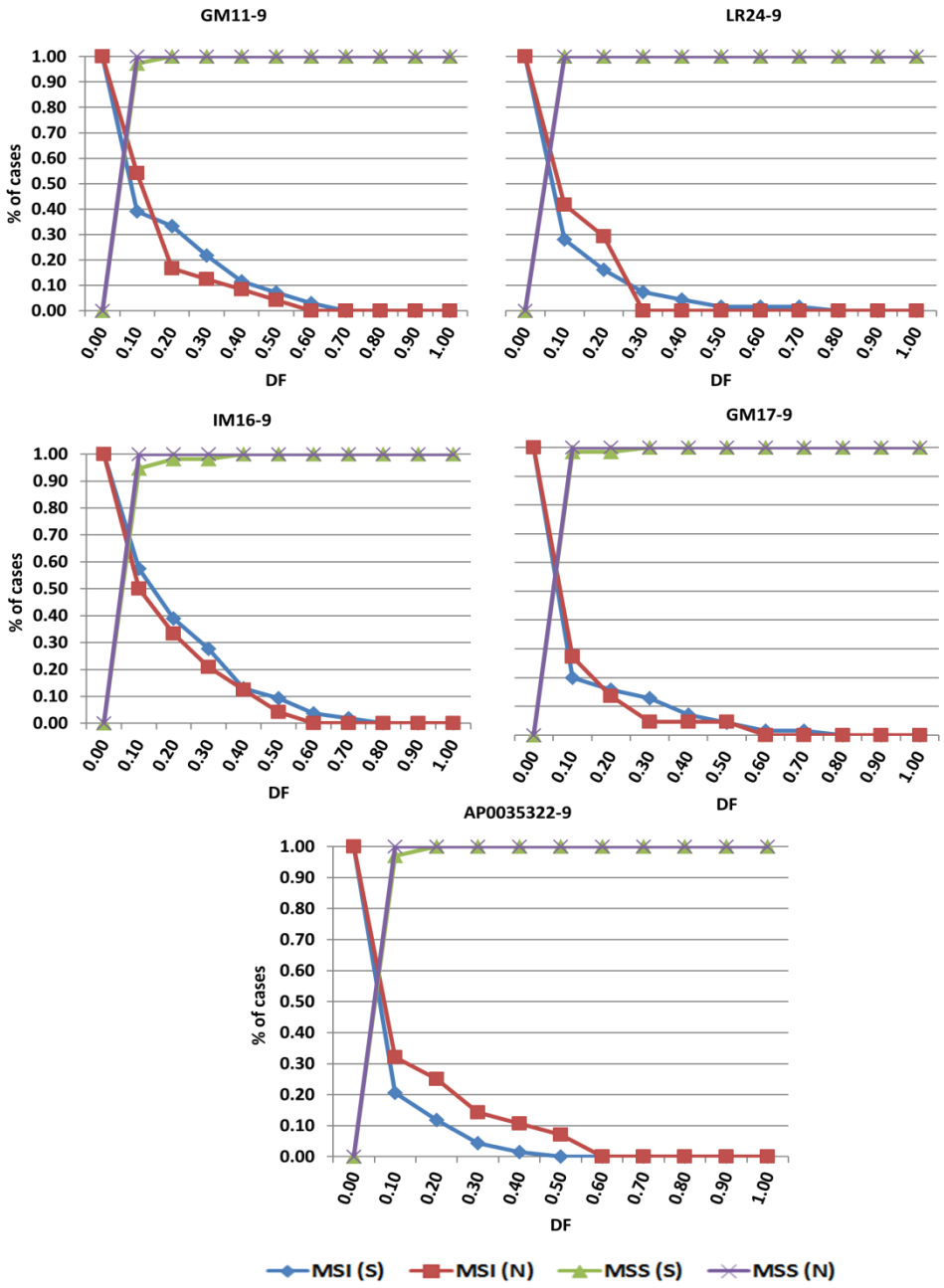


Figure 4-12: Threshold curves of the 9bp markers in both Newcastle and Spanish cohorts. Deletion frequency in the X-axis across the MSI-H samples (blue and red lines) and for MSS samples (purple and green lines). S= Spanish and N= Newcastle cohorts. For MSS samples, 1-proportion of cases was used. S= Spanish and N= Newcastle cohorts.

In the Spanish cohort, none of the 11bp markers showed a 100% specificity at a threshold value of 10% deletion, therefore, the threshold designed so to be at least 0.19 (i.e. 19%). This was the threshold value used in the previous assay (Redford, 2016). At that threshold value (i.e. 0.19), only the marker LR48-11 showed a 100% specificity while specificity for other markers was ranging between 72- 99% with the least specificity was observed in the marker GM14-11 as shown in Figure 4-13. The other threshold values used in T7 is 0.30 (30%) and at that value, additional marker

(LR11-11) approached 100% specificity, while even with this relatively high threshold, specificity for the marker GM14-11 did not reach 90%. The marker GM14-11 showed sensitivity of 75% and specificity of 89% at a deletion frequency of 30%.

When deletion profiles for the 11bp markers were compared against those of the Newcastle cohort, at a threshold value of 0.30, all markers (GM7-11, LR48-11, LR11-11 and GM14-11) showed a specificity of 100% with the highest sensitivity amongst them was for the marker GM7-11 where it reached to 57% at that deletion frequency in the Newcastle cohort. Interestingly, the deletion curve for the marker GM14-11 was significantly different (*p value* <0.05) between the 2 tested cohorts as shown in Figure 4-13 and Table 4-6.

To investigate the possible reason for that significant change in the deletion curve between the 2 tested cohorts, I have retrospectively checked the sequence nature of the GM14-11 amplicon and found that there is a SNP (with a minor allele frequency of 0.06) within the primer binding site. Another SNP is located immediately adjacent to the homopolymer but this is unlikely to be the reason as this SNP was also included in the amplicons of the previous run. Another possible reason for this inter-cohort variability is the ethnic difference between the backgrounds of the 2 cohorts. However, other factors like DNA quality could be contributory factors for this anomalous behaviour of the marker GM14-11.

For 12bp markers in the Spanish cohort, 0.19, 0.25 and 0.30 values were set as cutoff thresholds. Both markers LR44-12 and IM49-12 performed well at a threshold of 0.30 with a specificity of 100% and 99% for both, respectively, and a sensitivity of 77% and 57% respectively as shown in Figure 4-13. The marker LR36-12 showed a sensitivity of 86% and specificity of 92%. Interestingly, about half of the samples of the Spanish cohort had sequencing reads below 100 per/amplicon (which is the cutoff number of reads we used to assess instability) and the average number of sequencing reads (for those which achieved >100 per/amplicon) was relatively low (188 per/ amplicon) as mentioned in section 4.2.1.

By comparing deletion frequencies with those of the Newcastle cohort, there was a high consistency for the marker IM49-12 where both sensitivity and specificity curves looks almost similar as shown in Figure 4-13. For both LR44-12 and LR36-12 markers, there is a notable difference between deletion curves in the different cohorts

with the maximal difference was observed in the marker LR36-12. This is most likely to be due to the relatively low number of sequencing reads obtained from that marker.

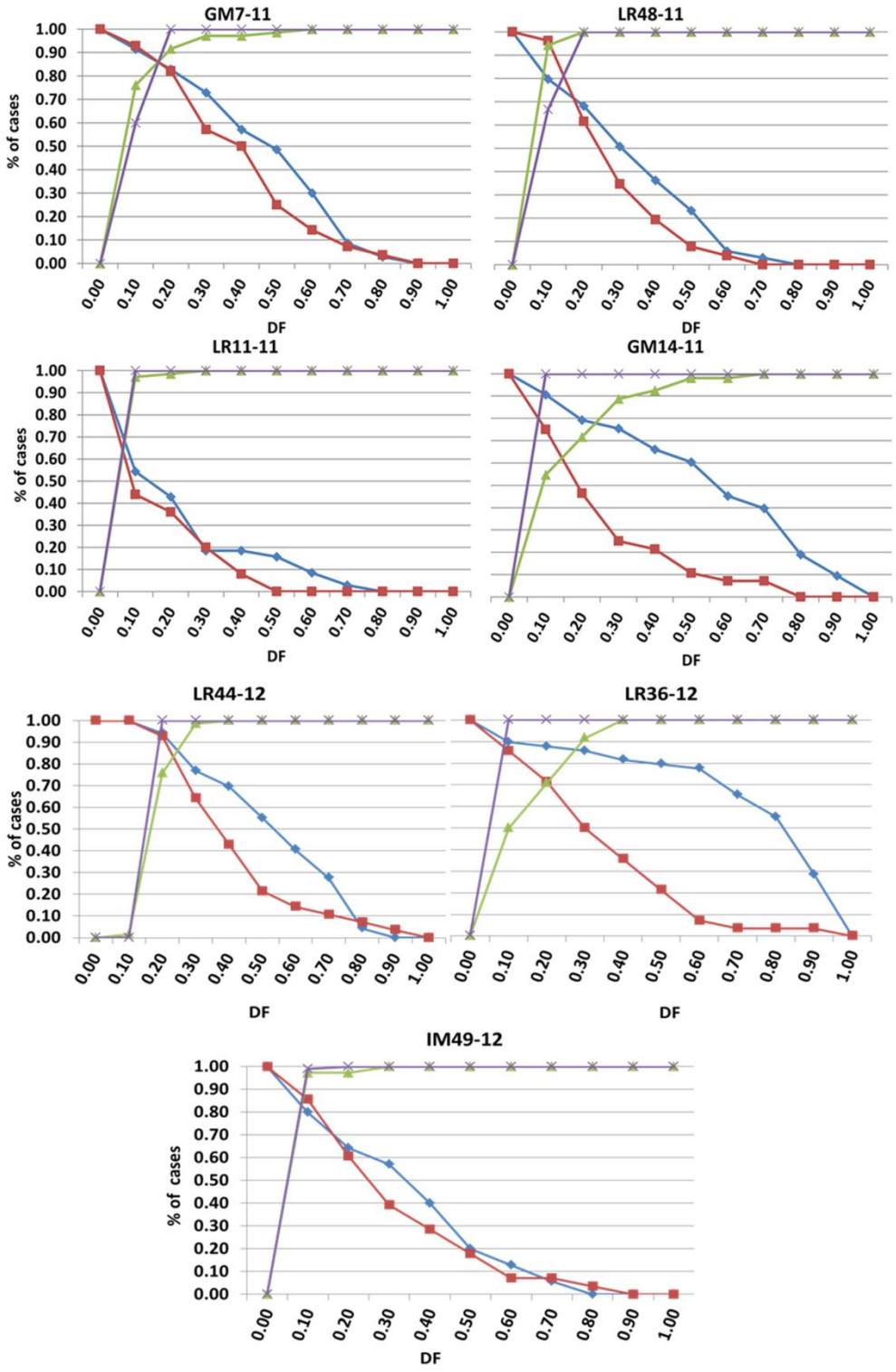


Figure 4-13: Threshold curves of the 11bp and 12bp markers in both Newcastle and Spanish cohorts. Deletion frequency in the X-axis across the MSI-H samples (blue and red lines) and for MSS samples (purple and green lines). S= Spanish and N= Newcastle cohorts. For MSS samples, 1-proportion of cases was used. In the 11bp group, the marker GM14-11 showed a notable difference between the 2 cohorts. In the 12bp group, the marker LR36-12 showed difference in both sensitivity and specificity between the 2 cohorts.

4.2.7. Assessment of DNA fragmentation of a selected subset of Spanish samples

As explained in the previous section, the marker GM14-11 showed the lowest specificity (due to 4 MSS samples exhibited deletion frequency more than 30% for that particular marker). To check for possible reasons behind such anomalous results, these 4 samples were selected for a further investigation of DNA integrity and compared with another group of concordant samples (MSS and MSI-H samples). In addition, the 3 false negative samples that caused the overall sensitivity to be 96%, and a fresh tissue sample, were tested. Results are shown in Figure 4-14. For that purpose, the percentage of DNA that have a size of <100bp (named as DV100) was calculated to be used as an indicator of DNA fragmentation.

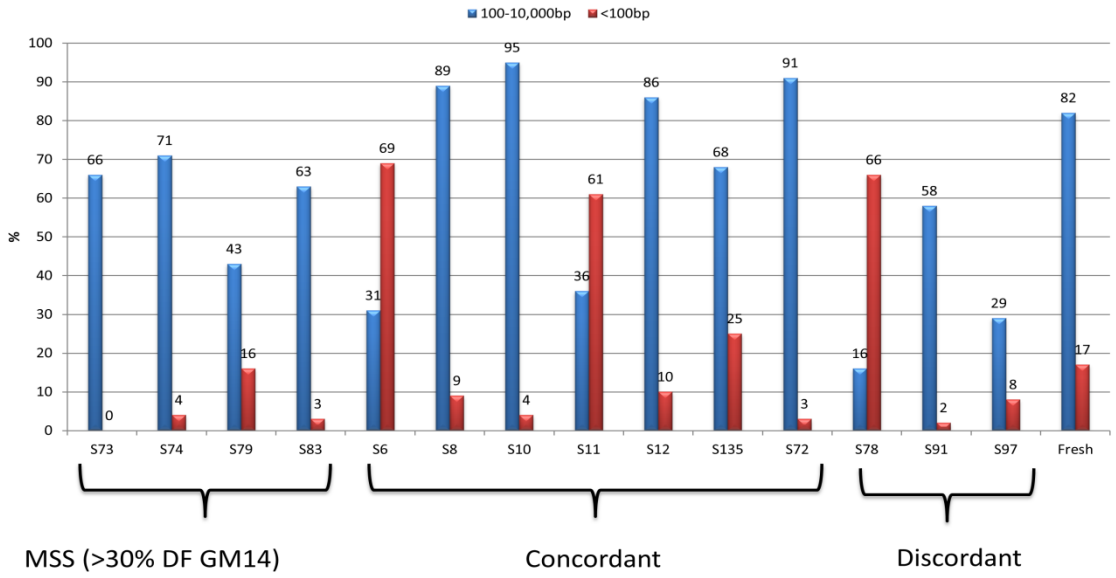


Figure 4-14: The DNA integrity test for a subset of Spanish samples. Samples are plotted in the X-axis and the percentage of DNA in the Y-axis. 3 groups of samples were tested, the 4 MSS samples that showed deletion frequency >30% for the marker GM14-11, 7 concordant samples (MSS and MSI-H) and the 3 discordant false negative samples compared to a fresh normal tissue sample.

The results of that assay showed that 17% of the tested DNA from the fresh tissue sample was <100bp in size and all samples in the first group (the 4 MSS samples with a deletion frequency of >30% in the marker GM14-11) had DV100 <17% of the tested DNA sample. On the other hand, 3 from concordant and 1 from discordant groups showed DV100>17%. These results indicate that there is no clear impact of DNA fragment size on the amplifiability and proper calling of the tested samples.

4.3. Discussion

It has been shown that threshold sets can be set such that an accurate discrimination between MSI-H and MSS samples can be obtained when applied to the 17 marker panel. In the work outlined in this Chapter, this panel was assessed using a large cohort composed of 141 CRCs with mixed phenotypes. Deletion frequency and allelic bias were used as parameters for the classification process and different thresholds were tested to find out the optimal cutoff values. For purposes of assessment, the initial analysis was done while I was blind to the MSI status of all samples in the tested cohort, therefore, thresholds were increased arbitrarily and the samples were categorized based on whether instability exceeds the length-specific threshold or not. In first 3 threshold sets (T1, T2 and T3), the cutoff values of short markers were changed, while for the other sets (T4, T5 and T6), the cutoff values for longer markers (11 and 12bp) were changed and those for short markers were kept constant.

Based on the assumption that allelic bias is a more reliable sign of instability than deletion frequency alone, the difference in the number of samples that have one or more marker with a deletion above the specified threshold and the number of samples that have one or more of the markers that show deletion frequency and allelic bias was closely observed. The number of samples that showed one or more deletion- only marker (DF subgroup) was significantly more than AB subgroup for the first group of threshold sets (T1, T2 and T3). When threshold values increased for the longer repeats, DF and AB numbers started to approximate to each other and the least difference was achieved in T5 where it became 8 samples only (this means those 8 samples have shown a deletion above the threshold in at least one marker but none of these markers show allelic bias). The very consistent and slightly changing values in the AB subgroup across the different threshold sets indicate that the existence of AB is a more confident sign of instability (compared to DF only) as it has been marginally changed with higher threshold values (from T3 upwards).

When the reported phenotype was unlocked, I became able to test the quality of the classification. Sensitivity and specificity for the first 3 threshold sets were low compared to those of T4, 5 and 6. This indicated that with increasing threshold values for the longer homopolymers, most of the faulty results were discarded. This also suggests that the PCR induced errors are far more associated with the longer

homopolymers (11bp and 12bp) compared to short homopolymers (7,8 and 9bp), which is consistent with what was stated by others (Fazekas et al., 2010, Redford, 2016). These results are also consistent with other studies where they found that the longer the repeat, the higher is the degree of length variation (Vilkki et al., 2002, Sammalkorpi et al., 2007, Clarke et al., 2001).

It was obvious from the analysis of threshold sets that the measuring of the allelic bias (AB) was more informative as it changed less with different cutoff values, and had better sensitivities and specificities than their corresponding DF subgroup. This indicates that allelic bias is likely to be a reliable parameter that can be collated in the analysis of our panel. Therefore, we sought combining both deletion frequency and the allelic bias in the scoring would strengthen the conclusion. However, it was evident that low level of microsatellite instability could be found in all CRCs (Laiho et al., 2002), so it was important to create a cautious threshold set in order to avoid misclassification. For that purpose, a weighted scoring system (called as T7) was adopted in which the marker that shows a deletion above the length-specific threshold was given a score of 1 and the marker which shows deletion and allelic bias was given a score of 2. According to T7, samples that have an overall score of equal or more than 3 were called as MSI-H and those with a score below 3 were called as MSS. With the new scoring, a cutoff value of 3 seems more cautious than being 2 only (where a single marker with deletion and allelic bias could achieve that score), as there must be at least 2 or 3 unstable markers for a sample to be called as unstable. Applying that scoring scheme has clearly improved the specificity (100%) while 3 MSI-H samples were miscalled as MSS in our panel (false negatives) giving rise to 96% sensitivity.

Further interrogation of the MMR IHC status of all samples, showed that the concordance rate of MSI scoring system was higher (=95%) than that of the reported MSI phenotype (=93.6%). One sample (out of the 3 false negative cases) has had normal IHC results, consistent with the possibility of being stable rather than MSI-H (as they were reported). All the equivocal samples were retested in an independent MiSeq run, and all of them yielded the same predicted phenotype. These results confirm the initial prediction and it might be worthy to test matched normal tissue for them (if it became available) in order to be 100% sure about the correct phenotype.

Furthermore, gaining the same prediction in 2 separate MiSeq runs give an additional evidence of the repeatability and reproducibility of our assay.

A single sample was called as polymorphic for a single marker (sample 132, marker DEPDC2-G), and this polymorphism was not found in other samples for the same marker. However, testing of a matched normal tissue (if it became available in future) is recommended for such a case to confirm the polymorphism.

The deletion curves were generated for all markers across MSI-H and MSS samples to examine the sensitivity and specificity respectively. Furthermore, deletion curves of all markers were compared to their curves in the Newcastle cohort. There was an overall consistency (no significant changes) in the marker behaviours in both Spanish and Newcastle cohorts. The only noticeable exception was the marker GM14-11. The sensitivity and specificity of the marker GM14-11 in the Spanish cohort were 75% and 89%, respectively. This profile is different (based on both sensitivity and specificity) from what was observed in the Newcastle cohort, with a significant difference (*p value* <0.05) between sensitivities in the 2 tested cohorts. In addition, 4 MSS samples showed deletion frequency >30%. To further investigate the reason behind that difference, the amplicon sequence has been checked and a SNP (with 0.06 MAF) was found in the primer annealing site. DNA integrity has been tested for a selected subset of samples and showed that there is no clear impact of the DNA fragment size on the subsequent analysis in these samples. The marker LR36-12, showed a difference in both sensitivity and specificity between the 2 tested cohorts. However, the sequencing read depth for that marker was relatively low compared to the overall depth (average depth of the LR36-12 amplicons was 188 per/amplicon compared to the 2900 per/amplicon coverage for the overall MiSeq run). In the re-analysis done for the 6 discordant samples, the overall average depth was higher than that in the initial analysis (= 4499 per/amplicon). The average depth for the marker LR36-12 in the re-analysis was higher (617 per/amplicon) than that in the initial analysis. This improvement in the coverage is likely to be due to the use of relatively higher library concentration (10 pM in the re-analysis run compared to 4 pM in the initial analysis run) as explained in section 4.2.5.

In this Chapter, 141 CRC samples were used to test the 17 marker panel. All of these samples were extracted from FFPE blocks and provided in the form of extracted DNA. Although there were some samples that have failed to amplify certain

amplicons, I successfully amplified at least 15 markers in 86% of samples. This is likely to be due to the designation of primers so that to amplify small sized amplicons (100-150bp).

4.4. Conclusions

It was possible for the combined panel to discriminate successfully between MSS and MSI-H samples after adjusting the threshold sets. Different threshold sets were tested and a weighted MSI scoring system was assessed. The weighted MSI scoring system yielded the highest sensitivity and specificity. Moreover, the modifications we made in the library preparation protocol and primer designing were shown to be working, adding more privileges to the overall approach. This suggests that our panel can be used efficiently in the routine diagnostic work for MSI testing. However, further validation would enforce and consolidate the hypothesis of the panel's employability in the clinical practice.

Chapter 5. Analytical validation of the weighted MSI score using an independent cohort of CRCs

5.1. Introduction and aims

5.1.1. Introduction

Prior to being implemented in the clinical laboratories, NGS- based genetic tests need to be extensively validated to ensure they perform efficiently in solving the target problem (discrimination between MSI-H and MSS in case of my test). Because of its recent and fast development, NGS approaches require worldwide standardised quality control guidelines. However, in 2013, the American College of Medical Genetics published the first practice guidelines for clinical laboratories adopting NGS approach in their routine work (Rehm et al., 2013). Recently, the Association of Clinical Genetics Sciences (ACGS) has approved practice guidelines for target NGS guidelines (Deans et al., 2015). Although both of these guidelines are relevant to mutation detection rather than MSI analysis and because there is no consensus guidelines to NGS- based MSI assay, these guidelines could be used to guide the quality requirements for the current assay as our assay represents an example of the targeted NGS assay.

5.1.1.1. Validation of the technical steps involved in the NGS test workflow

In the validation, all technical levels that are involved in the sequencing process should be evaluated, these include:

- **Clinical cases**: The clinical criteria of sample selection need to be carefully optimised and it is essential to use the same kind of samples that will be utilised in practice when the test is implemented.
- **Sample processing**: The preservation, transportation, macrodissection of tissue blocks, DNA extraction and storage of extracted DNA, all these steps need to be valid and follow careful instructions.
- **Targeting technique**: There are 2 main approaches currently available for targeting the region of interest (ROI), these are:
  1) PCR based techniques: In this approach, sequence specific oligonucleotide primers that flank the region of interest were utilised to amplify the ROI (also known as amplicon targeted NGS).

2) PCR-free protocols: in which, the target DNA directly subjected to library preparation and sequencing, thus cutting the cost and time and obliterate the PCR induced sequence artefacts. Hybridisation based assays is an outstanding example of the PCR free approaches.

- **Library preparation**: it has been recommended that all steps of the library preparation need to be carefully monitored, including pooling, clean up, barcoding and normalisation.
- **Sequencing**: it is essential to choose the sequencing platform that fits the purposes for which the test has been developed.
- **Data analysis**: during data analysis, all run metrics need to be interrogated and registered. These include quality score, cluster density, coverage depth and variant calling.

5.1.1.2. Assessment of reproducibility

It is recommended to achieve a reproducible test that fit with the clinical laboratory's requirements. It is **recommended** to test the NGS based assay (NGS based MSI assay in case of the current study) in at least 3 independent sequencing runs and then to document the concordance rate of results among cohorts (Rehm et al., 2013).

5.1.1.3. Calculation of sensitivity and specificity

One of the **essential** requirements for the NGS based test is to check the sensitivity and specificity of the target test. Sensitivity can be established by comparing the results obtained from the new test with results from a gold standard test. Analytical sensitivity is defined as the proportion of cases that tested positive by the assay and reported as positives by the gold standard assay (i.e. true positive) (Rehm et al., 2013), whereas analytical specificity is defined as the proportion of cases that were predicted as negative by the assay test and reported as negatives by the gold standard assay (i.e. true negative). Determination of the optimal sensitivity depends on the downstream application, for instance, for an ideal targeted NGS test designed for mutation detection, an error rate of a heterozygous/ homozygous mutation is recommended to be ≤ 5% (with 95% confidence) (Mattocks et al., 2010,

Deans et al., 2015). As our assay is still under development and there is no recommended sensitivity and specificity for such an assay, the aim is to achieve the highest concordance rate with the gold standard test during validation.

5.1.1.4. Establishing an optimal read depth (Coverage)

Read depth and coverage depth are used interchangeably in the literature, but they refer to the same definition. Coverage (read depth) is the number of times to which a base has been sequenced in a sequencing reaction. Read depth varies depending on the sequencing chemistry, sequencing approach, quality of the template DNA and other factors. The larger the read depth, the bigger the confidence the base is called correctly. It is essential to assess the read depth of all amplicons in the next generation sequencing run in addition to the minimum depth. Determination of the minimum read depth should be established during the validation assay and should meet the required criteria for the specified aim for instance, the coverage depth was recommended to be high in cancer samples in order to make it possible to detect those variants with low prevalence (Deans et al., 2015).

Based on this, the NGS based assay need to be validated to ensure the ability of the test to solve the target problem for which it has been designed. In the previous 2 chapters, the microsatellite panel of short homopolymers (which was composed of 17 markers) was tested across 2 different cohorts. The assessment of instability was done by testing different threshold sets and the final MSI scoring system showed the highest sensitivity and specificity.

In this chapter, the weighted MSI scoring system will be validated using a new independent cohort to consolidate the initial findings and further check the overall performance of the assay as a requirement for validation.

5.1.2. Aims

In the previous chapter, the short mononucleotide panel (which is composed of 17 short mononucleotide markers) was assessed across a large cohort represented by 141 CRC samples, which was composed of 68 MSI-H and 73 MSS samples. The tested approach showed a high sensitivity and specificity (96% and 100%, respectively) in that part of the study using the weighted MSI scoring system. To further validate the developed panel and the weighted MSI scoring system, a new

cohort of 100 CRC samples was used. These 100 CRCs were obtained from Department of Molecular Pathology, University of Edinburgh, UK. This chapter will outline the assessment and validation of the weighted MSI scoring system using this cohort. The overall aims of the work presented in this chapter are to:

- Assess a cohort of 100 previously analysed CRCs, blinded to MSI status using the optimal threshold set defined previously (weighted MSI scoring system).
- Assess the role of allelic bias as an additional parameter can be used in MSI calling, by assessing the sensitivity and specificity of allelic bias subgroups in each threshold set.
- Compare threshold curves of each individual marker with those from different cohorts (that were done in chapter 3 and 4) to compare their behaviour in different cohorts and detect if there is anomaly for each marker individually.
- Assess the quality metrics (Depth, quality score, cluster density and others) and provide bases for suggested values in an ideal MiSeq run.

5.2. Results

5.2.1. Amplification and sequencing of the short mononucleotide panel using a cohort of 100 CRC samples

For purposes of validation, a new independent cohort of 100 CRCs of a mixed MSI phenotype referred from Edinburgh (Dr Mark Arends, Department of Molecular Pathology, University of Edinburgh, UK). These cases were extracted, quantified and MSI tested in the original laboratory, and therefore, they were provided in the form of an extracted DNA. As a prerequisite for the validation, I have analysed the cohort blindly (without knowing their MSI status).

All samples were amplified using the 17 short mononucleotide markers panel. The primers were the same as those used in chapter 4. PCR amplifications were performed using the Herculase II Fusion DNA Polymerase (Agilent Technologies, CA, USA) in 35 PCR cycles. Out of those 100 samples, 70 were amplified using the 17 marker panel and a total of 1167 amplicons were generated. The library concentration used in the current MiSeq run was 10 pM.

Out of the 70 samples, 20 samples were included in the previous MiSeq run alongside with the Spanish cohort. The remaining 50 Edinburgh samples were

analysed in a separate run. The cluster density was 1450 k/mm^2 and the Q30 of the MiSeq run was 55.5% as shown in Figure 5-1.

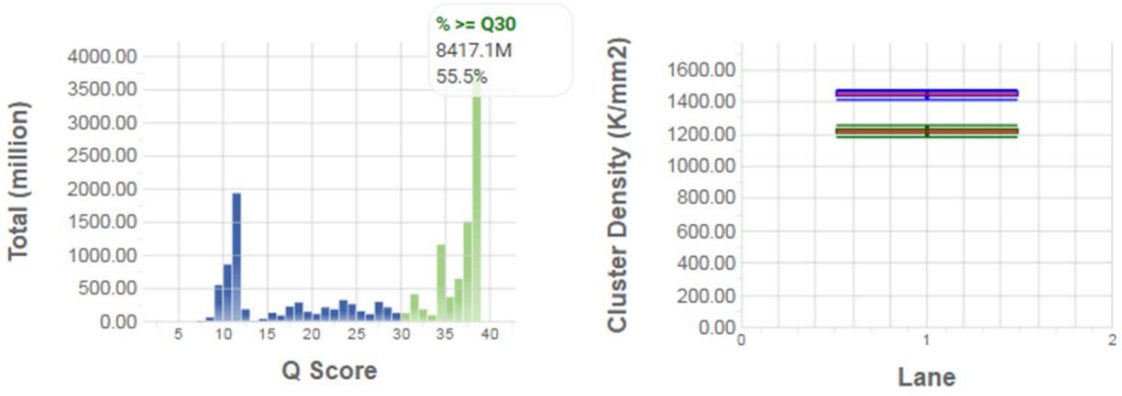


Figure 5-1: Q score (on the left) and cluster density (on the right) of the MiSeq run. The green box in cluster density graph is near to the blue box indicates that most of the reads passed filter.

The data outputs of the MiSeq run were generated in the form of FASTQ files which were analysed as explained in Chapter 2. Although all DNA samples were extracted from FFPE tissue curls, the vast majority of the 70 samples (99%) was successfully amplified, sequenced and called to adequate depth (i.e. ≥ 100 paired end read/ amplicon) for at least 15 markers of the tested 17 markers as shown in Table 5-1. Notably, a single sample was called for only 4 markers to ≥100 per/ amplicon. This sample was excluded from the downstream analysis. So the overall number of samples that were included in the downstream analysis was 69 samples.

No. of markers	No. of cases		
	Amplified	**Sequenced**	**Called ≥100 per/amplicon**
17	55	34	34
16	8	30	29
15	6	5	6
14	1	0	0
13	0	0	0
12	0	0	0
11	0	1	0
10	0	0	0
9	0	0	0
8	0	0	0
7	0	0	0
6	0	0	0
5	0	0	0
4	0	0	1
3	0	0	0
Total	70	70	70

Table 5-1: Number of amplicons that were amplified, sequenced and called to adequate depth (i.e. ≥100 reads) in Edinburgh cohort. The majority of samples amplified and sequenced ≥ 15 markers. **per**= paired end reads.

5.2.2. Analysis of the sequencing data and assessment of the weighted MSI scoring system (Threshold 7)

As explained in previous chapters, the COPReC data were generated to allow recognition of sequencing reads that generated from both SNP alleles. The two main parameters used in the subsequent analysis were deletion frequency and allelic bias. Allelic bias was calculated for heterozygous amplicons by Fisher's Exact test to assess its significance. The initial analysis was done while I was unaware of the MSI status of the tested samples. After the assessment of threshold sets, the phenotype key was unlocked in order to make a direct comparison between the predicted phenotype (by our panel) and the reported phenotypes (which is provided by the original lab) and calculate the sensitivity, specificity as explained in Chapter 2 section 2.10.6.

As done in Chapter 4, the weighted MSI scoring system (threshold 7) was designed by merging both deletion frequency (DF) and allelic bias (AB) features for

all markers in each sample. The score was calculated by giving a score of 1 for each marker with only deletion frequency more than cutoff values and a score of 2 for each marker with DF+AB. Those samples with an overall score of ≥3 were called as MSI-H and those with a score of <3 were called as MSS. Applying this scoring resulted in 36 samples to be called as MSI-H and 33 as MSS as shown in Figure 5-2.

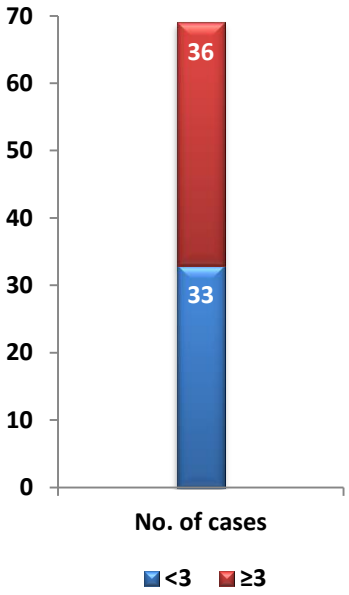


Figure 5-2: The distribution of samples, according to cutoff values in T7,where the new MSI score was used. Samples that showed an overall score of ≥3 are represented in red bar and those with a score less than 3 are presented as blue bar.

In this MSI scoring, the sensitivity and specificity were both approached 100% indicating a high efficiency in discrimination between MSI and MSS samples. By applying T7 across the tested cohort, all MSI-H samples have had a score more than 3 while all MSS samples have got an overall score of less than 3 as shown in Figure 5-3. These results consolidate the initial assessment of this system that done in Chapter 4.

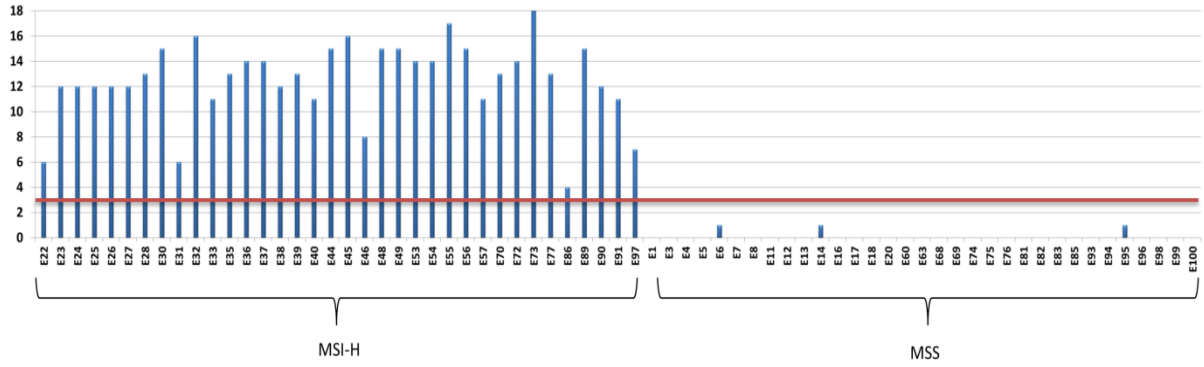


Figure 5-3: The overall MSI score for all tested Edinburgh samples (69 samples). Y axis represents the overall score. All MSI-H samples have got a score above 3 compared to MSS samples which have not.

