Short mononucleotide repeat detection of MSI: towards high throughput diagnosis

Lisa Redford

071275910

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

February 2016

Abstract

Microsatellites are short repetitive DNA sequences, which are liable to replication errors. Microsatellite instability (MSI) is controlled by the mismatch repair system, and accumulation of microsatellite mutations is used as a diagnostic criterion for tumours where this repair system is compromised, such as those which develop in Lynch Syndrome (HNPCC) patients. Currently, the Amsterdam II screening criteria and revised Bethesda Guidelines are used to identify tumours for MSI testing using both immunohistochemistry and fragment analysis tests. However, because Lynch Syndrome patients are being missed, testing for all colorectal and endometrial cancers is now being recommended. Faster and cheaper MSI testing methods are therefore desirable. Although PCR and sequencing error compromise sequence based typing of the repeats currently used for diagnosis, some short mononucleotide repeats have been identified which show low level instability, suggesting that sequence typing of short repeats may be possible. Here, I investigate the utility of high throughput sequencing (HTS) as the basis for MSI testing.

As an initial assessment of the method, I used the MiSeq platform to type 22 previously published short mononucleotide repeats in 4 microsatellite unstable (MSI-H) tumours, and showed that MSI could be detected above background noise in 7-12bp repeats. To identify the most variable short repeat markers for MSI testing, I then analysed MSI in whole genome sequence data from The Cancer Genome Atlas network, and identified a panel of 120 7-12bp informative mononucleotide repeats which were subsequently evaluated on a panel of 5 MSI-H tumours and controls. The most informative 20 markers were further tested on a panel of 58 colorectal tumours to define thresholds for instability calling. Using a panel of eighteen 8-12bp mononucleotides it was possible to distinguish between MSI-H and microsatellite stable (MSS) tumours with a sensitivity and specificity of 100%.

Flanking SNPs were also evaluated and identified an excess of allelic bias among MSI-H tumours compared to MSS tumours, a feature that could be integrated into the MSI test. Finally, short mononucleotide repeats with flanking SNPs were assessed for their potential to identify clonal variation in MSI-H tumours. Using a multiple biopsy approach evidence of different sub-clones was found in three MSI-H tumours, suggesting that these markers could be used for analysis of clonal variation and evolution.

Declaration

I certify that no part of the material documented in this thesis has previously been submitted for a degree or other qualification in this or any other university. I declare that this thesis represents my own work, except where it is acknowledged otherwise in the thesis text.

Lisa Redford

Acknowledgements

I would like to thank my supervisors Professor Sir John Burn, Dr Michael Jackson and Dr Mauro Santibaez-Koref for their help and support throughout this PhD. In particular John Burn for the inspiration he provided and for starting this project and promoting it so the project will carry on for years to come, Michael Jackson for his guidance and help with keeping my work on track, and Mauro Santibaez-Koref for teaching me Perl and Shell scripting, and helping me with the analysis of sequencing data.

I would like to thank the company QuantuMDx for funding this PhD project and allowing me to work with their technology. Especially Jonathan O'Halloran, Dr Stephen Osborne, Dr Sam Whitehouse and Dr John Tyson who welcomed me into QuantuMDx and introduced me to QuantuMDx's technology at the beginning of my PhD project, and Ryan Wetherell who taught me how to calculate match probability.

I would like to thank Dr John Tyson and the CAPP team for helping me with the ethics application for this project. Without their help, applying for ethics would have been a lot more time consuming and exhausting.

I would also like to thank Julie Coaker, Dr Stephanie Needham and Ottie O'Brien for their help with procuring samples for this project. They also taught me a lot about sample preparation and procedures for this within the Newcastle Hospitals NHS Foundation Trust. I would also like to thank Ottie O'Brien for providing expertise with the interpretation of fragment analysis traces.

I would like to thank the students Ghanim Alhilal and Iona V. Middleton for their assistance and contributions.

Finally, I am also grateful for the support from family and friends over the past 4 years I have spent working on my PhD project.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AUC: area under the receiver operating characteristic curve

bp : base pair

BLAST: Basic Local Alignment Search Tool

BLAT: BLAST-like alignment tool

BSA: bovine serum albumin

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

dH_2O: deionised water

DNA: deoxyribonucleic acid

dsDNA: double stranded DNA

EMAST: elevated microsatellite alterations at selected tetranucleotide repeats

EXO1: Exonuclease I

FAP: familial adenomatous polyposis

FFPE: formalin fixed paraffin embedded

Fst: fixation index

gDNA: genomic DNA

HDI: human development index

HNPCC: hereditary non polyposis colorectal cancer

HTS: high throughput sequencing

IGV: Integrative Genomics Viewer

Indel: insertion and deletion

LINES: Long Interspersed Nuclear Elements

MMR: mismatch repair

MSI: microsatellite instability

MSI-H: high levels of microsatellite instability

MSI-L: low levels of microsatellite instability

MSS: microsatellite stable

NHS: National Health Service

NSAIDs: non-steroidal anti-inflammatory drugs

PCR: polymerase chain reaction

POC: point of care

PVP: polyvinylphenol

RNA: ribonucleic acid

ROC: receiver operator characteristics

SelTarBase: Selective Targets in Human MSI-H Tumorigenesis Database

SINES: Short Interspersed Nuclear Elements

SNP: single nucleotide polymorphism

TBE: tris-acetate-EDTA

TCGA: The Cancer Genome Atlas

TE: Tris-EDTA

Tm : melting temperature

UCSC Genome Browser: University of California Santa Cruz Genome Browser

VCF: variant call file

5-FU: 5-fluorouracil

Chapter 1. Introduction

1.1. The global burden of cancer

Cancer is a leading cause of death worldwide. In 2012 there were estimated to be 8.2 million deaths caused by cancer around the world (Ferlay et al., 2015). In the USA, cancer is currently the second largest cause of death and it is estimated that it will become the most common cause of death over the next few years (Siegel et al., 2015). The world's cancer burden is expected to increase over the next few years. In 2012 there was estimated to be 14.1 million new cases of cancer, and estimated projections for 2025 total 20-24 million new cases of cancer (Ferlay et al., 2015, Gulland, 2014). It is therefore important to prioritise the treatment and prevention of cancer. Factors which it is believed will contribute to a future rise in cancer burden include life style changes, increased life expectancy, and an increase in the world's population.

Lung cancer is the most common form of cancer with an estimated 1.8 million new cases and ~1.6 million deaths worldwide in 2012 (Ferlay et al., 2015). This accounts for 13% of new cancer cases (Gulland, 2014). The number of lung cancer cases show a large positive correlation with the prevalence of tobacco smoking (Bray et al., 2012). In countries with a high human development index (HDI) the lung cancer rates in men are decreasing while there is an increase in the lung cancer rates in women (Bray et al., 2012, Siegel et al., 2015), reflecting the changes in smoking habits of men and women (Bray et al., 2012). Lung cancer is the most common form of cancer in countries with a high or medium HDI and is set to rise steeply in low HDI countries with rising use of tobacco (Bray et al., 2012).

Breast cancer is the second most common cancer with an estimated 1.7 million new cases in 2012 and ~522,000 deaths worldwide (Ferlay et al., 2015). This constitutes 11.9% of new cancer cases in 2012 (Gulland, 2014).

The third most common cancer type in 2012 was colorectal cancer with ~1.4 million new cases and ~694,000 deaths (Ferlay et al., 2015). Colorectal cancers therefore constitute 9.7% of the world's cancer burden (Gulland, 2014). There is an increasing rate of colorectal cancers in high and middle HDI areas (Bray et al., 2012). The reason for this is that many types of colorectal cancer can largely be attributed to lifestyle factors. For example, there is an increased risk of colorectal cancer associated with alcohol

consumption, smoking, obesity, diabetes, the consumption of large amounts of meat, and little physical activity (Huxley et al., 2009).

The fourth most common form of cancer in 2012 was prostate cancer which accounted for around 7.9% of new cancer cases (1.1 million) and ~307,000 deaths worldwide (Ferlay et al., 2015, Gulland, 2014).

Colorectal cancers are among the types associated with high socio-economic development; 40% of the global incidence of this cancer can be found in regions with a very high HDI, which only contain 15% of the world's population (Bray et al., 2012). In the future, it is likely that reduction in infection related cancers as a result of development in less developed countries will be offset by an increase cancers associated with a Western life style and increased life expectancy (Bray et al., 2012, Pourhoseingholi et al., 2015).

1.2. Colorectal cancer

There are different types of colorectal cancer (CRC) which are traditionally divided into two groups, those with chromosome instability and those with mismatch repair gene defects (Umar, 2004). Chromosome instability is the most common cause of colon cancer accounting for approximately 85% of CRCs (Sinicrope and Sargent, 2012). These cancers are characterized by the gain or loss of chromosomes and chromosome parts, the amplification of genes, and chromosome translocations (Grady, 2004). Chromosome instability can occur due to defects that affect the mitotic checkpoint (Pino and Chung, 2010). The mitotic checkpoint ensures that all chromatids are aligned properly before anaphase commences. A failure of this can lead to unequal chromosome segregation. Another cause of chromosome instability is abnormal centrosome function, which can also lead to unequal chromosome segregation (Pino and Chung, 2010). Other mechanisms that can cause chromosome instability include telomere dysfunction which can lead to chromosomes breaking and fusing during mitosis, and problems with the mitotic cell cycle arrest response which can lead to DNA damage not being repaired (Pino and Chung, 2010).

The other 15% of CRCs have mismatch repair gene defects and are characterized by microsatellite instability (MSI) (Grady, 2004), which can be defined as somatic changes in the length of microsatellites. Microsatellites are repetitive regions of DNA which are scattered throughout the genome. Because of their repetitive nature,

polymerases are more likely to cause slippage in the form of insertions and deletions while replicating microsatellites compared to other regions of DNA. Defects in mismatch repair genes cause MSI because errors during DNA replication are not rectified by the cell's compromised mismatch repair system. The DNA mismatch repair system is also a part of the mechanism which causes cell death when the mutation burden becomes too high (Boland, 2007). This function is also lost with a compromised mismatch repair system. A compromised mismatch repair system can, through these two mechanisms, lead to a high mutation burden which can cause cancer. MSI will cause tumorigenesis through mutations in genes which contain coding microsatellites (Grady, 2004). Two examples of such genes are *TGFBR2* and *BAX* (Grady, 2004).

Based on microsatellite status, colorectal tumours can be divided into 3 the categories; tumours with high levels of microsatellite instability (MSI-H), tumours with low levels of microsatellite instability (MSI-L), and tumours which are microsatellite stable (MSS). Tumours with mismatch repair defects have high levels of microsatellite instability and are categorised as MSI-H tumours. MSS tumours are usually tumours associated with chromosome instability. MSI-L tumours also appear to arise as a result of chromosome instability (Pawlik et al., 2004). The MSI-L category has been widely used, but there is debate over whether there is a qualitative difference between MSI-L and MSS tumours and if MSI-L tumours can be considered a discrete group (Tomlinson et al., 2002).

A recent molecular classification has identified four molecular sub groups (Guinney et al., 2015). The distinction of tumours with a breakdown in mismatch repair is still evident; they demonstrated marked inter connectivity across 6 different classification systems and distilled the groups into four consensus molecular subtypes:

CMS1 Microsatellite instability, immune (14%)

CMS2 Canonical (37%)

CMS3 Metabolic (13%)

CMS4 Mesenchymal (23%)

Tumours which could not be classified into one of these groups were deemed to represent a transitional phenotype or intratumoural heterogeneity. The focus of this thesis, in this new classification, is on CMS1.

1.2.1. Lynch Syndrome

Lynch syndrome, formerly known as hereditary non-polyposis colorectal cancer (HNPCC), is a hereditary form of autosomal dominant colon cancer which results from inherited mismatch repair gene defects and is characterized by high levels of microsatellite instability (Schofield et al., 2009). Lynch Syndrome constitutes 20% of MSI-H CRCs (Sinicrope and Sargent, 2012). Mutations in the *MLH1*, *MSH2*, *MSH6*, *PMS2* and *PMS1* genes can cause Lynch Syndrome (Silva et al., 2009). A deletion in the *EPCAM* gene upstream of *MSH2* can cause the knockout of *MSH2* and has also been shown to be a pathogenic mutation in some Lynch Syndrome patients (Kempers et al., 2011). Patients with Lynch Syndrome develop their first cancer early, on average in their mid forties, unlike patients with sporadic MSI-H cancers where the average age is over seventy (Boland, 2007). In addition to an increased risk of CRC, Lynch Syndrome is associated with an elevated risk of endometrial cancer (Aarnio et al., 1999, Hendriks et al., 2004), bladder cancer (van der Post et al., 2010), and tumours of the small intestine, ovary, urinary tract, stomach, biliary tract, pancreas, brain, and sebaceous glands (Balmana et al., 2010). The risk of developing CRC by the age of 70 years has been estimated at 66% for men, and for women the risk of developing a colorectal or endometrial cancer is estimated at 73% (Stoffel et al., 2009).

1.2.2. Sporadic microsatellite unstable tumours

Sporadic MSI-H tumours are usually caused by the epigenetic silencing of *MLH1* caused by promoter methylation (Boland, 2007, Sinicrope and Sargent, 2012). Whereas Lynch Syndrome tumours have been thought to arise from adenomas, sporadic MSI-H CRCs arise from serrated polyps (McGivern et al., 2004). More recently, the sessile serrated adenoma with its indistinct edges, mucus cap and characteristic “saw tooth” histology has become the primary suspect for the high prevalence of ascending colon “interval cancers” arising between frequent screening colonoscopies (Crockett et al., 2015).

Approximately 80% of MSI-H tumours are sporadic tumours. Sporadic MSI-H tumours, in addition to having on average a later age of onset compared to Lynch Syndrome tumours, also have a predisposition for the proximal colon and are more common in women than men (Jass, 2004).

1.3. Mismatch repair system in human cells

During DNA replication, DNA polymerases make replication errors at a rate of one error per 10^4-10^5 nucleotides (Iyer et al., 2006). The human DNA polymerases, polymerase δ and polymerase ε which perform the bulk of our DNA replication have proofreading exonuclease activity (Korona et al., 2011). The proofreading ability of these polymerases means that post-editing, the number of replication errors is further reduced to an error rate of one error per ~10^7 nucleotides (Iyer et al., 2006, Kunkel, 2004). In addition, the cells mismatch repair system corrects replication errors further reducing the error rate to one in ~10^9-10^{10} (Hsieh and Yamane, 2008).

The mismatch repair complex consists of the protein complexes MutSα, MutSβ and MutLα. MutSα is a heterodimer made up of MSH2 and MSH6, and this heterodimer repairs base mismatches including indels up to ~10 nucleotides in length (Iyer et al., 2006). MutSβ is a heterodimer made up of MSH2 and MSH3, which repairs indels of 2bp to ~10bp in length (Iyer et al., 2006). The MutS heterodimers form a clamp that moves along the DNA double helix examining around ~700bp for mismatches at a time before dissociating (Martin-Lopez and Fishel, 2013). When a MutSα or MutSβ protein identifies a mismatch they undergo a conformation change (Qiu et al., 2012). MutSα or MutSβ then recruits and mediates the binding of MutLα (Iyer et al., 2006, Qiu et al., 2012). The MutLα heterodimer is composed of the proteins MLH1 and PMS2. The function of MutLα is to assist the mismatch repair by interacting with either MutSα or MutSβ (Hsieh and Yamane, 2008).

Before a mismatch can be corrected, the strand with the mistake needs to be identified. The mechanisms that allows the recognition of the newly synthesised nascent strand by the mismatch repair complex is uncertain in eukaryotes (Lujan et al., 2013). In prokaryotes, the nascent strand is recognised by the mismatch repair system due to the lack of methylation on that strand (Pukkila et al., 1983). An endonuclease MutH incises the unmethylated strand creating a break, which initiates the removal of bases from the break to and including the mismatch (Kunkel and Erie, 2005). For prokaryotes, there is a delay between the synthesis of a new DNA strand and the methylation of that strand (Lyons and Schendel, 1984). This gives the mismatch repair system the opportunity to locate replication mistakes while the lack of methylation functions as a marker to identify the nascent strand. In Eukaryotes methylation is not used as a strand specific marker, but single strand breaks are sufficient to direct the mismatch repair system to hydrolyse the

strand with the break (Iyer et al., 2006). It is therefore possible that strand discontinuities that have occurred during DNA replication could be used by the mismatch repair system to identify the newly synthesised strand. For the lagging strand, it is believed that DNA discontinuities between Okazaki fragment may be used by the mismatch repair system to identify the newly replicated DNA strand (Nick McElhinny et al., 2010). For the leading strand, other mechanisms must be involved in marking the nascent strand. During DNA replication in eukaryotes, including humans, a ribonucleotide is erroneously incorporated on average every 1250 bases by DNA polymerase ε which catalyses leading strand replication (Dalgaard, 2012). Erroneously incorporated ribonucleotides are removed by ribonucleotide excision repair, which is initiated by the enzyme RNase H2 creating a nick in the newly replicated strand (Lujan et al., 2013). The strand breaks created by RNase H2 are one likely mechanism by which the mismatch repair system is directed to the nascent strand (Lujan et al., 2013).

Exonuclease I (EXO1) interacts with MSH2, MSH3 and MLH1, and performs the hydrolysis of the nascent strand from a strand break to ~ 150bp beyond the mismatch (Kunkel and Erie, 2005). Strand breaks used to initiate the process can be located either 5' or 3' to the mismatch. EXO1 is essential for both 5' and 3' directed mismatch repair (Constantin et al., 2005). EXO1 is a 5'- 3' endonuclease and therefore the endonuclease activity of MutLα is thought to play an essential role 3'-5' directed mismatch repair (Martin-Lopez and Fishel, 2013). DNA polymerase δ resynthesizes the part of the nascent strand that has been removed by Exonuclease I, repairing the DNA mismatch (Iyer et al., 2006).

Figure 1.1 gives simplified representation of the mismatch repair system.

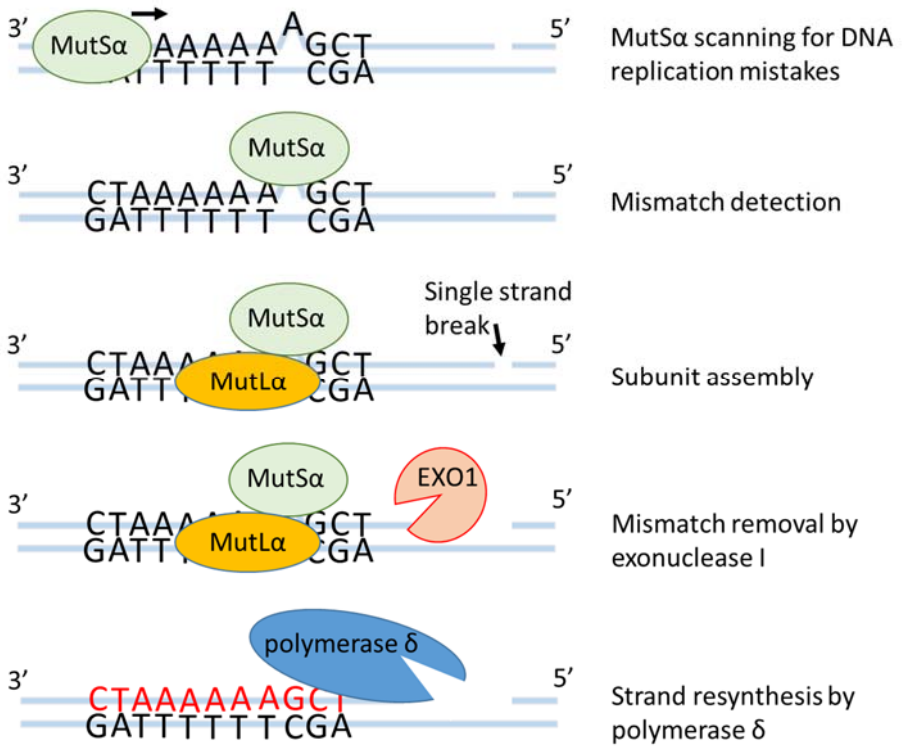

Figure 1.1: A simplified diagram of mismatch repair.

Knockout of mismatch repair function means that DNA replication errors in the form of base pair mismatches and indels up to 10bp are not repaired and can therefore accumulate. Microsatellite DNA is especially prone to replication mistakes in the form of insertions and deletions (Rose and Falush, 1998). Insertions and deletions in microsatellites are therefore used as an indicator of mismatch repair deficiency in tumours.

1.4. Mismatch repair and microsatellite instability

Microsatellites are repetitive sequences with repeat units of 1 - 6bp. Microsatellites are known to be unstable during meiotic and mitotic replication in eukaryotes and prokaryotes (Strauss, 1999). Factors which affect the susceptibility of microsatellites to slippage events include the length of the microsatellite, repeat unit length, base composition, and the sequence surrounding a microsatellite (Chung et al., 2010). Microsatellite instability is a failure to correct DNA replication errors as a result of defects in mismatch repair (MMR) genes. Testing for MSI in tumours is therefore used to identify MMR gene defects. Traditionally mononucleotide and dinucleotide repeats have been used in MSI tests (Boland et al., 1998, Umar, 2004). Tri-, tetra-, and

pentanucleotide repeats are less desirable in an MSI test because they show a low mutability in MSI-H tumours (Umar et al., 2004). Also, one cause of tetra nucleotide repeat instability, also know as Elevated Microsatellite Alterations at Selected Tetranucleotide repeats (EMAST), is believed to be a consequence of inflammation, and research suggests that this instability is reversible in tumours (Devaraj et al., 2010, Haugen et al., 2008).

1.5. Markers used for determining the MSI status of tumours

BAT26 is a quasimonomorphic microsatellite which is used as a marker for determining the MSI status of colorectal tumours. Bocker et al. (1997), Hoang et al. (1997), and Zhou et al. (1998) gave strong evidence that BAT26 could be used to classify the MSI status of tumours on its own without the use of normal DNA (Salahshor et al., 1999). There has, however, been some controversy about this over the years. Bubb et al. (1996) identified tumours with other unstable microsatellites where BAT26 was unaffected (Salahshor et al., 1999). Tumours which test positive for only BAT26 have also been observed (Salahshor et al., 1999). More recently, it has been suggested that tumours classed as MSI-H in the absence of BAT26 mutations may be falsely classified and are actually MSI-L or MSS. However, in MSI diagnostic testing, such confusion is avoided as several microsatellites are used. A reference panel of two mononucleotide microsatellites (BAT25 and BAT26), and three dinucleotide microsatellites (D5S346, D2S123, and D17S250) were proposed as consensus sequences for MSI testing at a National Cancer Institute workshop in 1997 (Boland et al., 1998).

In 2002 an international consensus recommended that dinucleotide repeats be replaced by mononucleotide repeats (Buhard et al., 2006). Since then there has not been an updated consensus over which new microsatellite markers to use. Buhard et al. (2006) suggested using the markers BAT26, BAT25, NR-27, NR-24, and NR-21 because these are quasimonomorphic in the Caucasian population. Using quasimonomorphic markers has the advantage that a comparison with a patients normal DNA is not needed. Buhard et al. (2006) have previously shown that a test they have compiled using these five mononucleotide repeats was 100% sensitive and specific. Buhard et al. (2006) show that using a cut off of 3 out of 5 markers with an abnormal length is a good test for a compromised mismatch repair system in CRCs worldwide, with few exceptions. The exceptions consist of Sub-Saharan African populations (the Biaka Pygmies and San

populations) where these markers are polymorphic, therefore 3 out of 5 markers with abnormal length does not necessarily indicate a MSI-H tumour. For these populations this test would require comparing normal DNA and the tumour DNA. In the Caucasian and Asian patients, 2 out of 5 unstable markers was a sufficient cut-off for calling a tumour MSI-H without any tumour being misclassified due to polymorphisms (Buhard et al., 2006).

The Northern Genetics Service in Newcastle Upon Tyne, England uses the following mononucleotide microsatellites: BAT25, BAT26, NR-21, NR-24 and MONO-27. If two or more of these microsatellite markers are unstable, then the cancer is classified as MSI-H. If the tumour has been classified as having MSI then the Northern Genetics Service will also test for the BRAFV600E mutation.

1.6. BRAF and KRAS mutations in CRCs

Other mutations frequently occurring in CRCs include mutations in the two genes *BRAF* and *KRAS*. These mutations are important to consider as they impact the prognosis of CRCs. In CRCs, activating mutations in the proto-oncogene *BRAF* are clustered in exon 15 creating the *BRAF* V600E amino acid substitution (Sinicrope and Sargent, 2012). The *BRAF* V600E mutation is found in sporadic MSI-H CRCs, but rarely in familial MSI-H CRCs caused by Lynch Syndrome. It can therefore be used to help distinguish hereditary from sporadic cancer (Schofield et al., 2009). The *BRAF* V600E mutation is associated with a worse prognosis for microsatellite stable CRCs, but it is still associated with a good prognosis in MSI-H CRCs (Samowitz et al., 2005).

KRAS is a component of the EGFR signalling pathway which regulates cell migration and proliferation among other cellular processes, and activating mutations in *KRAS* are found in 30-45% of CRCs (Heinemann et al., 2009, Kikuchi et al., 2009). An activating KRAS mutation leads to resistance to drugs like cetuximab and panitumumab, which are used in EGFR monoclonal inhibitory antibody chemotherapy (Heinemann et al., 2009). This is because EGFR monoclonal inhibitory antibody chemotherapy targets the EGFR signalling pathway upstream of *KRAS* (Heinemann et al., 2009). This targeting mechanism fails when *KRAS* has an activating mutation and therefore will reactivate the signalling pathway preventing effective treatment. Activating *KRAS* mutations occur mainly in codons 12 and 13, with up to 90% of these mutations found in these two codons

(Heinemann et al., 2009). These mutations are important to test for as they determine if EGFR monoclonal inhibitory antibody chemotherapy will be an effective treatment.

1.7. MSI as an indicator of prognosis

The literature generally agrees that MSI-H is a predictor of a better prognosis in CRCs compared to MSS (Popat et al., 2005, Sinicrope and Sargent, 2012). This is true for both sporadic and inherited CRCs. Popat et al. (2005) analysed the survival benefits of MSI in several different papers and concluded that MSI-H colorectal tumours had on average a 15% better overall survival rate. Figure 1.2 from the study of Popat et al. (2005) shows the correlation between MSI-H and survival benefit.

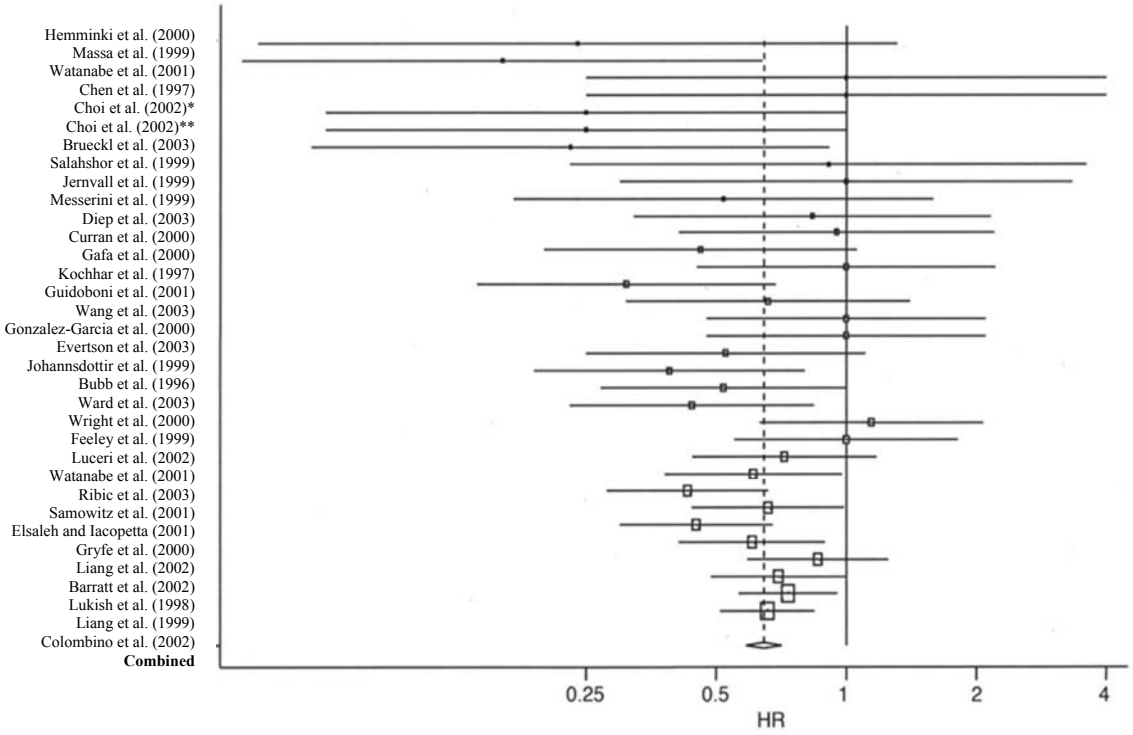

Figure 1.2: Forest plot of hazard ratios for overall survival of patients with MSI-H CRCs from different studies. These studies include patients with colorectal cancer stages 1, 2, 3, and 4. The forest plot has been taken from Popat et al, 2005 and shows that MSI CRCs have a hazard ratio of 0.65 compared to microsatellite stable CRCs which is a 0.35 reduction in hazard ratio. * Patients with stage 2 CRC. ** Patients with stage 1 CRC.

Although the literature is in overall agreement that MSI-H is associated with improved prognosis in CRCs there are exceptions. For example, the study by Salahshor et al. (1999) did not find a significant correlation between MSI-H colorectal tumours and better prognosis. They concluded that "MSI status is not an independent prognostic

factor", and the authors theorised that other studies may be finding a false correlation due to not taking into account the correlation between MSI and Dukes' stage. However the study had a very low number of positive controls (n=22) and their figures do show a weak trend towards better survival for the patients with MSI-H tumour even if it is not statistically significant. It could also be argued that if MSI-H tumours tend to have a lower tumour stage than MSS tumours, then this could be a characteristic of MSI-H tumours and Dukes' stage should not be corrected for. More recent studies, such as Popat et al. (2005) have also tried correcting for the stages of CRCs and found that this still results in improved prognosis in the presence of MSI-H. Despite MSI-H being a good predictor of prognosis MSI testing is only routinely used to identify patients with Lynch Syndrome (Sinicrope and Sargent, 2012).