Then, I tested the same threshold sets that were proposed and assessed in Chapter 4 with their cutoff values shown in Table 5-2. This assessment was to observe the performance of these different cutoff values in this particular cohort.

Threshold set	7bp	8bp	9bp	PolyG/C	11bp	12bp	GM14
Threshold 1	0.05	0.05	0.05	0.1	0.19	0.19	
Threshold 2	0.05	0.05	0.05	0.1	0.19	0.19	- GM14
Threshold 3	0.05	0.05	0.08	0.1	0.19	0.19	- GM14
Threshold 4	0.05	0.05	0.08	0.1	0.19	0.25	- GM14
Threshold 5a	0.05	0.05	0.08	0.1	0.19	0.30	- GM14
Threshold 5b	0.05	0.05	0.08	0.1	0.19	0.30	+GM14
Threshold 6a	0.05	0.05	0.08	0.1	0.30	0.30	-GM14
Threshold 6b	0.05	0.05	0.08	0.1	0.30	0.30	+GM14

Table 5-2: Threshold sets and their cutoff values. In each set, a single new event was introduced (red coloured) compared to the preceding threshold set.

When these threshold sets were applied, the number of samples that were called as unstable, especially in the DF subgroup was relatively higher in the initial threshold sets (T1- T3) compared to higher threshold sets (T4- T6), with the lowest number of samples observed in the T6 threshold set. As noted in Chapter 4, the specificity of the DF subgroup in the first 3 threshold sets (i.e. T1-T3) was generally low, with the lowest specificity (= 59%) observed in T1 DF set. The sensitivity was 100% for all threshold sets (DF subsets), as shown in Table 5-3.

The specificity was higher in the AB subgroup compared to the DF subgroup across all threshold sets. This likely justifies the inclusion of the allelic bias as an additional parameter to call instability compared to deletion only.

Threshold set	≥1	0	Sensitivity	Specificity
T1.DF	50	19	100%	59%
T1.AB	33	36	92%	97%
T2.DF	49	20	100%	62%
T2.AB	33	36	97%	97%
T3.DF	49	20	100%	62%
T3.AB	33	36	92%	97%
T4.DF	41	28	100%	85%
T4.AB	33	36	92%	100%
T5a.DF	41	28	100%	85%
T5a.AB	33	36	92%	97%
T5b.DF	43	26	100%	79%
T5b.AB	33	36	89%	97%
T6a.DF	40	29	100%	88%
T6a.AB	32	37	89%	97%
T6b.DF	40	29	100%	88%
T6b.AB	32	37	89%	97%

Table 5-3: The sensitivity, specificity for all threshold sets in both DF and AB subgroups. DF, deletion frequency subgroup and AB= allelic bias subgroup.

5.2.3. Comparison of threshold curves for all markers in the 3 independent cohorts

As done before, deletion frequencies for each marker were plotted across all MSS (specificity) and MSI-H (sensitivity) samples to construct the deletion curve (or threshold curve). Furthermore, deletion curves for each marker were compared to those curves from previous cohorts (Newcastle and Spanish) to investigate the difference between the values in different cohorts at a specific deletion frequency as shown in Table 4-5. The significance of difference (*p value*) between sensitivities and specificities across the different cohorts were calculated as explained in Chapter 2, section 2.10.6, with *p values* <0.05 considered as significant. Threshold curves for MSS samples were constructed in the form of 1- the percentage of MSS samples.

Markers	Cutoff value	Sensitivity			*p* value (Sensitivity)		Specificity			*p* value (Specificity)	
		E	N	S	E *vs* N	E *vs* S	E	N	S	E *vs* N	E *vs* S
LR49-7	0.05	56%	57%	44%	1.00	0.5796	100%	100%	98%	1.00	1.00
LR20-8	0.05	36%	44%	32%	0.6267	0.8385	97%	99%	98%	1.00	1.00
GM9-8	0.05	26%	29%	24%	1.00	1.00	100%	100%	99%	1.00	1.00
IM66-C	0.10	75%	29%	22%	0.0690	0.0070	100%	100%	99%	1.00	1.00
DEPDC2-G	0.10	30%	17%	39%	0.5497	0.6874	100%	100%	100%	1.00	1.00
GM11-9	0.10*	58%	54%	39%	1.00	0.2834	100%	100%	97%	1.00	1.00
LR24-9	0.10*	60%	42%	28%	0.4975	0.0560	100%	100%	100%	1.00	1.00
IM16-9	0.10*	56%	50%	57%	0.4837	1.00	100%	100%	95%	1.00	0.5480
GM17-9	0.10*	46%	27%	20%	0.3105	0.0560	100%	100%	99%	1.00	1.00
AP-9	0.10*	36%	32%	21%	1.00	0.2642	97%	100%	97%	1.00	1.00
GM7-11	0.30	94%	57%	73%	0.2462	0.4505	100%	100%	97%	1.00	1.00
LR48-11	0.30	83%	35%	51%	0.0574	0.1458	100%	100%	100%	1.00	1.00
LR11-11	0.30	64%	20%	19%	0.0521	0.0030	97%	100%	100%	1.00	1.00
GM14-11	0.30	82%	25%	75%	0.0158	0.8691	100%	100%	89%	1.00	0.093
LR44-12	0.30	92%	64%	77%	0.4441	0.6499	100%	100%	99%	1.00	1.00
LR36-12	0.30	88%	50%	86%	0.2267	1.00	100%	100%	92%	1.00	0.2123
IM49-12	0.30	78%	39%	57%	0.1442	0.3398	100%	100%	100%	1.00	1.00

Table 5-4: Sensitivity and specificity of all markers in the 3 cohorts (E= Edinburgh, N= Newcastle and S= Spanish cohorts) at the cutoff values specified in T7.*P values* <0.05 were highlighted in dark red boxes. Cutoff values of the 9bp group markers were shown as 0.10 while it was 0.08 in the T7 set (marked with asterisks). Among the 17 markers, 3 markers (IM66-C, LR11-11 and GM14-11) showed significant difference in sensitivity among cohorts.

By comparing the deletion profile of the marker LR49-7 with those in other cohorts (i.e. Newcastle and Spanish), the specificity was 100% for both Newcastle and Edinburgh cohorts at a deletion frequency of 5%, while it was 98% in the Spanish cohort. The sensitivity, on the other hand, was approximately the same for both Newcastle and Edinburgh cohorts (56% and 57%, respectively), while it was 44% for the Spanish cohort at the same deletion frequency as shown in Figure 5-4. However, there was no significant difference in both sensitivity and specificity between the 3 tested cohorts as shown in Table 4-5.

By comparing deletion profiles of the 8bp markers in the 3 different cohorts, the marker GM9-8 showed 100% specificity at the deletion frequency of 5% for both Newcastle and Edinburgh cohorts and 99% in the Spanish cohort. The sensitivity was ranging between 24-29%, with the lowest was observed in the Spanish cohort at 5% deletion frequency. There was no significant difference, neither in sensitivity nor in specificity across the 3 tested cohorts.

For the marker LR20-8, at the same deletion frequency (i.e. 5%), the specificity was ranging between 97% (in the Edinburgh cohort) to 99% (in the Newcastle cohort) and sensitivity was ranging between 32% (in the Spanish cohort) and 44% (in the Newcastle cohort) as shown in Figure 5-4. 100% specificity for the marker LR20-8 was achieved at a deletion frequency of 10% (in Newcastle), 20% (in Spanish cohort) and 40% in the Edinburgh cohort.

Comparison of deletion profiles for both Poly G/C markers across the different cohorts shows that, for the marker DEPDC2-G, the specificity was 100% at a deletion frequency of 10% for all cohorts. The least sensitivity (17%) was observed in the Newcastle cohort while the highest (39%) was observed in the Spanish cohort at the same deletion frequency.

For the marker IM66-C, the specificity at a deletion frequency of 10% in both Newcastle and Edinburgh cohorts was 100% while it was 99% in the Spanish cohort. At the same deletion frequency, the highest sensitivity (75%) was observed in the Edinburgh cohort while the lowest (22%) was observed in the Spanish cohort as shown in Figure 5-4. Interestingly, there was a significant difference in sensitivity between Edinburgh and Spanish cohorts at deletion frequency of 10% as shown in Table 4-5.

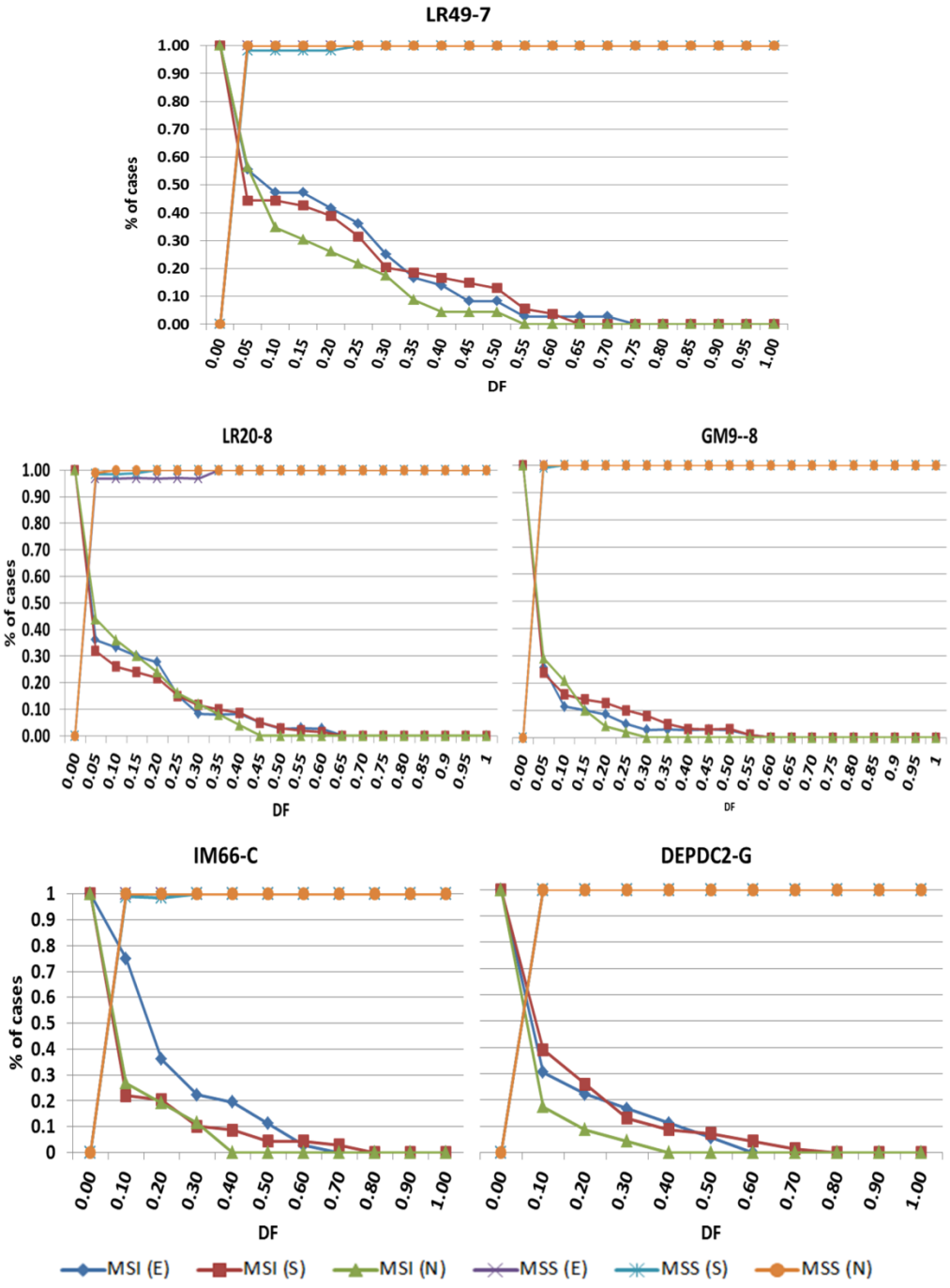


Figure 5-4: The deletion curve of the 7bp, 8bp and poly G/C markers in both MSS and MSI-H samples in the 3 independent cohorts (Newcastle, Spanish and Edinburgh). Y- axis refers to the proportion of samples, X- axis refers to the deletion frequency. Deletion curves of MSS group were plotted in 1- of MSS cases. E= Edinburgh, S=Spanish and N= Newcastle cohorts.

By comparing the deletion profile of the 9bp markers across the different cohorts, the marker LR24-9 was the only marker that showed 100% specificity across all cohorts at a 10% deletion frequency. At the same deletion frequency, 3 other markers (GM11-9, IM16-9 and GM17-9) have shown 100% specificity both in Newcastle and Edinburgh cohorts, but not in the Spanish cohort. The marker

AP0035322-9, have shown a 100% specificity at 10% deletion frequency only in the Newcastle cohort as shown in Figure 5-5 and there was no significant difference neither in sensitivity nor in specificity across all cohorts at deletion frequency of 10% as shown in Table 4-5.

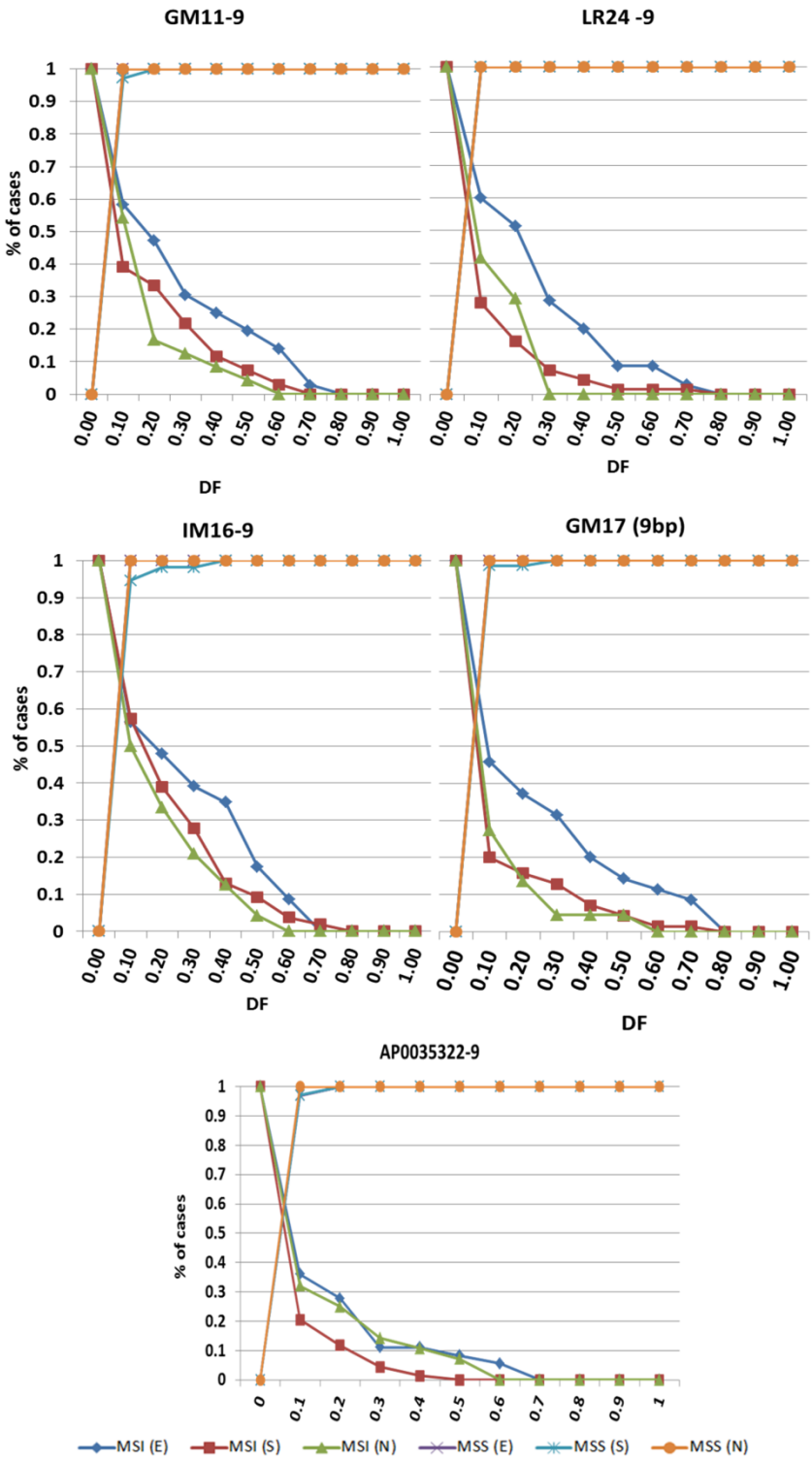


Figure 5-5: The deletion curve of the 9bp markers in both MSS and MSI-H samples in 3 independent cohorts (Newcastle, Spanish and Edinburgh). Y-axis refers to the proportion of cases; X- axis refers to the deletion frequency. Deletion curves of MSS group were plotted in 1- of MSS cases. E= Edinburgh, S=Spanish and N= Newcastle cohorts.

By comparing the deletion profiles of the 11bp markers across the different cohorts, a consistency in the specificity was observed in the marker LR48-11 at 30%

deletion frequency (100% specificity in all cohorts), while a mild difference was observed in the specificity curves of the markers LR11-11 (specificity was 99% in the Edinburgh cohort). For both markers (i.e. LR48-11 and LR11-11), the sensitivity at 30% deletion frequency was higher in the Edinburgh cohort (83% and 64%, respectively). For the marker LR11-11, there was a significant difference between sensitivities in Edinburgh and Spanish cohorts at a deletion frequency of 30% as shown in Table 4-5.

For the marker GM7-11, the specificity was the same for both Newcastle and Edinburgh at 30% deletion frequency (100% for both), while it was 97.2% in the Spanish cohort at the same deletion frequency. However, the specificity for the marker GM7-11 approached 100% at 60% deletion frequency. There was no significant difference among various cohorts, neither in sensitivity nor in specificity as shown in Table 4-5.

At 30% deletion frequency, the specificity of the marker GM14-11 was 100% for both Newcastle and Edinburgh cohorts, while it was 89% at the same deletion frequency in the Spanish cohort. At the same deletion frequency (30%), sensitivity was 82%, 75% and 25% in the Edinburgh, Spanish and Newcastle respectively. In the Spanish cohort, the specificity persisted suboptimal until 70% deletion frequency as shown in Figure 5-6. There was a significant difference in sensitivity between Edinburgh and Newcastle cohorts as shown in Table 4-5.

By comparing the deletion profiles of the 12bp markers in the different cohorts, there was a clear consistency (no significant difference) in the deletion profiles of the marker IM49-12 across the three cohorts, where the specificity was 100% at 30% deletion frequency for all cohorts and the sensitivity was ranging between 39- 78%.

For the marker LR44-12, specificity was 100% in both Newcastle and Edinburgh cohorts while it was 99% in the Spanish cohort at 30% deletion frequency. The sensitivity was ranging between 64- 92% in the three cohorts.

Interestingly, the marker LR36-12 showed a difference in the specificity curve between Spanish and Edinburgh cohorts at 0.30% deletion frequency as shown in Figure 5-6 and Table 4-5. Notably, the LR36-12 amplicon was sequenced to a relatively low depth (average depth was 188 paired end reads per amplicon) in that

cohort as mentioned in Chapter 4, section 4.2.1. This might be the underlying reason behind this alteration in the deletion curve among cohorts.

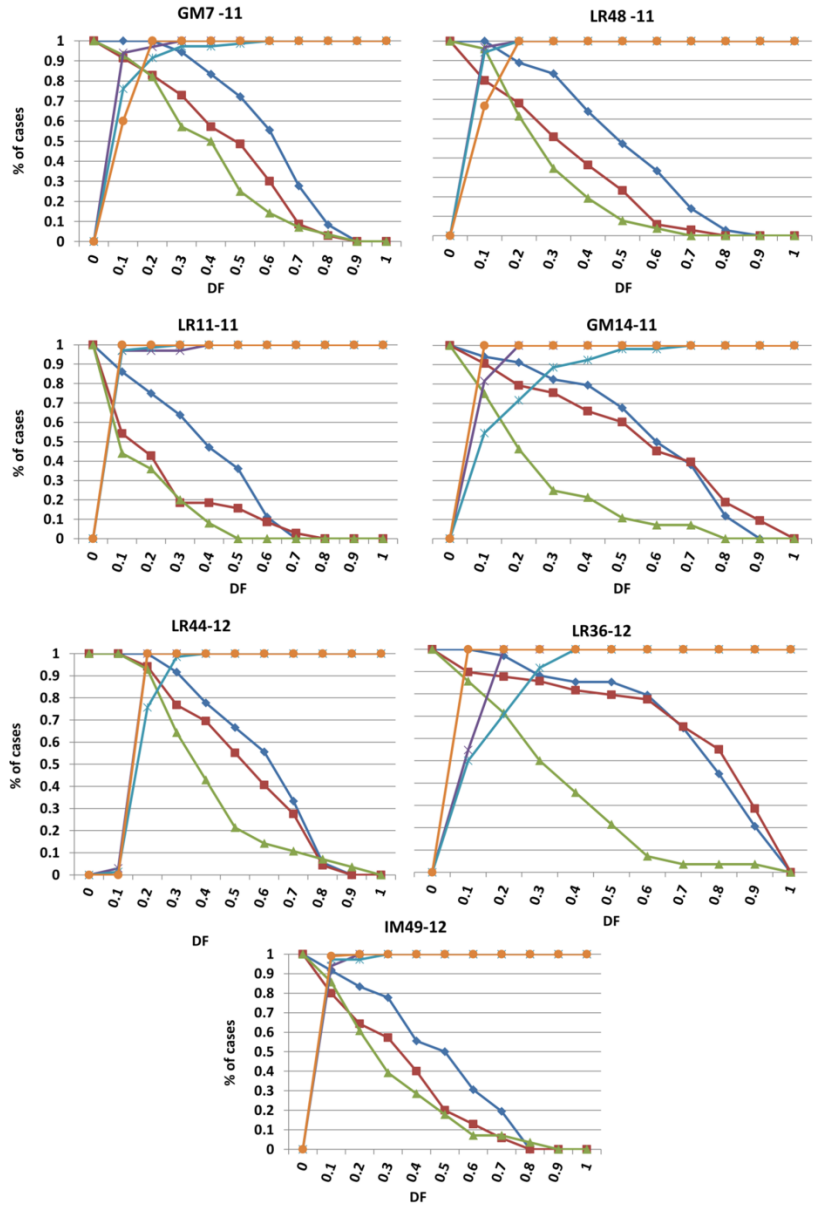


Figure 5-6: The deletion curve of the 11 and 12bp markers in both MSS and MSI-H samples in 3 independent cohorts (Newcastle, Spanish and Edinburgh). Y-axis refers to the proportion of cases; X- axis refers to the deletion frequency. Deletion curves of MSS group were plotted in 1- of MSS cases. The most obvious differences are observed in markers GM14-11 and LR36. E= Edinburgh, S=Spanish and N= Newcastle cohorts.

In conclusion, there was a clear consistency in the deletion curves of almost all markers in the Newcastle and Edinburgh cohorts. However, in the Spanish cohort, the most prominent variation was observed in the marker GM14-11, where it showed

the lowest specificity at 30% deletion frequency. Therefore, a further assessment of this marker would be useful to find out the possible reason underpin that variation.

5.2.4. Assessment of the inter-cohort inconsistency of marker GM14-11 deletion

5.2.4.1. Investigating the nature of the nucleotide sequence of the marker GM14-11

As shown in chapter 4, the marker GM14-11 showed the lowest specificity amongst other markers in the Spanish cohort. In this chapter, a comparison of the deletion curves of the marker GM14-11 across the 3 different cohorts showed that there is a significant difference between deletion curves in Edinburgh and Newcastle cohorts. Furthermore, the marker GM14-11 showed the lowest specificity in the Spanish cohort at 30% deletion frequency as explained in Table 5-4. To check for the possible reasons behind the different behaviour of that marker, the primer set was tested by the *in silico* PCR tool in UCSC genome browser. There were 3 noteworthy findings that might contribute to the variation of the marker, these are:

1) There was a SNP at the primer annealing site (rs539119173); however, that SNP has a very low minor allele frequency (0.06%).
2) There is a high MAF frequency SNP immediately adjacent to the target homopolymer (rs6804861) with a minor allele frequency of 0.37.
3) By direct observation of the nucleotide sequence of the GM14-11 amplicon, another shorter homopolymer (8bp (A) homopolymer) was found in the close vicinity of the target 11bp (A) homopolymer. This might generate an interrupted microsatellite tract, giving rise to compound instability. However, this is unlikely to be the reason, as the whole amplicon was included in the other cohorts.

In order to check the origin of the variant reads, BAM files of the GM14-11 amplicons for a selected group of MSI-H and MSS samples from the Spanish cohort were visualised by the Integrative Genome Viewer (IGV) (Robinson et al., 2011). The selected samples were composed of 4 MSI-H and 4 MSS samples, all of them showed a high deletion frequency of the marker GM14-11. For the MSI-H samples, re-alignment showed that all deletions were observed in the target homopolymer and none of them came from the adjacent homopolymer as shown in Figure 5-7.

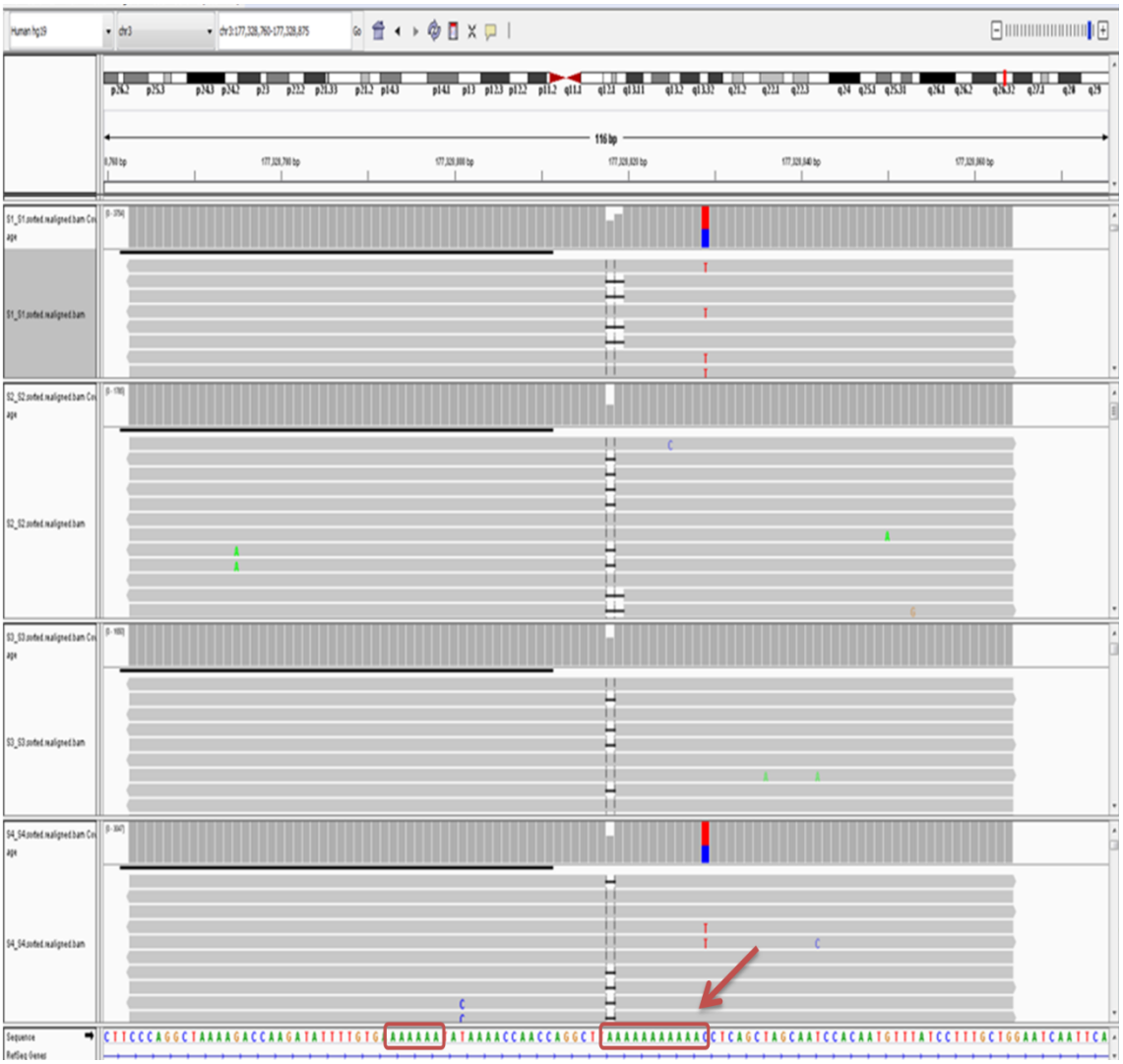


Figure 5-7: The alignment of the sequencing reads of the marker GM14-11 in 4 MSI-H samples. The target (A) homopolymer is red squared and marked by the red arrow, another adjacent poly A homopolymer can be observed few bases away from the target homopolymer (red rectangle). Obviously, all deletions were registered for the target homopolymer.

For the MSS group, none of the examined 4 MSS samples showed any deletion in the adjacent homopolymer as shown in Figure 5-8, indicating that all deletions registered in both groups (MSS and MSI-H) were originated from the same target homopolymer.

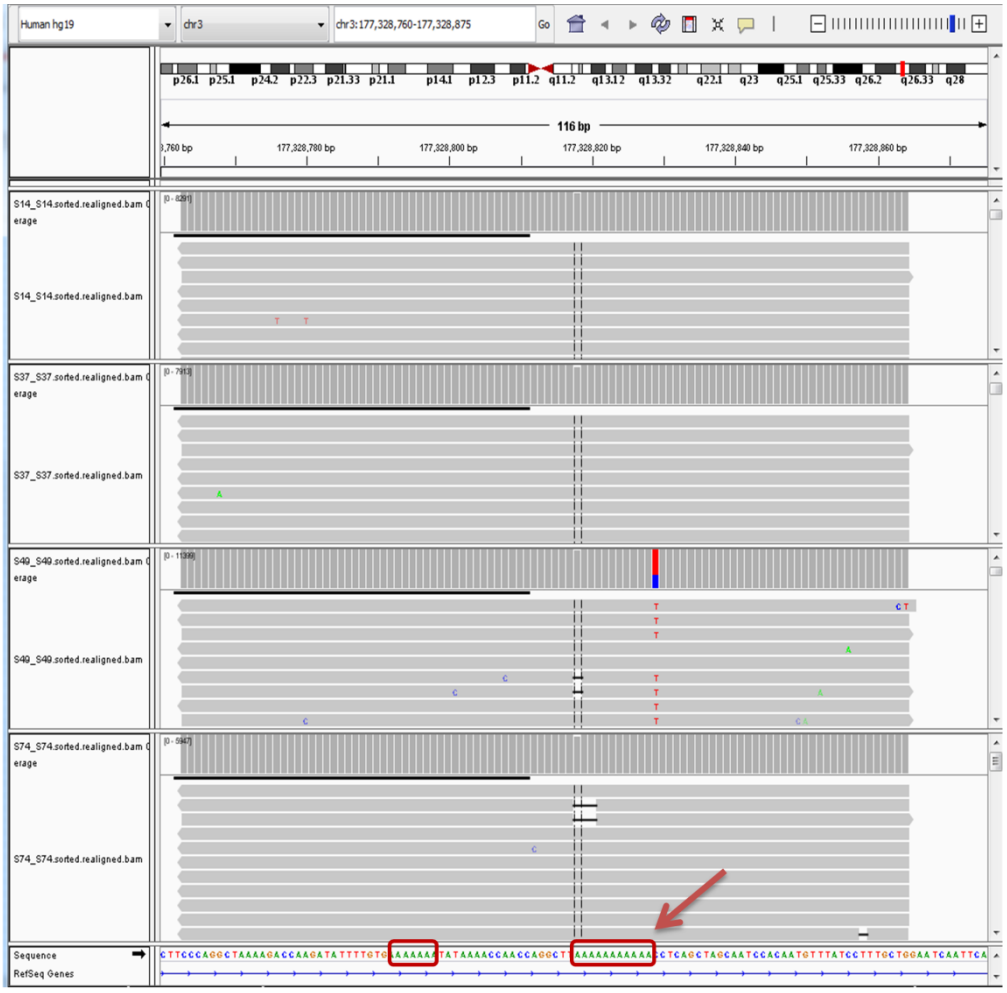


Figure 5-8: The alignment of the sequencing reads of the marker GM14-11 in 4 MSS samples. The target (A) homopolymer is red squared and marked by the red arrow, another adjacent poly A homopolymer can be observed few bases away from the target homopolymer (red rectangle). All deletions were registered for the target homopolymer.

Given the possible variability in the microsatellite tracts, the possibility of marker being a polymorphic cannot be excluded taking in account that the kind of populations is a major difference among the tested cohorts. However, to confirm that, matched normal tissue samples need to be tested.

5.2.4.2. Re-amplification of the marker GM14-11 using samples from Newcastle and Spanish cohorts

To examine the performance of the marker across different cohorts, GM14-11 was tested against a selected subset composed of 8 Spanish samples (7MSS and 1 MSI-H), and 16 Newcastle (8 MSS and 8 MSI-H) samples. All these samples were previously tested with the marker GM14-11, so a comparison between deletion frequencies from all samples would be possible. The new analysis was done using a

library concentration of 10 pM, which is different from both of the previous MiSeq runs (8 pM and 4 pM concentration in Newcastle and Spanish Miseq runs respectively). The 16 samples from the Newcastle cohort were tested previously for the marker GM14-11 using the long amplicon primers. This represented an additional testing of repeatability of the tests and the reproducibility of the results. Furthermore, this can be used to compare the results from the 2 different sets of primers that target the same homopolymer (300bp primers were used in Newcastle cohort, while ~150bp primers were used in the Spanish cohort). The deletion frequencies in the re-analysis test were broadly consistent with the initial results (reported from previous testing) as shown in Figure 5-9 with an average DF proportion (= DF re-analysis/ DF initial) = 1.1.