1.8. Chemotherapy response in MSI-H tumours

MSI-H CRCs, whether sporadic or inherited, respond similarly to many different drugs. This will be due to the mismatch repair system being knocked out in both cancer types. In the literature there is wide support for the theory that the drug 5-fluorouracil (5-FU) does not provide benefit to patients with MSI-H CRCs (Popat et al., 2005, Sargent et al., 2010). There is, however, some controversy over the effect 5-FU on CRCs with a compromised mismatch repair system. Hutchins et al. (2011) found that there was a benefit of using 5-FU which was independent of mismatch repair status, in their analysis of 2 year cancer recurrence rates using data from the Quick and Simple and Reliable (QUASAR) trial. There is also evidence that the HSP110ΔE9 mutation in the *HSP110* gene makes patients susceptible to the benefits of 5-FU. This mutation occurs as a result of a compromised mismatch repair system in around 53% of MSI-H CRCs (Dorard et al., 2011). MSI-H tumours with a high expression of HSP110ΔE9 mRNA appear to respond well to 5-FU (Dorard et al., 2011). The drug irinotecan shows promise as an MSI-H cancer drug. Data from preclinical studies suggest that it could be more effective for MSI-H CRC compared to MSS CRCs (Sinicrope and Sargent, 2012). *MRE11A* mutations, which are found in 70-85% of cancers with MSI-H, may be the reason for the susceptibility of these cancers to irinotecan (Sinicrope and Sargent, 2012). However, irinotecan requires further study (Sinicrope and Sargent, 2012). Although there is currently not much information about the drug oxaliplatin used on its own, evidence suggests that this drug works just as well for both CRCs with mismatch repair gene defects and those without (Sinicrope and Sargent, 2012). Another drug that may confer

survival benefit to patients with MSI-H colorectal cancers is the drug bevacizumab (Pogue-Geile et al., 2013). This drug does not appear to give any survival benefit to patients with MSS tumours. There are also other drugs which appear to work well for MSS CRCs but don't work well for MSI-H CRCs. For example, evidence suggests that the drugs cisplatin and carboplatin do not work well on cancers with a compromised mismatch repair system (Sinicrope and Sargent, 2012).

There are a lot of drugs being tested that might prove effective against MSI-H CRCs (Sinicrope and Sargent, 2012). As the knowledge about this type of cancer grows, more effective treatments will be devised. For example evidence suggests that immune system activity can be seen as a positive indicator of prognosis (Galluzzi et al., 2012). Immune system activity could also possibly be used to predict patient's responses to different drugs (Galluzzi et al., 2012). The highly active immune response seen in MSI-H CRCs could be taken advantage of in this way to improve patient outcome. MSI-H CRCs have a pronounced immune response in the form of tumor-infiltrating T cells (Schwitalle et al., 2008). This immune response is most likely a response to carboxy-terminal frameshift peptides produced by these cancers (Schwitalle et al., 2008).

In 2015 a major study of the drug pembrolizumab showed startling benefits in MMR deficient colorectal cancer with a highly significant beneficial effect in cases of metastatic disease when compared to MMR proficient tumours (Le et al., 2015). In this study 40% of the patients with a MMR colorectal cancer had an immune related objective response and the progression free survival rate at 20 weeks was 78% for the patients with a MMR colorectal cancer. If the benefits of pembrolizumab are confirmed, MMR functional testing of all colorectal cancers is likely to become mandatory.

1.9. Economic evaluation of testing for mismatch repair defects in all colorectal tumours

Identifying patients with Lynch Syndrome is important because they have a high risk of developing second primary tumours and many of their relatives are also likely to be affected (Hampel et al., 2008, Barrow et al., 2013). These patients would therefore benefit from regular follow up. If tumours are detected earlier this has a significant improved prognosis for the patients. Regular colonoscopies will also save lives through identifying pre-cancerous polyps, which can be removed (Barrow et al., 2013). It has been estimated that more than 60% of Lynch Syndrome cancer deaths could be saved with the

proper follow up (Hampel et al., 2008). In addition to early tumour identification, identification of Lynch Syndrome enables prophylactic medication (like aspirin) to be used. Burn et al. (2011) showed that taking 600mg of aspirin a day for > 2 years gave a ~60% reduction in Lynch Syndrome cancer rates. A combination of aspirin and regular colonoscopies could prove to be quite an effective treatment for Lynch Syndrome cancers, but the majority of sufferers remain undiagnosed until disease presents in either themselves or a close family member. The challenge is identifying Lynch Syndrome patients so that they can receive the right follow up and preventative treatment.

Patients who present with a colorectal cancer are usually only screened for Lynch Syndrome in NHS England if they have a family history of cancer (Vasen et al., 2010). Many other Western European countries also use this approach (Vasen et al., 2010). A strategy of screening all colorectal cancers for MSI, then Lynch Syndrome, could be used to detect many of the patients and families with Lynch Syndrome that currently go undetected. The main reason why not all CRCs are screened for Lynch Syndrome is the costs this would incur. Identifying instances of sporadic MSI-H CRCs is also important because, as mentioned earlier, the cancer causing mechanism present in these cancers differs from MSS CRCs. This also means that sporadic MSI-H CRCs have a different prognosis to MSS CRCs and also respond differently to chemotherapy. It may be beneficial to identify these patients so they can receive a personalised treatment. Here the cost of tests is also an issue. Because MSI tests are expensive it would, with the current methods, be very expensive to test all CRCs in order to identify MSI-H cancers though this would be cost effective according to the recent major health economic assessment (Snowsill et al., 2015). Testing of cases where the cancer occurs in patients under 50 years is now becoming routine in the NHS in England but relies on the labour intensive immunohistochemistry.

The knockout of mismatch repair genes can be screened for using an MSI test or by performing immunohistochemistry staining. In an MSI test, somatic changes in microsatellite lengths can be used to infer mismatch repair defects, and this is the basis of the fragment analysis based MSI test currently in use. In an immunohistochemistry test the mismatch repair genes are evaluated by looking at the expression of the MLH1, MSH2, MSH6, and PMS2 proteins. *BRAF* V600E mutation screening of MSI-H tumours can be used to narrow down which patients may have Lynch Syndrome and save screening costs because the *BRAF* mutation rarely occurs in Lynch Syndrome patients but is very common in sporadic MSI-H colorectal cancers (Jin et al., 2013). For example

the study by Domingo et al. (2004) analysed 206 sporadic MSI-H tumours and 111 Lynch Syndrome tumours and found that 40% of the sporadic tumours had a *BRAF* V600E mutation while none of the Lynch Syndrome tumours had the mutation.

The financial costs of screening all colorectal cancers for mismatch repair defects and Lynch Syndrome using current methods would be high, but the money that could be saved through the early identification of cancers is also high. The average cost of a MSI test in England is £202 (Snowsill et al., 2014). The average costs of an immunohistochemistry test and a BRAF test are £238 and £118 respectively (Snowsill et al., 2014). Snowsill et al. (2014) based these prices on numbers reported directly from laboratories for costs of NHS England provided tests, and where possible these prices also include the cost of administration, equipment wear and tear costs, training time and the costs of repeated tests. Whyte et al. (2012) report that the lifetime costs for treating colorectal cancer are dependent on the stage the cancer has reached. For a Dukes stage A cancer the lifetime treatment costs are estimated at £12455, while for a Dukes stage B,C and D colorectal cancer the lifetime treatment cost are £17137, £23502, and £25703 respectively (Whyte et al., 2012). This highlights the importance of discovering cancers early from an economic perspective. Identifying Lynch Syndrome patients and following them up by regular monitoring not only saves lives, but will also decrease the treatment costs of those patients as cancers are identified earlier, or prevented through prophylactic use of drugs such as Aspirin.

1.10. The history of Lynch Syndrome identification

1.10.1. The discovery of Lynch Syndrome

Dr. Aldred Warthin published the first study on this hereditary disorder in 1913 after becoming intrigued by the number of bowel and endometrial cancers in one family, designated family G (1985). Warthin concluded that there was an inherited increased susceptibility to cancer in this family, but at the time the tumour causing mechanism was unknown. At the time there was scepticism that there could be a hereditary component to cancer. The significance of this new hereditary form of cancer was not fully understood until Henry T. Lynch documented the cancer syndrome in more detail including family G, and it was established that the disease followed a Mendelian pattern of inheritance

(Lynch, 1985, Lynch and Smyrk, 1996). As a result of the work conducted by Henry T. Lynch the term Lynch Syndrome was proposed by Boland and has gained widespread acceptance. The connection between Lynch Syndrome/HNPCC and mismatch repair gene mutations was not discovered until 1993 (Fishel et al., 1993, Leach et al., 1993).

1.10.2. The Amsterdam criteria for Lynch Syndrome identification

The first screening criteria developed for Lynch Syndrome identification, the Amsterdam criteria, were developed in 1991 (Boland et al., 1998, Umar, 2006), and had an estimated sensitivity of 60% and specificity of 70% (Lipton et al., 2004). They were primarily established to provide uniformity across studies to aid linkage studies in HNPCC (Vasen et al., 1999). The Amsterdam criteria were criticized when used in clinical practice for a lack of sensitivity as they assessed only colorectal cancers, resulting in the exclusion of patients who presented with other cancer types associated with Lynch Syndrome (Vasen et al., 1999). As a result, the criteria were updated so as to include cancers of the large bowel, endometrium, small bowel, ureter, and renal pelvis (Amsterdam II criteria, see Table 1).

All of the Following Must Apply for a Putative Diagnosis of HNPCC to be Made in a Family
There are at least three relatives with an HNPCC-associated cancer (large bowel, endometrium, small bowel, ureter, or renal pelvis, though not including stomach, ovary, brain, bladder, or skin)
One affected person is a first-degree relative of the other two
At least one person was diagnosed before the age of 50 years
At least two successive generations are affected
Familial adenomatous polyposis has been excluded Tumors have been verified by pathologic examination

This table was modified from Lipton et al, (2004)

Table 1.1: Amsterdam II Criteria

1.10.3. The Bethesda Guidelines for Lynch Syndrome identification

The link between microsatellite instability and Lynch Syndrome was not discovered until 1993 (Aaltonen et al., 1993, Ionov et al., 1993, Peltomaki et al., 1993, Thibodeau et al., 1993). This, and subsequent research, led to the idea that MSI testing could be used to more accurately identify which patients had Lynch Syndrome. The Bethesda Guidelines, which are used to identify cancers to be tested for MSI, were

originally released in 1996 and resulted from the "The Intersection of Pathology and Genetics in the Hereditary Nonpolyposis Colorectal Cancer (HNPCC) Syndrome" workshop (Boland et al., 1998). According to Terdiman et al. (2001) the first version of the Bethesda Guidelines had a sensitivity and specificity of 96% and 27% respectively. In 2002 the Bethesda Guidelines were updated to reflect the target audience of clinicians and pathologists, and so the guidelines could easily be disseminated to the public (revised Bethesda Guidelines, see Table 2) (Umar et al., 2004, Umar, 2006). There was also more emphasis placed on the testing of relatives because of a fear that the importance of this was not being recognized (Umar et al., 2004).

Tumours from individuals should be tested for MSI in the following situations:
1. Colorectal cancer diagnosed in a patient who is less than 50 years of age.
2. Presence of synchronous, metachronous colorectal, or other HNPCC associated tumors,[1] regardless of age.
3. Colorectal cancer with the MSI-H histology[2] diagnosed in a patient who is less than 60 years of age.
4. Colorectal cancer diagnosed in one or more first-degree relatives with an HNPCC-related tumor, with one of the cancers being diagnosed under age 50 years.
5. Colorectal cancer diagnosed in two or more first- or second-degree relatives with HNPCC-related tumors, regardless of age.

[1]Hereditary nonpolyposis colorectal cancer (HNPCC)-related tumors include colorectal, endometrial, stomach, ovarian, pancreas, ureter and renal pelvis, biliary tract, and brain (usually glioblastoma as seen in Turcot syndrome) tumors, sebaceous gland adenomas and keratoacanthomas in Muir–Torre syndrome, and carcinoma of the small bowel.

[2]Presence of tumor infiltrating lymphocytes, Crohn's-like lymphocytic reaction, mucinous/signet-ring differentiation, or medullary growth pattern.

This table was modified from (Umar 2006)

Table 1.2: Revised Bethesda Criteria

1.10.4. MSI testing using fragment analysis

The National Cancer Institute Workshop on MSI in 1997 proposed a consensus panel of 5 markers for MSI testing which included 2 mononucleotide repeats and 3 dinucleotide repeats (Boland et al., 1998, Umar, 2004). Tumours were classified as MSI-H if instability was detected in two or more markers, MSI-L if instability was only present in 1 marker. If all five markers were found to be stable, the tumour could be classed as MSS. Prior to this there had been no agreement on which microsatellites to use in an MSI test.

In 2002 an international consensus recommended that five mononucleotide repeats be used instead of a panel of mononucleotide and dinucleotide repeats (Buhard et al., 2006). The change was proposed to improve the sensitivity of the panel. It was feared that the original panel would miss-classify MSI-H tumours and overestimate the number of MSI-L tumours due to the low sensitivity of the dinucleotide repeats (Umar et al., 2004). Swapping the 3 dinucleotide repeats for mononucleotide repeats was done both the increase the sensitivity of the panel and because the use of 5 quasimonomorphic markers would allow tests to be performed in the absence matched normal tissue (Umar et al., 2004). See Figure 1.3 for an example of a MSI test using a panel of 5 mononucleotide repeats.

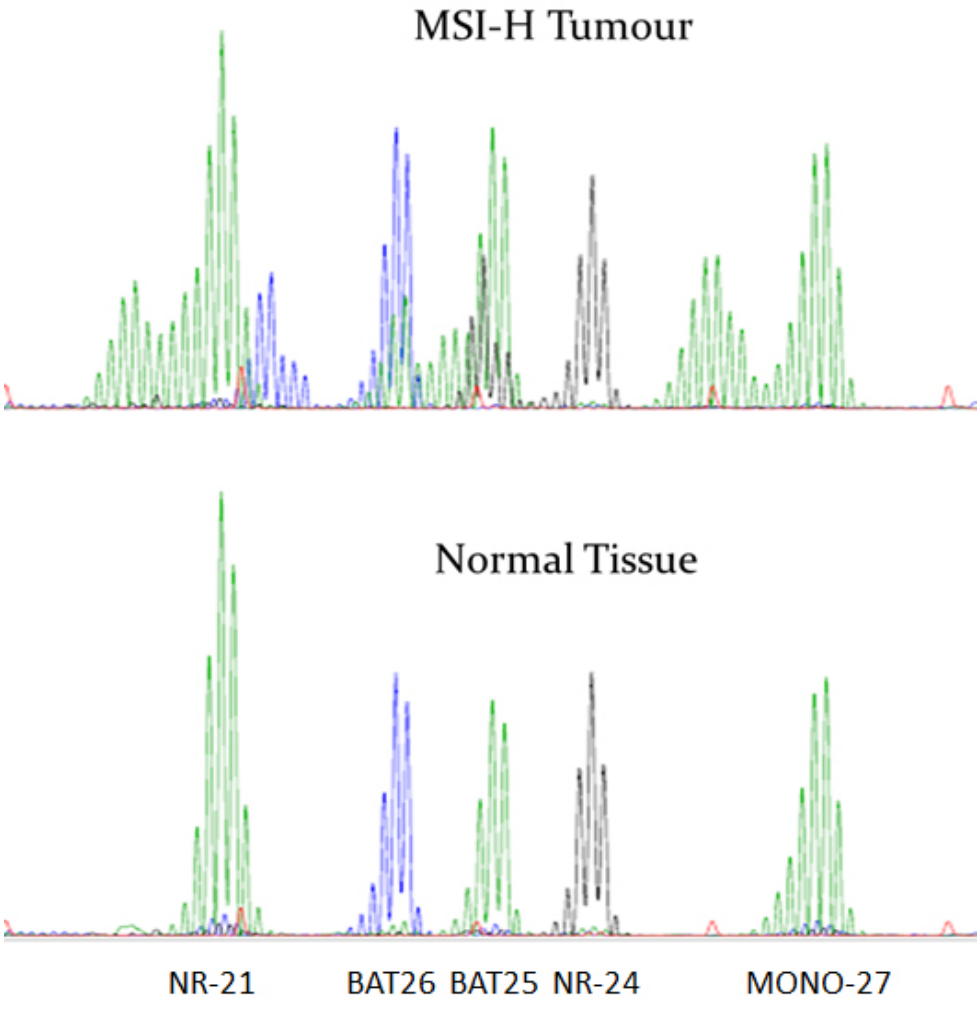

Figure 1.3: Standard MSI test with a panel of 5 mononucleotide repeats: This figure shows an example of a standard MSI test where tumour and matched normal have analyzed. The Promega MSI Analysis System, Version 1.2 kit was used to amplify the markers NR-21, BAT26, BAT25, NR-24, and MONO-27.