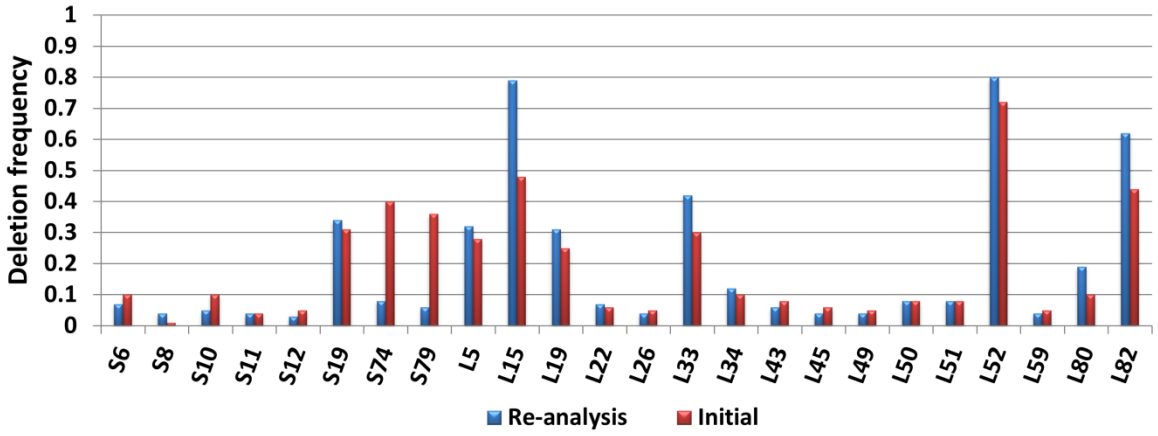


Figure 5-9: Deletion frequencies of the marker GM14-11 in 24 samples from 2 different cohorts. Re-analysis= refers to the deletion frequency from the repeated (re-amplification) analysis, Initial= the deletion frequency in the initial analysis. The frequencies are generally consistent except for 2 samples, S74 and S79.

Two notable exceptions can be observed in samples S74 and S79, where there is a notable difference in the deletion frequencies between the initial and the re-analysis deletion frequency (DF proportion= 0.2 for each). Notably, this alteration changed the prediction for both samples, as they both showed deletion frequency >30% in the initial analysis (which is the cutoff value used for that particular marker). In both S74 and S79, the overall number of sequencing reads in the initial analysis was significantly lower than that in the re-analysis (initial reads/ Re-analysis reads = 509/36487 and 253/37874 for both amplicons respectively), this might be the reason behind this alteration in deletion frequency proportion between the 2 analyses in these 2 particular samples. Interestingly, the predicted phenotype from the re-

analysis assay is consistent with the reported phenotype for both samples (S74 and S79), and thus, resolved the ambiguity in both samples.

5.3. Discussion

5.3.1. Validation of the MSI scoring system

In chapter 4, a panel of 17 mononucleotide markers was assessed using a large group of samples and different threshold sets were tested, and finally a weighted MSI score was proposed. In this chapter, an independent cohort was used to assess the weighted MSI score or T7, where an additional score was added to those markers that have instability with evidence of allelic bias. According to that score, samples with an overall score of ≥3 are called MSI-H and those with an overall score of less than 3 are called MSS. With that MSI score, 36 samples were called as MSI-H and 33 are MSS in the Edinburgh cohort with an overall sensitivity and specificity of 100%. This strongly consolidates the initial results and paves the way to use this MSI score as a main classifying system for the short mononucleotide panel.

The different threshold sets proposed and tested in Chapter 4 (T1-T6), were assessed in this Chapter as well. In the 3 earlier threshold sets (T1-T3), where relatively low cutoff values were used, a relatively high number of samples were called as MSI-H especially in the DF subgroup. In the latter 3 threshold sets (i.e. T4-T6), the number of samples that were called as MSI-H was considerably lower in the DF subgroup. Both sensitivity and specificity were higher towards the higher threshold sets, a finding that is consistent with that from the previous chapter. This indicates that most variant reads in longer homopolymers in the MSS group of samples were discarded by elevating cutoff values in T4-T6.

5.3.2. Fulfilment of the recommended requirements for validation of targeted NGS assay

5.3.2.1. Clinical samples

To fulfil the requirements of validation, all the samples tested here (and even the previous cohorts) were FFPE tissue samples (which are the same kind of samples that would be tested in practice when the test is implemented in diagnostic lab). However, sample storage, transportation and processing should also be assessed. As most of the samples (both Spanish and Edinburgh samples) were

referred from different places, the assessment of these factors was beyond our capability.

5.3.2.2. Sensitivity and specificity

It was possible for our test to achieve a high sensitivity, and with the weighted MSI scoring system, both sensitivity and specificity approached 100% in the tested cohort. This provides subjective evidence that this test performs perfectly in our laboratory.

5.3.2.3. Assessment of the test using multiple cohorts

The 17 marker panel was assessed across 3 independent cohorts, 2 of them (Spanish and Edinburgh cohorts) were assessed blindly.

5.3.2.4. Optimising library preparation and quality metrics

In the previous experiment done in Chapter 4 (Spanish cohort), a 4 pM library concentration was used as this is the recommended concentration in the 16S metagenomics protocol. However, that concentration resulted in a relatively low cluster density and consequently, a relatively low coverage depth (~2900 per/amplicon) as explained in Chapter 4. I, therefore, thought to increase the concentration to 10 pM in the MiSeq run of the Edinburgh cohort. This resulted in a much higher cluster density (1450 k/mm^2) and higher coverage (~5500 per/ amplicon). However, the improvement in the cluster density, and hence in the coverage, was associated with a drop in the Q30 score to 55.5% compared to 69.1% in the Spanish cohort MiSeq run. An interesting contributory factor to the overall performance of the run is the overall number of amplicons being analysed. These values could be used as a basis to conclude the final recommended workflow, and it would be worthy to try 8 pM library concentration aiming to optimal Q30 score and average depth.

5.3.3. Assessment of the run performance

Although the used primers were targeting small amplicons (~150bp), not all samples investigated in this part of the study were successfully amplified (70 amplified vs 30 failed). The failure of some samples to be amplified is possibly due to

severe FFPE-induced fragmentation. However, other factors like storage of samples, transportation of samples and the existence of PCR inhibitors could be possible reasons. Generally, 70% of samples included in this cohort were successfully amplified, of them, 99% of samples have amplified 15 markers or more as explained in Table 5-1.

5.3.3.1. Inter-cohort assessment of deletion curves

Comparison of deletion profiles of all the 17 markers across the 3 independent cohorts was important to compare the behaviour of individual markers, and the overall panel as well, in different cohorts. Only four out of the 17 tested markers (DEPDC2-G/C, LR24-9, LR48-11 and IM49-12) showed a clear consistency in the specificity across the 3 different cohorts (all of them showed 100% specificity across all the 3 tested cohorts). The marker GM14-11 showed the lowest specificity in the Spanish cohort (89% specificity at 30% deletion frequency) while it showed 100% specificity in both Newcastle and Edinburgh cohorts. This observable difference likely indicates that there is a cohort- specific reason. Assessment of the performance of marker GM14-11by retesting samples from different cohorts, has allowed a direct comparison of the deletion profile. In that assessment, 8 samples from the Spanish cohort and 16 from the Newcastle cohort were re-amplified and sequenced. The results showed a clear consistency of the deletion profile with the exception of 2 samples. However, both of these 2 samples were sequenced to significantly lower coverage in the initial investigation compared to the latest re-analysis. This might increase the possibility that the variability in the marker GM14-11 in the Spanish samples is likely to be restricted to that cohort. A possible explanation, is the difference in the ethnic background of that cohort (Spanish cohort) compared to both other UK cohorts (Newcastle and Edinburgh cohorts), as this might affect the MAF of the SNP in the primer annealing site (rs539119173). Furthermore, the number of samples in each cohort could be an additional reason for that inter-cohort variability. The number of samples in the Spanish cohort is more than the combined number of samples for both Newcastle and Edinburgh cohorts.

5.3.3.2. Assessment of allelic bias as an additional parameter for calling instability

During the assessment of the performance of the 6 threshold sets, it was evident that both sensitivity and specificity of the AB subgroup were higher than the corresponding DF subgroup for the same threshold set. The weighted MSI scoring system (T7) was constructed by collating both features of DF and AB together. The application of the MSI scoring system resulted in optimal sensitivity and specificity as explained above. This strongly suggests that using both parameters rather than deletion frequency alone would enforce the pickup rate of the proposed MSI panel. Interestingly, the number of samples that was called as MSI-H in the AB subgroup was the same across almost all threshold sets. This indicates that markers, which show evidence of allelic bias in this cohort, have had a high deletion frequency relative to their specified cutoff values.

However, the detection of allelic bias is limited to heterozygous amplicons only. In this cohort, it was not possible to assess the allelic bias in some samples, and out of the 36 samples that have a score of ≥3 (i.e. called as MSI-H), 3 samples were lacking evidence of allelic bias. These samples were E31 (MSI score=6), E46 (MSI score= 8) and E91 (MSI score= 11). In this particular cohort, this provides substantial evidence that the adoption of the MSI scoring system can eliminate the possibility of miscalling a sample that lacks an evidence of allelic bias.

5.4. Conclusions

The weighted MSI scoring system performed very well in discrimination between MSI-H and MSS samples in a cohort composed of 69 CRC samples. This system was able to raise both sensitivity and specificity of the test to 100% in this cohort. The role of allelic bias has been assessed and AB subgroup performed better in terms of sensitivity and specificity than their mate DF subgroups. This suggests that the existence of allelic bias adds more confidence to phenotype calling. It was possible to assess the quality metrics of the MiSeq run in this cohort and a better quality score and coverage were achieved compared to the previous run. The extensive validation done in this chapter provides an evidence of employability of the proposed 17 marker panel and the weighted MSI scoring system in MSI detection.

Chapter 6. Analysis of clonal characteristics of MSI-H CRC using short mononucleotide markers

6.1. Introduction and aim

6.1.1. Introduction

Cancer is a heterogeneous disease that results from uncontrolled cell division. Although it is heterogeneous, all cancers share common features collectively known as hallmarks of cancer (Hanahan and Weinberg, 2011). During evolution of cancer, cells acquire many genetic and epigenetic changes that give them the ability to divide out of control. The cancer- associated genetic changes can affect specific genomic sequences and/ or can be a stochastic process affecting any sequences. Different groups of genes can be considered as targets during carcinogenesis (which is the process of cancer development). Genetic mutations that affect a specific gene or group of genes and impose a positive selective advantage of the affected cells are called driver mutations. On the other hand, mutations that could randomly affect any gene or intergenic sequences and have no growth advantage are considered as passenger mutations (Stratton et al., 2009). The spatial distribution of the cancer associated mutations is not homogenous, thus it might happen in specific cells (and hence in subsequent clones) and not in others within the same tumour. The existence of specific mutations in specific cells could possibly add more features or make the bearing cells lack specific features. Because of the heterogeneous distribution of the genetic mutations, tumours are said to be genetically heterogeneous. From the molecular perspective, tumour heterogeneity (or genetic heterogeneity) can be classified into 4 subgroups:

1) **Intratumoral heterogeneity:** This refers to the variable genetic changes in different cells within the same tumour mass. Despite this notable variation, all cancer cells still share the most common somatic mutations (these mutations are called “trunk” mutations) while the majority of differences are “branch” or even “leaves or private” mutations that confer the intratumor heterogeneity. This variability provides the bases for the consequent intermetastatic variability (Vogelstein et al., 2013).
2) **Intermetastatic heterogeneity:** This kind of heterogeneity refers to the genetic variability in different metastases of the same tumour within the same patient. This kind of heterogeneity originates from the preceding intratumour

heterogeneity. The existence of this kind of heterogeneity gives rise to variable response to chemotherapy and, thus, eradication of a single clone will be unlikely to improve the long-term survival (Yachida et al., 2010). This kind of heterogeneity depends largely on the high number of passenger mutations, however all metastasis still share “positively selected” mutations from the main tumour mass.

3) **Intrametastatic heterogeneity:** When metastases grow from different clones and due to the acquisition of further mutations during development, cancer cells become more heterogeneous giving rise to possible resistance to anticancer therapy. Cancer recurrence after surgery or therapy might partly be explained by this phenomenon. It has been shown that at the time of diagnosis, about thousands of cells from each metastatic lesion are resistant to any anticancer drug. Thus, by using a single–agent regimen, the recurrence of cancer is just a matter of time. To overcome this issue, multidrug regimens are employed in the clinical practice (Komarova and Wodarz, 2005).
4) **Interpatient heterogeneity:** This kind of heterogeneity could be the underlying reason for the medical observation stating that; there are no 2 cancer patients have identical clinical course. This might be due to the variation in somatic mutations among patients. Different somatic mutations, even within the same gene, might have different consequences (Vogelstein et al., 2013)

The existence of different genetic mutations in different cells and clones resulted in intratumour variation in tumour characteristics and their response to therapy (Linnekamp et al., 2015, Hardiman et al., 2016). Moreover, intratumour heterogeneity (ITH) was suggested to have a role in prognosis and management of colorectal adenocarcinoma (Baisse et al., 2001). Thus, assessing ITH has clinical benefits. In addition, assessment of ITH helps to illuminate the evolution history of the different clones within the same tumour (Naxerova et al., 2014).

X-chromosome inactivation (also called Lyonisation) refers to the process of inactivation of a single X chromosome during embryonic life in the female fetus, leaving a single active X chromosome (Lyon, 1961). This phenomenon has been used to assess the clonal characteristics in cancer. Tumour cells that do not carry the same X chromosome inactivated cannot be clonal. However, this approach does not

give more information about the other genetic changes that are existed in tumour clones (Wang et al., 2009). Alternatively, a genome and exome wide approaches using multiple specimens from the same tumour would be an ideal way to assess the ITH. Such an approach has many limitations, as currently, only the large genomic centers are capable of adopting this approach. Furthermore, an exome-wide approach would limit the search to the coding sequences and thus focusing mainly on the driver mutations (Naxerova et al., 2014).

The proliferation of cancer cells is associated with increased DNA replication. DNA replication is usually accompanied by the generation of errors (up to 10,000 errors per cell per day) due to the inefficiency of the enzymatic machinery (Loeb, 2011). When mismatch repair (*MMR)* genes mutate, the mutation rate of the vulnerable sites (i.e. microsatellites) increases from 100-1000 folds (Bhattacharya et al., 1994, Shibata et al., 1994). These errors will accumulate with subsequent cell proliferation. Ideally, the time from the recent common ancestor cell can be documented by examining the microsatellite events in target tumour samples (Shibata et al., 1996). Both deletions and insertions of microsatellites could result in the generation of new alleles with different lengths that can be used to discriminate between different cells (and hence, the subsequent different clones) within the same tumour. It has been suggested that recently developed clones are likely to have similar alleles in adjacent cells, while the older clones are likely to show higher allelic diversity owing to the accumulations of more microsatellite mutational events (Shibata et al., 1996). Moreover, microsatellite mutations common to all specimens from a particular tumour were suggested to represent an early event during tumourigenesis.(Nagel et al., 1995).

As microsatellites mutate after *MMR* loss and as they represent a neutral markers (passenger mutations), it is possible to use microsatellites on a genome wide approach to investigate the clonal characteristics of MSI-H tumours. It has been suggested that one of the initial microsatellite changes which happen following the loss of *MMR* genes is the mononucleotide alteration (Ionov et al., 1993, Blake et al., 2001). Furthermore, the mutation frequency of polyguanine mononucleotides in mammals was estimated to be as high as 10^{-4} per cell per generation in both *MMR* deficient and proficient cells. Mutation rates of polyG repeats were found to be higher

than poly A repeats in both *MMR* proficient and deficient tumours (10-25 folds and 7-15 folds respectively) (Boyer et al., 2002).

Microsatellite (and minisatellite) markers were used to assess the intratumour heterogeneity using 20 different gastrointestinal tumours (includes 13 CRC tumours) (Nagel et al., 1995). It was concluded from that study that microsatellite mutations can be used to assess the clonal history of tumours irrespective to patient's sex and type of tumour being tested. Tsao et al (2000), has quantatively analysed microsatellite changes (dinucleotide markers) in 13 *MMR* deficient CRCs and found that each tumour has a unique history since *MMR* loss. In 2004, dinucleotide microsatellite markers were used to assess the lineage relationship in different stages of cutaneous T-cell lymphoma. In that study, the microsatellite allelic profiles from multiple confined tumour lesions (Mycosis Fungoides) were compared to those from a benign condition (Lichen Planus). It was possible, from that study, to construct a lineage tree between different tumour lesions including those from early stages of the disease (Rübben et al., 2004).

Poly G microsatellite markers were used to assess the cell fate in cultured mouse cells. Taking the advantage of change in length of polyguanine tracts after mitosis, it was possible to trace the phylogenetic tree of the cultured mouse cells (Salipante and Horwitz, 2006). Furthermore, alterations in polyG repeat length were found to be more abundant when there is associated neoplasia as opposed to ulcerative colitis with no neoplasia (Salk et al., 2009). In a study conducted at 2014, 20 polyG markers were used to assess the intratumour heterogeneity between spatially different samples of different cancers including colorectal cancers. It was possible with that approach to construct the phylogenetic tree for the tumours involved in that study. This provided evidence of utility of mononucleotide markers to assess ITH and lineage relationship (Naxerova et al., 2014). However, these studies used the conventional fragment analysis approach to assess the allele length alterations. Quite recently, a panel of 20 short (8-14bp in length) mononucleotide markers was used to assess the clonal characteristics of 3 MSI-H CRCs using a next generation sequencing approach (Redford, 2016). In that panel, 8 markers were included in the 17 marker panel that was developed and validated in the previous Chapters. It was possible, according to that study, to use the short mononucleotide markers to infer the clonal characteristics in spatially different samples derived from

MSI-H CRCs. However, that assay assessed a limited number (3 tumours with a total of 27 specimens) of fresh tissue samples rather than FFPE samples.

In this chapter, the panel of short mononucleotide markers that was developed and validated in previous chapters in addition to extra 6 markers, will be used to investigate whether short repeats can be used to analyse clonality of CRCs. For this purpose, 2 groups of MSI-H CRC samples were prepared for analysis; FFPE and fresh frozen tumour samples. For both groups, specimens from different positions within the same tumour were collected to investigate the clonal variability between them in terms of microsatellite profiles.

6.1.2. Aim

In this chapter, the aim is to:

- Assess the utility of short mononucleotide (7-12bp) markers to study the clonal characteristics of different samples from the same primary CRC or between primary tumour and its secondary metastasis.

6.2. Results

6.2.1. Collection of MSI-H CRC fresh tissue samples

Twelve groups of fresh CRC tissue samples were obtained from the Department of Cellular Pathology (Royal Victoria Infirmary, Newcastle Upon Tyne Hospitals NHS Foundation Trust, UK). Each group was composed of 8 fresh tissue specimens retrieved from different locations within the same CRC tumour as explained in Chapter 2 section 2.2.2 and shown in Figure 6-1.

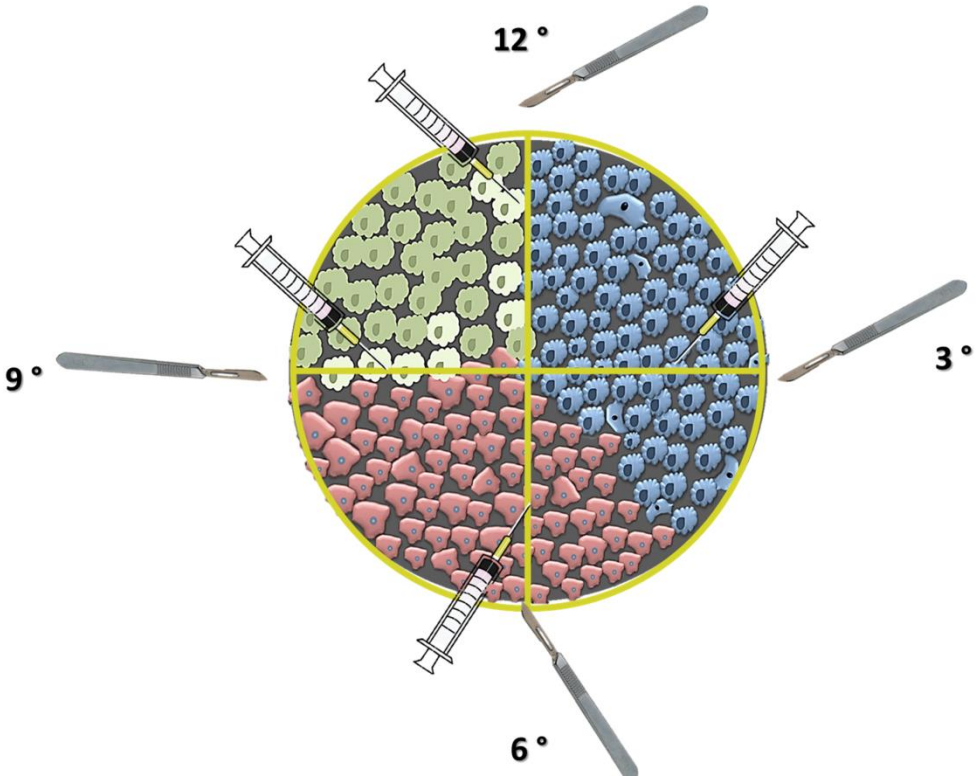


Figure 6-1: The orientation of specimens retrieved from fresh CRC tumours. All specimens were retrieved in a clockwise orientation. The nearest tumour edge to the antimesentric border was considered as the 12 'clock. Specimens retrieved by needle were denoted as "N", while specimens retrieved by scalpel were denoted as "S".

All fresh tissue samples were tested for MSI status using the Promega MSI test (MSI Analysis System, Version1.2: Promega, Madison, WI, USA) as explained in Chapter 2 section 2.8. A single tumour, out of the 12 tested CRC fresh tissue samples, was found to be MSI-H as shown in Figure 6-2.

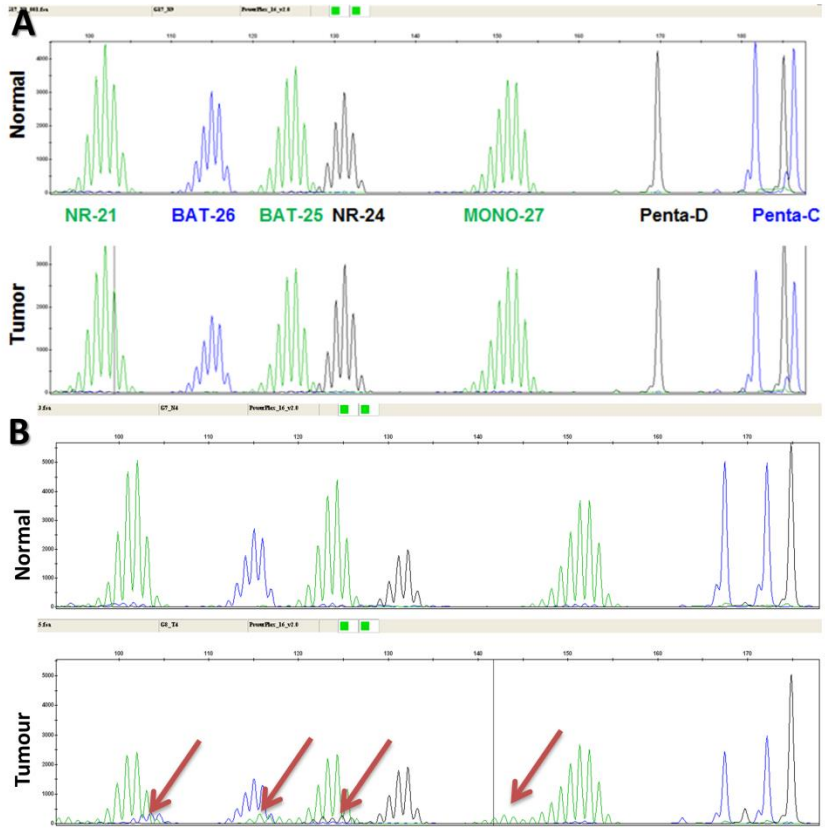


Figure 6-2: Electropherogram showing the fragment analysis of 2 fresh CRC tumours. (**A**) MSS tumour and (**B**) MSI-H tumour (PR32516/14). For each tumour, matched normal tissue sample was tested in parallel. **X-axis** refers to mononucleotide markers that are included in the panel and **Y-axis** refers to relative fluorescence intensity. The panel of markers was composed of 5 mononucleotide markers and 2 pentanucleotide markers. Names of markers are labelled next to each individual group of peaks. By careful visual comparison of the allelic profile, very similar profiles are observed in tumour **A** with its corresponding normal sample, while, 3 makers show different profiles in the tumour **B** (arrows).

Three additional MSI-H samples were obtained from Dr Lisa Redford (Newcastle University, UK), so the overall number of fresh MSI-H tumours included in the assay was 4, PR32516/14, PR10654/14, PR51896/13 and PR17848/14. From each tumour, 9 spatially different specimens were retrieved by a pathologist from different orientations as explained in Figure 6-1.

6.2.2. Collection of FFPE MSI-H CRC samples

I searched the NHS database looking for all MSI-H colorectal cancers, which have more than one FFPE specimen and diagnosed in the time interval from 2000-2015. The search was limited to those MSI-H CRC tumours, which have been referred from the Royal Victoria Infirmary (RVI) and have more than one tissue source (e.g. Tumour and lymph node or 2 or more tumours from different positions). Eighteen MSI-H CRC tumours were found to fulfil these criteria. Out of those 18, 7

tumours were collected and examined as explained in Chapter 2 section 2.2.2 and their details are shown in Table 6-1

							IHC for MMR				
No.	**Path. Ref. No.**	**Age**	**Block**	**Position**	**Diagnosis**	**Stage**	**MLH1**	**MSH2**	**MSH6**	**PMS2**	**MMR mutation**
1	**PR53139/13**	73	1J	Caecal	Moderately differentiated Adenocarcinoma	T2N0M0					NA
			1H	Caecal							
			1M	Ascending colon							
			1N	Ascending colon							
2	**PR7146/13**	53	1P	Lymph node	Medullary AC of Sigmoid	T4N1M0					NMD
			1W	Tumour + serosa							
3	**PR34630/03**	68	3J	Tumour	Moderately differentiated AC of Sigmoid	T4N1M0	NA	NA	NA	NA	NA
			3P	Apical lymph node							
4	**PR049276/12**	68	2I	Tumor + serosal invasion	Medullary AC of right colon	T3N0M0					NMD
			2H	Tumor with Gerota's fascia							
5	**PR53996/14**	67	1J	Tumour	Colonic Adenocarcinoma	T1N0M0	NA	NA	NA	NA	NA
			1K	Tumour							
6	**PR45703/14**	36	1H	Tumour	Caecal Adenocarcinoma	T1N0M0	NA	NA	NA	NA	*MSH2* mutation
			1J	Tumour							
7	**PR32079/14**	71	1P	Tumour close to serosa	Caecal poorly differentiated Adenocarcinoma	T3N0M0					NA
			1Q	Tumour+ LN close to CRM							
			1R	Tumour + LN close to CRM							
			1S	Tumour close to CRM							
			1T	Tumour							
			1O	Tumour + mesocolic fat							

Table 6-1: The FFPE CRC tumours that were used to assess ITH. For each tumour, at least 2 samples from different positions were requested. AC= adenocarcinoma, **TNM**= is the international staging system (Edge and Compton, 2010). **Age**= is the age at time of pathological report. **IHC for MMR**= results

of Immunohistochemistry for Mismatch repair proteins. Dark red boxes indicate absent protein, faint green boxes indicate the presence of protein. **NMD**= no mutation detected. **NA** indicates no available information.

6.2.3. PCR amplification and MiSeq sequencing of the collected samples to assess the clonal composition

All the 36 fresh tissue specimens (9 specimens from each tumour mass, 4 tumours in total) and all the 20 FFPE specimens (from 7 FFPE tumours) were amplified using a batch of 23 markers (shown in Table 6-2).

The primers used in PCR amplification were composed of those that were used in the previous chapters (the 17 markers panel) and additional 6 markers. Of those 6 additional markers, 3 markers were retrieved from Selective Targets in Human MSI-H Tumorigenesis Database website (SelTarbase) (SelTarBase, http://www.seltarbase.org). The search in the SelTarbase was limited to those markers with a length of 7-12bp that have a high percentage (more than 80%) of instability in CRCs. The markers that were retrieved from SelTarbase are ASTE1 (Woerner et al., 2003), AVIL and IRS2 (Woerner et al., 2010). Another 3 markers were chosen from a parallel assay conducted by Dr Lisa Redford (Redford, 2016) as they were variable in the majority of the tested specimens. These markers were GM29-10, LR17-11 and LR43-12.

	Marker	Repeat Size	SNP 1	SNP2	Repeat Position
1	LR49-7	7	rs12903384		Chr15:93619048
2	IM66-C	7	rs4794136	rs141474571	Chr17:48433967
3	DEPDC2-G	8	rs4610727		Chr8:68926683
4	LR20-8	8	rs146973215	rs217474	Chr1:64029634
5	GM9-8	8	rs79878287		Chr20:6836977
6	GM11-9	9	rs347435		Chr5:166099891
7	LR24-9	9	rs192329538		Chr1:153779429
8	IM16-9	9	rs73367791		Chr18:1108767
9	GM17-9	9	rs666398		Chr11:95551111
10	AP003532-2-9	9	rs138081624		Chr11:127625067
11	AVIL-10	10	rs141859389		Chr12:58202497
12	GM29-10	10	rs2687195		Chr3:70905560
13	GM7-11	11	rs2283006		Chr7:93085748
14	GM14-11	11	rs6804861		Chr3:177328818
15	ASTE1-11	11	rs753405495		Chr3: 130733047
16	LR48-11	11	rs368641323	rs11105832	Chr12:77988097
17	LR11-11	11	rs13011054		Chr2:217217871
18	LR17-11	11	rs1009977	rs1009978	Chr14: 55603031
19	LR36-12	12	rs17550217		Chr4:98999723
20	LR44-12	12	rs7905384	rs7905388	Chr10:99898286
21	IM49-12	12	rs7642389		Chr3:56682066
22	IRS2-12	12	rs57032199		Chr13: 110407562
23	LR43-12	12	rs6881835	rs10051666	Chr5:86199061

Table 6-2: The list of primers that were used in the clonal assessment of CRC samples alongside with the included SNPs in the amplicons.All tumour specimens were then amplified and sequenced and deletion frequency calculated as explained in Chapter 2. In fresh CRC tumours, for each specimen, deletion frequencies for all markers were compared to those in the associated normal specimen. This baseline calculation was done for all amplicons which showed an overall sequencing reads of ≥ 100 paired end reads. The amplicons with less than 100 sequencing reads were excluded from the baseline calculations. Out of the tested 36 fresh CRC specimens, 94% (34/36 samples) were amplified and sequenced to adequate depth (i.e. >100 sequencing reads) in ≥ 20 markers. A single specimen was sequenced for only 17 markers and

another specimen sequenced for <20 markers. On the other hand, 90 % of FFPE specimens (18/20) were sequenced to adequate depth in ≥ 20 markers.

The average depth to which all tumours in both groups were sequenced was approximately the same (4971 per/ amplicon for fresh tissue samples and 5119 per/ amplicon for FFPE samples). When a specific marker failed to be amplified in a particular specimen from a particular tumour, the marker was excluded from further analysis for all specimens of that tumour. To establish instability of each marker, cutoff values that used in threshold set 7 (developed and validated in Chapter 4 and Chapter 5, respectively) were used as shown in Table 6-3. The only exceptions are the 10bp markers, where there is no validated cutoff value for such a repeat length (as there was no 10bp marker in the 17 marker panel). For the 10bp markers, the cutoff value was set to a deletion frequency of 0.14 as this value yielded the least false positive rate (= 0 %) in a previous analysis (Redford, 2016).

Marker group	7bp	8bp	9bp	PolyG/C	10bp	11bp	12bp
Cutoff value	0.05	0.05	0.08	0.1	0.14	0.30	0.30

Table 6-3: Cutoff values for calling instability in the clonality assay. The same cutoff values used in threshold set 7 were used with the addition of new cutoff value for the 10bp markers.

For purposes of phylogenetic tree construction, Mesquite software (http://mesquiteproject.wikispaces.com/) was used as explained in Chapter 2 section 2.10.8.

6.2.4. Clonal composition of the FFPE tumours

6.2.4.1. Clonal composition of the tumour PR53139/13

The tumour PR53139/13 was examined by a certified pathologist (Dr Helen Turner, Department of Cellular Pathology, RVI, Newcastle Hospitals NHS Foundation Trust, UK) and reported as moderately differentiated adenocarcinoma. Four specimens from 4 different blocks were retrieved from that tumour (1H, 1J, 1M and 1N). The specimens 1H and 1J were originated from the Caecum while both 1M and 1N were from the ascending colon. The histopathological diagnosis for all specimens was the same. The malignant cell population was reported to be <5% in both 1J and 1H specimens while it was estimated to be 30% in specimen 1M and 40% in specimen 1N. The four specimens were successfully amplified and sequenced

across 22 markers. All the tested specimens have shown instability, but in different number of markers for each specimen as shown in Table 6-4.

Sample	No. of sequenced amplicons	Coverage (per/amplicon)	Tumour cell (%)	No. of unstable markers
1J	22	2442	<5%	5
1H	22	3106	<5%	9
1M	22	4437	30%	15
1N	22	7587	40%	16

Table 6-4: Number of unstable markers and sequencing coverage for the 4 specimens of the tumour PR53139/13. The coverage depth expressed in paired end read (per) / amplicon.

The least number of unstable markers was observed in the specimen 1J, where only 5 markers (out of the 22 successfully sequenced markers) showed a deletion frequency above threshold values. The marker IRS2-12 showed the highest deletion frequency among the unstable markers in that specimen (= 70%), most of deletions for that marker (67%) were 2bp deletion while 3bp deletion was observed in 3% and no 1bp deletion was recorded as shown in Appendix Figure 8-4. The existence of the low number of unstable markers in the specimen 1J can either be due to low tumour cell contents (reported as <5%) or it could represent the nearest specimen to the original trunk from which other specimens (branches) were originated.

In the specimen 1H, 4 additional markers (GM11-9, AP0035322-9, GM14-11 and LR44-12) showed deletion above the threshold values. The markers GM11-9 and AP0035322-9 showed only 1bp deletions (39% and 10%, respectively), while marker GM14-11 showed 2bp deletion (44%) in addition to 1bp deletion (3%). The deletion frequency in all the remaining 5 unstable markers was higher than the frequencies observed in the 1J specimen, except for the marker IRS2-12 where the 2bp deletion is lower than that observed in the 1J specimen (48% compared to 76% in the 1J sample).

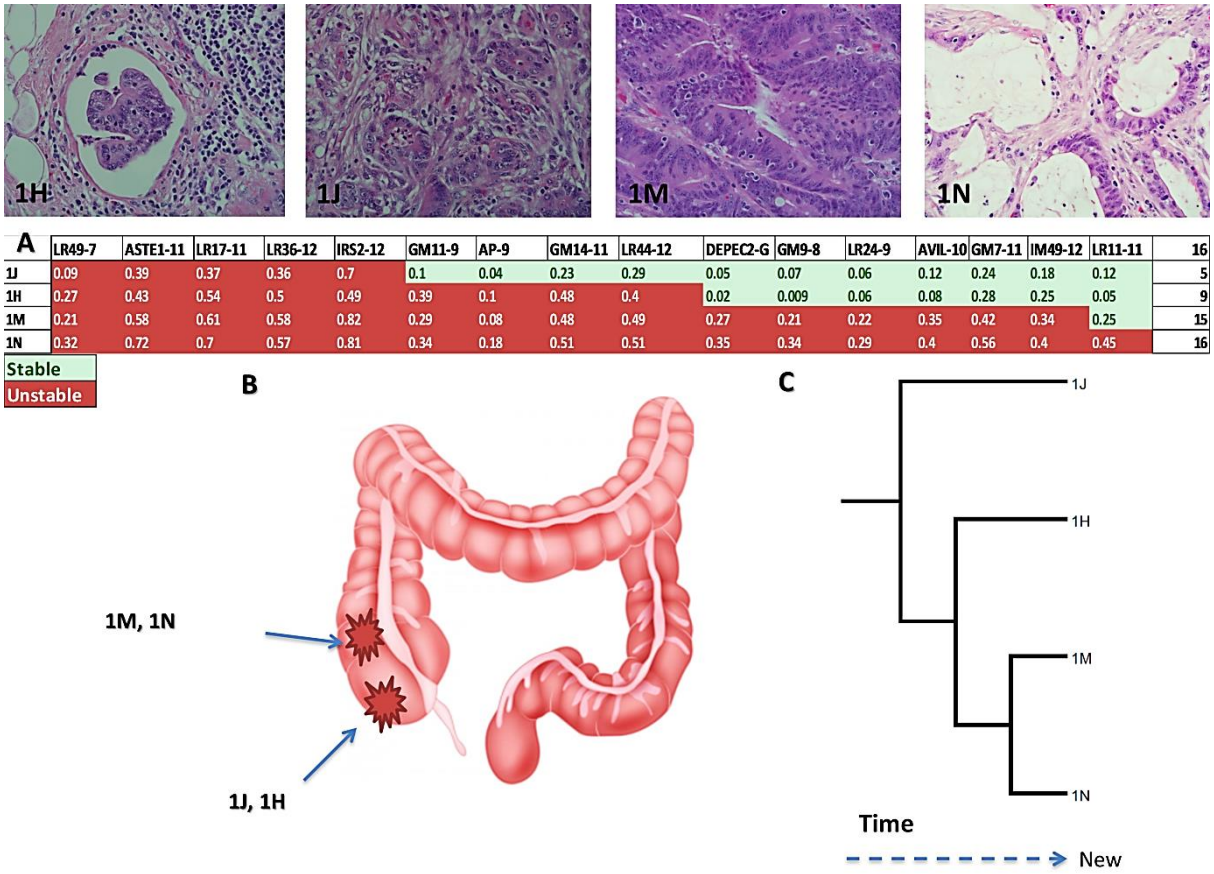
The specimen 1M showed deletion frequency above the length specific threshold in 15 markers. Compared to specimen 1H, the markers DEPDC2-G, GM9-8, LR24-9, AVIL-10 and IM49-12, showed deletion frequency above the threshold while they were stable in both 1H and 1J as shown in Figure 6-3. In the specimen

1N, an additional marker (LR11-11) showed a deletion frequency above threshold, giving rise to a unique deletion in that particular sample.

According to the number of unstable markers in each of the 4 tested specimens, a phylogenetic tree was constructed as shown in Figure 6-3. One possible conclusion is that the specimen 1J contains cells that belong to the nearest clone to the original ancestral tumour clone (as it has the lowest number of unstable markers). The specimen 1H stands out as a branch as it has more unstable markers than 1J but lower than both 1M and 1N. Furthermore, it shares 4 mutations with other specimens. This indicates that the specimen 1H contains cells of a descent clone from 1J but developed before the emergence of the clone whose cells are existed in specimens 1M and 1N. Both 1M and 1N specimens are likely to contain cells of a descent clone from the clone existed in specimen 1H as shown in Figure 6-3.

Although the recorded tumour cell content in both specimens 1J and 1H was very low (<5% for both), it was possible to detect at least 5 unstable markers in both specimens with relatively high deletion frequency (up to 70%). One possible explanation is that the tumour cell population was underestimated. Another possibility is that the coverage depth to which these specimens were sequenced (2000- 3000 per/amplicon) was high enough to detect such low prevalence mutations.

Overall, this indicates that the origin of the tumour PR53139/13 is likely to be located in the Caecum than in the ascending colon. However, the reported low tumour cell population in both 1J and 1H specimens (both are <5%) might contribute to the relatively low number of unstable markers in both specimens compared to specimens 1M and 1N. This assumption is supported by the clear positive relation between the number of unstable markers and the estimated tumour cell population as shown in Figure 6-3.



	LR49-7	ASTE1-11	LR17-11	LR36-12	IRS2-12	GM11-9	AP-9	GM14-11	LR44-12	DEPEC2-G	GM9-8	LR24-9	AVIL-10	GM7-11	IM49-12	LR11-11	16
1J	0.09	0.39	0.37	0.36	0.7	0.1	0.04	0.23	0.29	0.05	0.07	0.06	0.12	0.24	0.18	0.12	5
1H	0.27	0.43	0.54	0.5	0.49	0.39	0.1	0.48	0.4	0.02	0.009	0.06	0.08	0.28	0.25	0.05	9
1M	0.21	0.58	0.61	0.58	0.82	0.29	0.08	0.48	0.49	0.27	0.21	0.22	0.35	0.42	0.34	0.25	15
1N	0.32	0.72	0.7	0.57	0.81	0.34	0.18	0.51	0.51	0.35	0.34	0.29	0.4	0.56	0.4	0.45	16

Figure 6-3: The clonal characteristics of the tumour PR53139/13. Images of the histological sections were shown in the upper pane with their corresponding specimen identifier. All slides have the same diagnosis (moderately differentiated adenocarcinoma). **(A)** Deletion frequencies of 16 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 4 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Schematic representation of specimen locations in this tumour, where samples 1J and 1H were derived from the Caecum while specimens 1M and 1N were derived from the ascending colon. **(C)** The phylogenetic tree of the tested specimens. The newest clones are likely to be represented in specimens 1M and 1N. The horizontal timeline shows that the further the branching, the newer is the specimen.

6.2.4.2. Clonal composition of the tumour PR32079/14

The tumour PR32079/14 was reported as a poorly differentiated adenocarcinoma of the caecum with some medullary and focal neuroendocrine features. Six specimens were retrieved from different positions of the tumour (1O, 1P, 1Q, 1R, 1S and 1T). Of them, two specimens contained an associated lymph node (1Q and 1R). Of these 6 specimens, 2 specimens (1O and 1Q) were reviewed by the pathologist and the malignant cell population was estimated to be 40% for both specimens. Among the 23 tested markers, 3 specimens (1P, 1S and 1T) were successfully amplified and sequenced for 23 markers, 2 specimens (1Q and 1R) for

22 markers and a single specimen (1O) for 21 markers with average coverage of 4699 per/ amplicon as shown in Table 6-5.

Sample	No. of sequenced amplicons	Coverage (per/amplicon)	Tumour cell (%)	No. of unstable markers
1O	21	2873	40%	14
1P	23	7748		14
1Q	22	5801	40%	13
1R	22	3665		12
1S	23	4487		9
1T	23	3623		13

Table 6-5: Number of unstable markers in the 6 specimens of tumour PR32079/14. The tumour cell population was assessed for only 2 samples. The coverage depth expressed in pairs end read (per) / amplicon.

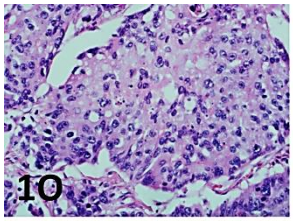
Based on the number of unstable markers for all specimens, the phylogenetic tree was constructed as shown in Figure 6-4. All specimens showed at least 9 unstable markers (i.e. Have a deletion frequency above the length-specific cutoff values) as shown in Table 6-5. The least number of unstable markers was observed in the specimen 1S, indicating that this specimen is likely to contain cells from the nearest clone to the ancestral tumour clone.

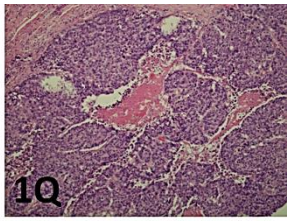
The specimen 1R have had an extra 3 unstable markers compared to the specimen 1S, while, specimen 1Q showed 13 unstable markers. In the specimen 1Q, the marker DEPDC2-G showed 38% deletion frequency, making it the only observed difference from specimen 1R. The specimen 1P exhibited 14 unstable markers and shares all the 13 unstable markers in the specimen 1Q. It has an additional unique instability in the marker IM66-C (13% deletion frequency).

The specimen 1T had 13 unstable markers, sharing all the 12 unstable markers that existed in the specimen 1R. This specimen had an additional unstable marker (45% deletion frequency in the marker LR43-12). The marker LR36-12 shows a new event represented by the emergence of a new 5bp deletion in the specimen 1T as shown in Appendix Figure 8-3.

The specimen 1O showed instability in 14 markers, sharing all the unstable markers in the specimen 1T. This specimen shows an additional instability in the marker DEPDC2-G with a 15% deletion frequency.

Overall, the above characteristics of individual specimens indicate that the specimen 1S contains cells that likely belong to the oldest clone (the nearest one to the common ancestor) among the tested specimens. Specimens 1R, 1T, 1Q, 1O and 1P contain cells that likely belong to downstream branches from the 1S specimen as shown in the phylogenetic tree in Figure 6-4. However, variability in the percentage of tumour cell population among specimens is another possibility for the variation in deletion profile.





A

	LR49-7	LR24-9	GM7-11	LR36-12	LR44-12	GM14-11	ASTE1-11	GM29-10	IRS2-12	AVIL-10	LR48-11	LR17-11	DEPEC2-G	LR43-12	IM66-C	15
1S	0.18	0.11	0.3	0.48	0.37	0.31	0.57	0.15	0.7	0.13	0.19	0.25	0.05	0.2	0.005	9
1R	0.13	0.18	0.5	0.55	0.5	0.42	0.67	0.22	0.93	0.23	0.4	0.46	0.08	0.15	0.001	12
1T	0.09	0.25	0.6	0.78	0.69	0.63	0.82	0.4	0.93	0.32	0.57	0.63	0.04	0.45	0.01	13
1Q	0.19	0.21	0.64	0.76	0.43	0.38	0.67	0.26	0.88	0.31	0.42	0.61	0.38	0.27	0.07	13
1O	0.39	0.52	0.78	0.96	0.76	0.9	0.91	0.54	0.98	0.43	0.77	0.81	0.15	0.87	0.02	14
1P	0.46	0.15	0.35	0.68	0.33	0.6	0.61	0.25	0.97	0.18	0.4	0.52	0.21	0.16	0.13	14

Stable
Unstable

B

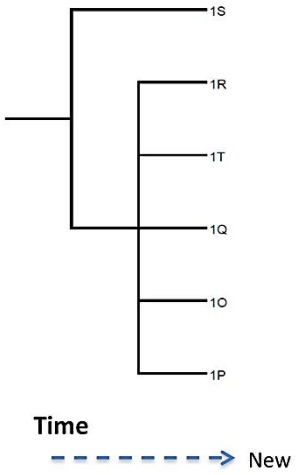


Figure 6-4: The clonal characteristics of the tumour PR32079/14. Upper images represent the histological sections of corresponding specimens (1O and 1Q). Both slides showed the same histological features (poorly differentiated adenocarcinoma of Caecum). **(A)** Deletion frequencies of 15 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** The phylogenetic tree based on the instability profile of the tested specimens. The horizontal timeline shows that the further the branching, the newer is the specimen.

6.2.4.3. Clonal composition of tumour PR7146/13

The tumour PR7146/13 was reported as a stage III medullary adenocarcinoma arose from the Sigmoid with evidence of lymph node involvement (T4N1Mx). Two specimens were retrieved from that tumour, 1P and 1W. The

specimen 1P was from an involved lymph node while the specimen 1W was from the tumour itself. Out of the 23 tested markers, specimen 1P was amplified for 22 markers with an average coverage of 7400 per/amplicon while the specimen 1W has amplified only 17 markers with an average coverage of 8400 per/amplicon. Both specimens (1P and 1W) showed instability of 9 markers as shown in Figure 6-5. This indicates that both specimens (the tumour and the lymph node) likely contain cells that belong to a single clone with a similar age.

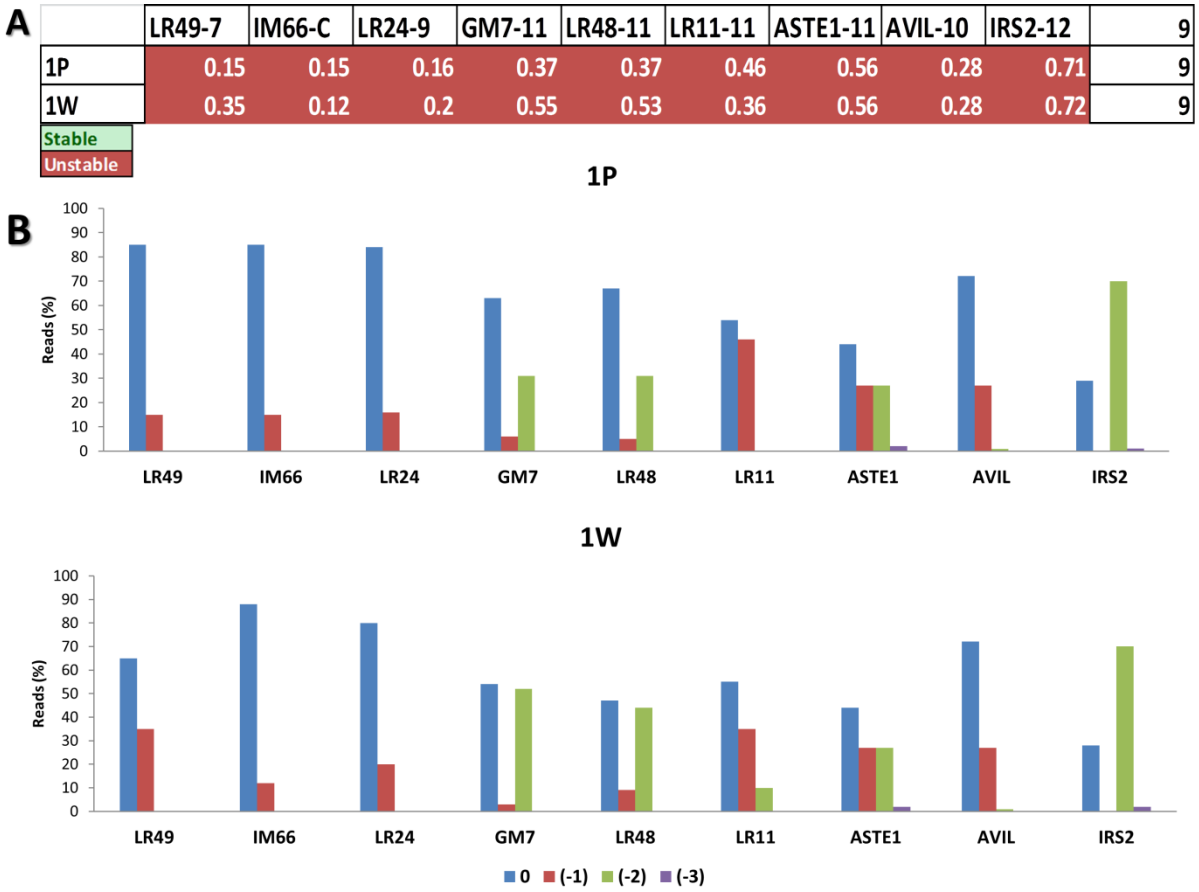


Figure 6-5: The clonal characteristics of the tumour PR7146/13.(A) Deletion frequencies of 9 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Bar charts represent the variant allele frequencies of the 9 markers in both specimens.

6.2.4.4. Clonal composition of the tumour PR34630/03

The tumour PR34630/03 was reported as stage III moderately differentiated adenocarcinoma of the Sigmoid with lymph node involvement (T4N1Mx). Two specimens were retrieved from that tumour, 3J which obtained from the tumour mass itself and 3P which obtained from an involved apical lymph node. Out of the 23 tested

markers, 21 markers were successfully amplified and typed in both specimens. Seven markers were unstable in at least one specimen of that tumour. In the specimen 3P (the lymph node), 7 markers showed a deletion frequency above the threshold while the specimen 3J (the tumour) showed instability in 11 markers as shown in Figure 6-6. Interestingly, almost all markers showed deletion frequencies in the specimen 3J higher than in the specimen 3P. This indicates that specimens 3P and 3J contain cells likely belong to clones that are distant from each other and the clone whose cells are existed in the specimen 3P represent the nearest between them to the ancestral clone.

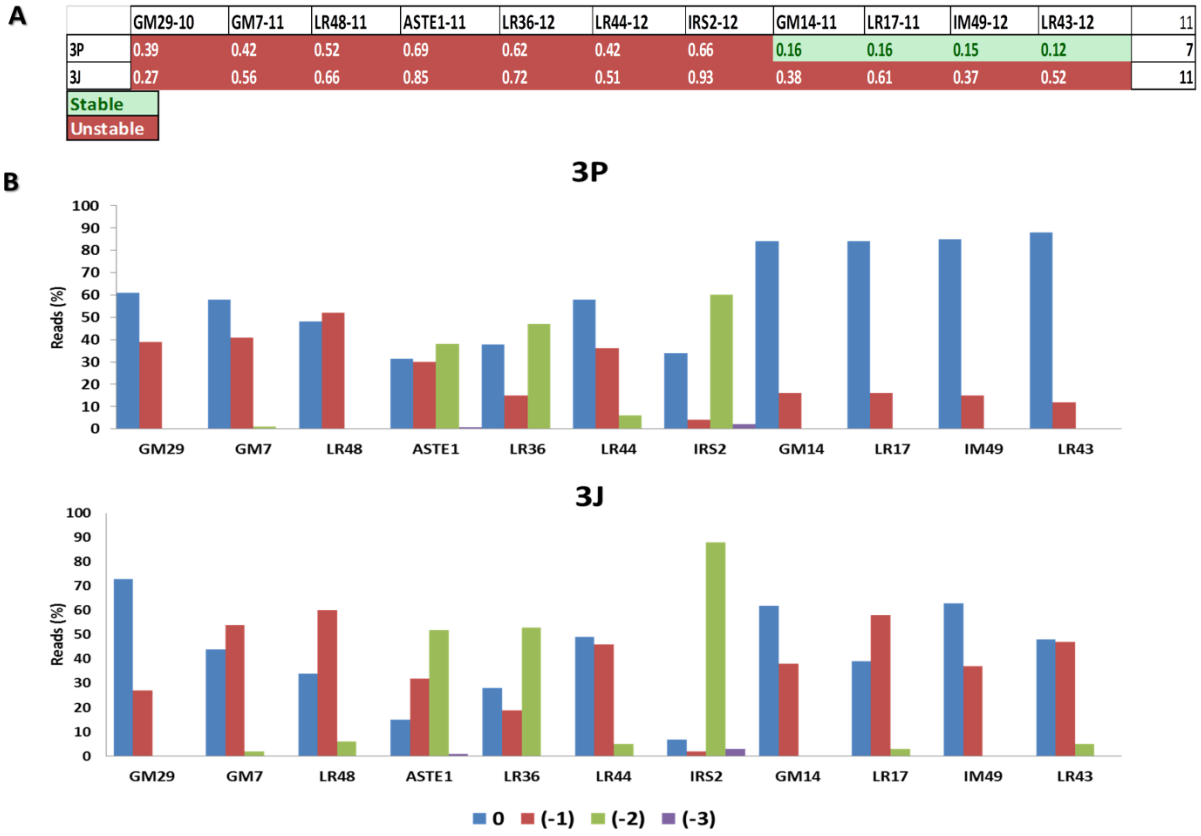


Figure 6-6: The clonal characteristics of the tumour PR34630/03.(A) Deletion frequencies of 11 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Variant allele frequencies of the 11 markers in both specimens.

6.2.4.5. Clonal composition of the tumour PR049276/12

The tumour PR049276/12 was reported as a stage II moderately differentiated adenocarcinoma of the caecum. Two specimens were retrieved from that tumour, 2I that represents tumour with serosa and 2H, which represents tumour

with Gerota's fascia (renal fascia). The malignant cell population was reported to be 40% in the specimen 2I and 60% in specimen 2H. The specimen 2I was successfully amplified and sequenced across 21 markers while specimen 2H was sequenced across 19 markers with approximately double sequencing depth compared to specimen 2I (6655 per/ amplicon for specimen 2H compared to 3524 per/amplicon for specimen 2I).

Both specimens have shown instability in 11 markers, of them, 10 markers were unstable in both specimens. The specimen 2I showed instability in the shared 10 markers and an additional single marker (the marker IM49-12). In the other specimen (2H), in addition to the 10 shared unstable markers, one additional marker was unstable (GM14-11) as shown in Figure 6-7. There was new 3bp deletion alleles observed in 3 markers (GM14-11, LR36-12 and IM49-12) in the specimen 2I. Thus, these specimens contain cells that likely belong to separate clones originated from a single trunk.

A

	IM66-C	DEPEC2-G	AVIL-10	GM7-11	LR48-11	ASTE1-11	LR17-11	GM14-11	LR36-12	LR44-12	IM49-12	IRS2-12	12
2I	0.13	0.22	0.2	0.64	0.35	0.53	0.51	0.28	0.85	0.49	0.36	0.85	11
2H	0.35	0.39	0.25	0.43	0.67	0.36	0.54	0.52	0.81	0.49	0.16	0.82	11

Stable
Unstable

B

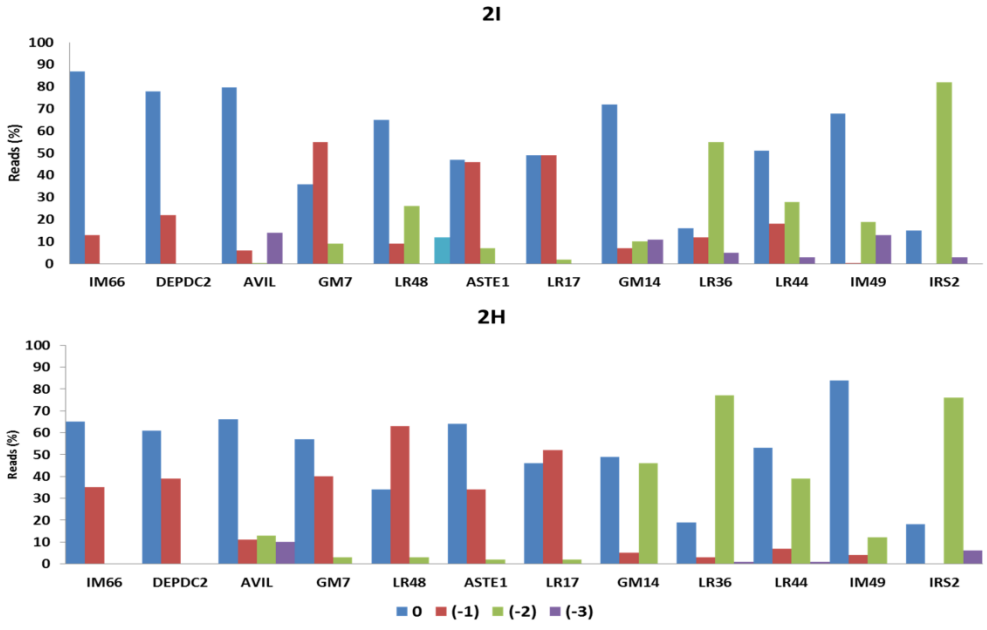


Figure 6-7: The clonal characteristics of the tumour PR049276/12. (A) Deletion frequencies of 12 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 2 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Bar charts represent variant allele frequencies of the 13 markers of both specimens.

6.2.4.6. Clonal composition of the tumour PR53996/14

The tumour PR53996/14 was reported as stage I adenocarcinoma of the colon, and 2 specimens (1J and 1K) were retrieved from the archived tumour blocks. The malignant cell population was reported to be 30% for both specimens. Out of the 23 tested markers, the number of the markers that were successfully amplified and sequenced is 22 for specimen 1J and 20 for specimen 1K. Both specimens (1J and 1K) showed a shared instability in 9 markers while each specimen showed instabilities in additional 2 different markers.

In the specimen 1J, in addition to the 9 shared markers, 2 unstable markers (LR24-9 and LR43-12) were observed. In the specimen 1K, on the other hand, 11 unstable markers were observed, including the 9 shared markers and additional 2 markers (LR49-7 and LR17-11). The specimen 1K showed new events represented by the existence of 3bp deletion allele in 2 markers (ASTE1-11 and AVIL-10) and 4bp deletion allele in the marker IRS2-12 (Figure 6-8).

As the additional (non- shared) unstable marker were not the same in both specimens, instabilities of these markers could indicate that both specimens contain cells that belong to 2 independent branches from a common trunk. However, the normal cell contamination is still a possible reason for the difference in instability between the tested 2 specimens.

A

	GM11-9	AVIL-10	GM29-10	GM7-11	ASTE1-11	LR36-12	LR44-12	IM49-12	IRS2-12	LR49-7	LR24-9	LR17-11	LR43-12	13
1J	0.18	0.32	0.2	0.67	0.66	0.34	0.39	0.57	0.66	0.015	0.15	0.19	0.34	11
1K	0.18	0.32	0.21	0.62	0.66	0.22	0.47	0.51	0.77	0.07	0.02	0.3	0.14	11

Stable
Unstable

B

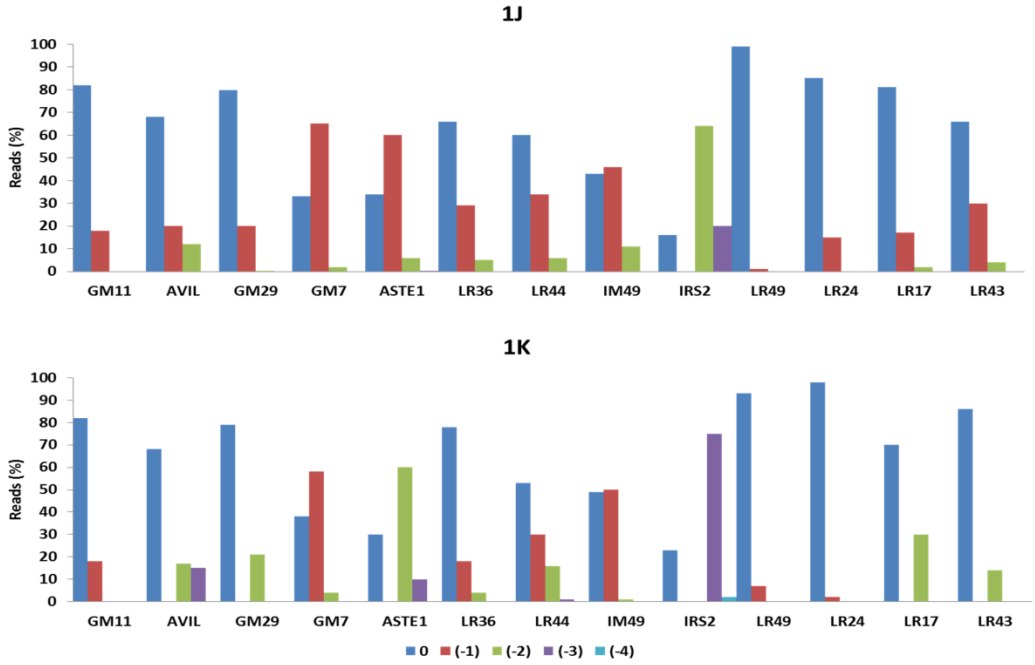


Figure 6-8: The clonal characteristics of the tumour PR53996/14. (A) Deletion frequencies of 13 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 2 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Variant allele frequencies of the 13 unstable markers of both specimens.

6.2.4.7. The clonal composition of the tumour PR45703/14

The tumour PR45703/14 was reported as a stage I adenocarcinoma of the caecum and 2 specimens from 2 different tissue blocks were retrieved, 1H and 1J. Of the 23 tested markers, the number of markers that were amplified and sequenced is 18 markers for specimen 1H and 20 markers for specimen 1J. Both specimens shared a deletion in 10 markers. The specimen 1H showed an additional unique mutation in the marker IM16-9. Furthermore, this specimen showed a higher 3bp deletion frequency in 2 markers (ASTE1-11 and IRS2-12) compared to those in the specimen 1J as shown in Figure 6-9. This mutational profile is likely to indicate that both specimens contain cells that belong to 2 clones bifurcated from the same trunk. Another possible scenario is that cells within both specimens belong to a single clone, but contamination with normal cell resulted in the observed difference in the IM16-9 instability between the two specimens. This conclusion is supported by the

marginal difference between deletion frequencies for the marker IM16-9 in both samples.

A

	AP-9	AVIL-10	GM7-11	LR48-11	ASTE1-11	LR17-11	IM49-12	LR36-12	LR43-12	IRS2-12	IM16-9	11
1J	0.26	0.32	0.49	0.69	0.77	0.7	0.56	0.72	0.31	0.67	0.06	10
1H	0.37	0.35	0.53	0.72	0.85	0.81	0.68	0.73	0.38	0.88	0.08	11

Stable
Unstable

B

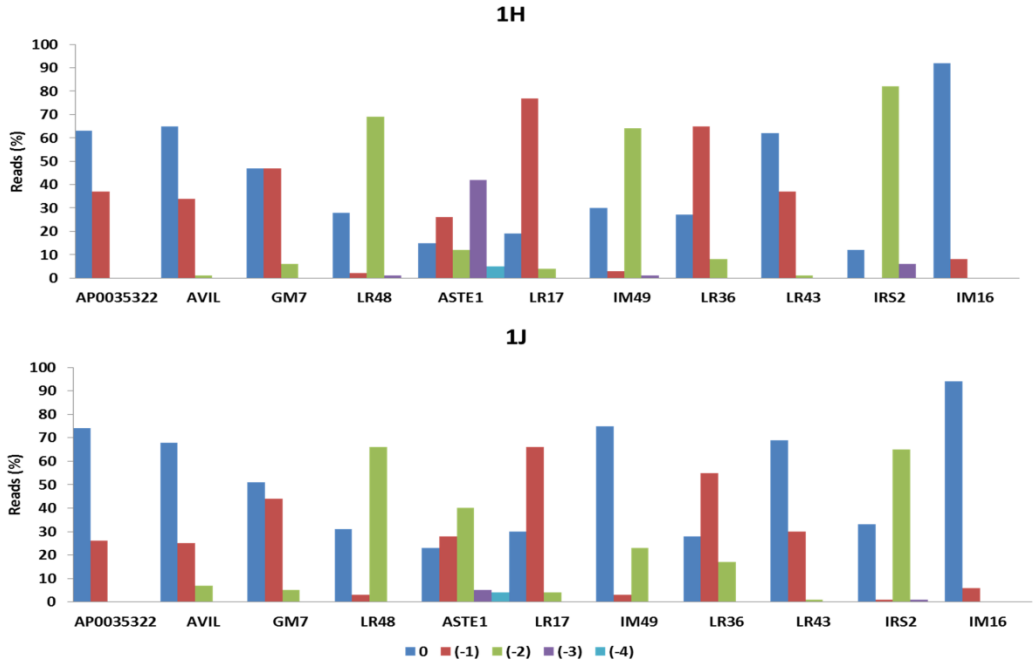


Figure 6-9: The clonal characteristics of the tumour PR45703/14. (A) Deletion frequencies of 11 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 2 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** Variant allele frequencies of the 11 markers in both specimens.

6.2.5. Clonal composition of the Fresh MSI-H CRC samples

Thirty six specimens were tested from 4 fresh MSI-H tumours using the 23 short repeats. Details of deletion frequencies of all individual specimens are shown in Appendix Figure 8-5, Figure 8-6, Figure 8-7and Figure 8-8.

6.2.5.1. Clonal composition of the tumour PR10654/14

The tumour PR10654/14 involves the Ileocaecal junction. It was tested by fragment analysis and found to be MSI-H.

Of the 23 tested markers, normal tissue specimen did not shown deletions in any of the tested markers. For this tumour, 15 markers exhibited deletion frequencies above the length specific threshold values. Of those 15 markers, 5 markers were

found to be unstable in all the tested tumour specimens. In the 6 O'clock scalpel (6° S) specimen, only 5 markers were unstable, indicating that this specimen likely contain cells belong to a clone which is the nearest to the original ancestral clone. In the scalpel 3 O'clock (3° S) specimen, 6 additional markers showed instability. The 3°, 6° and 12° O'clock needle biopsies showed a similar deletion profile as all these specimens showed instability in 13 markers as shown in Figure 6-10. In the 12° O'clock scalpel (12°S) specimen, a new 1bp deletion was observed in the marker AP0035322-9. This deletion was not observed in any other specimen, indicating that the 12° S specimen likely contains cells belong to a clone derived from a preceding branch.

In the specimen 9 O'clock needle (9°N), an additional deletion in the marker LR11-11 was observed. This alteration was not observed in any other specimen, indicating this specimen contain cells belong to a clone originates from previous clones. A very similar deletion profile was noticed for specimen 9° S, where the same markers showed deletion frequencies above the length specific cutoff values in the specimen 9° N. The only difference between the 2 specimens (i.e. 9°N and 9° S) was the marker LR48-11, where a deletion below the cutoff value in the specimen 9° S was found. However, deletion frequency of that marker in specimen 9° S (=29%) was very close to the cutoff value (which is 30%). The likely scenario is that contamination with normal cells could be the underlying reason for the reduction in deletion frequency of this marker in that particular specimen and both specimens contain cells that, in fact, belong to a single clone as shown in Figure 6-10.