1.11. Drawbacks of the Amsterdam II criteria and revised Bethesda Guidelines

While the Amsterdam II criteria and revised Bethesda Guidelines are effective at identifying patients for further screening for Lynch Syndrome, a large number of Lynch Syndrome patients are being missed by the current approach. Perez-Carbonell et al. (2012) screened 2093 colorectal cancer patients for Lynch Syndrome, and found that 14.3% of the Lynch Syndrome patients did not meet the revised Bethesda Guidelines. These Lynch Syndrome patients would not have been discovered if the revised Bethesda Guidelines had been used to identify which patients should receive molecular testing. Another study by Canard et al. (2012) tested 1040 colorectal cancer and identified 25 patients with Lynch Syndrome. Out of these 25 patients 11 would have been missed if the Amsterdam II criteria were used, and 3 would have been missed using the Bethesda Guidelines.

Similarly, identifying patients with Lynch Syndrome based on the Bethesda guidelines has been reported to miss around 28% of cases (Hampel et al., 2008). Other studies have also shown that the Amsterdam criteria fail to identify a large number of patients with Lynch Syndrome (Boland et al., 1998, Hampel et al., 2008) and give false positives when faced with other familial cancers (Boland, 2007). Hampel et al. (2008) conclude that in the future as family sizes decrease, using family based methods for detecting Lynch Syndrome will get harder. Testing all CRCs for Lynch Syndrome is therefore important, as it will save many lives. As mentioned before, MSI testing is expensive (Umar, 2004). Current international testing varies, but due to the costs of tests, most places rely on the Bethesda guidelines and Amsterdam criteria to find high risk patients and only test these. Because the Bethesda guidelines and Amsterdam criteria fail to identify a significant number of Lynch Syndrome patients (Canard et al., 2012, Mills et al., 2014, Perez-Carbonell et al., 2012). This has led to suggestions that all CRC and endometrial tumours should receive molecular testing (Vasen et al., 2013, Canard et al., 2012, Mills et al., 2014, Julie et al., 2008).

1.12. The drawback of using immunohistochemistry and MSI testing to identify loss of MMR function

Both the MSI test and immunohistochemistry require expert interpretation as the stutter peaks of an MSI test and the staining patterns of an immunohistochemistry test can both be tricky to interpret in some cases. One example of this is shown in a study where 7 pathologists evaluated 100 cases using immunohistochemistry (Overbeek et al., 2008). Only in 82% of cases did 5 or more pathologists reach the same conclusion, but the 2 experienced pathologists identified all MSI-H tumours correctly. This example shows that highly trained personnel are vital for immunohistochemistry interpretations. Without highly trained personnel mistakes can be made.

A recent US based analysis of results from a biannual proficiency test for MSI testing, involving between 42 - 104 laboratories from 2005-2012, established that the average correct classification rate of samples by participating labs was 95.4% (Boyle et al., 2014). The standards of the MSI tests between different laboratories is currently high, but there is some variation which could indicate that moving to a more automated MSI test where simpler interpretation is required could be an improvement.

1.13. Testing for MSI using next generation sequencing

The advent of high throughput sequencing technologies has enabled the potential for sequence based MSI classification to be investigated at the genome level. The potential utility of a next generation sequence based approach was established by a study from The Cancer Genome Atlas (TCGA) project. They analysed the exomes of 224 matched CRC / normal pairs looking at 6-10bp mononucleotide repeats, and established that MSI could be detected using next generation sequencing. (Cancer Genome Atlas Network, 2012). This result was later confirmed in gastric cancers and gastric cancer cell lines where mononucleotide repeat sizes >4bp were analysed (Yoon et al., 2013). Since then, software has been developed to analyse whole genome, exome, whole transcriptome, and capture panel data (Lu et al., 2013, Niu et al., 2014, Salipante et al., 2014).

Currently, such genome-wide approaches are not cost effective, but suggest that systematic assessment of shorter repeats for sequence-based MSI detection is warranted. Sequence based MSI typing could be advantageous in terms of cost, with high throughput

enabling the sequencing of many tumours simultaneously, and ease of interpretation through automation. However, long microsatellites are not amenable to sequence analysis, and although some short (6-14bp) mononucleotide repeats have been identified which exhibit instability, the frequencies of instability are highly variable (Sammalkorpi et al., 2007, Vilkki et al., 2002, Woerner et al., 2003, Zhang et al., 2001). A panel of short repeats that are highly variable in MSI-H tumours and amenable to sequencing could, however, prove effective in an MSI test.

Personalised oncology will in the future be used to prescribe the best treatment for each individual's cancer (Sinicrope and Sargent, 2012). Personalised oncology involves prescribing chemotherapy based on a tumours molecular signature. As mentioned previously, MSI-H and MSS tumours respond differently to different drugs such as 5-fluorouracil, irinotecan, bevacizumab and pembrocizumab. Future CRC treatment strategies could therefore take into a tumour's MSI status in addition to other tumour biomarkers to enable the prescription of a more personalised treatment.

The advent of future technologies targeted to the market of personalised medicine, such as point of care devices would enable a test to be performed cheaply with a fast turnaround time. One company currently developing a point of care device aimed at cancer diagnostics, among other things, is the Newcastle Upon Tyne, UK based company QuantuMDx. The QuantuMDx Q-POC platform, which is currently under development, should be capable of detecting 64 genetic features per disposable cassette at a price of ~£20 (Burn, 2013). It is estimated that genetic tests using the Q-POC will take as little as 15 minutes or less (Burn, 2013). QuantuMDx's device has four main components; a tissue lysis chamber, a DNA extraction cassette containing a proprietary sorbent filter that allows DNA to pass through while binding cellular material such as proteins and lipids, a microfluidic based PCR cassette, and a silicon nanowire field effect transistor which can be used as a nanosensor for detecting base incorporation in a sequencing by synthesis reaction. The Q-POC device being developed by the company QuantuMDx, should lend itself to short amplicon sequencing based assays which are both cheap and quick. Developing an MSI assay compatible with QuantuMDx's device could enable a sequencing based MSI test that is fast and affordable, which would bring us one step further towards the goal of testing all CRCs and endometrial cancers for MSI without adding an extra financial burden on health systems. On the Q-POC platform, MSI testing could be performed in conjunction with testing for other cancer biomarkers such as *BRAF* and *KRAS* mutations. A deposable cassette that allows the testing of many different cancer

biomarkers, including microsatellite instability, would eliminate the need for multiple separate diagnostic tests.

1.14. Project aims and outline of results chapters

The primary aim of this project has been to identify markers for a sequencing based MSI test, and test these on CRCs to create a panel of markers that can differentiate between MSI-H and MSS CRCs and is compatible with the QuantuMDx Q-POC technology. To address this aim, mononucleotide repeats obtained from the literature were analysed to assess the suitability of using short repeats and Illumina sequencing to detect MSI. Whole genome analysis of MSI-H CRCs was performed to identify highly unstable mononucleotide repeats, which could be used as markers in a sequencing based MSI test. 120 of the identified mononucleotide repeats were analysed on a small panel of tumours to confirm that these repeats could be used as markers for identifying MSI. Finally, a larger panel of colorectal tumours were analysed using 20 of the most informative repeats to find out if a small number of repeats, which are highly susceptible to deletions in MSI-H tumours, could be used to differentiate between MSI-H and MSS tumours. Short mononucleotide repeats may also be good for assessing clonality in MSI-H tumours so a subsidiary aim was to test this hypothesis on tumours where multiple biopsies had been procured.

In the first results chapter (Chapter 3) I evaluate the feasibility of using Illumina sequencing of short mononucleotide repeats for differentiating between MSI-H and MSS tumours. The amount of noise in the form of sequencing and PCR error produced from sequencing different lengths of mononucleotide repeats is also evaluated, to assess if real mutations can be detected over background noise levels.

In the second results chapter (Chapter 4) I evaluate the ability of different variant callers to identify indels in mononucleotide repeats, and whole genome sequences of MSI-H CRCs are analysed to determine the distribution of variant reads in 7-12bp mononucleotide repeats, and identify candidate markers for an MSI test.

In the third results chapter (Chapter 5) I select the most variable markers with neighbouring SNPs, identified in the whole genome analysis, and test their levels of instability in a small panel of tumours to enable the selection of the most variable repeat for use in a future sequencing based MSI test. In this chapter the results of an analysis of

allelic bias in microsatellite unstable repeats is also presented, which aims to evaluate if this can be used to differentiate between real mutations and sequencing/PCR error.

In the fourth results chapter (Chapter 6) I use previously tested markers with neighbouring SNPs to analyse MSI-H tumour biopsies for clonal variations across different tumour regions.

In the fifth results chapter (Chapter 7) I evaluate 20 of the most variable short mononucleotide repeats identified previously, on a panel of over 50 CRC to determine thresholds for calling instability, and evaluate if the panel can correctly classify all MSI-H and MSS tumours. This chapter also contains my contributions towards the development of a potential platform for a sequencing based MSI test; the QuantuMDx Q-POC.

Chapter 2. Methods

2.1. Clinical work

2.1.1. Ethical approval

Tissue and blood samples from patients enrolled in the CAPP2 study were obtained after ethical review (REC reference MREC/98/3/24). The CAPP2 patents were anonymised for the purpose of this work, but by using the CAPP patient U numbers included in this thesis the samples can be linked back to the patient details by someone with authorised access to the CAPP study files.

Blood samples were collected as part of the DISC study: Diet related biomarkers of colorectal cancer risk. These samples were covered by ethics ref. 09/H0907/77 granted by the Newcastle and North Tyneside 2 REC. The blood samples were anonymised prior to use in this work.

All other human tissue samples collected and used as part of this PhD project were covered by ethical approval as part of the study "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" (IRAS project ID: 99148, REC reference: 13/LO/1514). All tissue samples were anonymised for the purpose of this PhD project, but patient data could be retrieved by someone with the proper authorised access.

2.1.2. Tissue collection

2.1.2.1. Lynch Syndrome patient samples

For patients enrolled in the CAPP2 study, tumour, blood and normal mucosa samples were collected previously by the CAPP study team. Criteria for inclusion in the CAPP2 study included a diagnosis of Lynch Syndrome, an intact colon or only a segmental resection. Exclusion criteria for participation in the CAPP2 study included medical contraindications for aspirin, and patients already on NSAIDs or steroids. Tumour and normal mucosa samples were supplied as Formalin-fixed, paraffin-

embedded (FFPE) tissue in the form of wax curls (A list of tumour samples can be found in Appendix Table 9.2. DNA from blood samples had previously been extracted by the CAPP study group.

2.1.2.2. Normal control blood samples

Blood to be used as normal controls was supplied by the DISC study group.

2.1.2.3. Samples MSI tested by the Northern Genetics Service

DNA and wax curls were obtained from tumours previously tested for MSI by the Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust using the Promega MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, United States of America).

2.1.2.4. Tumour sampling techniques for the clonality analysis

Tumour and tissue samples for clonality analyses were obtained from the Pathology Department, Newcastle Hospitals NHS Foundation Trust. Biopsies were taken from fresh colorectal tumours shortly after resection by Dr Stephanie Needham (Pathology department, Newcastle Hospitals NHS Foundation Trust). Biopsies were taken from each tumour at intervals using the hours of a clock face as a reference point. The side of the tumour closest to the antimesenteric border was defined as 12 o'clock. In some of the tumours it was impractical to use the antimesenteric border as 12 o'clock, for example because the tumours had grown across the antimesenteric border. In these cases the proximal orientation of the tumour was defined as 12 o'clock. Where possible four scalpel biopsies of external tumour tissue were taken from the 3, 6, 9 and 12 o'clock positions round the tumour followed by four fine needle aspiration biopsies taken from the 3, 6, 9 and 12 o'clock positions from deeper within the tumours using BD Microlance 21-gage needles (BD, New Jersey, United States of America). If the tumour was too small for this sampling technique then not all 8 biopsies were collected. Normal mucosa was sampled using a scalpel 7-10cm away from the tumour to ensure the normal mucosa biopsies were not contaminated by any tumour tissue.

2.2. DNA extraction

2.2.1. DNA extraction from FFPE tissues using the Promega DNA ReliaPrep™ FFPE gDNA Miniprep System kit

DNA extractions from wax embedded tissue were performed using the ReliaPrep™ FFPE gDNA Miniprep System kit (Promega, Madison, WI, United States of America) according to the manufacturer's protocol, with the exception of a prolonged deparaffination step. Briefly, samples were deparaffinised using mineral oil incubated at 80^{o}C for ~1 hour. Following cooling, proteinase K digestion was performed at 56^{o}C for an hour then 80^{o}C for an hour, RNase was added to the samples, and then the DNA was purified using spin columns. Elution volumes used consisted of either 30µl or 40µl.

2.2.2. DNA extraction from FFPE tissues using the BiOstic® FFPE Tissue DNA Isolation Kit

DNA extractions from wax embedded tissue were performed using the BiOstic® FFPE Tissue DNA Isolation Kit (Mo Bio, Carlsbad, CA, USA) according to the manufacturer's protocol. Elution volumes of 50µl -100µl were used.

2.2.3. DNA extraction from fresh tissue

DNA was extracted from fresh and fresh frozen tissue using the ReliaPrep™ gDNA Tissue Miniprep System kit (Promega, Madison, WI, United States of America) according to the manufacturer's protocol. Elution volumes consisted of 50µl -80µl.

2.2.4. DNA extraction from blood

DNA was extracted from blood using a QIAamp DNA Investigator Kit (Qiagen, Venlo, Netherlands) according to the manufacturer's protocol. The volume of blood used was 85µl. DNA was eluted using 65µl of deionised water (dH_2O).

2.2.5. DNA extraction using QuantuMDx's DNA extraction cassette

Experiments on both blood and wax embedded tissue have been performed using QuantuMDx's DNA extraction cassette. The extraction of DNA from samples consisted of a lysis step, activating the sorbent filter in the DNA extraction cassette with prep buffer, and DNA extraction by passing the lysis mixture through the activated sorbent filter.

Lysis of whole blood consisted of adding 5µl of whole blood to 95µl of proprietary lysis solutions and incubating this at 60ºC for 2 hours. For FFPE tissues the wax removal and tissue lysis was performed using reagents taken from the ReliaPrep™ FFPE gDNA Miniprep System kit as detailed in section 2.2.1. The DNA extraction program consisted of first wetting the filter with ~200µl of the proprietary prep buffer to activate the filter. After a 5 minute activation time the sample was loaded onto the filter at a speed of 100µl/min. Then the cassette was refilled with prep buffer. This second lot of buffer was run through the cassette to elute the DNA at a speed of 50µl/min. Small pauses can be programmed in at intervals determined by the user so that the DNA can be collected in elute fractions of varying volumes. The filter retains the cellular components that are passed through it, with the exception of DNA, which is passed out through the collection channel together with the buffer. A photo showing the layout of the DNA extraction cassette can be found in Figure 2.1. The sample and buffer are pushed through the DNA extraction cassette using the syringes of QuantuMDx's prototype machine (the MiniChemLab) (see Figure 2.2).