	LR49-7	GM11-9	LR24-9	AVIL-10	ASTE1-11	LR20-8	IM66-C	GM7-11	GM14-11	LR48-11	LR44-12	LR43-12	IM49-12	LR11-11	AP-9	15
Normal	0.04	0.03	0.13	0.08	0.2	0.03	0.02	0.16	0.11	0.09	0.2	0.13	0.08	0.09	0.05	0
6°S	0.07	0.09	0.3	0.14	0.38	0.04	0.07	0.18	0.24	0.17	0.25	0.16	0.15	0.08	0.02	5
3°S	0.6	0.65	0.4	0.43	0.89	0.37	0.35	0.82	0.77	0.72	0.75	0.18	0.17	0.09	0.05	11
12°N	0.86	0.45	0.31	0.33	0.78	0.23	0.31	0.67	0.63	0.37	0.61	0.38	0.59	0.09	0.04	13
3°N	0.57	0.53	0.33	0.31	0.84	0.31	0.3	0.68	0.77	0.4	0.73	0.39	0.62	0.1	0.07	13
6°N	0.51	0.5	0.3	0.4	0.83	0.26	0.29	0.67	0.73	0.42	0.63	0.31	0.64	0.28	0.05	13
12°S	0.5	0.51	0.29	0.41	0.9	0.25	0.29	0.74	0.78	0.48	0.76	0.41	0.67	0.18	0.12	14
9°S	0.28	0.35	0.22	0.42	0.74	0.2	0.18	0.6	0.56	0.29	0.55	0.3	0.47	0.43	0.03	13
9°N	0.57	0.58	0.37	0.58	0.85	0.28	0.45	0.76	0.75	0.41	0.73	0.43	0.65	0.62	0.02	14
Stable																
Unstable																

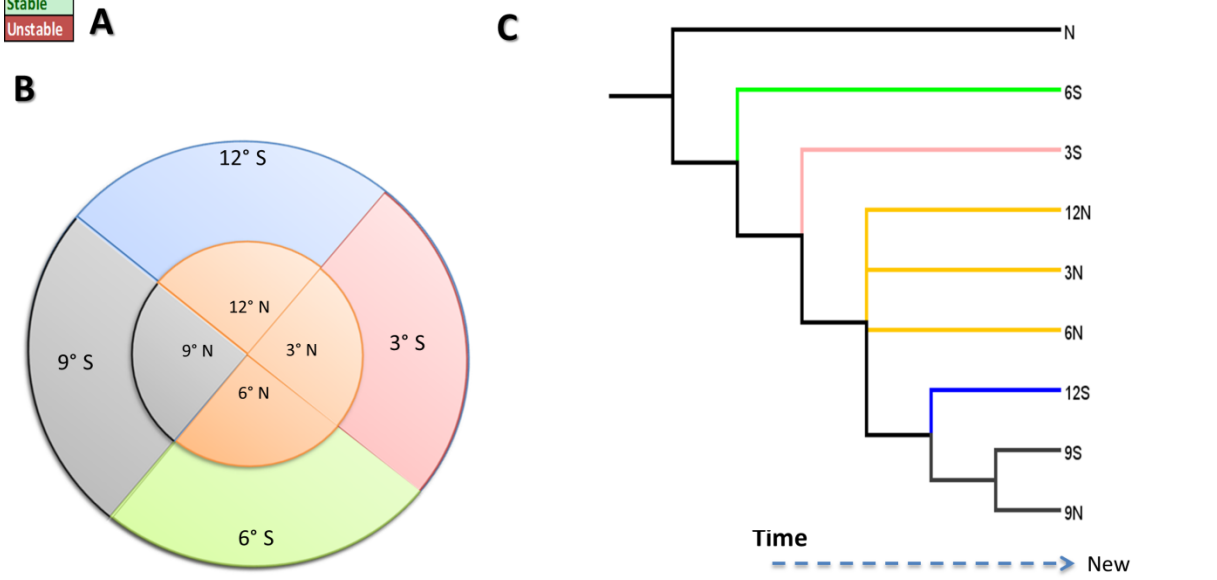


Figure 6-10: The clonal characteristics of the tumour PR10654/14.(A) Deletion frequencies of 15 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 9 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** A schematic representation of the distribution of clones according to the orientation from which they were obtained. Sections with the same colour indicate that they have the same deletion profiles and, thus, contain cells that belong to the same clone. **(C)** The phylogenetic tree based on the instability of markers in different specimens rooted to the normal specimen. Colours of the branches are corresponding to their clone colours in (B). The horizontal timeline shows that the further the branching, the newer is the specimen.

Overall, the instability profile of the tested specimens indicates that the specimen 6° S contains cells that belong to the nearest clone to the ancestral tumour cells, while the newest clones would likely be presented in specimens 9° N and 9° S.

6.2.5.2. Clonal composition of the tumour PR17848/14

The tumour PR17848/14 was tested by fragment analysis and found to be MSI-H, and 8 specimens were obtained in the same locations as described in Chapter 2 section 2.2.2. The normal specimen was retrieved from normal looking mucosa and it did not show any degree of deletion in the tested markers. All tumour specimens showed deletion frequencies above the cutoff values and each specimen showed at least 5 unstable markers. The specimen 6° S showed the least number of

unstable markers (5 markers), and thus it is likely to contain cells that belong to the nearest clone to the common ancestor.

The specimen 9° S contains cells that belong to clone branching from the clone presented in the specimen 6° S as it harbors 12 unstable markers. The specimen 12° N showed instability in 14 markers, sharing all mutations in the specimen 9° S and an additional 2 markers (LR43-12 and AVIL-10). This likely to indicate that the specimen 12° N contain cells of a descendent clone from the specimen 9° S as shown in Figure 6-11.

	LR49-7	GM11-9	LR24-9	GM17-9	ASTE1-11	LR17-11	LR20-8	IM66-C	AP-9	GM29-10	LR11-11	LR44-12	LR36-12	IM49-12	LR43-12	AVIL-10	GM7-11	GM14-11	18
Normal	0.009	0.02	0.07	0.02	0.19	0.15	0.004	0.008	0.02	0.03	0.06	0.15	0.22	0.06	0.13	0.07	0.09	0.05	0
6°S	0.08	0.08	0.09	0.08	0.36	0.29	0.04	0.08	0.06	0.07	0.16	0.23	0.25	0.15	0.16	0.07	0.07	0.05	5
9°S	0.18	0.16	0.13	0.13	0.47	0.46	0.08	0.17	0.16	0.17	0.32	0.31	0.2	0.22	0.19	0.11	0.29	0.22	12
12°N	0.63	0.58	0.21	0.37	0.88	0.75	0.29	0.64	0.33	0.61	0.76	0.78	0.21	0.16	0.43	0.14	0.08	0.17	14
3°S	0.31	0.32	0.11	0.22	0.67	0.53	0.15	0.34	0.2	0.34	0.51	0.5	0.57	0.37	0.26	0.07	0.07	0.18	14
12°S	0.53	0.5	0.12	0.35	0.83	0.79	0.29	0.6	0.35	0.57	0.76	0.71	0.62	0.5	0.39	0.1	0.08	0.08	15
3°N	0.8	0.75	0.17	0.48	0.94	0.82	0.43	0.77	0.59	0.77	0.89	0.84	0.73	0.55	0.5	0.1	0.11	0.1	15
6°N	0.45	0.49	0.19	0.32	0.82	0.7	0.26	0.33	0.28	0.46	0.6	0.58	0.3	0.43	0.37	0.14	0.21	0.1	16
9°N	0.91	0.86	0.26	0.51	0.42	0.74	0.5	0.89	0.87	0.87	0.91	0.94	0.38	0.8	0.53	0.13	0.43	0.89	17
Stable																			
Unstable																			

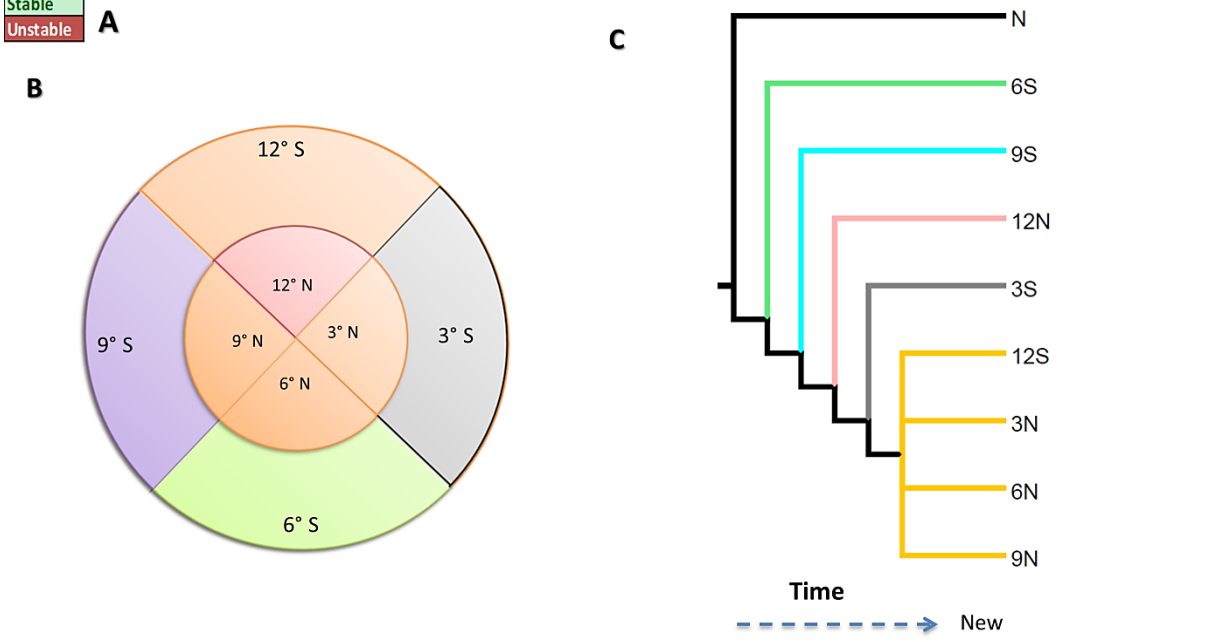


Figure 6-11: The clonal characteristics of the tumour PR17848/14.(A) Deletion frequencies of 18 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 9 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** A schematic representation of the distribution of clones according to the orientation from which they were obtained. Sections with the same colour indicate that they have the same deletion profiles and, thus, contain cells that belong to the same clone. **(C)** The phylogenetic tree of the tested specimens. Colours of branches are corresponding to their clone colours in (B). The horizontal timeline shows that the further the branching, the newer is the specimen.

The specimen 3° S showed instability in 14 markers, sharing all the deletions observed in specimen 9° S and additional 2 branch mutations (LR36-12 and IM49-12). The specimen 3° N and 12° S share the same instability profile (have 15

unstable markers) and they contain cells that are likely belonging to a clone that is derived from the same upstream clone (3° S).

There is an additional mutation in a single marker (AVIL-10) in specimen 6° N and 2 unique mutations in the specimen 9° N, these are in the markers GM7-11 and GM14-11. These mutations were not observed in other samples and thus, are likely to indicate that this specimen (i.e. 9° N) contains cells that are likely to belong to the newest clone among the tested specimens.

The specimens 12° N and 6° N are the only specimens that have shown instability in the marker AVIL-10, however, specimen 12° N lacks instability in 2 markers (LR36-12 and IM49-12) compared to sample 6° N.

6.2.5.3. Clonal composition of the tumour PR51896/13

The tumour PR51896/13 was obtained from a tumour mass involving the Ileocaecal valve and assessed by fragment analysis and found to be MSI-H. Of the 23 tested markers, all specimens showed deletion in at least 2 markers. Specimen 6° S showed the smallest number of unstable markers (LR24-9 and AVIL-10), thus this specimen is likely to contain cells that belong to the nearest clone to the common ancestral clone. The specimen 3° N showed a deletion in additional 9 markers, so this specimen lies further away from the parental tumour clone compared to specimen 6° S. The specimen 6° N showed a deletion in 12 markers with additional marker (GM11-9) compared to the sample 3° N.

Specimens 12° N and 3° S showed instability in 13 markers and both of them have had identical deletion profiles, so they contain cells that belong to a single clone originated from the upstream clone.

The specimen 12° S showed deletions in 13 markers with the absence of a deletion of a single marker that was unstable in the preceding specimen (LR36-12) as shown in Figure 6-12. The absence of LR36-12 instability is likely to be explained by contamination of the sample with normal cells. The specimens 9° N and 9° S exhibit the same number of unstable markers.

There is a high similarity in the deletion profiles among specimens 9° N, 9° S and 12° S, with the only exception is the marker LR36-12 which is stable in one

specimen while unstable in the other two. However, this stability is likely to be due to contamination with normal cells.

	LR24-9	AVIL-10	LR20-8	IM16-9	IM66-C	GM7-11	GM14-11	ASTE1-11	LR17-11	LR44-12	LR36-12	GM11-9	AP-9	IM49-12	14
N	0.05	0.07	0.002	0.02	0.005	0.06	0.05	0.18	0.14	0.14	0.13	0.02	0.01	0.05	0
6°S	0.11	0.14	0.04	0.05	0.05	0.13	0.09	0.28	0.23	0.23	0.21	0.02	0.03	0.08	2
3°N	0.32	0.39	0.32	0.26	0.34	0.5	0.55	0.64	0.51	0.47	0.36	0.05	0.07	0.22	11
6°N	0.23	0.28	0.21	0.17	0.25	0.33	0.38	0.49	0.3	0.35	0.45	0.09	0.02	0.13	12
12°N	0.35	0.44	0.34	0.23	0.34	0.49	0.55	0.69	0.47	0.51	0.5	0.04	0.2	0.37	13
3°S	0.28	0.37	0.29	0.32	0.29	0.4	0.5	0.61	0.45	0.45	0.47	0.05	0.08	0.35	13
9°S	0.55	0.38	0.59	0.37	0.65	0.7	0.76	0.83	0.63	0.7	0.6	0.1	0.34	0.64	14
12°S	0.31	0.21	0.32	0.23	0.29	0.48	0.37	0.62	0.41	0.44	0.16	0.32	0.19	0.4	13
9°N	0.33	0.43	0.35	0.24	0.36	0.45	0.58	0.64	0.49	0.52	0.34	0.2	0.19	0.41	14
Stable															
Unstable															

A

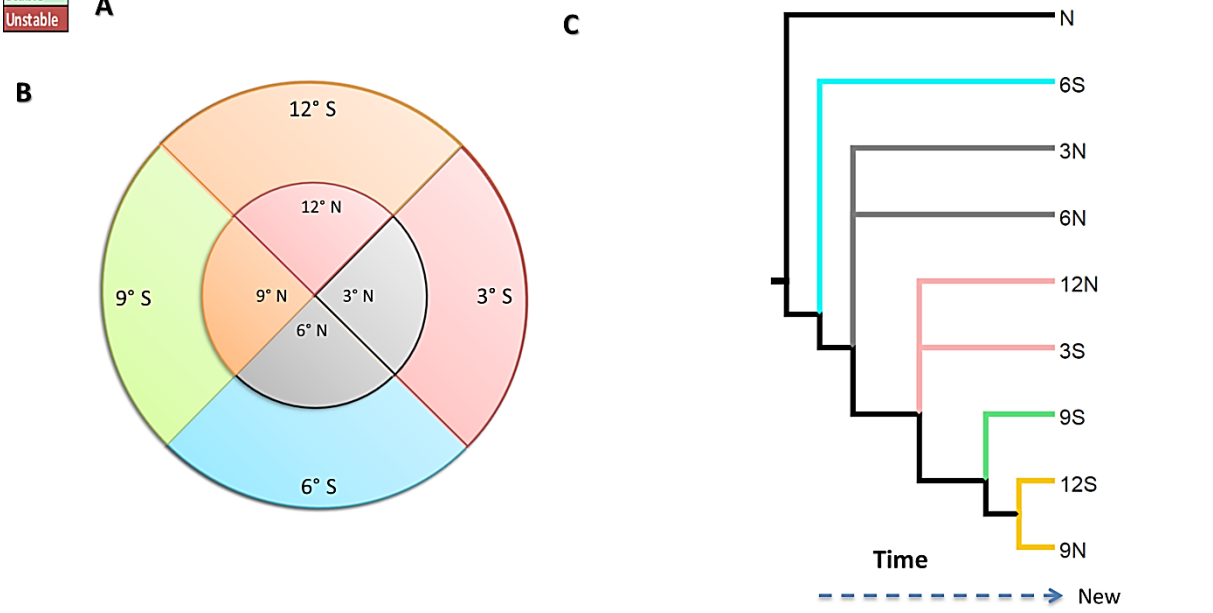


Figure 6-12: The clonal characteristics of the tumour PR51896/13. (A) Deletion frequencies of 14 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 9 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** A schematic representation of the distribution of clones according to the orientation from which they were obtained. The sections with the same colour indicate that they have the same deletion profiles. The samples 9° N and 12° S seems to represent a single clone **(C)** The phylogenetic tree of the specimens based on deletion profiles. Colours of branches are corresponding to their clone colours in (B). The horizontal timeline shows that the further the branching, the newer is the specimen. The specimens 9° N and 12° S seem to represent the newest clone.

6.2.5.4. Clonal composition of the tumour PR32516/14

The tumour PR32516/14 was assessed by fragment analysis and found to be MSI-H. Out of the 23 tested markers, the specimens showed instability in 7-13 markers, with the least number of unstable markers observed in the specimen 6° N. The specimen 6° N showed instability in 7 markers and these markers were unstable in all descendant samples.

The specimens 12° N and 12° S showed similar profile and are likely to contain cells belong to a clone that originate from the same upstream clone present

in specimen 6° N, as they showed an identical deletion profile across markers (both have 9 unstable markers) as shown in Figure 6-13.

Both of the specimens 9° N and 9° S show a new deletion in marker LR43-12, so they are likely to contain cells that belong to a clone located downstream to the clone presented in specimens 12° S and 12° N.

Ten markers in the specimen 3° S showed deletion frequency above the cutoff values. The specimen 3° N had 12 unstable markers with a single additional marker (IM49-12) compared to the sample 9° N. Based on the similarity of deletion profiles, specimens 3° N and 3° S seem to contain cells belong to a single clone.

There was a unique deletion in the marker GM17-9 in specimen 6° S. However, this specimen lacks the deletion in the marker GM29-10 which was observed in both 3° S and 3° N samples.

	LR49-7	GM11-9	LR24-9	IM16-9	AP-9	AVIL-10	ASTE1-11	IM66-C	GM14-11	LR43-12	GM29-10	IM49-12	GM17-9	13
N	0.005	0.1	0.1	0.02	0.01	0.09	0.23	0.003	0.09	0.11	0.02	0.06	0.02	0
6°N	0.19	0.08	0.33	0.1	0.15	0.18	0.45	0.09	0.23	0.16	0.03	0.14	0.02	7
12°N	0.16	0.15	0.33	0.12	0.17	0.2	0.49	0.11	0.3	0.23	0.02	0.12	0.01	9
12°S	0.26	0.13	0.16	0.15	0.28	0.19	0.52	0.21	0.49	0.2	0.03	0.23	0.02	9
9°N	0.26	0.16	0.47	0.19	0.27	0.31	0.66	0.14	0.3	0.35	0.12	0.15	0.03	10
9°S	0.38	0.18	0.25	0.24	0.39	0.29	0.72	0.18	0.44	0.36	0.02	0.25	0.03	10
6°S	0.67	0.36	0.39	0.41	0.71	0.46	0.87	0.35	0.65	0.41	0.03	0.43	0.33	12
3°S	0.21	0.15	0.21	0.26	0.39	0.27	0.88	0.3	0.76	0.23	0.29	0.28	0.001	10
3°N	0.51	0.25	0.41	0.3	0.48	0.36	0.8	0.35	0.72	0.43	0.42	0.32	0.02	12

Stable
Unstable

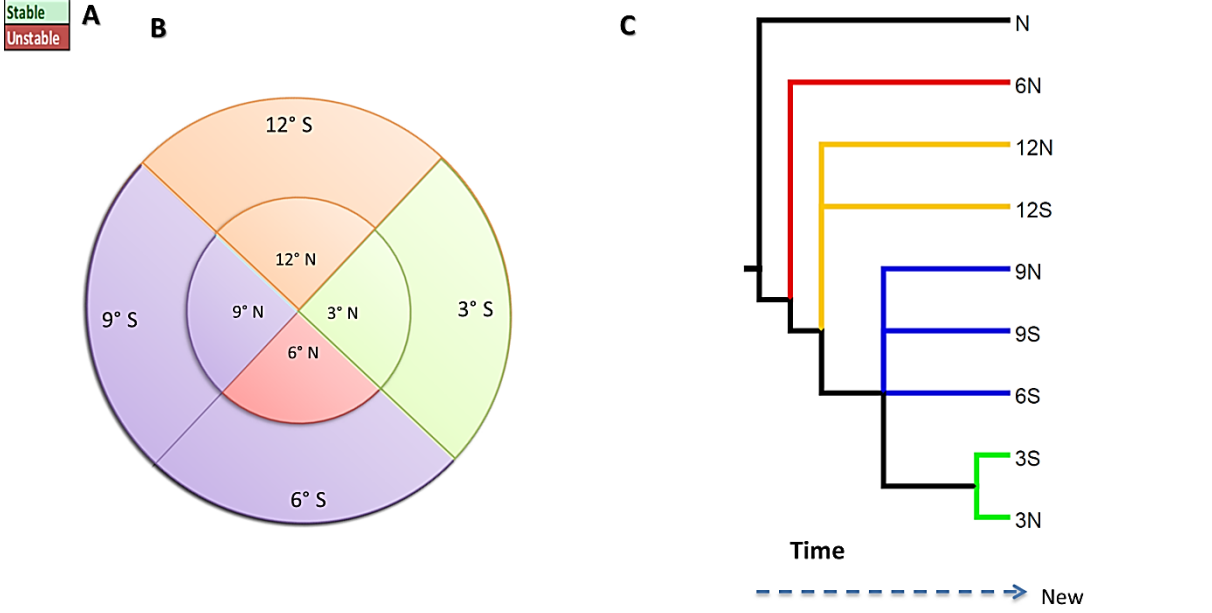


Figure 6-13: The clonal characteristics of the tumour PR32516/14. (A) Deletion frequencies of 13 markers (the upper raw), dark red boxes refer to unstable markers and green boxes refer to stable markers in the 9 tested specimens (leftmost column). The total number of unstable markers in each specimen is shown in the rightmost column. **(B)** A schematic representation of the distribution of clones according to the orientation from which they were obtained. Places with the same colour indicate that they have the same deletion profiles. The samples 12° N and 12° S seems to represent a single clone **(C)** The phylogenetic tree based on the instability of the tested specimens. The colours of branches are corresponding to their clone colours in (B). The horizontal timeline shows that the further the branching, the newer is the specimen.

6.3. Discussion

Intratumour heterogeneity has its impact on diagnosis, prognosis and management of different human tumours (Baisse et al., 2001, Linnekamp et al., 2015, Hardiman et al., 2016). Microsatellite markers were used previously to analyse the ITH and clonal characteristics of different human cancers (Rübben et al., 2004, Salipante and Horwitz, 2006, Redford, 2016). The development of the new NGS based MSI panel investigated and validated by our group (in previous chapters), made it possible to assess the intratumour heterogeneity on a large scale basis. In this Chapter, variant allele frequency was used as the main classifier for assessing the relationship between different specimens from the same tumour. The markers were called unstable based on the criteria and cutoff values proposed and validated

in previous chapters. Then, the number of unstable markers for each specimen was calculated and compared with that of other specimens for the same tumour.

Mutations in microsatellites accumulate with time following *MMR* gene knockout, thus the higher the number of mutations the longer the time since *MMR* loss (Ionov et al., 1993, Shibata et al., 1996). Based on the assumption that microsatellite instability events are cumulative, mutations common to all specimens (of a particular tumour) represent an early event (Tsao et al., 2000). Tumours associated with a higher number of unstable markers are likely to represent clones which have arisen more recently. It was possible to construct the phylogenetic tree for each individual tumour by applying the derived shared characters as a basis for constructing trees using the software package Mesquite (version 3.04).

During the search of the NHS database, a notable challenge of the search process was the lack of adequate pathological data in the database, especially for old samples; however, it was possible to retrieve 8 MSI-H CRCs with at least 2 specimens for each. The higher the number of specimens collected for each tumour, the better the ability to construct an evolutionary tree.

The possibility to disseminate to local or distant lymph nodes in MSI-H tumours is considerably lower compared to the MSS samples (Hu et al., 2011, Birgisson et al., 2015); this was reflected in the difficulty to find an MSI-H tumour with involved lymph node during the search process. However, it was possible to collect lymph node specimens for 3 tumours (PR34630/03, PR7146/13 and PR32079/14) out of the tested 8 FFPE tumours. In the tumour PR7146/13, both tumour and lymph node specimens exhibited the same number of unstable markers (9 markers for each). In the other tumour (PR34630/03), lymph node specimen exhibited a different number of unstable markers compared to the tumour specimen (7 markers for the lymph node vs 11 markers for the tumour specimen). In the 3 tumours, lymph node specimens share the unstable markers with the original tumour. This provides evidence that microsatellites can be used to assess the clonal characteristics in the involved lymph nodes in MSI-H tumours.

All tumours (FFPE and fresh CRCs) were previously tested by fragment analysis and all of them were classified as MSI-H. The instability status was

confirmed in our assay based on the previously proposed criteria for calling instability in the NGS based MSI assay.

A notable weakness of the overall approach is the inability to have an exact estimation of the tumour cell count within the tested specimen. An approximate estimation was made by the examining pathologist. It has been predicted that for 99.9% sensitivity detection of a heterozygous mutation in a heterogeneous tumour sample with 10% malignant cell population, 400-500 high quality sequencing reads are required (Fisher et al., 2016, Lin et al., 2014). Assuming this prediction is applicable for MSI detection, it is likely for our assay with the relatively high average coverage (of ~5000 per/amplicon) to detect microsatellite mutations even with the existence of a small percentage of tumour cells. This assumption is supported by the ability to detect instability in specimens 1J and 1H of the tumour PR53139/13, although both specimens had a very low tumour cell count (less than 5%).

For all fresh CRC tumours, 8 specimens from a clockwise orientation in addition to a normal specimen were provided. It was possible to construct the phylogenetic tree for all fresh tumours. In all fresh tumours, normal specimens were stable for all markers.

In 3 out of the 4 tested fresh MSI-H CRC tumours (PR10654/14, PR17848/14 and PR51896/13), the specimen 6° S exhibited the least number of unstable markers, while it showed the highest number of unstable markers in the 4th tumour (PR32516/14). Given these specimens were retrieved from all the 4 tumours in the same orientation and by the same pathologist, this reduces the possibility that the sampling technique could be the underlying reason for the low number of unstable markers in that particular biopsy in these 3 tumours.

Results obtained from this Chapter can be used as additional evidence of the employability of the panel of markers (as well as the calling system) developed and validated in the previous Chapters. It was possible to detect instability in all specimens from all tumours, while normal specimens did not exhibit any unstable marker. These findings confirm the initial reported phenotypes of the tested tumours and strongly support the previous validation of the NGS based MSI assay.

The development and validation of short amplicons (~ 150bp) has improved the success rate of analysing FFPE samples (90% of samples were successfully

amplified and sequenced with ≥20 markers). The ability to investigate the FFPE samples adds more advantage as these samples represent the major biological resource available in the pathological archives. Furthermore, being NGS based, the current approach looks more reliable than other studies (Fisher et al., 2016, Blake et al., 2001) as they were using fragment analysis to assess the clonal microsatellite alterations. However, the notable shortcoming of the current assay is that it is applied to MSI-H CRCs only. The exact estimation of tumour cell percentage and contamination with normal cells are additional limitations. However, tumour enrichment techniques (e.g. Laser capture microdissection) could be applied in future to reduce normal cell contamination.

6.4. Conclusions

The microsatellites can be used as evolutionary markers to assess the intratumour heterogeneity. The newly developed microsatellite markers proposed and validated by our group could be used to assess the clonality of the CRC tumours. The cutoff values that were suggested previously worked very well in discrimination between MSI-H and MSS CRC samples. The results of this chapter can be used to confirm the utility of microsatellites as convenient markers to assess the tumour age in MSI-H CRCs as evident in phylogenetic trees. Moreover, these results provide additional evidence for the reproducibility of the NGS based MSI assay developed and validated in the previous Chapters.

Chapter 7. General Discussion and Future work

7.1. General discussion

Due to the importance for diagnosis, prognosis and treatment, microsatellite instability testing has been recommended to be performed for all newly diagnosed CRC cases (de la Chapelle and Hampel, 2010, Vasen et al., 2013). Recently, a group of scientific societies (Association for Molecular Pathology, American Society of Clinical Oncology, American Society for Clinical Pathology and College of American Pathologist) released a guideline draft for the assessment of molecular biomarkers in CRCs. In that draft, MSI test was recommended for all newly diagnosed CRCs (Medscape, 2015). The most widely adopted MSI testing methodology is by multiplex PCR of 5 mononucleotide markers (20- 27bp in length) followed by fragment analysis and observation of allelic profile of the amplified markers in tumour compared to matched normal sample (Suraweera et al., 2002, Boyle et al., 2014). Although this approach is the most widely used one, it has limitations. It is laborious and the interpretation of results is mainly based on subjective inspection of the allelic peaks in addition to suboptimal specificity (Nguyen et al., 2013, Berg et al., 2009). For all the above reasons, this test is inconvenient to test all the newly diagnosed CRCs in line with the recent recommendations. Therefore, there is a need to develop a high throughput approach capable of satisfying the increasing need to test more CRCs being referred for MSI analysis.

One of the reported caveats of NGS is the high error rate in sequencing long homopolymers (Minoche et al., 2011). Long homopolymers were reported to generate more sequencing errors (~60% error rate for 13bp repeats), and these errors are reduced as the length of repeat decreases (down to error free for 9bp repeats) (Clarke et al., 2001). Short mononucleotide repeats (7-12bp), therefore, offer a good option to implement such an approach as they are less prone to generate sequencing errors compared to longer ones (Redford, 2016). I have developed a panel of 17 short (7-12bp) repeats derived from an *in silico* search of the length variable repeats in MSI-H CRCs. Observation of deletion curves for all the tested markers in that part of the assay showed that the sensitivity of a given marker is always less than specificity at a specific value of deletion frequency (see Chapter 3 section 3.2.2.1). This indicated that the inclusion of more markers in the final panel would be justified to improve the sensitivity. A threshold set of length-specific cutoff

values were determined. Variant reads in the MSS samples were more prevalent with the longer (11 and 12bp) markers than shorter markers (7-9bp), therefore, rising cutoff values for these long markers improved specificity. This also confirmed that longer markers are more variable (length variation) compared to short repeats (7-10bp). This observation is consistent with other studies (Clarke et al., 2001, Fazekas et al., 2010, Redford, 2016). Another finding is that including the allelic bias information (which is the other parameter used in this assay to identify instability) increases the reliability of MSI calling compared to using deletion frequency alone. Therefore, a combined score for calling instability (referred to as weighted scoring system) was proposed. In the weighted MSI scoring system, a specific score was given for each feature (i.e. deletion frequency and allelic bias) and an overall score of ≥3 was used as cutoff for calling a sample as MSI-H. The use of a cutoff value as high as ≥3, appears justifiable as this means that any given sample can only be called as MSI-H when at least 2 markers exhibited deletion frequency above the length specific threshold, and one of them should have evidence of allelic bias. In the absence of allelic bias, 3 markers with deletion frequency above threshold (none with allelic bias) are required for a sample to be called as MSI-H. This would reduce the possibility of mis-calling a stable sample as MSI-H when the cutoff value is low, as there is evidence that all CRC cases (including MSS cases) have a low level of instability (Laiho et al., 2002). This weighted scoring system was then validated across an independent cohort and yielded a 100% sensitivity and specificity. This provides evidence for employability of the 17 markers panel alongside the weighted scoring system. Although this system yielded 100% specificity and 96% sensitivity, the reliance on heterozygous SNPs to determine the allelic bias represents the main drawback of this system.

7.1.1. Comparison of the current assay with other methods

In 2012, TCGA published a seminal paper about the genome wide somatic alteration in CRCs (Cancer Genome Atlas, 2012). In that study, researchers used next generation sequencing to detect mutations and classify CRCs. In addition to investigating somatic mutations, they also utilised exome data to analyse MSI (using a set of 36 coding mononucleotide markers 6-10bp in length) by quantitative comparison of the altered allele to the wildtype allele (variant allele frequency). It was possible, from that study, to detect high variant allele frequency of mononucleotide

markers in selected genes in tumour samples compared to the corresponding normal tissue. Tumours that lack the *MLH1* gene function were found to show up to 50-fold increase in frameshift mutations in these target genes compared to those with wildtype *MLH1*. Furthermore, it was concluded that the degree of instability of a given marker affected by the length of that marker. That study provided the first evidence of the possibility to test MSI in short repeats using next generation sequencing. Since then, several groups have tried to develop new approaches for MSI testing robust enough to cope with the increased number of cases being referred. In 2013, an MSI assay based on RNA sequencing data was investigated (Lu et al., 2013). In that study, RNA-seq data were gathered and analysed from 20 different cancer cell lines with known MSI status. The proportion of insertions in microsatellite loci over all insertions (named as PI) and the proportion of deletions in microsatellite loci over all deletions (named as PD) were calculated. The ratio between them (PI/PD) was referred to as MSI- seq index and used to assess the instability of each microsatellite locus. A significant increase in the proportion of indels was observed in the MSI samples compared to those in control samples (HapMap samples in that particular study), while there was no significant difference between MSS and the control samples. The disadvantages of this approach are the dependence on the expressed microsatellites, the lack of analytical validation for the proposed MSI-Seq index and being an RNA- seq based assay, which is relatively expensive. The work presented in this thesis has the advantages on Lu et al's work in that short repeats were analysed in amplicon based MiSeq sequencing and the calling system was validated using an independent cohort.

In 2014, a further NGS based MSI analysis pipeline (called mSING), was proposed and assessed (Salipante et al., 2014). NGS data from a total of 324 different tumours (colorectal, endometrial, ovarian, breast and others) from 3 independent assays (TCGA exome, ColoSeq UW and Oncoplex UW) were retrospectively examined to assess instability in 2957 (for TCGA samples), 146 (for ColoSeq UW) and 15 (for Oncoplex UW) mononucleotide markers. These markers were 3-36bp in length. The analysis was based on calculation of the variant allele frequency. Each variant allele with sequencing reads exceed 5% of the most abundant allele was tallied and the variant allele frequency for each marker was calculated in the MSS group to create a baseline reference value. Then, mean

number of alleles for each microsatellite marker and standard deviation were calculated for the MSS group. The same calculations were done for the MSI-H group of samples and the marker was said to be unstable if the number of alleles is more than (mean number of alleles + (3xSD) of the baseline reference) of the MSS group. They called this approach mSING (MSI by NGS) and a combined sensitivity and specificity of 97.8% and 98.3% was obtained in the tested cohorts. One disadvantage of that study is that it lacks a validation assay and the vast majority of the analysed markers were derived from the TCGA exome analysis (which has used fresh frozen CRC samples). Furthermore, being dependent on the MSS reference value, this means that MSS baseline reference value needs to be set in each run separately. Advantages of the work presented in this thesis on Salipante's work are that, all CRC samples (during both development and validation) were FFPE, which is the same kind of samples that would be referred in routine work, making the assay more compatible with current practice in clinical laboratories. In addition, a fixed set of length specific cutoff values of deletion frequencies was proposed for calling instability, which can be used permanently across different MiSeq runs, without the need for normal tissue to be routinely analysed.

More recently, Gan *et al*. (2015) used Illumina MiSeq platform to assess the utility of both mononucleotides (3 markers 25-34bp in length) and dinucleotides (2 markers 40bp in length) in microsatellite instability using 2 independent cohorts collectively composed of 52 CRC samples. For each marker, the most prevalent allele in both tumour and a corresponding normal sample was calculated. Then, the marker was classified as unstable if the deviation in the tumour sample compared to normal was ≥2bp for mononucleotides and ≥ 4bp for dinucleotide markers. Quantitative analysis of the MiSeq data from that study showed that the sensitivity and specificity of these mononucleotides were both 100%, while for dinucleotides, the sensitivity was 47- 59% and specificity was 96- 100%. Although this study used the ultra-deep sequencing of microsatellites, notable disadvantages of this analysis were the use of long mononucleotide and dinucleotide markers, the unblinded analysis and the comparison with matched normal tissue. Advantages of the work presented in this thesis over Gan *et al*'s work are that MSI analysis was done using a relatively high number of samples (266 FFPE CRC samples in total), of them, 211 samples were analysed blindly using short (7-12bp) mononucleotide markers.

Furthermore, the analysis was done without interrogation of matched normal samples which will reduce the cost. In addition, our assay has utilised another feature to assess instability and to increase the confidence that the detected alteration is a real instability rather than a sequencing artefact. This additional tool called allelic bias and to our knowledge, this is the first time this tool has been deployed in detection of microsatellite instability.

7.1.2. Comparison of deletion curves suggests further assessment of robustness and reproducibility is required

Deletion curves are the graphical plotting of deletion frequencies for a given marker across all MSI-H and MSS samples. The comparison of deletion curves of the 17 markers across the 3 different cohorts indicated that 2 markers (LR36-12 and GM14-11) showed low specificity in the Spanish cohort as explained in Chapter 5 section 5.2.3. The marker LR36-12 showed a low specificity in the Spanish cohort compared to other cohorts (Specificity= 92%) at 30% deletion frequency. This anomaly is likely due to the relatively low depth to which that marker was sequenced (average sequencing depth was 188 per/amplicon) compared to other markers (ranging from 1619 for marker GM7-11 to 5096 for the marker LR24-9) in that particular cohort. This assumption is supported by the fact that there was no such a problem with this marker either in Newcastle or in the Edinburgh cohorts, where the average depth to which this marker was sequenced in both cohorts was more than 188 per/amplicon.

Marker GM14-11, showed a significantly lower sensitivity in the Newcastle cohort (= 25%) and showed the lowest specificity in the Spanish cohort (= 89%) at a deletion frequency of 30%. For this marker, 4 MSS samples in the Spanish cohort were found to be discordant as they showed deletion frequency above 30% (which is the cutoff value used for that particular marker). I investigated the possible underlying reason for this difference by inspecting the nature of surrounding sequences and checking for the existence of SNPs at the primer annealing site. There was a low minor allele frequency (MAF) SNP (rs539119173 with MAF= 0.06%) at the primer annealing site (as explained in Chapter 5 section 5.2.4.1), so this could be one possible reason for the inter-cohort difference.

I re-tested the marker GM14-11 in 24 samples from 2 different cohorts (Spanish and Newcastle cohorts) in an independent MiSeq run with higher library

concentration (= 10 pM) compared to the initial test (= 4 pM). Out of the 24 retested samples, only 2 samples (S79 and S74) were called differently in the re-analysis assay, and both of them were called as stable in the re-analysis while they were unstable in the initial assay. For both amplicons with discordant reads, the deletion frequency was lower in the re-analysis assay compared to the initial assay as shown in Table 7-1. This difference in deletion frequency is more likely owing to the change in the library concentration, and consequently sequencing depth (as explained in Chapter 5 section 5.2.4.2)

Parameter	S79		S74	
	Initial	**Re-analysis**	**Initial**	**Re-analysis**
MSI status	Unstable	Stable	Unstable	Stable
Library Concentration	4 pM	10 pM	4 pM	10 pM
Deletion Frequency	36%	6%	40%	8%
Depth (per/amplicon)	253	1756	509	36487

Table 7-1: Specifications of the marker GM14-11 with discordant reads in 2 different samples (S74 and S79). There is a notable difference in deletion frequencies between the initial and re-analysis in both amplicons, which could be explained by the notable difference in sequencing depth for both amplicons.

Both amplicons were included in the DNA integrity assay shown in Chapter 4 section 4.2.7 and both of them showed DV100<17% of the tested DNA samples, indicating that DNA degradation is an unlikely reason. This supports the notion that a specific depth is required to ensure that a correct calling of instability is made.

7.1.3. Suspicion of polymorphism in 2 independent markers in 2 different samples

All the 17 markers included in the panel were defined as monomorphic according to dbSNP 144 build. However, the discovery of private polymorphisms is always possible. During the extensive analysis of these markers against the 3 geographically distinct cohorts used here, 2 markers showed an unusually strong allelic bias in 2 different samples from 2 independent cohorts. The marker DEPDC2-G showed 90% of sequencing reads from +1 genotype (i.e. 1bp insertion) in sample S129 (from the Spanish cohort). The other marker is LR48-11 showed 98% of sequencing reads from -2 genotype (i.e. 2bp deletion) in the sample E82 from Edinburgh cohort. However, I was unable to confirm the polymorphism in both instances because this needs a corresponding germline DNA to be tested. Therefore,

it is recommended that, whenever there is a suspicion of polymorphism which could influence an MSI classification, a normal tissue from the same patient need to be tested to confirm or exclude polymorphism. This indicates that allelic bias should be carefully interrogated in order to differentiate it from polymorphism.

7.1.4. Determination of the optimal quality metrics for an NGS based MSI approach

Three library concentrations were tried in the 3 independent MiSeq runs, 4, 8 and 10 pM. In the first Miseq run, I followed the instructions of the Nextera library preparation protocol and a library concentration of 8 pM was used. That run resulted in 1700 k/mm^2 cluster density and Q30 score (99.9% probability of a base being called correctly) of 66.7% as shown in Table 7-2

When primers are modified (tagged primers targeting ~150bp amplicons), the 16S metagenomics protocol was used for library preparation. According to that protocol, it is recommended to start with 4 pM as initial library concentration. Using that concentration resulted in a relatively low cluster density and low overall sequencing reads (510 k/mm^2 cluster density and 12,610,764 sequencing reads). However, using that concentration resulted in the highest quality score (Q30= 69.1%), which is expected in the view of such a low cluster density. In order to improve both data output and cluster density, I tried higher library concentration (10 pM) in the subsequent MiSeq run. That run yielded a better cluster density and overall sequencing reads (1450 k/mm^2 and 34,774,582), but at the cost of Q30 score where it dropped down to 55.5% as shown in Table 7-2

Parameter	MiSeq runs		
	Newcastle	**Spanish**	**Edinburgh**
Library Concentration (pM)	8	4	10
No. of amplicons	1,200	2,267	2,595
Amplicon size (bp)	~300	~150	~150
Sequencing Success rate (%)	97.5%	80%	99%
Total No. of reads	41,383,441	12,610,764	34,774,582
% PF	92.7%	94 %	88.8%
Average Coverage (per/amplicon)	10,400	2,900	5,100
Q30	66.7%	69.1%	55.5%
Cluster density (k/mm^2)	1,700	510	1,450

Table 7-2: Quality metrics for the 3 MiSeq runs using 3 different library concentrations. per= paired end reads. **Sequencing success rate**= the percentage of amplicons that were sequenced to a depth of ≥100 reads. **%PF**= percentage of sequencing reads that have passed the quality filter.

Based on the information summarized in Table 7-2, it would be possible to recommend specific inputs for an optimal assay. A run composed of 96 samples amplified across the 17 markers (all have amplicon size ~150bp) in a library concentration of 8 pM could therefore be suggested. In theory, the average sequencing depth (coverage) expected from such a MiSeq run is 5,000-10,000 per/amplicon. It would be possible to increase the number of samples being analysed, but of course at the cost of average depth. However, it was possible to successfully call instability in most amplicons (80%) even in the MiSeq run with the lowest depth (the MiSeq run of Spanish samples). Performing such a run, would be useful to determine the optimal number of samples that can be analysed in a single run at a specific library concentration in the future.

7.1.5. Cost analysis and turnaround time

Microsatellite instability becomes increasingly important, especially for CRCs. Recently, the cost effectiveness of 8 strategies for detection of Lynch syndrome in all early onset (<50 years) CRCs was assessed (Snowsill et al., 2015). In that systematic review, the 8 testing strategies ranged from no genetic testing (strategy 1) to direct mutation testing (strategy 8) for all newly diagnosed CRCs before the age of 50. Interestingly, strategy 5 (which represents MSI testing (if positive) > *BRAF* testing (if negative) > mutation testing) was found to be the most cost effective strategy

(costing 5,491 GBP/ Quality adjusted life year gained over no testing strategy). This strongly supports the cost effectiveness of MSI testing of all newly diagnosed CRC patients and increases the demand to develop an accurate and cost effective MSI assay. However, further analysis is underway to inform NHS policy and the cost of MSI testing is likely to be a pivotal issue in that debate (Tristan Snowsill personal communication).

In order to reduce the overall cost of our assay, primers were designed to be incorporated with tag sequences, thus reducing the cost and time required for the subsequent library preparation. Moreover, the tagged primers were designed to amplify shorter sequences (~150bp) rather than the initial 300bp amplicons that are recommended by the Nextera library preparation protocol. Multiplexing the amplicons in a single PCR reaction would streamline the assay, reduce the time, effort and cost. Previous trials of duplexing 2 amplicons were successful and ongoing trials of multiplexing showed the feasibility to amplify the 17 markers in down to 4 separate multiplexed reactions.

A cost analysis was done (shown in Appendix Table 8-3) and according to that analysis, the cost of the recommended MiSeq run is expected to be 26.2 GBP/ sample and the test results are expected to be reported in 7 days as explained in Appendix Table 8-3 and Table 8-4. However, Automation of pre and post PCR steps together with use of automated pipeline for data analysis would shorten this turnaround time to a single week (i.e. 5 working days only).

We are in a position now to liaise with the clinical laboratories to perform the MSI test developed and validated in this thesis in parallel to their routine test as a further step to fulfil the clinical monitoring of the MSI assay. This monitoring would be useful to assess the concordance rate of our test with the currently used test and, ultimately, making it ready to be commercialized. As our test is designed to be a high throughput approach, the collection of 96 samples (that was recommended in section 7.1.4) from a single laboratory would be likely to take a long time, so it would be worth to setting up the test in a reference genetic laboratory and asking all regional laboratories to refer their CRC samples. This will allow gathering the required number of samples in a comparatively short time.

7.1.6. Future improvements in assay design

As this test is developed and tested in an NGS platform, it would also be possible to include other relevant variants (e.g. *BRAF*, *MMR*, *KRAS* genes) with this MSI assay to be done at the same run. This will facilitate the overall approach by doing these tests in a single run while they are currently done separately. In addition, this will minimise the cost per case dramatically compared to the current multistep approach. Recently, both MSI and mutational hot spots of relevant genes (*KRAS*, *NRAS*, and *BRAF*) were combined in a single NGS assay in an approach called MSIplus (Hempelmann et al., 2015). MSI was assessed in 81 tumours using mononucleotide markers (12- 28bp in length) while mutational analysis was done for 61 tumours. Among them, 15 samples were tested for both MSI and mutational hotspots in the target genes mentioned above using the Illumina MiSeq platform. The MSIplus approach used the software mSINGS for calling instability developed by (Salipante et al., 2014) and the assay was reported to be 97% sensitive and 100% specific in the detection of MSI. Advantages of this assay are: the inclusion of additional target genes in the same sequencing run, the omission of the need for normal tissue control and the high sensitivity and specificity. Disadvantages of the MSIplus approach are the use of long repeats and the use of mSINGS that (as explained earlier) depend on the establishment of MSS baseline values of variant read length. This means that, a baseline needs to be established for each run independently. The work in this thesis has the advantages on MSIplus in that: short mononucleotides were used and a clearcut threshold set of cutoff values was developed and validated.

The development and validation of the NGS based MSI test done in this thesis would provide a basis for further development of the assay to be able to detect microsatellite instability in the existence of very low number of tumour cells or become a non-invasive approach. In 2011, microsatellite instability was detected in saliva in higher abundance than in peripheral blood in *MMR* gene mutations carriers (Hu et al., 2011). This means the test could be offered in a non-invasive approach by targeting the cellular components of saliva. However, detection of such a low prevalent DNA needs a more sensitive methodology (high sensitivity to detect cells existed in a low population). To improve detection of low prevalence variants (e.g. in saliva or circulating tumour cells), researchers tried an ultrasensitive targeted

sequencing using a modified probe called single molecule molecular inversion probe (smMIP) (Hiatt et al., 2013). Such an approach would streamline the assay and make the detection of low prevalent variants possible. Furthermore, this approach is highly accurate (error rate down to 1 error/10,000 bases) and reduces the impact of low quality DNA. It would be worth to try to incorporate the 17 primer pairs developed and validated in this thesis with smMIPs to enable the assay to be multiplexed and able to detect low prevalence mutations.

7.1.7. Using short mononucleotide markers to assess intratumour heterogeneity

In chapter 6, 23 short mononucleotides were used to assess the clonal characteristics of 56 MSI-H tumour specimens (from 4 fresh and 7 FFPE). It was possible to detect intratumour variation in microsatellite instability. Lymph node specimens were tested in 3 FFPE tumours (PR7146/13, PR34630/03 and PR32079/14). In all the 3 tumours, the lymph node showed shared instability (instability in some or all markers) with the corresponding primary tumours. This provided evidence of employability of short repeats to assess lineage relationship of both primary tumour and its associated metastasis.

In the tumour PR53139/13, 4 samples (1H, 1J, 1M and 1N) were investigated. All the 4 tested specimens showed at least 5 unstable markers. Interestingly, both 1J and 1H specimens of that tumour were reported to have <5% tumour cell population and exhibited at least 5 unstable markers (in the specimen 1J). A possible explanation for the detection of microsatellite instability in spite of the low tumour cell content in those samples is that this tumour might be too old (long time since loss of *MMR* genes), thus resulted in a high level of instability.

In the fresh CRC tumours, 8 specimens were retrieved from each tumour together with normal tissue sample from the same patient. It was possible to assess the clonal characteristics of all tumours. Furthermore, this enabled me to construct the phylogenetic tree for all of them. Three of the 4 tested tumours were examined previously for clonal characteristics by Dr Lisa Redford (Redford, 2016). However, in that assay, these tumours were investigated across 20 mononucleotide markers (8-14bp in length), 12 of them were included in my assay. The average depth in Redford's work (1672 per/amplicon) was relatively lower than the depth achieved in my work (4971 per/ amplicon). In Redford's work, the specimen 6° S of the tumour

PR51896/13 did not show instability in any of the 20 tested markers, while the same specimen showed 2 unstable markers (LR24-9 and AVL-10) in my analysis. This difference is, perhaps, due to the relatively low sequencing depth of that specimen (1100 per/amplicon) compared to the depth to which the same specimen was sequenced in my analysis (5036 per/amplicon).

Overall, results from the clonality analysis showed that mononucleotide markers can be used to identify intratumour heterogeneity and help to identify cells from different clones, thus providing a tool to construct a lineage relationship among different specimens from the same tumour. Furthermore, the incorporation of primers with smMIPs would make the approach more sensitive and enables detection of variants in tumour cell population down to 1% (Hiatt et al., 2013).

Chapter 8. Appendix

Case No.	MS Status	Homopolymers tested
G1	MSI-H	All 25 markers
G2	MSI-H	All 25 markers
G3	MSI-H	All 25 markers
G4	MSI-H	All 25 markers
G5	MSI-H	All 25 markers
G6	MSI-H	All 25 markers
G7	MSI-H	All 25 markers
G8	MSI-H	All 25 markers
G9	MSI-H	All 25 markers
G10	MSI-H	All 25 markers
G11	MSI-H	All 25 markers
G12	MSI-H	All 25 markers
G13	MSI-H	All 25 markers
G14	MSI-H	All 25 markers
G15	MSI-H	All 25 markers
G16	MSI-H	All 25 markers
G17	MSI-H	All 25 markers
G18	MSI-H	All 25 markers
G19	MSI-H	All 25 markers
G20	MSI-H	All 25 markers
G21	MSI-H	All 25 markers
G22	MSI-H	All 25 markers
G23	MSI-H	GM9,GM11,GM21,GM28,LR15,IM14,LR46,IM41, LR24,IM43,IM59,LR49,LR51,GM17,IM16,LR40
G24	MSI-H	GM9,GM11,GM21,GM28,LR15,IM14,IM41, LR24,IM43,LR49,LR51,GM17,IM16,LR40
G25	MSI-H	IM19,IM55,IM66,IM67,LR10,GM23,LR20,LR8,LR21
6	MSS	All 25 markers
9	MSS	All 25 markers
11	MSI-L	All 25 markers
12	MSI-L	All 25 markers
13	MSI-L	All 25 markers
14	MSI-L	All 25 markers
15	MSI-L	All 25 markers
16	MSS	All 25 markers
17	MSS	All 25 markers
18	MSS	All 25 markers
19	MSS	All 25 markers
20	MSS	All 25 markers
21	MSS	All 25 markers
22	MSS	All 25 markers
23	MSS	All 25 markers
24	MSS	All 25 markers
25	MSS	All 25 markers

Case No.	MS Status	Homopolymers tested
26	MSS	All 25 markers
27	MSS	All 25 markers
28	MSS	All 25 markers
29	MSS	All 25 markers
30	MSS	All 25 markers
31	MSS	All 25 markers
32	MSS	All 25 markers
33	MSS	All 25 markers
34	MSS	All 25 markers
35	MSS	All 25 markers
48	MSS	All 25 markers
49	MSS	All 25 markers
50	MSS	All 25 markers

Table 8-1: Samples that were included in the Newcastle cohort and the markers analysed for each sample. Not all MSI-H DNA samples were amplified against equal number of markers.

Marker	Sequence
LR49- 7 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAATTTGGGAAAGGGGCACAA
LR49- 7 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTGATGGCCAAGTCCCC
IM66 - C FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGGAGGTGCTGGAAATCC
IM66 - C RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATCAGCCGCGTCGTAGG
DEPDC2- G FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTTCACACACATGCAAGCTG
DEPDC2- G RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGAAGGGTAGGGAGATGCAGA
GM9- 8 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCGTATTCCAGGAGTAAGAGT
GM9- 8 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCTCAGAGGGAAGGTGGCA
LR20- 8 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGCATTGCCCCTATATACTGT
LR20- 8 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTCCCAGTTCTGAATCTAGAAAGA
GM11- 9 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCCTACGTATCTAAGTATTCTCCAGC
GM11- 9 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACAGTGGGTTTCAAATGTCACTTC
LR24- 9 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTAACCAAAGCAGGAAAACATT
LR24- 9 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCTCTCTCCCTGGAATAAGT
IM16- 9 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAATCAGCAGTGTTCATACCTTC
IM16- 9 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTTGTTCACTTTAGTAGGAACTGGT
GM17- 9 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGAGAAGTCAGTGCATGTGTCTT
GM17- 9 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCCACCAAGATTGTAAAATGTGA
AP0035322- 9 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGACTGTGGTTTTAATTTGCATTTCCC
AP0035322- 9 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGTGTGCCTTTAAAGTGACCTT

Marker	Sequence
GM7- 11 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTGGCTTGTTTTCATTTTGTC
GM7- 11 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCATATGGGGTTTGGTCACATTTT
GM14- 11 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCTCAATGACTTCCCAGGCTAA
GM14- 11 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGAAACATTGTGGATTGCTAGCTG
LR48- 11 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGGAGGAAGTATCTGGTCTTCT
LR48- 11 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGCACATTTACTTAAGCCCTGG
LR11- 11 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCTGTGGTCTGTGAAGCTA
LR11- 11 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGTGCATTTGAACATCGCCTC
LR36- 12 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGTGGTGACCCTGAACGTTAA
LR36- 12 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCCTGGGTGTAAATGATGGGAA
LR44- 12 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGGCCAAGAGTTCAAGACCA
LR44- 12 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGGATGAGAATTAGCATACCTTCCA
IM49- 12 FP	TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGGTAGTTGGATCGCTTCAGG
IM49- 12 RP	GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGCCTCTTGAGTAGCTTGG

Table 8-2: The 17 primer sets that were used in the analysis of both Spanish and Edinburgh cohorts. All primers were incorporated with 5' tag oligos. The average primer length is ~55 bases (include both the primer and the incorporated tag sequence).

MiSeq based assay					Fragment Analysis assay				
	No.	**Price/ item (£)**	**No of reactions**	**Cost/sample (£)**		**No.**	**Price/item (£)**	**No. of reactions**	**Cost/sample (£)**
Primers (synthesis scale 0.04 µmol)	17	17	400	0.04 x 17 (0.68)	**MSI Analysis system**	1	757	50 pairs	19 (40 pairs + cotrols for 5 runs)
Herculase (800 rcs for 25 µl)	1	281	800	0.35 x 20 (7)	**AmpliTaq Gold DNA polymerase (100 rcs)**	1	38	100	0.42 (90 rcs + controls)
MiSeq kit v3 (600 cycles)	1	1,035	96	10.8	**HiDi formamide (25ml)**	1	32	200	0.16
QIAxcel screening kit (2400)	1	517	2,400	4.3					
AMPure XP (60 ml)	1	721	500	1.4					
Qubit dsDNA HS assay kit	1	168	500	0.33					
Nextera XT Index kit (96 indexes for 384 samples)	1	662	384	1.7					
Overall cost/ sample (£)				26.2 (96 samples) 22.3(150 samples)					19.58 (40 pairs)

Table 8-3: Cost analysis and comparison between MiSeq based assay and the conventional fragment analysis based assay. The cost of MiSeq analysis of 96 samples is 26.2 GBP/ sample (red coloured text) while the cost could be reduced to 22.3 GBP/ sample when 150 samples are analysed in a single MiSeq run.

MiSeq analysis (96 samples)	
Step	**Time (days)**
DNA extraction and quantitation	2
PCR * and QIAxcel analysis	4 (1*)
Library preparation	1
MiSeq	2
Data analysis **	1**
Overall TAT	11 (7*) days

Table 8-4: Total turnaround time (TAT) expected from the MiSeq based assay. The expected TAT to analyse 96 samples is 11 working days. When the 17 markers are amplified in 4 multiplex PCR reactions (rather than 17 reactions) for each sample and data analysis is automated, the TAT would be reduced to 7 working days only (shown in asterisks). However, when both pre and post-PCR steps are automated, the TAT is expected to be as short as 5 working days only.

An MSI Test Suitable for MiSeq or Use in Local Pathology Departments

Lisa Redford (1), Mauro Santibanez Koref (1), Stephanie Needham (2), David Evans (2), Julie Coaker (1,4), Ottie O'Brien (3), Matthias Kloor (5), John Tyson (4), Ghanim Alhilal (1), Iona V. Middleton (1) Jonathan O'Halloran (4), Michael Jackson (1) and John Burn (1,3,4)

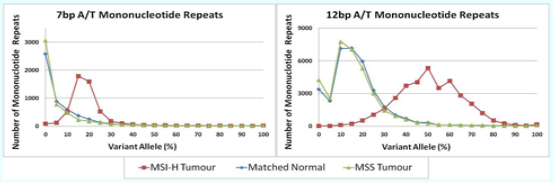
Background. Microsatellite Instability (MSI), a breakdown in mismatch repair (MMR) gene function, is observed in about 15% of colorectal cancers (CRCs)[(1)]. Lynch syndrome patients (3% of CRCs), are at high risk of a range of cancers, and regular monitoring is key to early cancer identification[(1,2)]. Currently, immunohistochemistry (IHC) and fragment analysis are used to detect MMR gene defects in CRC, but not all tumours are directly tested for MSI due to cost. As a result, many Lynch Syndrome patients go undetected.

A sequence based MSI test may be cheaper and faster than IHC and fragment analysis. We therefore mined CRC genome data (http://cancergenome.nih.gov/) to identify short mononucleotide repeats susceptible to MSI, and assessed both stability and sequencing error in microsatellite stable (MSS) and unstable (MSI-H) tumours and controls. We aim to develop a sequence based MSI test suitable for a point of care device currently in development by QuantuMDx (http://www.quantumdx.com/).

Methods A total of 35 MSI-H, matched normal tissue, and MSS low depth whole genome sequences[(3)] were mined for indels in all monomorphic 7-12bp mononucleotide repeats using BWA, Samtools, GATK and In house Perl scripts. Candidate repeats for inclusion were PCR amplified from samples collected within the Northern Genetics Service and Pathology department, Newcastle Hospitals NHS Foundation Trust*. Libraries were generated using the Illumina Nextera XT library prep and sequenced to an average read depth of 10590 per amplicon.

Results 1. Identifying Mononucleotide Repeats

179517 variable 7-12bp mononucleotide repeats were identified. MSI-H samples had more variant reads than MSS and normal samples (Figure 1), suggesting that PCR based sequence analysis of instability is viable.

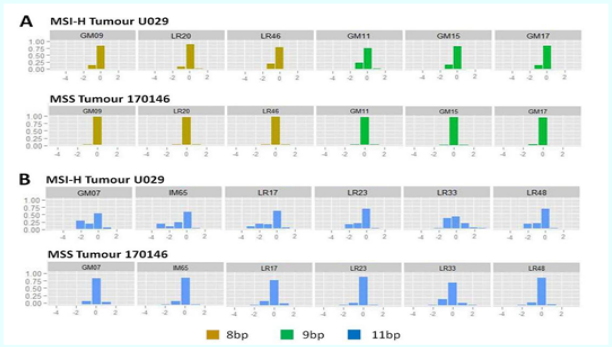


Frequencies of variant reads in all 7bp and 12bp repeats. Variant reads are more abundant in MSI-H samples, and in longer repeats.

Results 2. In depth analysis of Individual Repeats

To validate specific repeats for MSI detection, 119 of the most unstable repeats (7-12bp) were amplified from FFPE tissue and sequenced using the Illumina MiSeq. The FFPE tissues consisted of 5 MSI-H cancers, matching normal mucosa for 4 of the tumours, and 6 MSS cancers.

Of the 119 repeats sequenced, MSI was observed as an increase in deletion frequency or the emergence of new alleles in at least one MSI-H cancer for 66 repeats. 39% of the short repeats (7bp-9bp) showed MSI derived variability, compared to 80% of longer (10bp-11bp) repeats. However longer repeats showed more PCR/Sequencing error derived variability in control tissues (e.g. see panel B below).



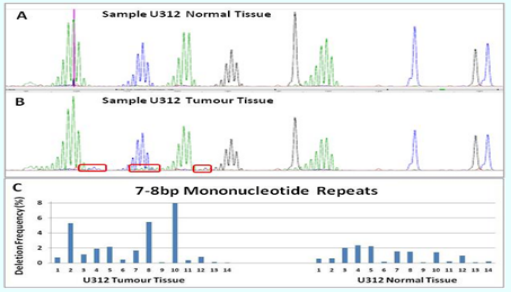
Read length frequency distributions in MSI-H and MSS tumours.

A: Six representative microsatellite unstable short repeats: MSI is manifest as an increase in the frequency of 1bp deletions in the MSI-H tumour.

B: Six representative microsatellite unstable long repeats. MSI is manifest as an increase in variant read frequency, and the appearance of multiple deletion alleles (e.g. -2bp, -3bp) in the MSI-H tumour. 0=wild type allele length.

Results 3. Fragment Analysis versus Sequencing

Despite the existence of variant reads from PCR based error, preliminary analysis (see panel top right) shows that use of multiple short repeats can readily identify MSI-H tumours that exhibit limited instability as assessed by fragment analysis.



A-B: Fragment analysis of a sample where expert interpretation was necessary. **A:** Normal Mucosa **B:** MSI-H Tumour previously classified as unstable. Variant alleles highlighted in red.
C: Results for the same tumour using 14 mononucleotide repeats. 3 out of the 14 repeats (repeats 2, 8, and 10) have a much higher deletion frequency in the MSI-H tumour than the normal tissue.

Conclusions

- Short mononucleotide repeats appear amenable to sequence based MSI testing.
- Short mononucleotide repeats show less sequencing error than longer mononucleotide repeats, but are less variable.
- We are currently assessing 45 of the most variable markers in a larger panel of tumours of known MSI status to optimise the marker panel and classification criteria.

References

(1) Sinicrope and Sargent (2012) *Clinical Cancer Research*, 18, pp 1506-1512
(2) Grover et al (2009) *Best Practice & Research Clinical Gastroenterology*, 23, pp 185-196
(3) Cancer Genome Atlas Network (19 July 2012) *Nature*, 487, pp 330–337

Affiliations

(1) Institute of Genetic Medicine, Newcastle University, International Centre for Life, Newcastle upon Tyne. NE1 3BZ. (2) Pathology department and ,3) Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust (. 4) QuantuMDx group ltd, International Centre for Life, Newcastle upon Tyne, UK. (5) University of Heidelberg, Heidelberg, Germany, DE.

* Under ethical approval REC reference: 13/LO/1514

Figure 8-1: Poster shows the initial selection of the markers. The poster exhibited in the BSGM meeting, 2015.

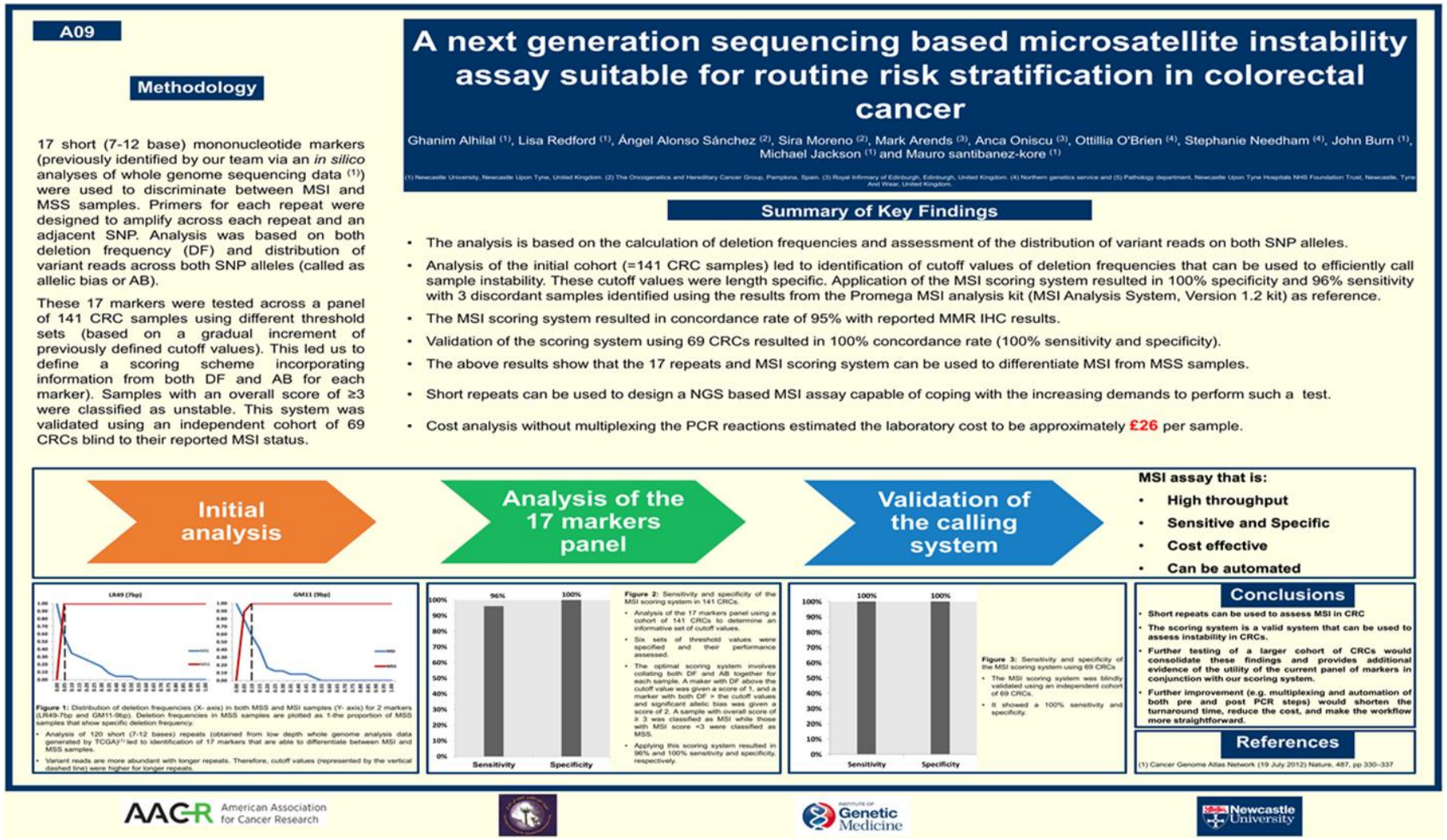


Figure 8- 2: Poster shows the overall workflow The poster exhibited in the AACR meeting, 2016.

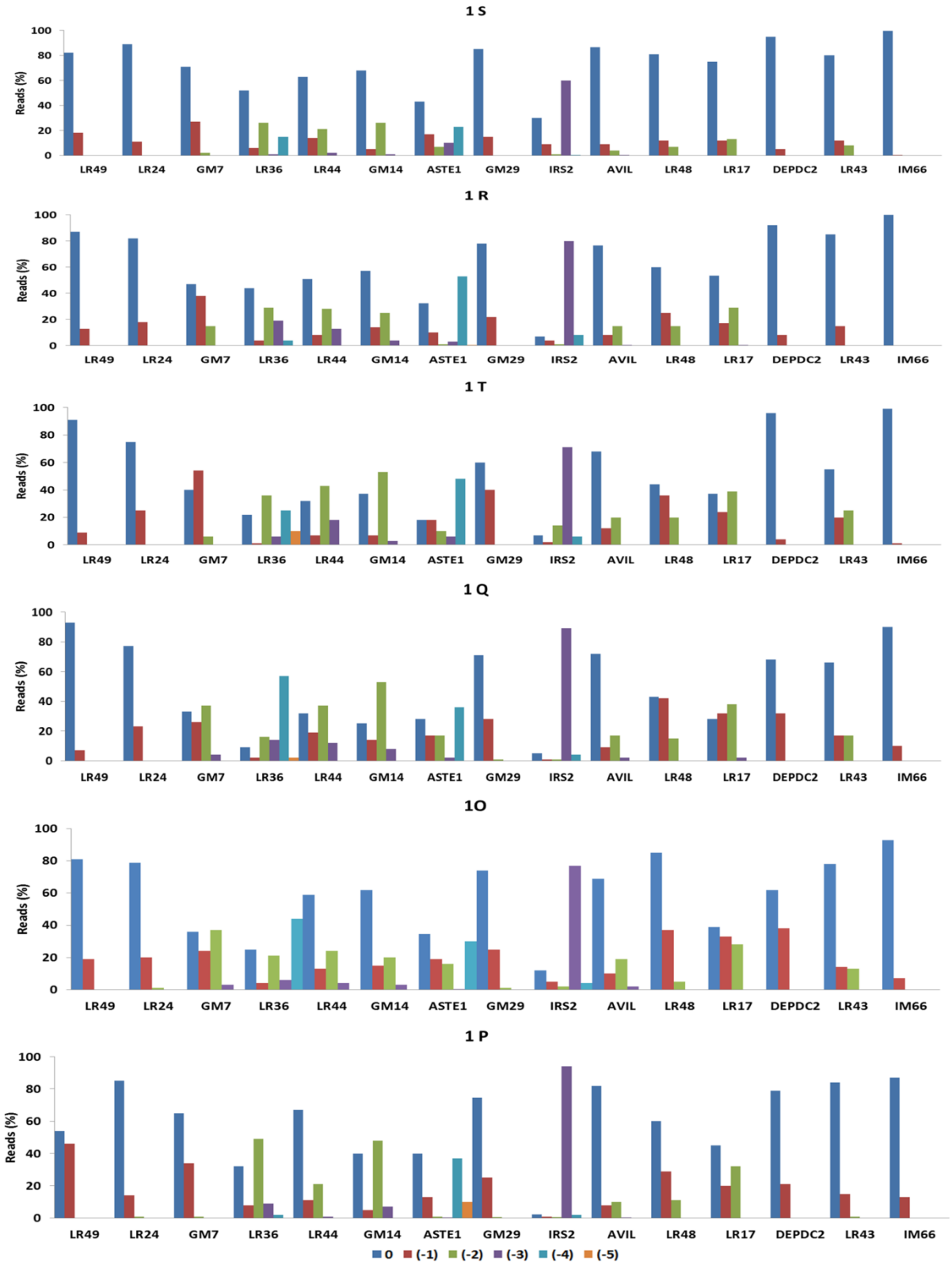


Figure 8-3: Deletion frequencies of a subset of 15 markers in the tumour PR32079/14. 1S, 1R, 1T, 1Q, 1O and 1P are the samples that retrieved from that tumour.