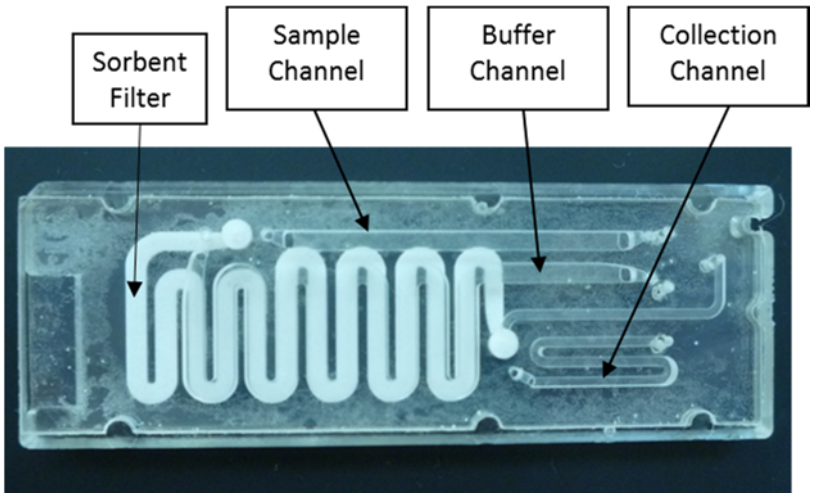

Figure 2.1: QuantuMDx's prototype DNA extraction cassette.

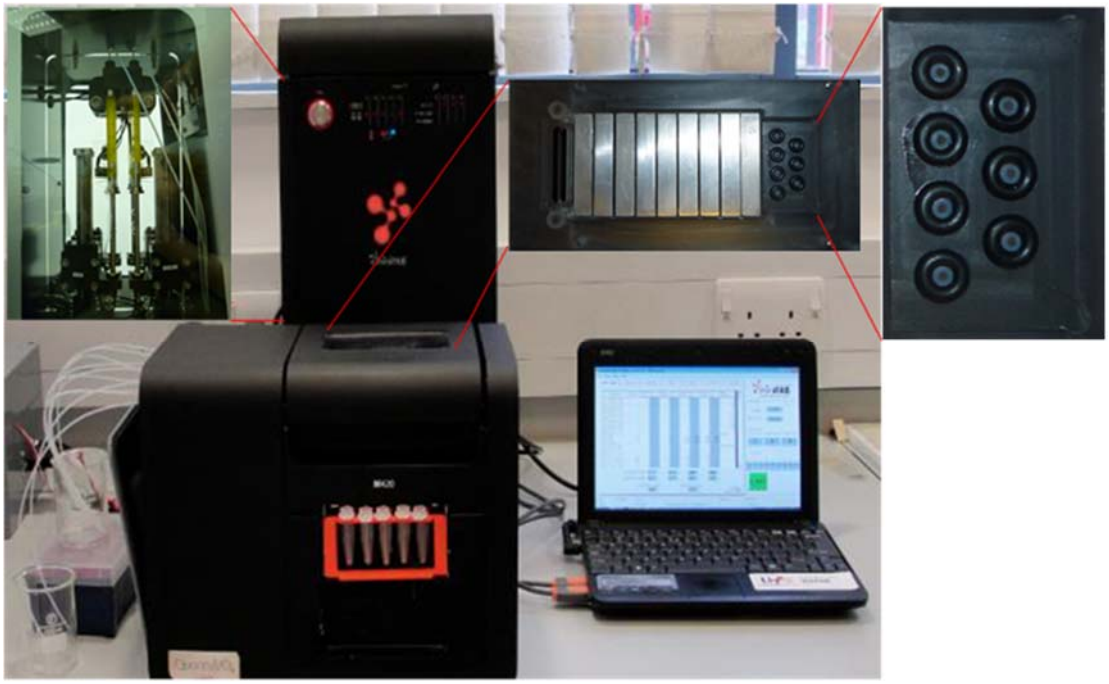

Figure 2.2: The QuantuMDx prototype (The MiniChemLab) with the cassette manifold (right corner) and the syringe pumps (left corner).

2.3. Polymerase chain reaction

2.3.1. PCR using QuantuMDx's PCR cassette

PCR cassette experiments were performed using the first generation of QuantuMDx PCR cassettes produced by MiniFab (MiniFab, Melbourne, Australia). Cassettes were run both with and without a surfactant coating the PCR channels. A surfactant coating was applied by soaking the PCR channels with solutions of 0.1mg/ml BSA and/or 2.5% PVP for between 10min to 3hours. PCR reactions were carried out using reagents from the HotStarTaq Plus kit (Qiagen, Venlo, Netherlands) and Deoxynucleotide (dNTP) Solution Mix (New England BioLabs, Ipswich, Massachusetts, USA). PCR reactions were made up of 1x reaction buffer, 0.2mM dNTP mix, 0.1-0.125U/µl polymerase, 1uM forward primer, 1uM reverse primer, DNA (master mix concentrations of 0.4ng/µl -1ng/µl gDNA or PCR product), and 0%-2.5% PVP. To activate the HotStartTaq the PCR mix was heated to 95^{o}C for 5min on a MultiGene II Thermocycler (Labnet International Inc, Edison, USA) prior to loading onto a PCR cassette. PCR mix volumes of 50-100µl were pumped through PCR cassettes at flow rates of either 10µl/min or 5µl/min. The PCR channel of QuantuMDx's cassettes provide 30 PCR cycles. All PCR reactions performed were two-step PCR reactions with denaturation

temperatures varying between 90-95ºC and an annealing temperature of 56ºC. See Figure 2.3 for a diagram of QuantuMDx's PCR cassette and testbed used for running the cassettes.

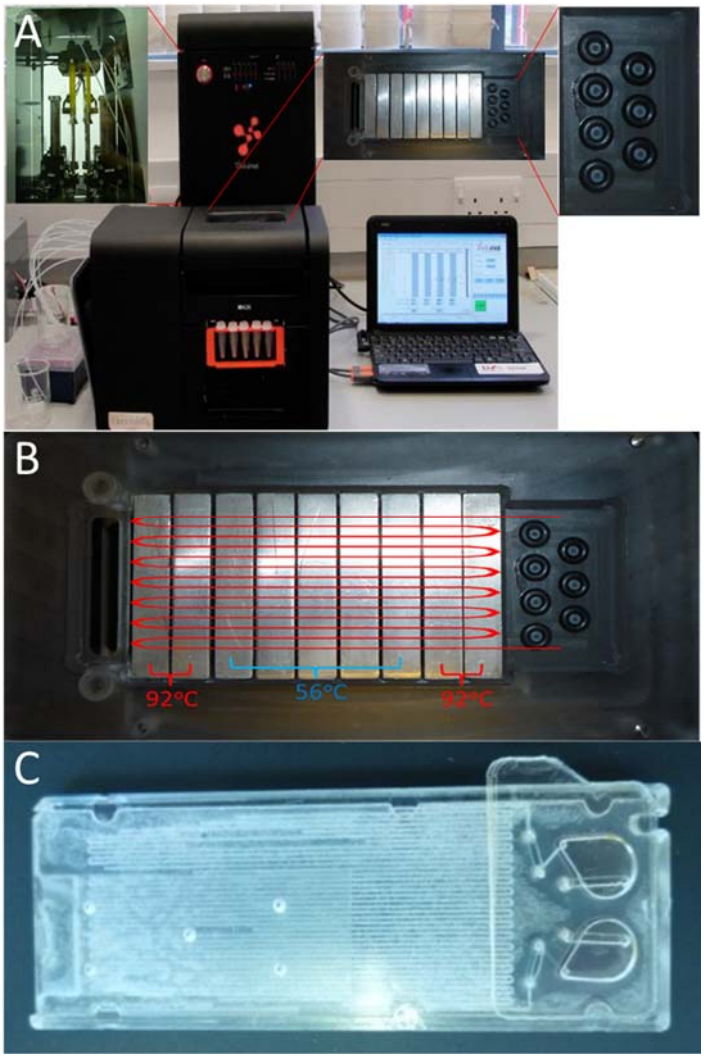

Figure 2.3: QuantuMDx's first generation cassette PCR using a prototype machine developed by MiniFab. Panel A: The testing platform for QuantuMDx's first generation PCR cassettes. Panel B: Simplified diagram showing how the PCR cassette works when it is placed on the heaters of QuantuMDx's prototype. Panel C: A used PCR cassette.

2.3.2. Tube based PCR

2.3.2.1. Positive and negative controls for PCR reactions performed on QuantuMDx's PCR cassettes

Negative controls for PCR consisted of adding an aliquot of the PCR mix prior to the addition of DNA, in a 0.2ml PCR tube, and adding dH_2O to dilute the master mix components to the correct concentrations. Positive controls for the cassette PCRs

consisted of placing an aliquot of the PCR mix with DNA in a 0.2ml PCR tube. Both positive and negative controls were run on a MultiGene II Thermocycler using the program found in Table 2.1.

PCR program	Temperature °C	Time	cycles
pre denaturation (initiation)	95	5min	1
Stage 1 (denaturation)	92	30sec	30
Stage 2 (Annealing)	56	30sec	
Delay (Post extension)	56	2min	1

Table 2.1: Thermocycler program used to amplify positive and negative controls for the PCR cassette experiments.

2.3.2.2. Amplicon production for sequencing based MSI detection

PCR reactions for the purpose of creating amplicons for sequencing on the Illumina MiSeq were created using the high fidelity Herculase II Fusion DNA polymerase (Agilent, Santa Clara, CA, USA). PCR reactions were carried out in a total volume of 25µl consisting of 17.25µl H_2O, 5µl 5x Reaction buffer, 0.25µl dNTP mix, 0.25µl polymerase, 0.63µl (10uM) forward primer, 0.63µl (10uM) reverse primer, 1µl DNA (10-40ng depending on DNA quality). PCR amplification was carried out on a Bio-Rad T100 Thermal Cycler (Bio-Rad, Hercules, CA, USA) according to the program found in Table 2.2.

PCR program	Temperature °C	Time	cycles
pre denaturation (initiation)	95	2min	1
Stage 1 (denaturation)	95	20sec	35* 28**
Stage 2 (Annealing)	58	20sec	
Stage 3 (Extension)	72	30sec	
Delay (Post extension)	72	3min	1
Hold	4	hold	1

Table 2.2: Primary PCR thermocycler program to produce amplicons for sequencing based MSI detection. * Number of Cycles used for DNA obtained from FFPE tissues. ** Number of cycles used for DNA obtained from fresh or fresh frozen tissue.

2.4. DNA quantification

2.4.1. Nanodrop assay

For quantification of DNA concentration, 1.5µl of DNA sample was loaded onto a Nanodrop ND-2000 Spectrophotometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA), which had been pre blanked with the same DNA suspension buffer used to elute the DNA samples being measured. Purity of DNA was measured by taking the A260/280 ratio.

2.4.2. Picogreen assay

Double stranded DNA concentration was measured using the Quant-iT™ PicoGreen® dsDNA Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA) according to manufacturer's recommendations. Briefly, a 1x PicoGreen working solution was prepared by diluting the 200x stock Invitrogen Quant-iT™ PicoGreen® solution in 1x Tris-EDTA (TE) buffer. Samples were measured in triplicate where possible and the standard curve was prepared in duplicate on each plate. Each reaction volume totalled 100µl. Absorbance readings were taken using a Fluoroskan Ascent FL (Thermo Fisher Scientific, Waltham, Massachusetts, USA). The Fluoroskan Ascent FL program included shaking the plate before absorbance readings were taken. The averages of the standard curves were plotted in Microsoft Excel and the linear regression equation obtained from this curve was used to convert the absorbance readings of the samples into DNA concentrations.

2.4.3. Qubit DNA quantification

DNA quantification using a Qubit 2.0 Fluorometer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) was performed using the Invitrogen dsDNA HS Assay Kit and Invitrogen dsDNA BR Assay Kit (Thermo Fisher Scientific, Waltham, Massachusetts, USA) according to manufacturer's instructions.

2.4.4. Bioanalyser

All Bioanalyser experiments were performed using the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA, USA). Bioanalyzer High Sensitivity Assays were performed according to the instructions in the Agilent High Sensitivity DNA Kit Quick Start Guide.

2.4.5. QIAxcel

The QIAxcel was used to quantify amplicons generated for the second and third MiSeq sequencing runs. Capillary Gel electrophoresis was performed on a QIAxcel System (Qiagen, Venlo, Netherlands) using a QIAxcel DNA Screening Kit (2400) (Qiagen, Venlo, Netherlands) according to manufacturer's protocol.

2.5. PCR product visualisation

2.5.1. Gel preparation and electrophoresis

Gels consisted of 2% Invitrogen E-Gels (Thermo Fisher Scientific, Waltham, Massachusetts, USA) or Gels made up by dissolving 1.5g SeaKem® LE Agarose (Lonza, Basel, Switzerland) in 100ml of 1% tris-acetate-EDTA (TBE) using a microwave oven. For TBE gels, 10µl of GelRed Nucleic Acid Gel Stain 10,000x (Biotium Inc, Hayward, CA, USA) was added to allow the visualisation of gels. Samples and ladders were mixed with BlueJuice™ Gel Loading Buffer (Thermo Fisher Scientific, Waltham, Massachusetts, USA) at a concentration of 1x prior to loading onto a Gel. The ladders were either Quick-Load® 2-Log DNA Ladder (0.1-10.0 kb) (New England BioLabs Inc, Ipswich, Massachusetts, USA) or Invitrogen 1kb Plus DNA Ladder (Thermo Fisher Scientific, Waltham, Massachusetts, USA). TBE gels were run at 95 volts until separation of DNA fragments was achieved, and E-Gels were run for 26 minutes on an E-Gel iBase Power System (Thermo Fisher Scientific, Waltham, Massachusetts, USA) before visualisation using a UVP GelDoc-It 310 Imaging System (UVP, Upland CA, USA).

2.6. DNA purification

2.6.1. AMPure magnetic bead purification of PCR product

Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA) PCR product purification was performed manually in accordance with the recommendations found in the Nextera XT DNA Sample Preparation Guide (revision C). Ampure cleanup was used to purify pooled PCR product to be used as input for the Nextera XT library prep, and as part of the Nextera XT library prep procedure.

2.7. Sequencing and fragment analysis

2.7.1. MSI testing using fragment analysis

Amplification of MSI markers was performed using a MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, United States of America). The PCR amplification mix was produced according to the manufacturer's specifications. The PCR reactions were run on a SensoQuest Labcycler Thermocycler (SensoQuest GmbH, Göttingen, Germany) using the program found in Table 2.3).

PCR program	Temperature °C	Time	cycles
pre denaturation (initiation)	95	11min	1
pre denaturation (initiation)	96	1min	1
Stage 1 (denaturation)	94	30sec	10
Stage 2 (Annealing)	58	ramp: 0.53°C/sec, hold: 30sec	
Stage 3 (Extension)	70	ramp: 0.24°C/sec, hold: 1min	
Stage 1 (denaturation)	90	30sec	22
Stage 2 (Annealing)	58	ramp: 0.53°C/sec, hold: 30sec	
Stage 3 (Extension)	70	ramp: 0.24°C/sec, hold: 1min	
Delay (Post extension)	60	30min	1
Hold	4	hold	1

Table 2.3: Thermocycler program used for the amplification of MSI markers from the Promega MSI Analysis System kit.

Fragment analysis was performed on an ABI PRISM 3130xl Genetic Analyzer (Life Technologies, Carlsbad, CA, United States of America) using 11µl Hi-Di

formamide, 1µl ILS 600, and 2µl PCR product per well. The MSI analysis was carried out using the GeneMapper Software (Life Technologies, Carlsbad, CA, United States of America). The interpretation of all fragment analysis traces were checked by Ottie O'Brien (Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust).