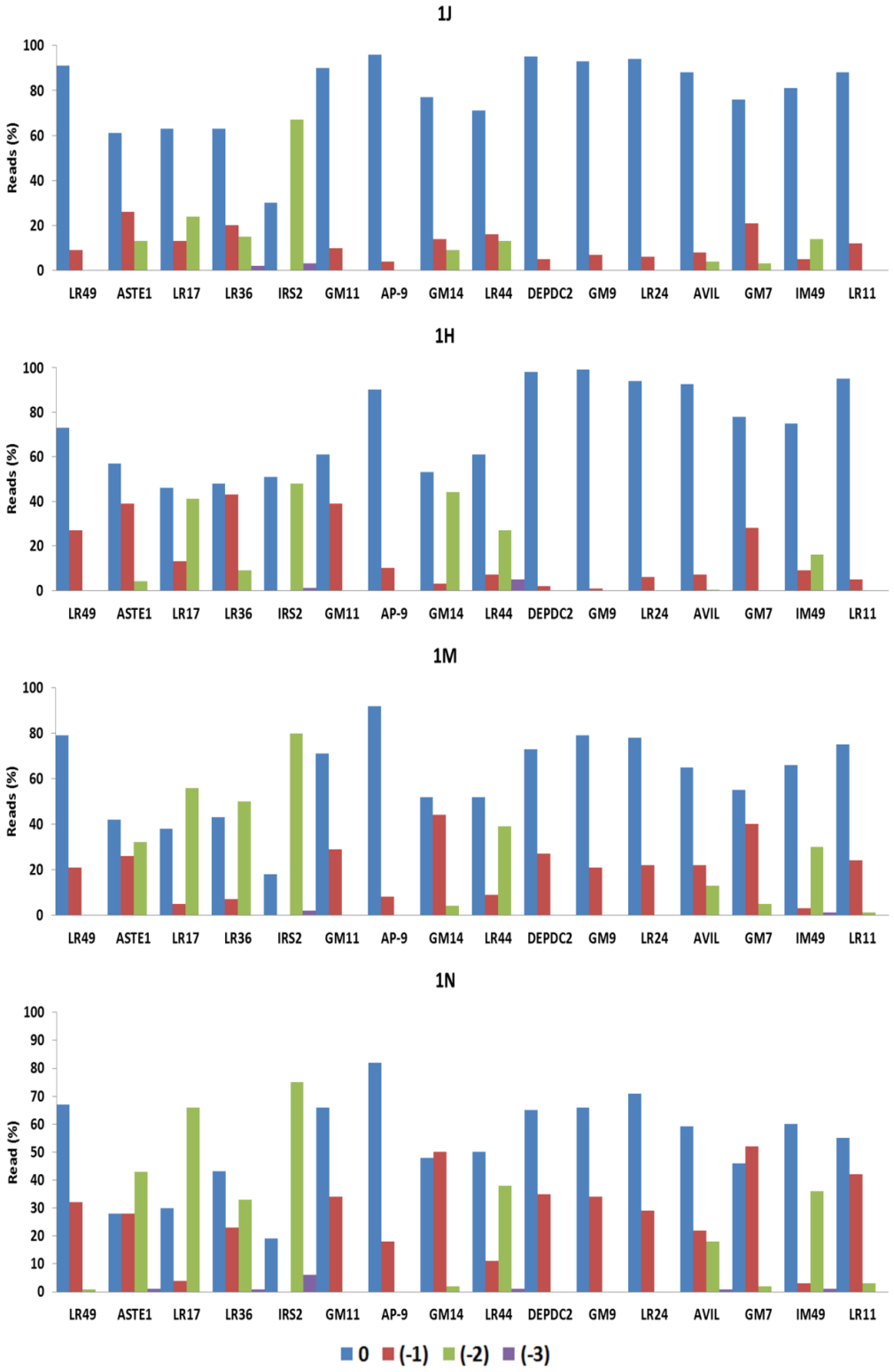


Figure 8-4: Deletion frequencies of a subset of 16 markers in the tumour PR53139/13. 1J, 1H, 1M and 1P are the samples that retrieved from that tumour.

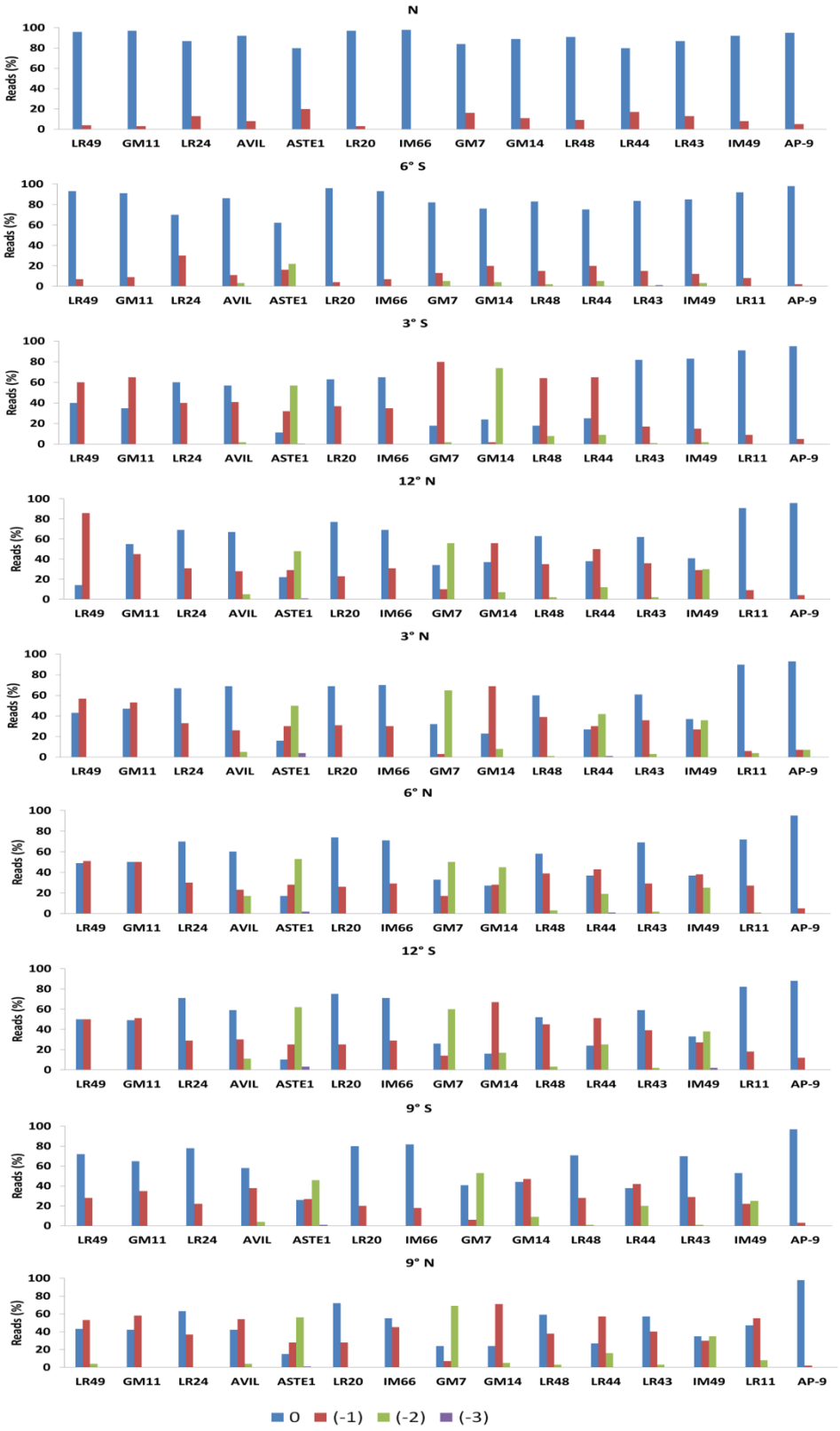


Figure 8-5: Deletion frequencies of a subset of 14 markers in the tumour PR10654/14. (N) refers to normal samples and all other are tumour samples. All markers were stable in the normal sample.

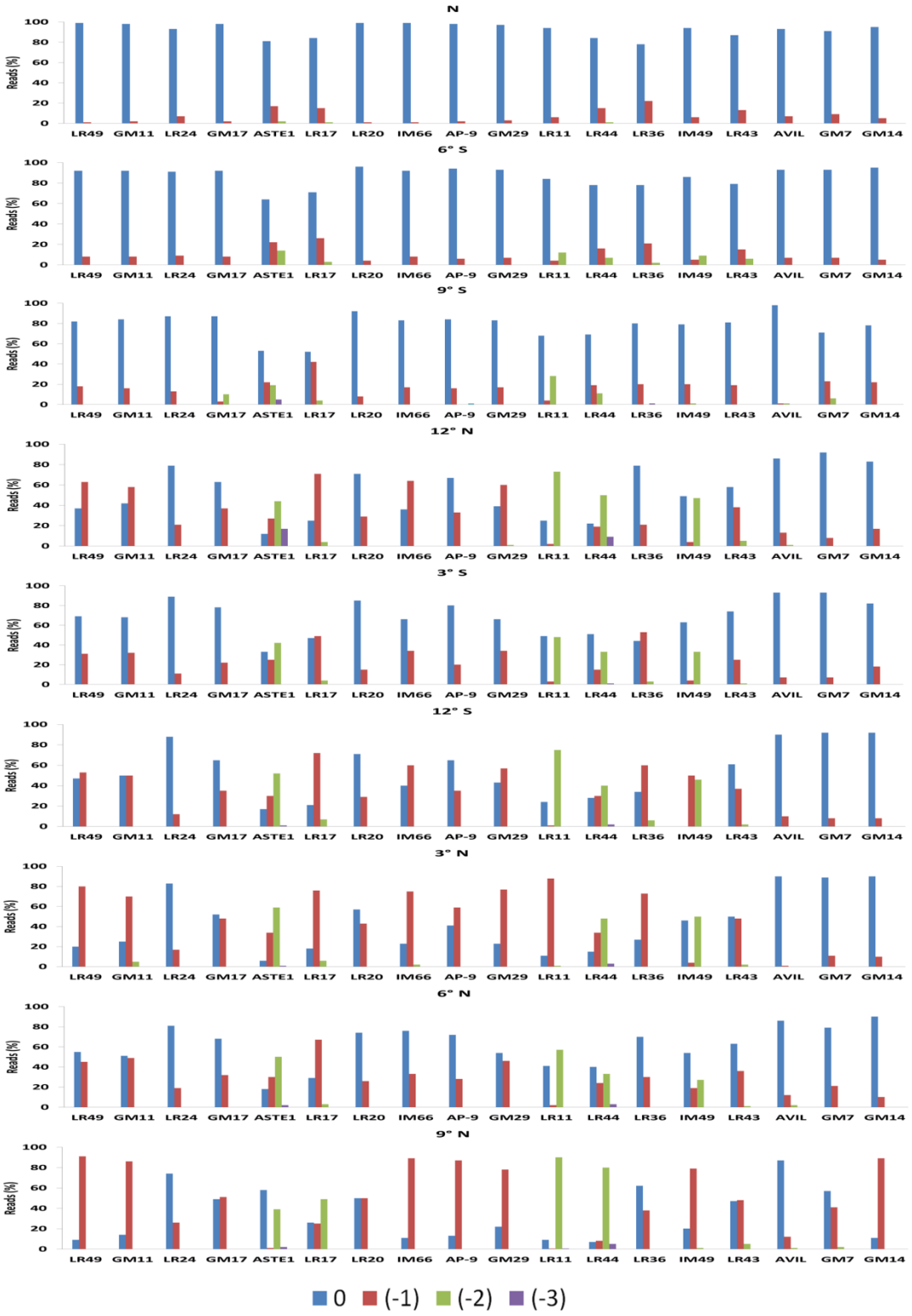


Figure 8-6: Deletion frequencies of a subset of 18 markers in the tumour PR17848/14. (N) refers to normal samples and all other are tumour samples. All markers were stable in the normal sample.

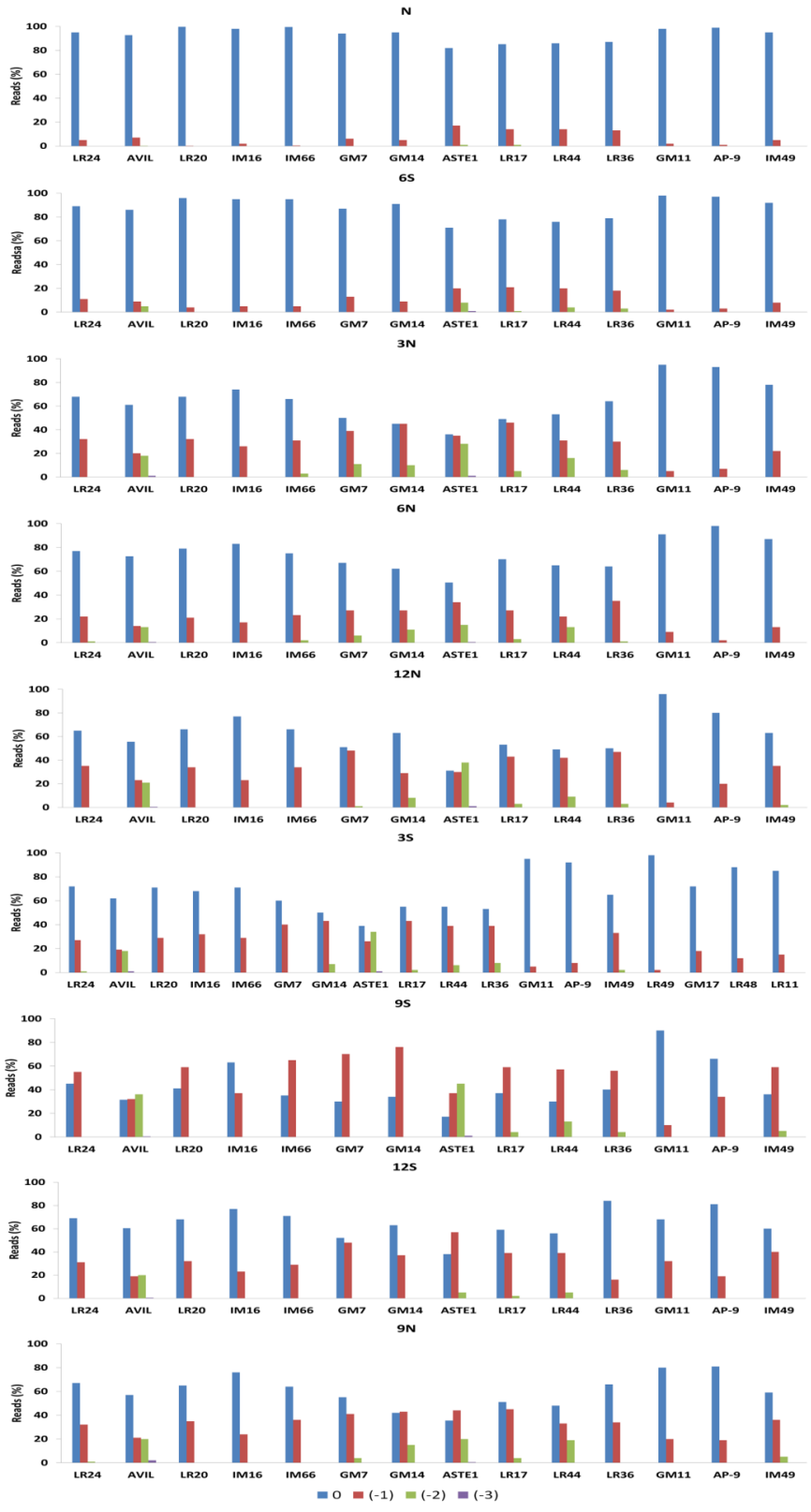


Figure 8-7: Deletion frequencies of a subset of 14 markers in the tumour PR51896/13. (N) refers to normal samples and all other are tumour samples. All markers were stable in the normal sample.

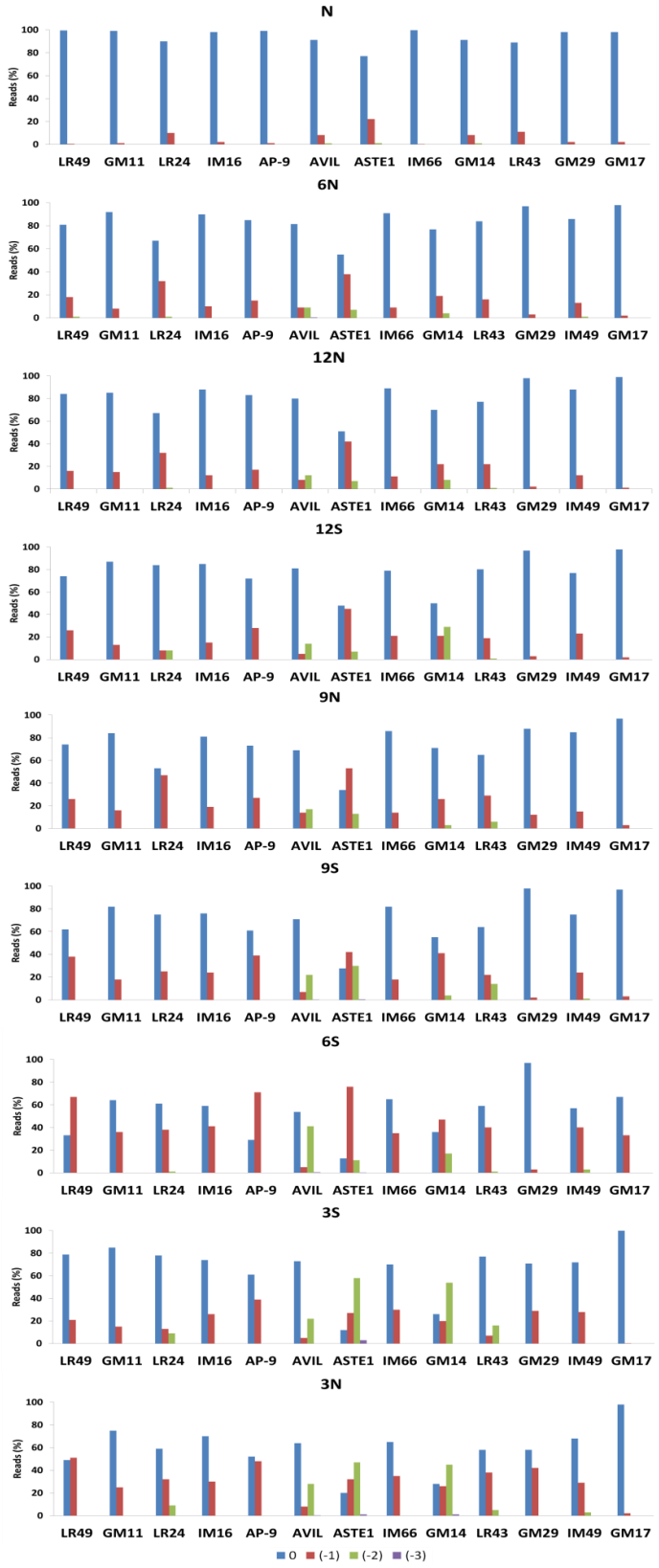


Figure 8-8: Deletion frequencies of a subset of 13 markers in the tumour PR32516/14. (N) refers to normal samples and all other are tumour samples. All markers were stable in the normal sample.

Chapter 9. References

ALHOPURO, P., SAMMALKORPI, H., NIITTYMÄKI, I., BISTRÖM, M., RAITILA, A., SAHARINEN, J., NOUSIAINEN, K., LEHTONEN, H. J., HELIÖVAARA, E. & PUHAKKA, J. 2012. Candidate driver genes in microsatellite-unstable colorectal cancer. *International journal of cancer,* 130**,** 1558-1566.

ARONSON, M., HOLTER, S., SEMOTIUK, K., WINTER, L., POLLETT, A., GALLINGER, S., COHEN, Z. & GRYFE, R. 2015. DNA Mismatch Repair Status Predicts Need for Future Colorectal Surgery for Metachronous Neoplasms in Young Individuals Undergoing Colorectal Cancer Resection. *Diseases of the Colon & Rectum,* 58**,** 645-652.

AXEL COURNAC, R. K., AND JULIEN MOZZICONACCI 2015. The 3D folding of metazoan genomes correlates with the association of similar repetitive elements. *Nucleic Acids Res.,* Nov.

BACHER, J. W., FLANAGAN, L. A., SMALLEY, R. L., NASSIF, N. A., BURGART, L. J., HALBERG, R. B., MEGID, W. M. A. & THIBODEAU, S. N. 2004. Development of a fluorescent multiplex assay for detection of MSI-High tumors. *Disease markers,* 20**,** 237-250.

BAISSE, B., BOUZOURENE, H., SARAGA, E. P., BOSMAN, F. T. & BENHATTAR, J. 2001. Intratumor genetic heterogeneity in advanced human colorectal adenocarcinoma. *International journal of cancer,* 93**,** 346-352.

BAUDHUIN, L. M., FERBER, M. J., WINTERS, J. L., STEENBLOCK, K. J., SWANSON, R. L., FRENCH, A. J., BUTZ, M. L. & THIBODEAU, S. N. 2005. Characterization of hMLH1 and hMSH2 gene dosage alterations in Lynch syndrome patients. *Gastroenterology,* 129**,** 846-854.

BEGGS, A., DOMINGO, E., ABULAFI, M., HODGSON, S. & TOMLINSON, I. 2013. A study of genomic instability in early preneoplastic colonic lesions. *Oncogene,* 32**,** 5333-5337.

BERG, A. O., ARMSTRONG, K., BOTKIN, J., CALONGE, N., HADDOW, J., HAYES, M., KAYE, C., PHILLIPS, K. A., PIPER, M. & RICHARDS, C. S. 2009. Recommendations from the EGAPP working group. *Genetics in Medicine,* 11**,** 35-41.

BHATTACHARYA, N. P., SKANDALIS, A., GANESH, A., GRODEN, J. & MEUTH, M. 1994. Mutator phenotypes in human colorectal carcinoma cell lines. *Proceedings of the National Academy of Sciences,* 91**,** 6319-6323.

BIRGISSON, H., EDLUND, K., WALLIN, U., PÅHLMAN, L., KULTIMA, H. G., MAYRHOFER, M., MICKE, P., ISAKSSON, A., BOTLING, J. & GLIMELIUS, B. 2015. Microsatellite instability and mutations in BRAF and KRAS are significant predictors of disseminated disease in colon cancer. *BMC cancer,* 15**,** 1.

BLAKE, C., TSAO, J.-L., WU, A. & SHIBATA, D. 2001. Stepwise deletions of polyA sequences in mismatch repair-deficient colorectal cancers. *The American journal of pathology,* 158**,** 1867-1870.

BOCKER, T., DIERMANN, J., FRIEDL, W., GEBERT, J., HOLINSKI-FEDER, E., KARNER-HANUSCH, J., VON KNEBEL-DOEBERITZ, M., KOELBLE, K., MOESLEIN, G., SCHACKERT, H.-K., HANS-CHRISTIAN, W., RICHARD, F. & JOSEF, R. 1997. Microsatellite instability analysis: a multicenter study for reliability and quality control. *Cancer Research,* 57**,** 4739-4743.

BOLAND, C. R., THIBODEAU, S. N., HAMILTON, S. R., SIDRANSKY, D., ESHLEMAN, J. R., BURT, R. W., MELTZER, S. J., RODRIGUEZ-BIGAS, M. A., FODDE, R. & RANZANI, G. N. 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**,** 5248-5257.

BOYER, J. C., YAMADA, N. A., ROQUES, C. N., HATCH, S. B., RIESS, K. & FARBER, R. A. 2002. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. *Human molecular genetics,* 11**,** 707-713.

BOYLE, T. A., BRIDGE, J. A., SABATINI, L. M., NOWAK, J. A., VASALOS, P., JENNINGS, L. J. & HALLING, K. C. 2014. 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**,** 363-370.

BRINKMANN, B., KLINTSCHAR, M., NEUHUBER, F., HÜHNE, J. & ROLF, B. 1998. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *The American Journal of Human Genetics,* 62**,** 1408-1415.

CANCER-RESEARCH-UK. 2015. *http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer/incidence* [Online]. [Accessed].

CANCER GENOME ATLAS, N. 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-337.

CARETHERS, J. M., CHAUHAN, D. P., FINK, D., NEBEL, S., BRESALIER, R. S., HOWELL, S. B. & BOLAND, C. R. 1999. Mismatch repair proficiency and in vitro response to 5-fluorouracil. *Gastroenterology,* 117**,** 123-131.

CARREIRA, S., ROMANEL, A., GOODALL, J., GRIST, E., FERRALDESCHI, R., MIRANDA, S., PRANDI, D., LORENTE, D., FRENEL, J.-S. & PEZARO, C. 2014. Tumor clone dynamics in lethal prostate cancer. *Science translational medicine,* 6**,** 254ra125-254ra125.

CHAKRABORTY, R., KIMMEL, M., STIVERS, D. N., DAVISON, L. J. & DEKA, R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences,* 94**,** 1041-1046.

CHERUKURI, D. P., DEIGNAN, J. L., DAS, K., GRODY, W. W. & HERSCHMAN, H. 2015. Instability of a dinucleotide repeat in the 3′-untranslated region (UTR) of the microsomal prostaglandin E synthase-1 (mPGES-1) gene in microsatellite instability-high (MSI-H) colorectal carcinoma. *Molecular oncology*.

CICEK, M. S., LINDOR, N. M., GALLINGER, S., BAPAT, B., HOPPER, J. L., JENKINS, M. A., YOUNG, J., BUCHANAN, D., WALSH, M. D. & LE MARCHAND, L. 2011. Quality assessment and correlation of microsatellite instability and immunohistochemical markers among population-and clinic-based colorectal tumors: results from the Colon Cancer Family Registry. *The Journal of Molecular Diagnostics,* 13**,** 271-281.

CLARKE, L. A., REBELO, C. S., GONCALVES, J., BOAVIDA, M. G. & JORDAN, P. 2001. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Molecular Pathology,* 54**,** 351-353.

CONTENTE, A., DITTMER, A., KOCH, M. C., ROTH, J. & DOBBELSTEIN, M. 2002. A polymorphic microsatellite that mediates induction of PIG3 by p53. *Nature genetics,* 30**,** 315-320.

DE BRUIN, E. C., TAYLOR, T. B. & SWANTON, C. 2013. Intra-tumor heterogeneity: lessons from microbial evolution and clinical implications. *Genome medicine,* 5**,** 1.

DE CUBA, E. M. V., SNAEBJORNSSON, P., HEIDEMAN, D. A. M., VAN GRIEKEN, N. C. T., BOSCH, L. J. W., FIJNEMAN, R. J. A., BELT, E., BRIL, H., STOCKMANN, H. & HOOIJBERG, E. 2015. Prognostic value of BRAF and KRAS mutation status in stage II and III microsatellite instable colon cancers. *International Journal of Cancer*.

DE GRASSI, A., IANNELLI, F., CEREDA, M., VOLORIO, S., MELOCCHI, V., VIEL, A., BASSO, G., LAGHI, L., CASELLE, M. & CICCARELLI, F. D. 2014. Deep sequencing of the X chromosome reveals the proliferation history of colorectal adenomas. *Genome biology,* 15**,** 1.

DE LA CHAPELLE, A. 2004. Genetic predisposition to colorectal cancer. *Nature Reviews Cancer,* 4**,** 769-780.

DE LA CHAPELLE, A. & HAMPEL, H. 2010. Clinical relevance of microsatellite instability in colorectal cancer. *Journal of Clinical Oncology,* 28**,** 3380-3387.

DE ROSA, N., RODRIGUEZ-BIGAS, M. A., CHANG, G. J., VEERAPONG, J., BORRAS, E., KRISHNAN, S., BEDNARSKI, B., MESSICK, C. A., SKIBBER, J. M. & FEIG, B. W. 2016. DNA Mismatch Repair Deficiency in Rectal Cancer: Benchmarking Its Impact on Prognosis, Neoadjuvant Response Prediction, and Clinical Cancer Genetics. *Journal of Clinical Oncology***,** JCO666826.

DEANS, Z., , C. M. W., , R. C., , S., ELLARD, , Y. W., , C. M. & , A. S. A. 2015. Practice guidelines for Targeted Next Generation

Sequencing Analysis and Interpretation. *Association for clinical genetics sciences*.

DES GUETZ, G., SCHISCHMANOFF, O., NICOLAS, P., PERRET, G.-Y., MORERE, J.-F. & UZZAN, B. 2009. Does microsatellite instability predict the efficacy of adjuvant chemotherapy in colorectal cancer? A systematic review with meta-analysis. *European journal of cancer,* 45**,** 1890-1896.

EDGE, S. B. & COMPTON, C. C. 2010. The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM. *Annals of surgical oncology,* 17**,** 1471-1474.

ELLEGREN, H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nature genetics,* 24**,** 400-402.

ELLEGREN, H. 2004. Microsatellites: simple sequences with complex evolution. *Nature reviews genetics,* 5**,** 435-445.

ELSAYED, F. A., KETS, C. M., RUANO, D., VAN DEN AKKER, B., MENSENKAMP, A. R., SCHRUMPF, M., NIELSEN, M., WIJNEN, J. T., TOPS, C. M. & LIGTENBERG, M. J. 2014. Germline variants in POLE are associated with early onset mismatch repair deficient colorectal cancer. *European Journal of Human Genetics*.

ESCHER, D., BODMER-GLAVAS, M., BARBERIS, A. & SCHAFFNER, W. 2000. Conservation of glutamine-rich transactivation function between yeast and humans. *Molecular and cellular biology,* 20**,** 2774-2782.

FAN, H. & CHU, J.-Y. 2007. A brief review of short tandem repeat mutation. *Genomics, Proteomics & Bioinformatics,* 5**,** 7-14.

FAZEKAS, A. J., STEEVES, R. & NEWMASTER, S. G. 2010. Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques,* 48**,** 277-285.

FISHER, K. E., ZHANG, L., WANG, J., SMITH, G. H., NEWMAN, S., SCHNEIDER, T. M., PILLAI, R. N., KUDCHADKAR, R. R., OWONIKOKO, T. K. & RAMALINGAM, S. S. 2016. Clinical Validation and Implementation of a Targeted Next-Generation Sequencing Assay to Detect Somatic Variants in Non-Small Cell Lung, Melanoma, and Gastrointestinal Malignancies. *The Journal of Molecular Diagnostics,* 18**,** 299-315.

FUTREAL, P. A., COIN, L., MARSHALL, M., DOWN, T., HUBBARD, T., WOOSTER, R., RAHMAN, N. & STRATTON, M. R. 2004. A census of human cancer genes. *Nature Reviews Cancer,* 4**,** 177-183.

GAN, C., LOVE, C., BESHAY, V., MACRAE, F., FOX, S., WARING, P. & TAYLOR, G. 2015. Applicability of next generation sequencing technology in microsatellite instability testing. *Genes,* 6**,** 46-59.

GAZZOLI, I., LODA, M., GARBER, J., SYNGAL, S. & KOLODNER, R. D. 2002. A hereditary nonpolyposis colorectal carcinoma case associated with hypermethylation of the MLH1 gene in normal tissue and loss of heterozygosity of the unmethylated allele in the resulting microsatellite instability-high tumor. *Cancer Research,* 62**,** 3925-3928.

GEBHARDT, F., ZÄNKER, K. S. & BRANDT, B. 1999. Modulation of epidermal growth factor receptor gene transcription by a polymorphic dinucleotide repeat in intron 1. *Journal of Biological Chemistry,* 274**,** 13176-13180.

GENSCHEL, J., BAZEMORE, L. R. & MODRICH, P. 2002. Human exonuclease I is required for 5′ and 3′ mismatch repair. *Journal of Biological Chemistry,* 277**,** 13302-13311.

GLENN, T. C., STEPHAN, W., DESSAUER, H. C. & BRAUN, M. J. 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Molecular Biology and Evolution,* 13**,** 1151-1154.

GOEL, A., NAGASAKA, T., HAMELIN, R. & BOLAND, C. R. 2010. An optimized pentaplex PCR for detecting DNA mismatch repair-deficient colorectal cancers. *Plos one,* 5**,** e9393.

GUARINOS, C., CASTILLEJO, A., BARBERÁ, V.-M., PÉREZ-CARBONELL, L., SÁNCHEZ-HERAS, A.-B., SEGURA, Á., GUILLÉN-PONCE, C., MARTÍNEZ-CANTÓ, A., CASTILLEJO, M.-I. & EGOAVIL, C.-M. 2010. EPCAM germ line deletions as causes of Lynch syndrome in Spanish patients. *The Journal of Molecular Diagnostics,* 12**,** 765-770.

GUO, Y., ZHOU, J., HUANG, A., LI, J., YAN, M., ZHU, Z., ZHAO, X., GU, J., LIU, B. & SHAO, Z. 2015. Spatially defined microsatellite analysis reveals extensive genetic mosaicism and clonal complexity in intestinal metaplastic glands. *International Journal of Cancer,* 136**,** 2973-2979.

HANAHAN, D. & WEINBERG, R. A. 2011. Hallmarks of cancer: the next generation. *cell,* 144**,** 646-674.

HARALDSDOTTIR, S., HAMPEL, H., TOMSIC, J., FRANKEL, W. L., PEARLMAN, R., DE LA CHAPELLE, A. & PRITCHARD, C. C. 2014. Colon and endometrial cancers with mismatch repair deficiency can arise from somatic, rather than germline, mutations. *Gastroenterology,* 147**,** 1308-1316. e1.

HARDIMAN, K. M., ULINTZ, P. J., KUICK, R. D., HOVELSON, D. H., GATES, C. M., BHASI, A., GRANT, A. R., LIU, J., CANI, A. K. & GREENSON, J. K. 2016. Intra-tumor genetic heterogeneity in rectal cancer. *Laboratory Investigation,* 96**,** 4-15.