2.7.2. Next generation sequencing on the Illumina MiSeq platform

Amplicons to be sequenced on the Illumina MiSeq platform (Illumina, San Diego, CA, United States of America) were generated using the PCR protocol listed in section 2.3.2.2. For the first MiSeq run, amplicons were run on a TBE gel to confirm that amplicons had the desired length. A selection of 5-7 amplicons of varying band intensities from each gel were quantified using a Qubit 2.0. The concentrations of the remaining amplicons were estimated based on the band intensities seen on the gel images by comparing them to the bands of known concentration. For subsequent MiSeq runs all amplicons were quantified on a QIAxcel System. For each sample, amplicons were pooled at a roughly equal concentration prior to PCR cleanup using Agencourt AMPure XP. After AMPure clean up all amplicon pools were quantified using a Qubit 2.0 Fluorometer and diluted to a concentration of ~0.2ng/µl, which is the recommended input DNA concentration for the Illumina Nextera XT kit.

2.7.2.1. Nextera XT adapter and barcoding

Library prep was performed using the Nextera XT DNA Library Prep kit (Illumina, San Diego, CA, United States of America). Briefly, sequencing adaptors were added using an enzymatic tagmentation step followed by a PCR reaction to add sample specific indexes. Following PCR cleanup, a magnetic bead based normalisation step was used to bring all samples to the same concentrations prior to pooling all samples ready for sequencing.

The library prep was performed according to the manufacturers protocol (Nextera XT DNA Sample Preparation Guide, revision C) with the exception of the following: The PCR plates and seals consisted of 96-Well PCR Plates, Non-Skirted Cuttable (Starlab, Milton Keynes, UK) and Aluminium StarSeal (Starlab, Milton Keynes, UK). A magnetic ring plate was used instead of the recommended magnetic plate. For the first MiSeq run a Vortex Mixer – WIZARD (VELP Scientifica, Usmate Velate MB, Italy) set to a speed

of 1700rpm was used instead of the recommended plate shaker. For the first MiSeq run the Bioanalyser was used to check samples after the AMPure cleanup step. For subsequent MiSeq runs the QIAxcel was used instead of the Bioanalyser.

2.7.2.2. MiSeq sequencing

Amplicons prepared for sequencing using the Nextera XT Library Prep kit were sequenced on the Illumina MiSeq platform using a MiSeq Reagent kit V2 (500 cycles) (Illumina, San Diego, CA, United States of America) for the first MiSeq run, and a MiSeq Reagent kit V3 (600 cycles) (Illumina, San Diego, CA, United States of America) for subsequent MiSeq runs. For the first MiSeq run, the sequenced library was made up of 24µl PAL (Pooled Amplicon Library), 546µl HT1, and 30µl 12.5pM PhiX creating a 12.5pM library with a 5% PhiX spike-in. For subsequent MiSeq runs the sequenced library consisted of 35µl PAL (Pooled Amplicon Library), 535µl HT1, and 30µl 20pM PhiX creating a 20pM library with a 5% PhiX spike-in.

Sequencing was performed using targeted resequencing and the PCR amplicon workflow with paired end read sequencing (251 cycles for both read 1 and read 2) and adaptor trimming. Sequencing was performed on the Illumina MiSeq to an average read depth of >10000 per amplicon.

2.8. Informatics

2.8.1. Literature review and homopolymer selection

To identify short homopolymers previously shown to be unstable in CRCs, a systematic literature review was carried out using PubMed (http://www.ncbi.nlm.nih.gov/pubmed), Google Scholar (http://scholar.google.co.uk) and the Selective Targets in Human MSI-H Tumorigenesis Database (SelTarBase, http://www.seltarbase.org). 6bp -16bp homopolymers were identified for potential inclusion in this study. Repeats were checked for common polymorphisms and neighbouring SNPs using the UCSC Genome Browser (Kent et al., 2002) and dbSNP (build 173) (Sherry et al., 2001). Repeats containing a known repeat length polymorphism were excluded from the study with the exception of polymorphisms where no frequency

data was available. These were not excluded because the polymorphisms were assumed to be very rare or not validated. To facilitate the investigation of allelic bias of MSI, homopolymers within 80bp of SNPs with a minor allele frequency between 0.05 – 0.95 were selected where possible.

2.8.2. Primer design

Primers were designed using Primer3 (Rozen and Skaletsky, 2000) or manually if Primer3 returned no suitable oligos. Primers designed manually had a Tm of 57ºC-60ºC. The Tm was calculated as follows: Tm = 4×(G+C)+2×(A+T). Primers were designed to create amplicons of ~300-350bp. All primers were checked for common SNPs using SNP Check (https://ngrl.manchester.ac.uk/SNPCheckV2/snpcheck.htm), off target binding using BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi) or BLAT (Kent, 2002), and appropriate melting temperatures and absence of secondary structures using OligoCalc (http://www.basic.northwestern.edu/biotools/oligocalc.html) or Primer3. The primers were produced by either Metabion (Metabion International AG, Steinkirchen, Germany) or Biobasic (Bio Basic Inc., Markham, Canada) and purified by desalting. A list of all primers can be found in Table 9.1 in the Appendix.

2.8.3. Sequence data files

The sequence data analysed in this thesis consisted of amplicon sequence data, whole genome sequences, and one exome sequence.

Amplicon sequence data were generated on the Illumina MiSeq as described in section 2.7.2. The data were retrieved from the MiSeq in the form of FASTQ files.

Whole genome sequences consisting of MSI-H colorectal cancers, matched normals, and MSS cancers were obtained from The Cancer Genome Atlas (TCGA) group in the form of BAM files (Data access request [#17798-1] approved by the TCGA Data Access Committee). The samples used consisted of 12 MSI-H tumours, 12 MSS tumours and matched normal tissue for 11 of the MSI-H tumours (see Appendix Table 9.3 for details of the samples used). The whole genome sequences generated by the TCGA had a ~3-4 fold sequence coverage for each sample (Cancer Genome Atlas Network, 2012).

FASTQ files for one exome sequence were provided by Dr Mauro Santibaez-Koref, (Institute of Genetic Medicine, Newcastle University). This sequence data were derived from the normal tissue of a patient unaffected by Lynch Syndrome.

2.8.4. Producing scripts

The text editor GNU nano 2.0.9 (Allegretta) was used to write and edit the shell and Perl scripts used to perform the work detailed in this thesis.

2.8.5. Visualization of sequence alignments

The Integrative Genomics Viewer (IGV) (Robinson et al., 2011) was used to visualise the aligned reads from BAM files.

2.8.6. DNA sequence analysis pipeline

2.8.6.1. Sequence alignment

BAM files, obtained from The Cancer Genome Atlas group, were converted to FASTQ files using bam2fastq (version 1.1.0) (bam2fastq software [http://gsl.hudsonalpha.org/information/software/bam2fastq]).

For sequence alignment the Burrows–Wheeler Aligner (BWA) (version 0.6.2) (Li and Durbin, 2009) was used. Input files consisted of FASTQ files and output files consisted of SAM files. Reads were aligned to the human genome sequence build GRCh37/hg19.

The conversion of SAM files to BAM files as well as indexing and sorting of BAM files was achieved using Samtools (version 0.1.18) (Li et al., 2009). Samtools was also used to create the pileup file needed for variant calling with VarScan.

Prior to indel calling using GATK duplicate sequences were removed from sorted and indexed BAM files using Picard (version 1.75) (PICARD [http://picard.sourceforge.net]).

2.8.6.2. Variant calling

The pileup2indel function from VarScan (version 2.2.2) (Koboldt et al., 2009) was used with default parameters. Variant calling was performed using the pileup file created by Samtools as an input file.

Dindel (version 1.01) (Albers et al., 2011) was run using default parameters. The sorted and indexed BAM file, processed by Samtools, was used as the input file.

Prior to variant calling from the whole genome sequences using GATK, sorted and indexed BAM files with duplicates removed were merged into a multisample BAM file using GATK (version 2.2.9) and realignment around indels was performed. This multisample BAM file was then used as the input file for variant calling with GATK. For the Variant caller comparison a sorted and indexed BAM file, processed by Samtools, was used as the input file for GATK.

The GATK UnifiedGenotyper (version 2.2.9) was first tried using default parameters, however due to the low complexity of the data, maxGaussians had to be lowered to 4 for the indel error model. The SNP error model was run with a maxGaussians value of 6 as recommended. For the indel caller comparison, the GATK HomopolymerRun annotation was used to allow the identification of variants in homopolymers. For the analysis of colorectal cancer whole genome sequences, the GATK UnifiedGenotyper was used to produce a raw variant call file annotated using the TandemRepeatAnnotator annotation for the ease of identifying indels in mononucleotide repeats. All homopolymers with known polymorphisms as of dbSNP (version 137, hg19) were also annotated.

An in-house variant caller, Concordant Overlapping Paired Reads Caller (COPReC), was designed and run by Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University). COPReC only reports indels in concordant overlapping reads. This caller uses SAM files as input files. The output from this variant caller consists of a table for each homopolymer and SNP combination, which contains the paired end read counts for each recorded homopolymer length, and the base at the SNP site for each read that contains both homopolymer and SNP. COPReC was used to analyse all of the amplicon sequence data.

2.8.6.2.a Scripts for the variant calling pipelines

Shell scripts for Varscan, Dindel and GATK can be found on the supplementary CD (see the folder "Variant Calling/Indel Caller Comparison").

Shell scripts for the GATK pipeline used for analysing whole genome sequences can be found in the folder "Variant Calling/TCGA Analysis" on the accompanying CD.

2.8.7. Data manipulation and analysis using in house Perl scripts

The following Perl scripts were written to parse data, and/or analyse read frequencies and indel size distributions (for details, see text in the result sections). All scripts are included as Supplementary Information in the folder "Sequence Analysis" on the accompanying CD.

2.8.7.1. Comparison of the variant callers Dindel and GATK

Dindel_GATK_compare.pl: Counts and lists the indels in homopolymers >7bp that the Dindel and GATK VCF files have in common, as well as counting and listing the indels that are unique to a VCF file using the chromosome and position data to determine if two indels are the same.

2.8.7.2. Analysis of indel frequencies in MSI-H samples and controls using whole genome sequence data

Perl_SelectVariants_RPA_RU.pl: Extracts all indels in homopolymers of 7-12bp from a VCF file created by the GATK UnifiedGenotyper and annotated using GATK's TandemRepeatAnnotator.

TCGA_AnnotationSelector.pl: Filters out unnecessary annotations from the file generated by Perl_SelectVariants_RPA_RU.pl, creating a smaller output file containing the variants of interest and useful annotations.

REF_ALT_AnnotationSelector.pl: Calculates the number of reference and variant reads in each sample for each homopolymer. This script uses the output of TCGA_AnnotationSelector.pl as an input file.

The output from REF_ALT_AnnotationSelector.pl was opened in Microsoft Excel where reference and variant reads for each sample group (MSI-H samples, MSS samples and matched normal for the MSI-H samples) were added up and the percentage of reference and variant reads for each group was calculated. All reads from each group were combined before analysis because of the low pass nature of the sequence data. The percentage of variant reads was rounded up to the nearest 5% prior to plotting graphs showing the number of homopolymers with different variant read frequencies. Separate graphs were produced for each homopolymer length, and G/C and A/T homopolymers were also analysed separately. All repeats with common polymorphisms (dbSNP version 173) were removed prior to any analysis and the creation of graphs.

2.8.7.3. Analysis of indel sizes in different homopolymer lengths using whole genome sequence data

TCGA_AnnotationSelector_ForIndividIndelPercentages.pl: Using the output of Perl_SelectVariants_RPA_RU.pl as an input file, selects useful annotations and adds a read count of zero for samples that have no reads spanning a homopolymer.

REF_ALT_AnnotationSelector_Percentages.pl: Using the output of TCGA_AnnotationSelector_ForIndividIndelPercentages.pl as an input file, calculates the number of reference alleles and number of reads corresponding to each indel size for each sample group (MSI-H samples, MSS samples and matched normal for the MSI-H samples). Then calculates the percentages of reads corresponding to each indel size.

IndelGaps_AnnotationSelector.pl: Adds the size of each indel to the end of the lines in the file produced by REF_ALT_AnnotationSelector_Percentages.pl.

IndelSizeSelector.pl: Extracts homopolymers of user specified length from the file generated by IndelGaps_AnnotationSelector.pl, so that different lengths of homopolymer can be analysed separately.

HomopolymerCount.pl: Counts the number of homopolymers with indels of each size so this can easily be plotted in Microsoft Excel. Separate counts are done for A/T homopolymers and G/C homopolymer. The script uses the output of IndelSizeSelector.pl as input.

HomopolymerCount_percent.pl: A variation of HomopolymerCount.pl, which allows thresholds to be set so that only indels with a frequency that passes the threshold will be counted. The script in Supplementary Information currently has a threshold set to 10%. This script uses the output of IndelSizeSelector.pl as input.

2.8.7.4. Annotating neighbouring SNPs

AnnotateCloseSNPs.pl: Using a tab delimited text file as input, annotates any SNPs from dbSNP (version 137, hg19) (Sherry et al., 2001) within 30bp of the start of repeats.

2.8.7.5. Analysis of allelic bias

AlleleicBias_IndividualIndels.pl: Using output from COPReC, identifies repeats that are heterozygous for a neighbouring SNP and calculates the percentage of reads corresponding to each variant repeat length, and reference repeat length, for both alleles. Repeats are defined as heterozygous if there are ≥100 paired end reads spanning both SNP and repeat for each allele, and one allele does not have less than 10% of the total read count.

ChangeIndelOrder_AllelicBias.pl: Uses the output from AlleleicBias_IndividualIndels.pl to print out a table containing the fractions of variant and reference reads in descending order of repeat length for each homopolymer to allow for the easy creation of graphs in Microsoft Excel.

2.9. Statistical analyses

2.9.1. Fisher's exact tests

The following Perl scripts were written to parse data, and perform two-tailed Fisher's exact tests (for details, see text in the result sections). All scripts are included in Supplementary Information on the accompanying CD in the folder "Fisher's Exact Test".

FisherTest_AllDeletions.pl: Using output generated by COPReC, this script identifies repeats that are heterozygous for a neighbouring SNP and performs a two-tailed Fisher's exact test to determine if the fraction of deletions are significantly different between the two alleles. Repeats are defined as heterozygous if there are ≥100 paired end reads spanning both SNP and repeat for each allele, and one allele does not have less than 10% of the total read count. This script calculates the number of reads that contain a deletion and the number of reads that do not contain a deletion for each allele, and then uses these values to perform a Fisher's exact test. The Fisher's exact test calculations were performed using an external module integrated into the Perl script (Pedersen T., https://metacpan.org/pod/Text::NSP::Measures::2D::Fisher::twotailed).

FisherTest_IndividualIndels.pl: Using output generated by COPReC, this script identifies repeats that are heterozygous for a neighbouring SNP and performs a two-tailed Fisher's exact test to determine if the fraction individual indels is significantly different between the two alleles. Repeats are defined as heterozygous if there are ≥100 paired end reads spanning both SNP and repeat for each allele, and one allele does not have less than 10% of the total read count. For each allele this script categorises reads as; reads containing the indel size under investigation, or reads that do not contain the indel size under investigation. Next, this script calculates the number of reads in each category for both alleles and uses this as the input in the Fisher's exact test 2 x 2 contingency table. The two-tailed Fisher's exact test calculations were performed using an external open source module integrated into the Perl script (Pedersen T., https://metacpan.org/pod/Text::NSP::Measures::2D::Fisher::twotailed).

2.9.2. Match probability calculations

A match probability calculation was used to determine if there had been a sample mix-up for the U303 tumour sample. For the calculations the NorthGene (NorthGene Ltd, Newcastle upon Tyne, UK) Caucasian allele database and an Fst of 2% were used. Below are the calculations for size bias, which take into account that the database used is only an estimation of the allele frequencies in the population.

Heterozygotes:

$$\frac{(N \times Allele\ Frequency)+2}{N+4}$$

Homozygotes:

$$\frac{(N \times Allele\ Frequency)+4}{N+4}$$

N=Size of allele database

Next, the match probability frequencies were calculated using the equations bellow.

Heterozygotes:

$$\frac{2(\theta+(1-\theta)\,fp)(\theta+(1-\theta)fq}{(1+\theta)(1+2\theta)}$$

Homozygotes:

$$\frac{(2\theta+(1-\theta)fp)(3\theta+(1-\theta)fq}{(1+\theta)(1+2\theta)}$$

θ=0.02 (the Fst value)
fp= the size bias for allele 1
fq= the size bias for allele 2

The match probability frequencies for all markers were multiplied together. 1 divided by the product of the match probability frequencies generates the final match probability figure. The chance of obtaining a match if the sample originated from someone other than, and unrelated to, the person being tested is 1 in whatever the final match probability figure is. If there are no mismatches between tissues and the match probability figure is over 1 billion, then it can be concluded that both samples belong to the same person.

2.9.3. Optimising thresholds for differentiating between MSI-H and MSS samples

Different deletion frequencies of between 0 and 1 at increments of 0.001 were used as potential thresholds. Initially, the deletion frequency that gave the lowest number of errors was identified and used as the threshold. If the lowest number of errors could be obtained at more than one deletion frequency, then the lowest of these deletion frequencies was used as the threshold. Frequency of errors was calculated using the following equation:

$$FPR \times Nbr\ MSS + FNR \times Nbr\ MSI\text{-}H = Number\ of\ errors$$

FPR= false positive rate
FNR= false negative rate
Nbr MSS= number of MSS tumours
Nbr MSI-H = number of MS-H tumours

The false positive rate was defined as the fraction of repeats with a deletion frequency of or above the threshold in the MSS samples, and the false negative rate was defined as the number of repeats with a deletion frequency below the threshold in MSI-H samples.

The weighting of different errors was also used to adjust the thresholds. A false positive error was weighted as 1.5x and 2x worse than a false negative error. This was achieved by multiplying the number of false positives by the weighting before adding up false positives and false negatives. The deletion frequency with the lowest number was used as the threshold. If the lowest number could be found at several different deletion frequencies, then the lowest of these deletion frequencies was used as the threshold. The weighting of false positive errors needed to achieve a deletion frequency threshold where there would be no false positive errors was also identified. The equation used can be found below:

$$W_{FP} \times FPR \times Nbr\ MSS + FNR \times Nbr\ MSI\text{-}H = Weighted\ errors$$

FPR= false positive rate
FNR= false negative rate
Nbr MSS= number of MSS tumours
Nbr MSI-H = number of MS-H tumours
W_{FP} = weighting of false positive errors

For each homopolymer size there would be a different deletion frequency that gave the lowest number of errors or weighted errors. For each set of deletion frequencies, the number of repeats classed as unstable for each tumour was calculated and plotted. Each set of deletion frequencies was also used to predict how many errors there would be for each repeat size given a panel of tumours, which conform to a division of 85% MSS tumours and 15% MSI-H tumours. This was achieved by multiplying the false positive rate by 85 to obtain the percentage of false positive errors and multiplying the false negative rate by 15 to obtain the percentage of false negative errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours.

All graphs and calculations in the section were drawn using Microsoft Excel.

2.9.4. Binomial classification

Area under the receiver operating characteristic curve (AUC) calculations, and the sensitivity and specificity curves were produced by Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University) using the statistical computing environment R (R Core Team).

Chapter 3. Assessing next generation sequencing of known short homopolymers in microsatellite unstable tumours

3.1. **Introduction and Aims**

3.1.1. Introduction

3.1.1.1. MSI in long and short homopolymers

Currently, fragment analysis is used for MSI testing. Fragment analysis is a capillary electrophoresis based method, which allows the measurement of DNA fragment lengths. The markers that are most frequently used for MSI testing by fragment analysis today are BAT26 $(A)_{26}$, BAT25 $(A)_{25}$, MONO-27 $(A)_{27}$, NR-21 $(A)_{21}$, and NR-24 $(A)_{24}$ (Boyle et al., 2014). Mononucleotide repeats of these lengths have the advantage that they are highly susceptible to slippage in tumours with mismatch repair defects (Umar et al., 2004). This means that a panel of as little as five markers is enough for an MSI test. For example using the mononucleotide repeats BAT-25, BAT-26, NR-21, NR-24 and NR- 27 with two unstable markers as the criteria for classifying a sample as MSI-H, Goel et al. (2010) achieved a sensitivity of 95.6% and a positive predictive value of 100% for a panel of 114 mismatch repair deficient tumours and 99 mismatch repair proficient tumours. Suraweera et al. (2002) have achieved a 100% sensitivity and specificity using the same panel of repeats on a different panel of tumours. On the other hand, the drawback of using microsatellites of these lengths is that they are also highly unstable in vitro (Fazekas et al., 2010). Due to the lengths of these mononucleotide repeats, commercially available polymerases are unable to faultlessly replicate them. The result of this is seen as a stutter pattern on the fragment analysis traces, even in repeats amplified from MSS tumours. For this reason, the repeat lengths used in MSI tests today would not be ideal for a sequencing based assay.

In the late 1990s and early 2000s there was evidence that short mononucleotide repeats were susceptible to MSI, but at a much lower frequency than the longer repeats used in current fragment analysis tests (Vilkki et al., 2002). Although short repeats could be used in an MSI test, there is data available which indicates that repeat length affects

stability and error rate. This should be considered when selecting mononucleotide repeats for an MSI test.

The susceptibility of different homopolymer lengths to MSI has been studied previously, which gives some indication of which repeat lengths to would be appropriate for a next generation sequencing based MSI test. Microsatellites as short as 7-13bp have been reported as being susceptible to MSI (Sammalkorpi et al., 2007, Vilkki et al., 2002, Woerner et al., 2003). Vilkki et al. (2002) screened fourteen intronic homopolymers of 6bp-9bp in up to 93 MSI-H tumours and identified instability in six of these markers. These markers consisted of one 7bp repeat showing instability in two tumours (2/93), two 8bp repeats showing instability in 2 tumours (2/81) and 5 tumours (5/93) respectively, and three 9bp repeats showing instability in 4 tumours (4/84), 5 tumours (5/88) and 4 tumours (4/93) respectively. They also screened eight coding homopolymers and found instability in one 9bp repeat for 22.9% of the MSI-H tumours analysed. Woerner et al. (2003) analysed 181 homopolymers of lengths 4bp-13bp in colorectal cancers and found 15 repeats with instability in over 40% of the MSI-H tumours analysed. These repeats consisted of two 8bp repeats, two 9bp repeats, six 10bp repeats, four 11bp repeats and one 13bp repeat.

Sammalkorpi et al. (2007) studied 114 intergenic repeats in up to 30 MSI-H tumours to assess their instability using Sanger sequencing. The repeats were 6-10bp in length. Repeats were classed as unstable if a variant with >10% of the relative fluorescent units compared to the wild type allele was detected. Only four out of the twenty-nine 6bp repeats showed instability for at least one sample, suggesting that 6bp repeats are not very susceptible to MSI. For the 7bp and 8bp repeats thirteen out of twenty-five repeats were classed as unstable and eighteen out of twenty-two were classed as unstable respectively. On average, the 7bp repeats were unstable in 3% of the samples and the 8bp repeats were unstable in 13% of the samples. For the 9bp repeats surveyed, all sixteen showed MSI in at least one tumour and on average repeats showed instability in 29% of the samples. Only one of the twenty-two 10bp repeats was not unstable in any sample and on average 10bp repeats were unstable in 50% of the samples. This data indicates that MSI rates increase with the length of the homopolymer and there are large differences in instability rates for homopolymers of different lengths.

Unfortunately, PCR and sequencing error is also expected to increase with homopolymer length. Clarke et al. (2001) found that a *Thermus aquaticus* based

polymerase (AmpliTaq) could correctly amplify a mononucleotide repeat of 9bp under standard PCR conditions, but for a mononucleotide repeat of 11bp there was a 10% error rate measured using sequencing individual clones after subcloning of the PCR product, and for a 13bp mononucleotide repeat there was a 66% error rate. Fazekas et al. (2010) showed that using the polymerase Herculase II Fusion improved replication of mononucleotide repeats to the point where DNA replication was nearly error free after 35 PCR cycles for homopolymers up to 13bp in length. After 13bp the error rate increases with homopolymer length. In theory, therefore, a panel of short homopolymers could be used to create a MSI test that is compatible with sequencing. However, because of the reduced susceptibility of 7-13bp homopolymers to MSI, a much larger panel of repeats would be need in an MSI test.

Generally, the length of a microsatellite and the susceptibility of a microsatellite to MSI are positively correlated (Sinicrope and Sargent, 2012, Vilkki et al., 2002, Woerner et al., 2003). On the other hand, the rate of PCR and sequencing error also increases with repeat length (Clarke et al., 2001, Fazekas et al., 2010). Because the optimal trade off between error rates and susceptibility to MSI has not been determined for a sequencing based MSI assay, more empirical data would be required to determine the appropriate repeat length and number of markers to use. PCR errors have the potential to occur during amplicon generation, library prep (unless a PCR free library prep is used), during cluster formation prior to next generation sequencing and during the sequencing by synthesis reaction. To develop a high throughput sequencing based MSI test it would, therefore, be necessary to investigate rates of instability and rates of error on the chosen sequencing platform, to determine both the optimal size and number of repeats to use, and also determine criteria for distinguishing between MSI-H and MSS samples.

3.1.1.2. Next generation sequencing of short homopolymers

Despite the discovery of homopolymers that were susceptible to MSI and short enough to sequence in the late 1990s and early 2000s a sequencing based MSI assay was not implemented because the sequencing technology at the time (Sanger sequencing) meant that it was impractical and not economically viable compared to fragment analysis. One reason for this is the number of amplicons that would need to be created and sequenced individually. With recent improvements in sequencing, and huge reduction in sequencing costs, it is now possible to consider high throughput screening approaches to

test for microsatellite instability. The monomolecular nature of next generation sequencing also provides a quantitative approach to measuring insertion and deletion (indel) frequencies, which would be useful for creating a sequencing based MSI test.

The first paper to illustrate the potential use of next generation sequencing for detecting MSI was a high throughput sequence analysis of colorectal cancers performed by the Cancer Genome Atlas Network (2012). Their main focus was mutation detection and classification of subtypes of CRCs. However, they did also analyse MSI using exome data. Using their pipeline for variant calling they were initially unable to distinguish between MSI-H tumours and controls. These difficulties were due to low mutation frequencies being detected in mononucleotide repeats in normal tissue. These mutated reads they concluded were most likely derived by errors occurring from PCR amplification. They therefore focused their analysis of MSI on a handful of mononucleotide repeats in selected genes. Manual inspection of the reads from the MSI-H tumours showed that some monunucleotide repeats had a higher variant read frequency compared to what was seen in the matched normal tissue. A variant read frequency difference of 20% between tumour and matched normal was defined as the cut off for calling a marker unstable. Using this criteria, mononucleotide repeats in 28 genes were analysed manually, the results of which showed, that tumours with *MLH1* silencing had a 50 fold higher rate of frameshift mutations in these genes compared to tumours with a mutation rate of $\leq$ 12 per 10^6 bases. This showed, as a proof of principle, that microsatellite instability in short repeats was detectable using next generation sequencing. Prior to the start of this project, this was the only work that had been conducted on MSI using next generation sequencing.

3.1.1.3. Sequencing platforms

For any sequencing based MSI test selecting a sequencing technology which can cope well with long homopolymers is important. Sequencing using chain termination would be more appropriate than a sequencing technology such as 454 sequencing or IonTorrent where the number of bases in a homopolymer is inferred by signal intensity. SOLiD sequencing was discounted because of the aim to develop an MSI test, which would ultimately be compatible with the sequencing technology being developed by the company QuantuMDx. QuantuMDx will be using a sequencing by synthesis approach. SOLiD sequencing on the other hand, uses a sequencing by ligation approach using di-

base probes. It would be easier to transfer a test developed using a similar sequencing technology to the QuantuMDx platform. Using SOLiD sequencing would also have the disadvantage that the sequencing would have to be outsourced.

Illumina sequencing would be the most appropriate because of its low error rate for homopolymers. Minoche et al. (2011) have reported average error rate for Illumina sequencing as 0.002% for 2bp homopolymers, rising to ~2% for 17bp homopolymers. Illumina sequencing has also shown promise in the paper by the Cancer Genome Atlas Network (2012), and was therefore the first choice for investigating MSI in homopolymers. The MiSeq should be an appropriate Illumina platform because it would give a sufficient read depth for investigating the suitability of using short homopolymers to detect MSI. Another advantage is that there is a MiSeq located at the Centre for Life allowing the sequencing to be performed locally hence avoiding the extra cost and delay in outsourcing the sequencing.

3.1.1.4. Allelic distributions of MSI

A sequence based approach may also enable the allelic origin of instability to be investigated through the analysis of single nucleotide polymorphisms (SNPs) located close to the repeat. Including these SNPs means that in heterozygous individuals it will be possible to identify which allele homopolymer length variants belong to on reads that span both SNP and homopolymer. It should therefore be possible to determine if a specific indel is more prevalent on one allele than the other. If microsatellite instability is caused by random errors in microsatellite replication, which are not corrected by a cells compromised MMR system, then instability events are unlikely to affect both alleles of a short homopolymer. This is because short homopolymers have a low susceptibility to replication errors in vivo and two errors in the same position on both chromosomes are therefore less likely to occur. SNPs may therefore be useful as it may provide a method by which instability could be distinguished from error, as PCR or sequencing error is unlikely to be allele specific because this type of error is likely to occur several times during a PCR reaction and both alleles will be susceptible.

3.1.2. Aims

The initial high throughput genome sequencing of CRC patients had established the potential for a sequence based MSI test, and the plummeting cost of sequencing suggests that it may be economically viable. Sanger based analyses of individual repeats had established that extensive variation in stability and error rates existed. The initial aim of this work was to investigate the suitability of the MiSeq platform for MSI detection using known variable short repeats. Specifically, this work aimed to:

- Determine the optimal homopolymer length for use in a sequencing based assay
- Determine how easy it is to distinguish between MSI-H samples and controls using short homopolymers.
- Evaluate the feasibility of a sequencing based MSI test.
- Evaluate the feasibility of using SNPs to distinguish between alleles

3.2. **Results**

3.2.1. Error frequencies for homopolymers in Illumina data

First, to check that next generation sequencing is capable of accurately sequencing short homopolymers, alignment files produced from one control exome were analysed. The aim was to examine how sequencing errors and PCR artefacts are influenced by homopolymer length. To identify unstable homopolymers of a suitable length for this initial assessment the Selective Targets in Human MSI-H Tumorigenesis Database (SelTarBase, http://www.seltarbase.org) was screened to identify homopolymers of lengths 7-16bp. SelTarBase is a database containing microsatellites that have shown instability in MSI-H tumours. The selected 7-16bp homopolymers were checked for common polymorphisms using the UCSC Genome Browser (Kent et al., 2002). If no polymorphisms were listed then the homopolymers were assumed to be monomorphic. A list of the 29 monomorphic homopolymers that were selected is presented in Table 3.1.