HEALE, S. M. & PETES, T. D. 1995. The stabilization of repetitive tracts of DNA by variant repeats requires a functional DNA mismatch repair system. *Cell,* 83**,** 539-545.

HEGDE, M., FERBER, M., MAO, R., SAMOWITZ, W. & GANGULY, A. 2013. ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genetics in Medicine,* 16**,** 101-116.

HEGDE, M., FERBER, M., MAO, R., SAMOWITZ, W., GANGULY, A. & WORKING GROUP OF THE AMERICAN COLLEGE OF MEDICAL, G. 2014. Genomics (ACMG) Laboratory Quality Assurance Committee. ACMG technical standards and guidelines for genetic testing for inherited colorectal cancer (Lynch syndrome, familial adenomatous polyposis, and MYH-associated polyposis). *Genet Med,* 16**,** 101-116.

HEMPELMANN, J. A., SCROGGINS, S. M., PRITCHARD, C. C. & SALIPANTE, S. J. 2015. MSIplus for Integrated Colorectal Cancer Molecular Testing by Next-Generation Sequencing. *The Journal of Molecular Diagnostics,* 17**,** 705-714.

HIATT, J. B., PRITCHARD, C. C., SALIPANTE, S. J., O'ROAK, B. J. & SHENDURE, J. 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome research,* 23**,** 843-854.

HOFFMEISTER, M., BLÄKER, H., KLOOR, M., ROTH, W., TOTH, C., HERPEL, E., FRANK, B., SCHIRMACHER, P., CHANG-CLAUDE, J. & BRENNER, H. 2013.

Body mass index and microsatellite instability in colorectal cancer: a population-based study. *Cancer Epidemiology Biomarkers & Prevention,* 22**,** 2303-2311.

HU, P., LEE, C. W., XU, J. P., SIMIEN, C., FAN, C. L., TAM, M., RAMAGLI, L., BROWN, B. W., LYNCH, P. & FRAZIER, M. L. 2011. Microsatellite Instability in Saliva from Patients with Hereditary Non-polyposis Colon Cancer and Siblings Carrying Germline Mismatch Repair Gene Mutations. *Annals of Clinical & Laboratory Science,* 41**,** 321-330.

IONOV, Y., PEINADO, M. A., MALKHOSYAN, S., SHIBATA, D. & PERUCHO, M. 1993. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature,* 363**,** 558-561.

JASCUR, T. & BOLAND, C. R. 2006. Structure and function of the components of the human DNA mismatch repair system. *International journal of cancer,* 119**,** 2030-2035.

JULIÉ, C., TRÉSALLET, C., BROUQUET, A., VALLOT, C., ZIMMERMANN, U., MITRY, E., RADVANYI, F., ROULEAU, E., LIDEREAU, R. & COULET, F. 2008. Identification in daily practice of patients with Lynch syndrome (hereditary nonpolyposis colorectal cancer): revised Bethesda guidelines-based approach versus molecular screening. *The American journal of gastroenterology,* 103**,** 2825-2835.

KANE, M. F., LODA, M., GAIDA, G. M., LIPMAN, J., MISHRA, R., GOLDMAN, H., JESSUP, J. M. & KOLODNER, R. 1997. Methylation of the hMLH1 promoter correlates with lack of expression of hMLH1 in sporadic colon tumors and mismatch repair-defective human tumor cell lines. *Cancer research,* 57**,** 808-811.

KANG, J., LEE, H. W., KIM, I.-K., KIM, N. K., SOHN, S.-K. & LEE, K. Y. 2015. Clinical Implications of Microsatellite Instability in T1 Colorectal Cancer. *Yonsei medical journal,* 56**,** 175-181.

KENT, W. J., SUGNET, C. W., FUREY, T. S., ROSKIN, K. M., PRINGLE, T. H., ZAHLER, A. M. & HAUSSLER, D. 2002. The human genome browser at UCSC. *Genome research,* 12**,** 996-1006.

KIM, H., JEN, J., VOGELSTEIN, B. & HAMILTON, S. R. 1994. Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences. *The American journal of pathology,* 145**,** 148.

KLINGBIEL, D., SARIDAKI, Z., ROTH, A. D., BOSMAN, F. T., DELORENZI, M. & TEJPAR, S. 2015. Prognosis of stage II and III colon cancer treated with adjuvant 5-fluorouracil or FOLFIRI in relation to microsatellite status: results of the PETACC-3 trial. *Annals of Oncology,* 26**,** 126-132.

KLOOR, M., REUSCHENBACH, M., KARBACH, J., RAFIYAN, M., AL-BATRAN, S.-E., PAULIGK, C., JAEGER, E. & VON KNEBEL DOEBERITZ, M. Vaccination of MSI-H colorectal cancer patients with frameshift peptide antigens: A phase I/IIa clinical trial. ASCO Annual Meeting Proceedings, 2015. 3020.

KLUMP, B., NEHLS, O., OKECH, T., HSIEH, C. J., GACO, V., GITTINGER, F. S., SARBIA, M., BORCHARD, F., GRESCHNIOK, A. & GRUENAGEL, H. H. 2004. Molecular lesions in colorectal cancer: impact on prognosis? *International journal of colorectal disease,* 19**,** 23-42.

KOMAROVA, N. L. & WODARZ, D. 2005. Drug resistance in cancer: principles of emergence and prevention. *Proceedings of the National Academy of Sciences of the United States of America,* 102**,** 9714-9719.

KORNBERG, A., BERTSCH, L. L., JACKSON, J. F. & 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 of the United States of America,* 51**,** 315.

KOVACS, M. E., PAPP, J., SZENTIRMAY, Z., OTTO, S. & OLAH, E. 2009. Deletions removing the last exon of TACSTD1 constitute a distinct class of mutations predisposing to Lynch syndrome. *Human mutation,* 30**,** 197-203.

KUISMANEN, S. A., MOISIO, A.-L., SCHWEIZER, P., TRUNINGER, K., SALOVAARA, R., AROLA, J., BUTZOW, R., JIRICNY, J., NYSTRÖM-LAHTI, M. & PELTOMÄKI, P. 2002. Endometrial and colorectal tumors from patients with hereditary nonpolyposis colon cancer display different patterns of microsatellite instability. *The American journal of pathology,* 160**,** 1953-1958.

KUNKEL, T. A. 1990. Misalignment-mediated DNA synthesis errors. *Biochemistry,* 29**,** 8003-8011.

LAI, Y. & SUN, F. 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. *Molecular biology and evolution,* 20**,** 2123-2131.

LAIHO, P., LAUNONEN, V., LAHERMO, P., ESTELLER, M., GUO, M., HERMAN, J. G., MECKLIN, J.-P., JÄRVINEN, H., SISTONEN, P. & KIM, K.-M. 2002. Low-level microsatellite instability in most colorectal carcinomas. *Cancer research,* 62**,** 1166-1170.

LE, D. T., URAM, J. N., WANG, H., BARTLETT, B. R., KEMBERLING, H., EYRING, A. D., SKORA, A. D., LUBER, B. S., AZAD, N. S. & LAHERU, D. 2015. PD-1 blockade in tumors with mismatch-repair deficiency. *New England Journal of Medicine,* 372**,** 2509-2520.

LENGAUER, C., KINZLER, K. W. & VOGELSTEIN, B. 1998. Genetic instabilities in human cancers. *Nature,* 396**,** 643-649.

LI, D., HU, F., WANG, F., CUI, B., DONG, X., ZHANG, W., LIN, C., LI, X., WANG, D. & ZHAO, Y. 2013a. Prevalence of pathological germline mutations of hMLH1 and hMSH2 genes in colorectal cancer. *PloS one,* 8**,** e51240.

LI, X., YAO, X., WANG, Y., HU, F., WANG, F., JIANG, L., LIU, Y., WANG, D., SUN, G. & ZHAO, Y. 2013b. MLH1 promoter methylation frequency in colorectal cancer patients and related clinicopathological and molecular features. *PLoS One,* 8**,** e59064.

LIGTENBERG, M. J. L., KUIPER, R. P., CHAN, T. L., GOOSSENS, M., HEBEDA, K. M., VOORENDT, M., LEE, T. Y. H., BODMER, D., HOENSELAAR, E. & HENDRIKS-CORNELISSEN, S. J. B. 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**,** 112-117.

LIN, M.-T., MOSIER, S. L., THIESS, M., BEIERL, K. F., DEBELJAK, M., TSENG, L.-H., CHEN, G., YEGNASUBRAMANIAN, S., HO, H. & COPE, L. 2014. Clinical validation of KRAS, BRAF, and EGFR mutation detection using next-generation sequencing. *American journal of clinical pathology,* 141**,** 856-866.

LINDBLOM, A., TANNERGÅRD, P., WERELIUS, B. & NORDENSKJÖLD, M. 1993. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature genetics,* 5**,** 279-282.

LINNEBACHER, M., GEBERT, J., RUDY, W., WOERNER, S., YUAN, Y. P., BORK, P. & VON KNEBEL DOEBERITZ, M. 2001. Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens. *International journal of cancer,* 93**,** 6-11.

LINNEKAMP, J. F., WANG, X., MEDEMA, J. P. & VERMEULEN, L. 2015. Colorectal cancer heterogeneity and targeted therapy: a case for molecular disease subtypes. *Cancer research,* 75**,** 245-249.

LLOSA, N. J., CRUISE, M., TAM, A., WICKS, E. C., HECHENBLEIKNER, E. M., TAUBE, J. M., BLOSSER, R. L., FAN, H., WANG, H. & LUBER, B. S. 2015. The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints. *Cancer discovery,* 5**,** 43-51.

LOEB, L. A. 2011. Human cancers express mutator phenotypes: origin, consequences and targeting. *Nature Reviews Cancer,* 11**,** 450-457.

LOUGHREY, M. B., QUIRKE, P. & SHEPHERD, N. A. 2014. Dataset for colorectal cancer histopathology reports. *Royal College of Pathologist.* *https://www.rcpath.org/resourceLibrary/dataset-for-colorectal-cancer-histopathology-reports--3rd-edition-.html*

LOUKOLA, A., SALOVAARA, R., KRISTO, P., MOISIO, A.-L., KÄÄRIÄINEN, H., AHTOLA, H., ESKELINEN, M., HÄRKÖNEN, N., JULKUNEN, R. & KANGAS, E. 1999. Microsatellite instability in adenomas as a marker for hereditary nonpolyposis colorectal cancer. *The American journal of pathology,* 155**,** 1849-1853.

LU, Y., SOONG, T. D. & ELEMENTO, O. 2013. A novel approach for characterizing microsatellite instability in cancer cells. *PloS one,* 8**,** e63056.

LYNCH, H. T. & DE LA CHAPELLE, A. 2003. Hereditary colorectal cancer. *New England Journal of Medicine,* 348**,** 919-932.

LYNCH, H. T., LYNCH, P. M., LANSPA, S. J., SNYDER, C. L., LYNCH, J. F. & BOLAND, C. R. 2009. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical genetics,* 76**,** 1-18.

LYON, M. F. 1961. Gene action in the X-chromosome of the mouse (Mus musculus L.).

MABY, P., TOUGERON, D., HAMIEH, M., MLECNIK, B., KORA, H., BINDEA, G., ANGELL, H. K., FREDRIKSEN, T., ELIE, N. & FAUQUEMBERGUE, E. 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**,** 3446-3455.

MAO, L., LEE, D. J., TOCKMAN, M. S., EROZAN, Y. S., ASKIN, F. & SIDRANSKY, D. 1994. Microsatellite alterations as clonal markers for the detection of human cancer. *Proceedings of the National Academy of Sciences,* 91**,** 9871-9875.

MARTIN, A. & SCHARFF, M. D. 2002. AID and mismatch repair in antibody diversification. *Nature Reviews Immunology,* 2**,** 605-614.

MARTIN, S. A., MCCARTHY, A., BARBER, L. J., BURGESS, D. J., PARRY, S., LORD, C. J. & ASHWORTH, A. 2009. Methotrexate induces oxidative DNA damage and is selectively lethal to tumour cells with defects in the DNA mismatch repair gene MSH2. *EMBO molecular medicine,* 1**,** 323-337.

MATTOCKS, C. J., MORRIS, M. A., MATTHIJS, G., SWINNEN, E., CORVELEYN, A., DEQUEKER, E., MÜLLER, C. R., PRATT, V. & WALLACE, A. 2010. A standardized framework for the validation and verification of clinical molecular genetic tests. *European Journal of Human Genetics,* 18**,** 1276-1288.

MAYRHOFER, M., KULTIMA, H. G., BIRGISSON, H., SUNDSTRÖM, M., MATHOT, L., EDLUND, K., VIKLUND, B., SJÖBLOM, T., BOTLING, J. & MICKE, P. 2014. 1p36 deletion is a marker for tumour dissemination in microsatellite stable stage II-III colon cancer. *BMC cancer,* 14**,** 1.

MCFARLAND, C. D., KOROLEV, K. S., KRYUKOV, G. V., SUNYAEV, S. R. & MIRNY, L. A. 2013. Impact of deleterious passenger mutations on cancer

progression. *Proceedings of the National Academy of Sciences,* 110**,** 2910-2915.

MEDSCAPE 2015. New Guidelines on Colorectal Cancer Molecular Testing.

MELONI, R., ALBANÈSE, V., RAVASSARD, P., TREILHOU, F. & MALLET, J. 1998. A tetranucleotide polymorphic microsatellite, located in the first intron of the tyrosine hydroxylase gene, acts as a transcription regulatory element in vitro. *Human Molecular Genetics,* 7**,** 423-428.

MELONI, R., LAURENT, C., CAMPION, D., BEN, H. B., THIBAUT, F., DOLLFUS, S., PETIT, M., SAMOLYK, D., MARTINEZ, M. & POIRIER, M.-F. 1995. A rare allele of a microsatellite located in the tyrosine hydroxylase gene found in schizophrenic patients. *Comptes rendus de l'Academie des sciences. Serie III, Sciences de la vie,* 318**,** 803-809.

MEROK, M. A., AHLQUIST, T., RØYRVIK, E. C., TUFTELAND, K. F., HEKTOEN, M., SJO, O. H., MALA, T., SVINDLAND, A., LOTHE, R. A. & NESBAKKEN, A. 2013. Microsatellite instability has a positive prognostic impact on stage II colorectal cancer after complete resection: results from a large, consecutive Norwegian series. *Annals of oncology,* 24**,** 1274-1282.

MESSICK, C. A., KRAVOCHUCK, S., CHURCH, J. M. & KALADY, M. F. 2014. Metachronous serrated neoplasia is uncommon after right colectomy in patients with methylator colon cancers with a high degree of microsatellite instability. *Diseases of the Colon & Rectum,* 57**,** 39-46.

MESSIER, W., LI, S.-H. & STEWART, C.-B. 1996. The birth of microsatellites. *Nature,* 381**,** 483.

MICHOR, F. & POLYAK, K. 2010. The origins and implications of intratumor heterogeneity. *Cancer prevention research,* 3**,** 1361-1364.

MINOCHE, A. E., DOHM, J. C. & HIMMELBAUER, H. 2011. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome analyzer systems. *Genome biology,* 12**,** 1.

MOSLEIN, G., TESTER, D. J., LINDOR, N. M., HONCHEL, R., CUNNINGHAM, J. M., FRENCH, A. J., HALLING, K. C., SCHWAB, M., GORETZKI, P. & THIBODEAU, S. N. 1996. Microsatellite instability and mutation analysis of

hMSH2 and hMLH1 in patients with sporadic, familial and hereditary colorectal cancer. *Human molecular genetics,* 5**,** 1245-1252.

MUELLER, J., GAZZOLI, I., BANDIPALLIAM, P., GARBER, J. E., SYNGAL, S. & KOLODNER, R. D. 2009. Comprehensive molecular analysis of mismatch repair gene defects in suspected Lynch syndrome (hereditary nonpolyposis colorectal cancer) cases. *Cancer research,* 69**,** 7053-7061.

NACHMAN, M. W. & CROWELL, S. L. 2000. Estimate of the mutation rate per nucleotide in humans. *Genetics,* 156**,** 297-304.

NADIR, E., MARGALIT, H., GALLILY, T. & BEN-SASSON, S. A. 1996. Microsatellite spreading in the human genome: evolutionary mechanisms and structural implications. *Proceedings of the National Academy of Sciences,* 93**,** 6470-6475.

NAGEL, S., BORISCH, B., THEIN, S. L., OESTREICHER, M., NÖTHIGER, F., BIRRER, S., TOBLER, A. & FEY, M. F. 1995. Somatic mutations detected by mini-and microsatellite DNA markers reveal clonal intratumor heterogeneity in gastrointestinal cancers. *Cancer research,* 55**,** 2866-2870.

NAKAGAWA, H., LOCKMAN, J. C., FRANKEL, W. L., HAMPEL, H., STEENBLOCK, K., BURGART, L. J., THIBODEAU, S. N. & DE LA CHAPELLE, A. 2004. Mismatch repair gene PMS2 disease-causing germline mutations are frequent in patients whose tumors stain negative for PMS2 protein, but paralogous genes obscure mutation detection and interpretation. *Cancer Research,* 64**,** 4721-4727.

NAXEROVA, K., BRACHTEL, E., SALK, J. J., SEESE, A. M., POWER, K., ABBASI, B., SNUDERL, M., CHIANG, S., KASIF, S. & JAIN, R. K. 2014. Hypermutable DNA chronicles the evolution of human colon cancer. *Proceedings of the National Academy of Sciences,* 111**,** E1889-E1898.

NGUYEN, T. T. M., LAKHAN, S. E. & FINETTE, B. A. 2013. Development of a cost-effective high-throughput process of microsatellite analysis involving miniaturized multiplexed PCR amplification and automated allele identification. *Human genomics,* 7**,** 1.

NIESSEN, R. C., HOFSTRA, R. M., WESTERS, H., LIGTENBERG, M. J., KOOI, K., JAGER, P. O., DE GROOTE, M. L., DIJKHUIZEN, T., OLDERODE-BERENDS, M. J. & HOLLEMA, H. 2009. Germline hypermethylation of MLH1 and EPCAM deletions are a frequent cause of Lynch syndrome. *Genes, Chromosomes and Cancer,* 48**,** 737-744.

NIK-ZAINAL, S., VAN LOO, P., WEDGE, D. C., ALEXANDROV, L. B., GREENMAN, C. D., LAU, K. W., RAINE, K., JONES, D., MARSHALL, J. & RAMAKRISHNA, M. 2012. The life history of 21 breast cancers. *Cell,* 149**,** 994-1007.

NILBERT, M., PLANCK, M., FERNEBRO, E., BORG, Å. & JOHNSON, A. 1999. Microsatellite instability is rare in rectal carcinomas and signifies hereditary cancer. *European Journal of Cancer,* 35**,** 942-945.

NOWELL, P. C. 1976. The clonal evolution of tumor cell populations. *Science,* 194**,** 23-28.

PAGIN, A., ZERIMECH, F., LECLERC, J., WACRENIER, A., LEJEUNE, S., DESCARPENTRIES, C., ESCANDE, F., PORCHET, N. & BUISINE, M. P. 2013. Evaluation of a new panel of six mononucleotide repeat markers for the detection of DNA mismatch repair-deficient tumours. *British journal of cancer,* 108**,** 2079-2087.

PARSONS, R., MYEROFF, L. L., LIU, B., WILLSON, J. K., MARKOWITZ, S. D., KINZLER, K. W. & VOGELSTEIN, B. 1995. Microsatellite instability and mutations of the transforming growth factor β type II receptor gene in colorectal cancer. *Cancer Research,* 55**,** 5548-5550.

PELTOMAKI, P., AALTONEN, L. A., SISTONEN, P., PYLKKANEN, L., MECKLIN, J.-P., JARVINEN, H., GREEN, J. S., WEBER, J. L. & LEACH, F. S. 1993. Genetic mapping of a locus predisposing to human colorectal cancer. *Science,* 260**,** 810-812.

POGUE-GEILE, K., YOTHERS, G., TANIYAMA, Y., TANAKA, N., GAVIN, P., COLANGELO, L., BLACKMON, N., LIPCHIK, C., KIM, S. R. & SHARIF, S. 2013. Defective mismatch repair and benefit from bevacizumab for colon cancer: findings from NSABP C-08. *Journal of the National Cancer Institute,* 105**,** 989-992.

POLYAK, K. 2008. Is breast tumor progression really linear? *Clinical Cancer Research,* 14**,** 339-341.

POPAT, S., HUBNER, R. & HOULSTON, R. S. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. *Journal of Clinical Oncology,* 23**,** 609-618.

R-CORE-TEAM R: A Language and Environment for Statistical Computing. Vienna,

Austria: R Foundation for Statistical Computing.

REDFORD, L. 2016. *Short mononucleotide repeat detection of MSI: towards high throughput diagnosis.* Newcastle university.

REHM, H. L., BALE, S. J., BAYRAK-TOYDEMIR, P., BERG, J. S., BROWN, K. K., DEIGNAN, J. L., FRIEZ, M. J., FUNKE, B. H., HEGDE, M. R. & LYON, E. 2013. ACMG clinical laboratory standards for next-generation sequencing. *Genetics in Medicine,* 15**,** 733-747.

RHEES, J., ARNOLD, M. & BOLAND, C. R. 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**,** 219-225.

RIBIC, C. M., SARGENT, D. J., MOORE, M. J., THIBODEAU, S. N., FRENCH, A. J., GOLDBERG, R. M., HAMILTON, S. R., LAURENT-PUIG, P., GRYFE, R. & SHEPHERD, L. E. 2003. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *New England Journal of Medicine,* 349**,** 247-257.

ROBERTS, S. A. & GORDENIN, D. A. 2014. Hypermutation in human cancer genomes: footprints and mechanisms. *Nature Reviews Cancer,* 14**,** 786-800.

ROBINSON, J. T., THORVALDSDÓTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nature biotechnology,* 29**,** 24-26.

RODRIGUEZ-BIGAS, M. A., BOLAND, C. R., HAMILTON, S. R., HENSON, D. E., SRIVASTAVA, S., JASS, J. R., KHAN, P. M., LYNCH, H., SMYRK, T. & PERUCHO, M. 1997. A National Cancer Institute workshop on hereditary nonpolyposis colorectal cancer syndrome: meeting highlights and Bethesda guidelines. *Journal of the National Cancer Institute,* 89**,** 1758-1762.

ROZEN, S. & SKALETSKY, H. 1999. Primer3 on the WWW for general users and for biologist programmers. *Bioinformatics methods and protocols***,** 365-386.

RÜBBEN, A., KEMPF, W., KADIN, M. E., ZIMMERMANN, D. R. & BURG, G. 2004. Multilineage progression of genetically unstable tumor subclones in cutaneous T-cell lymphoma. *Experimental dermatology,* 13**,** 472-483.

SALIPANTE, S. J. & HORWITZ, M. S. 2006. Phylogenetic fate mapping. *Proceedings of the National Academy of Sciences,* 103**,** 5448-5453.

SALIPANTE, S. J., SCROGGINS, S. M., HAMPEL, H. L., TURNER, E. H. & PRITCHARD, C. C. 2014. Microsatellite instability detection by next generation sequencing. *Clinical chemistry,* 60**,** 1192-1199.

SALK, J. J. & HORWITZ, M. S. Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Seminars in cancer biology, 2010. Elsevier, 294-303.

SALK, J. J., SALIPANTE, S. J., RISQUES, R. A., CRISPIN, D. A., LI, L., BRONNER, M. P., BRENTNALL, T. A., RABINOVITCH, P. S., HORWITZ, M. S. & LOEB, L. A. 2009. Clonal expansions in ulcerative colitis identify patients with neoplasia. *Proceedings of the National Academy of Sciences,* 106**,** 20871-20876.

SAMMALKORPI, H., ALHOPURO, P., LEHTONEN, R., TUIMALA, J., MECKLIN, J.-P., JÄRVINEN, H. J., JIRICNY, J., KARHU, A. & AALTONEN, L. A. 2007. Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer research,* 67**,** 5691-5698.

SARIDAKI, Z., SOUGLAKOS, J. & GEORGOULIAS, V. 2014. Prognostic and predictive significance of MSI in stages II/III colon cancer. *World journal of gastroenterology: WJG,* 20**,** 6809.

SCALLY, A. & DURBIN, R. 2012. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics,* 13**,** 745-753.

SCHUG, M. D., HUTTER, C. M., NOOR, M. A. F. & AQUADRO, C. F. 1998. Mutation and evolution of microsatellites in Drosophila melanogaster. *Mutation and Evolution.* Springer.

SHARMA, P. C., GROVER, A. & KAHL, G. 2007. Mining microsatellites in eukaryotic genomes. *Trends in biotechnology,* 25**,** 490-498.

SHIBATA, D., NAVIDI, W., SALOVAARA, R., LI, Z.-H. & AALTONEN, L. A. 1996. Somatic microsatellite mutations as molecular tumor clocks. *Nature medicine,* 2**,** 676-681.

SHIBATA, D., PEINADO, M. A., IONOV, Y. & MALKHOSYAN, S. 1994. Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis. *Nature genetics,* 6.

SHIN, K.-H., SHIN, J.-H., KIM, J.-H. & PARK, J.-G. 2002. Mutational analysis of promoters of mismatch repair genes hMSH2 and hMLH1 in hereditary nonpolyposis colorectal cancer and early onset colorectal cancer patients: identification of three novel germ-line mutations in promoter of the hMSH2 gene. *Cancer research,* 62**,** 38-42.

SHINDE, D., LAI, Y., SUN, F. & ARNHEIM, N. 2003. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis:(CA/GT) n and (A/T) n microsatellites. *Nucleic acids research,* 31**,** 974-980.

SINICROPE, F. A., FOSTER, N. R., THIBODEAU, S. N., MARSONI, S., MONGES, G., LABIANCA, R., YOTHERS, G., ALLEGRA, C., MOORE, M. J. & GALLINGER, S. 2011. DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. *Journal of the National Cancer Institute,* 103**,** 863-875.

SNOWSILL, T., HUXLEY, N., HOYLE, M., JONES-HUGHES, T., COELHO, H., COOPER, C., FRAYLING, I. & HYDE, C. 2015. A model-based assessment of the cost–utility of strategies to identify Lynch syndrome in early-onset colorectal cancer patients. *BMC cancer,* 15**,** 1.

SØREIDE, K., SLEWA, A., STOKKELAND, P. J., VAN DIERMEN, B., JANSSEN, E. A. M., SØREIDE, J. A., BAAK, J. & KØRNER, H. 2009. Microsatellite instability and DNA ploidy in colorectal cancer. *Cancer,* 115**,** 271-282.

STORMORKEN, A. T., BOWITZ-LOTHE, I. M., NORÈN, T., KURE, E., AASE, S., WIJNEN, J., APOLD, J., HEIMDAL, K. & MØLLER, P. 2005. Immunohistochemistry identifies carriers of mismatch repair gene defects

causing hereditary nonpolyposis colorectal cancer. *Journal of clinical oncology,* 23**,** 4705-4712.

STRAND, M., PROLLA, T. A., LISKAY, R. M. & PETES, T. D. 1993. Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. *Nature,* 365**,** 274-276.

STRATTON, M. R., CAMPBELL, P. J. & FUTREAL, P. A. 2009. The cancer genome. *Nature,* 458**,** 719-724.

SUBRAMANIAN, S., MISHRA, R. K. & SINGH, L. 2003. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome biol,* 4**,** R13.

SURAWEERA, N., DUVAL, A., REPERANT, M., VAURY, C., FURLAN, D., LEROY, K., SERUCA, R., IACOPETTA, B. & HAMELIN, R. 2002. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. *Gastroenterology,* 123**,** 1804-1811.

SUTTER, C., GEBERT, J., BISCHOFF, P., HERFARTH, C. & VON KNEBEL DOEBERITZ, M. 1999. Molecular screening of potential HNPCC patients using a multiplex microsatellite PCR system. *Molecular and cellular probes,* 13**,** 157-165.

TOMLINSON, I. P., NOVELLI, M. & BODMER, W. 1996. The mutation rate and cancer. *Proceedings of the National Academy of Sciences,* 93**,** 14800-14803.

TRAN, B., KOPETZ, S., TIE, J., GIBBS, P., JIANG, Z. Q., LIEU, C. H., AGARWAL, A., MARU, D. M., SIEBER, O. & DESAI, J. 2011. Impact of BRAF mutation and microsatellite instability on the pattern of metastatic spread and prognosis in metastatic colorectal cancer. *Cancer,* 117**,** 4623-4632.

TRAN, H. T., KEEN, J. D., KRICKER, M., RESNICK, M. A. & GORDENIN, D. A. 1997. Hypermutability of homonucleotide runs in mismatch repair and DNA polymerase proofreading yeast mutants. *Molecular and Cellular Biology,* 17**,** 2859-2865.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics,* 13**,** 36-46.

TSAO, J.-L., YATABE, Y., SALOVAARA, R., JÄRVINEN, H. J., MECKLIN, J.-P., AALTONEN, L. A., TAVARÉ, S. & SHIBATA, D. 2000. Genetic reconstruction of individual colorectal tumor histories. *Proceedings of the National Academy of Sciences,* 97**,** 1236-1241.

UMAR, A., BOLAND, C. R., TERDIMAN, J. P., SYNGAL, S., DE LA CHAPELLE, A., RÜSCHOFF, J., FISHEL, R., LINDOR, N. M., BURGART, L. J. & HAMELIN, R. 2004. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute,* 96**,** 261-268.

URQUHART, A., KIMPTON, C., DOWNES, T. & GILL, P. 1994. Variation in short tandem repeat sequences—a survey of twelve microsatellite loci for use as forensic identification markers. *International journal of legal medicine,* 107**,** 13-20.

VASEN, H. F., BLANCO, I., AKTAN-COLLAN, K., GOPIE, J. P., ALONSO, A., ARETZ, S., BERNSTEIN, I., BERTARIO, L., BURN, J. & CAPELLA, G. 2013. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut,* 62**,** 812-823.

VEGANZONES, S., MAESTRO, M. L., RAFAEL, S., DE LA ORDEN, V., VIDAURRETA, M., MEDIERO, B., ESPANTALEÓN, M., CERDÁN, J. & DÍAZ-RUBIO, E. 2015. Combined methylation of p16 and hMLH1 (CMETH2) discriminates a subpopulation with better prognosis in colorectal cancer patients with microsatellite instability tumors. *Tumor Biology,* 36**,** 3853-3861.

VILKKI, S., LAUNONEN, V., KARHU, A., SISTONEN, P., VÄSTRIK, I. & AALTONEN, L. A. 2002. Screening for microsatellite instability target genes in colorectal cancers. *Journal of medical genetics,* 39**,** 785-789.

VOGELSTEIN, B., PAPADOPOULOS, N., VELCULESCU, V. E., ZHOU, S., DIAZ, L. A. & KINZLER, K. W. 2013. Cancer genome landscapes. *science,* 339**,** 1546-1558.

WAHLS, W. P., WALLACE, L. & MOORE, P. D. 1990. The Z-DNA motif d (TG) 30 promotes reception of information during gene conversion events while stimulating homologous recombination in human cells in culture. *Molecular and cellular biology,* 10**,** 785-793.

WANG, G., CARBAJAL, S., VIJG, J., DIGIOVANNI, J. & VASQUEZ, K. M. 2008. DNA structure-induced genomic instability in vivo. *Journal of the National Cancer Institute,* 100**,** 1815-1817.

WANG, X., WANG, M., MACLENNAN, G. T., ABDUL-KARIM, F. W., EBLE, J. N., JONES, T. D., OLOBATUYI, F., EISENBERG, R., CUMMINGS, O. W. & ZHANG, S. 2009. Evidence for common clonal origin of multifocal lung cancers. *Journal of the National Cancer Institute,* 101**,** 560-570.

WANG, Y.-H. & GRIFFITH, J. 1995. Expanded CTG triplet blocks from the myotonic dystrophy gene create the strongest known natural nucleosome positioning elements. *Genomics,* 25**,** 570-573.

WARD, R., MEAGHER, A., TOMLINSON, I., O'CONNOR, T., NORRIE, M., WU, R. & HAWKINS, N. 2001. Microsatellite instability and the clinicopathological features of sporadic colorectal cancer. *Gut,* 48**,** 821-829.

WEBER, J. L. & WONG, C. 1993. Mutation of human short tandem repeats. *Human molecular genetics,* 2**,** 1123-1128.

WILSON, T. M., VAISMAN, A., MARTOMO, S. A., SULLIVAN, P., LAN, L., HANAOKA, F., YASUI, A., WOODGATE, R. & GEARHART, P. J. 2005. MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A: T mutations in antibody genes. *The Journal of experimental medicine,* 201**,** 637-645.

WOERNER, S. M., BENNER, A., SUTTER, C., SCHILLER, M., YUAN, Y. P., KELLER, G., BORK, P., VON KNEBEL DOEBERITZ, M. & GEBERT, J. F. 2003. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene,* 22**,** 2226-2235.

WOERNER, S. M., YUAN, Y. P., BENNER, A., KORFF, S., VON KNEBEL DOEBERITZ, M. & BORK, P. 2010. SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. *Nucleic acids research,* 38**,** D682-D689.

XIAO, Y. & FREEMAN, G. J. 2015. The microsatellite instable subset of colorectal cancer is a particularly good candidate for checkpoint blockade immunotherapy. *Cancer discovery,* 5**,** 16-18.

YACHIDA, S., JONES, S., BOZIC, I., ANTAL, T., LEARY, R., FU, B., KAMIYAMA, M., HRUBAN, R. H., ESHLEMAN, J. R. & NOWAK, M. A. 2010. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature,* 467**,** 1114-1117.

YASHIRO, M., HIRAKAWA, K. & BOLAND, C. R. 2010. Mutations in TGFbeta-RII and BAX mediate tumor progression in the later stages of colorectal cancer with microsatellite instability. *BMC cancer,* 10**,** 303.

YOON, K., LEE, S., HAN, T.-S., MOON, S. Y., YUN, S. M., KONG, S.-H., JHO, S., CHOE, J., YU, J. & LEE, H.-J. 2013. Comprehensive genome-and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome research,* 23**,** 1109-1117.

ZHANG, L., LU, H. H. S., CHUNG, W.-Y., YANG, J. & LI, W.-H. 2005. Patterns of segmental duplication in the human genome. *Molecular biology and evolution,* 22**,** 135-141.

ZHAO, H., THIENPONT, B., YESILYURT, B. T., MOISSE, M., REUMERS, J., COENEGRACHTS, L., SAGAERT, X., SCHRAUWEN, S., SMEETS, D. & MATTHIJS, G. 2014. Mismatch repair deficiency endows tumors with a unique mutation signature and sensitivity to DNA double-strand breaks. *Elife,* 3**,** e02725.

ZHOU, X.-P., HOANG, J.-M., LI, Y.-J., SERUCA, R., CARNEIRO, F., SOBRINHO-SIMOES, M., LOTHE, R. A., GLEESON, C. M., RUSSELL, S. H. & MUZEAU, F. 1998. Determination of the replication error phenotype in human tumors without the requirement for matching normal DNA by analysis of mononucleotide repeat microsatellites. *Genes Chromosomes and Cancer,* 21**,** 101-107.