Size bp	Base	Name	Variant Position	Read Depth
7	C	Axin2	chr17:63532585-63532591	9
7	C	XYLT2	chr17:48433967-48433973	33
7	C	RFX5	chr1:151318741-151318747	11
8	A	ACVR2	chr2:148683686-148683693	54
8	C	BAX	chr19:49458971-49458978	74
8	C	BRD1	chr22:50193070-50193077	1
8	A	CCKBR	chr11:6292451-6292458	52
8	A	LARP7	chr4:113570754-113570761	89
8	C	LIMK2	chr22:31672777-31672784	10
8	C	MAPRE3	chr2:27248517-27248524	18
8	C	MYH11	chr16:15802687-15802694	24
8	A	MYO1A	chr12:57422573-57422580	89
8	A	PA2G4	chr12:56505302-56505309	152
9	A	C4orf6	chr4:5527116-5527125	37
9	A	CLOCK	chr4:56336954-56336962	125
9	A	TTK	chr6:80751897-80751905	122
9	C	ELAVL3	chr19:11577605-11577613	2
10	A	TGFBR2	chr3:30691872-30691881	149
10	A	RFC3	chr13:34398063-34398072	97
11	A	ASTE1	chr3:130733047-130733057	129
11	A	MRE11A	chr11:94212931-94212941	92
11	A	SLC22A9	chr11:63149671-63149681	134
11	A	TAF1B	chr2:9989571-9989581	34
12	C	MRPL2	chr6:43021977-43021988	9
12	A	PCDHGA12	chr5:140812756-140812805	110
13	A	LGALS3	chr14:55612007-55612019	106
13	A	CCDC88A	chr2:47635524-47635536	124
16	A	FLJ20489	chr12:48174352-48174367	40
27	A	BAT26	chr2:47641560-47641586	8

Table 3.1: Monomorphic homopolymers that were used to investigate levels of sequencing error in Illumina sequencing.

The Integrated Genome Viewer (IGV) (Robinson et al., 2011) was then used to inspect these homopolymers for indels within a single exome sequence from a normal control subject. Reads spanning the homopolymers and at least 5bp either side of the homopolymer were counted. As the microsatellites are assumed to be monomorphic, any deviation from the reference sequence was counted as an error.

Analysis of the homopolymers showed that PCR/sequencing errors do increase with repeat length (see Figure 3.1). This is consistent with what has been reported by (Fazekas et al., 2010). For the 7bp – 10bp homopolymers, less than 3% of reads contained PCR/sequencing errors (see Figure 3.1). Homopolymers of lengths 11bp – 12bp were more prone to PCR and sequencing error, the fraction of reads containing errors being around 10% (see Figure 3.1). For the 13bp homopolymers analysed, 16% of the reads contained errors. At the time when this analysis was being performed, no studies had been published addressing whether it would be possible to distinguish between MSI and artefacts with such high background noise. It was therefore deemed a risk to focus on repeats of this length and longer. For the 16bp repeat analysed only 65% of the reads matched the reference sequence. With such a high error rate it was concluded that it would be very hard to detect indels caused by MSI in homopolymers of this size and longer. For comparison one of the repeats used in a standard fragment analysis, BAT26 (27bp), was also analysed. Only 50% of the reads corresponded to the reference sequence for this long repeat. Because the 7bp – 10bp homopolymers had error rates of less than 3%, it was concluded that microsatellites of these lengths would be possible to type in a MSI assay without much interfering background noise from PCR/Sequencing error.

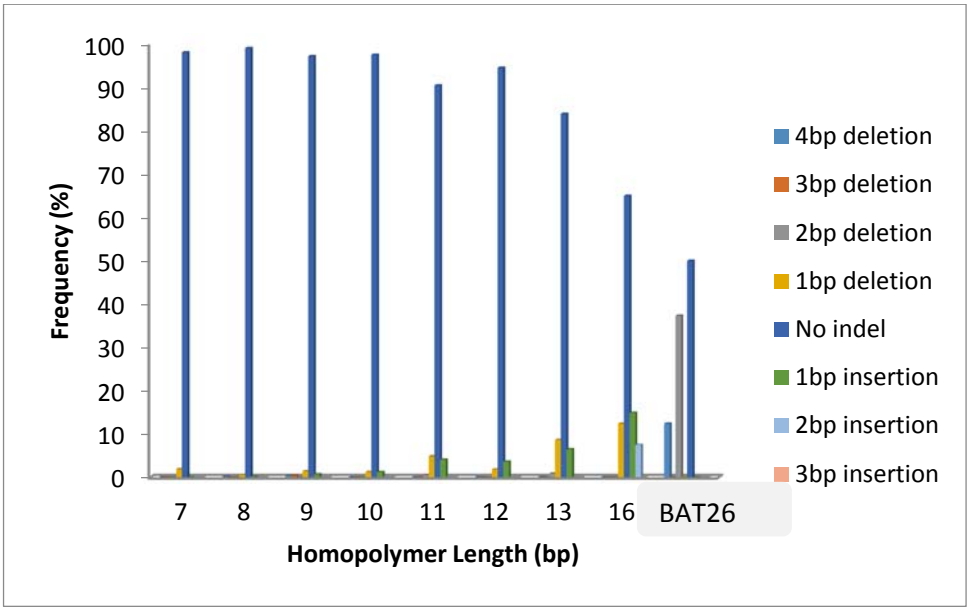

Figure 3.1: Effect of homopolymers size on error rates in Illumina sequencing. This figure shows indel frequencies in monomorphic homopolymers. Indel frequencies increase with homopolymer length indicating that sequencing/PCR errors increase with homopolymer length.

3.2.2. Selecting suitable known homopolymers for MSI identification

Having established that is possible to sequence homopolymers between 7-10bp using Illumina sequencing; these lengths were considered the most promising for use in an MSI test. Current literature and the MSI database SelTarBase (http://www.seltarbase.org) were used to identify short homopolymers (between 7 – 10bp) that have been reported to be affected by microsatellite instability. To facilitate investigation of allelic stability, homopolymers in close proximity to SNPs with a high minor allele frequency were selected where possible, by using the UCSC Genome Browser (Kent et al., 2002) and dbSNP (build 173) (Sherry et al., 2001) to identify suitable SNPs with a minor allele frequency between 0.05 – 0.95 (a minor allele frequency close to 0.5 being preferred). The UCSC Genome Browser and dbSNP (build 173) was also used to exclude any homopolymers with SNPs that could cause a length change of the repeat. Potential repeat length polymorphisms where no frequency data was available were not excluded because they were assumed to be very rare or not validated. Subsequent to the analysis three repeats were found to not conform to the criteria for selecting monomorphic repeats. The three repeats AL590078, SLC4A3, and AL390295 all have SNPs with a high minor frequency where the minor allele creates a length change in the repeat.

Twenty-two homopolymers were identified and primers were produced. These included five homopolymers, one of each size from 11-15bp, which were chosen to see if data from these repeat sizes might be of interest. In total 17/22 repeats analysed had neighbouring SNPs with a minor allele frequency >0.05. Primers were designed to create amplicons ~300bp. This was done so amplicons would be compatible with the requirements of the Nextera XT sample prep kit (Illumina, San Diego, CA, United States of America), and to allow for SNPs and homopolymer to be sequenced together in both the forwards and reverse direction (see methods section 2.8.2). Creating amplicons which allowed overlapping paired end reads meant that it will be possible to further reduce sequencing error by only analysing concordant reads.

Repeat Name	Repeat length (bp)	Repeat Unit	Instability in CRC (%) (SelTarbase release 201307)	SNP minor allele frequency and base (dbSNP build 173)	Repeat length polymorphism minor allele frequency (dbSNP build 173)	Reference
Axin2	7	C	14.4	A: 0.174	none	Thorstensen et al. (2005)
AL590078	8	A	10.7	C: 0.203	0.150	Sammalkorpi et al. (2007)
MX1	8	C	13	A: 0.260	NFD	Kloor (pers. Comm.)
HPS1	8	C	28	G: 0.053 G: 0.052	none	Alhopuro et al. (2012)
IL1R2	8	C	32.3	G: 0.227	none	Alhopuro et al. (2008)
DEPDC2	8	C	35	C: 0.407	none	Alhopuro et al. (2008)
APBB2	8	C	36.6	G: 0.138	NFD	Alhopuro et al. (2008)
SLC4A3	8	C	36.7	A: 0.038	0.038	Woerner et al. (2001)
AC079893	9	A	3.3	T: 0.298	none	Sammalkorpi et al. (2007)
AL390295	9	A	13.8	A: 0.222	0.222 0.251	Sammalkorpi et al. (2007)
AL359238	9	A	44	T: 0.062	none	Sammalkorpi et al. (2007)
AP003532_2	9	A	46.4	G: 0.111	none	Sammalkorpi et al. (2007)
TTK	9	A	50.2	A: 0.079	none	Williams et al. (2010)
C4orf6	9	A	60	A: 0.059 G: 0.0192	NFD	Woerner et al. (2001)
AL954650	9	C	63	T: 0.138	none	Sammalkorpi et al. (2007)
AL355154	10	A	66.7	T: 0.403	none	Sammalkorpi et al. (2007)
AVIL	10	A	70.2	A: 0.247	none	Woerner et al. (2010)
ASTE1	11	A	78	No SNP	none	Woerner et al. (2010)
MRPL2	12	C	91.5	T: 0.245	none	Woerner et al. (2010)
EGFR	13	A	72.1	No SNP	none	Yuan et al. (2009)
FBXO46	14	A	95.2	A: 0.027	NFD	Woerner et al. (2001)
FTO	15	A	81.8	-: 0.042 C: 0.016	none	Woerner et al. (2001)

Table 3.2: A list of the repeats sequenced in results chapter 3, the MSI rates reported in SelTarBase, and the minor allele frequency of neighbouring SNPs. Note: many of the homopolymers are named after the gene they are located in. NFD = no frequency data available in dbSNP build 173.

3.2.3. Data generation

Initially, material from 8 Lynch Syndrome patients was assessed to make sure there was enough material present to generate 22 amplicons. Samples of FFPE tumour material, FFPE normal mucosa, and blood were available for all 8 patients. Blood from four age matched normal controls, and FFPE microsatellite stabile (MSS) tumour tissue from four age matched and sex matched controls were also obtained. Having these controls means that it is possible to identify any artefacts that could be caused by imbedding the samples in wax, and it will be possible to control for PCR/sequencing artefacts in the tumours using the matched normal tissue. DNA extraction was carried out on each sample as described in methods sections 2.2.2 and 2.2.4. The DNA samples obtained were quantified and a PCR reaction was performed using FBXO46 to check the quality of the DNA. For three out of eight Lynch Syndrome patient samples, at least one of the FFPE samples failed to produce any PCR product. High failure rates from PCR reactions using DNA derived from FFPE tissues is a well-known problem (Gilbert et al., 2007). Formalin fixing causes a degradation of DNA. This degradation is dependent on factors such as the length of time a tissue sample is retained in formalin solution, temperature and pH during fixation and the age of wax blocks (Gilbert et al., 2007).

The samples that were used consist of 3 tissues (FFPE tumour sample, FFPE normal mucosa, and blood) from 5 Lynch Syndrome patients. Having these 3 matched tissue samples from each patient makes it possible to decipher the patient's genotype and compare it to the variants found in the cancer sample. The mutation status of each of the 5 patients can be found in Table 3.3.

Patient Number	Mutation
U096	MLH 1 exon 17, familial splice site mutation (c.1989+1G>A)
U179	MLH1 exon 18. single base pair deletion in codon 697, this is a frameshift mutation resulting in 61 novel amino acids at the 3' end of MLH1 protein
U184	Missense mutation (c.677G>T; p.Arg226Leu) in exon 8 of MLH1.
U303	MLH1 missense T117M in exon 4
U312	MSH2 - deleted exon 8

Table 3.3: Germline mutations in the five Lynch Syndrome patients who's tumours were analysed in this study.

Amplicons were created for each homopolymer (a total of 575 amplicons). The PCRs were performed using the Herculase II Fusion polymerase (Agilent, Santa Clara, CA, United States of America) as this polymerase had the lowest error rates when replicating homopolymers in a study by Fazekas et al. (2010). All products were

generated usinga Bio-Rad T100^{tm} thermal cycler (Bio-Rad, Hercules, CA, United States of America) using the same PCR program. PCR amplification for all DNA samples was performed on one plate for each amplicon set to minimise any differences in processing between samples. Products were visualised on an agarose gel to confirm that amplicons had the expected size and to check for miss-priming. If any samples failed PCR amplification they were repeated. As an example, amplicons created for the homopolymer FBXO46 can be found in Figure 3.2. Although the normal mucosa sample from patient 4 shows weak amplification, all reactions generated the expected amplicon of size 303bp.

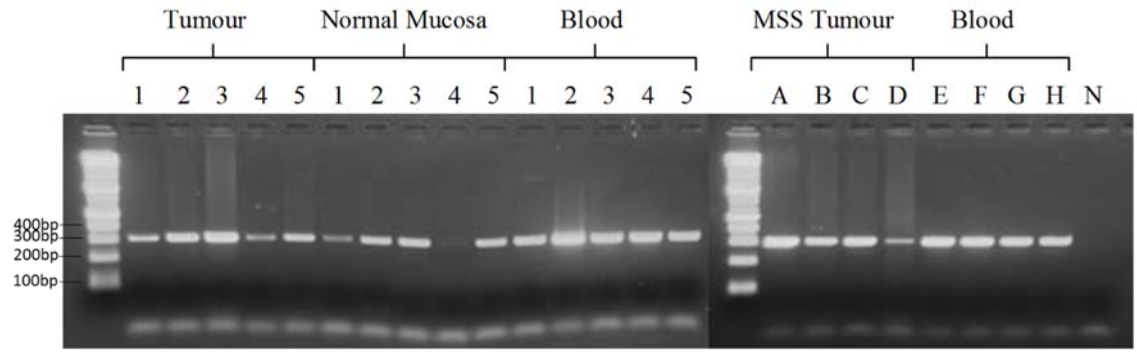

Figure 3.2: The amplicon set for one of the homopolymers (FBXO46). 1= patient U096, 2= patient U179, 3=patient U184, 4= patient U303, 5= patient U312. A-D= Normal control MSS Tumour Tissue, E-H= Normal control blood, N= Negative (no DNA) control

Once all amplicons were produced, between 5 and 7 amplicon from each gel image were selected for quantification using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, United States of America). Amplicons were chosen so that amplicons with a range of different band intensities on the gel were quantified. The concentrations of the remaining amplicons were estimated based on the band intensities seen on the gel images by comparing them to the bands of known concentration. Amplicons for each sample were pooled at a roughly equal concentration. The pooled amplicons were cleaned using Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, United States) before being diluted to ~0.2ng/µl, which is the recommended input DNA concentration for the Illumina Nextera XT kit. Illumina adapters were added to the amplicons using the Nextera XT library prep kit (Illumina, San Diego, CA, United States of America). Agilent Bioanalyser high sensitivity chips (Agilent, Santa Clara, CA, United States of America) were used to determine the quality of the library before the Nextera XT normalization step. Sequencing was then performed on an Illumina MiSeq.

For the MiSeq run a MiSeq Reagent Kit (500-cycles) (Illumina, San Diego, CA, United States of America) was used. A cluster density of 560000 clusters per mm^2 was obtained and a Q-score of over 30 was achieved for 64.57% of the bases sequenced (see Figure 3.4). A Q-score of 30 is equal to 99.9% probability of a base being called accurately. A drop of in Q-score was observed towards the latter cycles (see Figure 3.3). This is believed to be due to having reaching the end of many of the amplicons. A total of 11,236,567 reads were obtained from this MiSeq run and all samples were represented. Despite having the least reads, the U303 tumour sample had an average of 98,500 reads per amplicon.

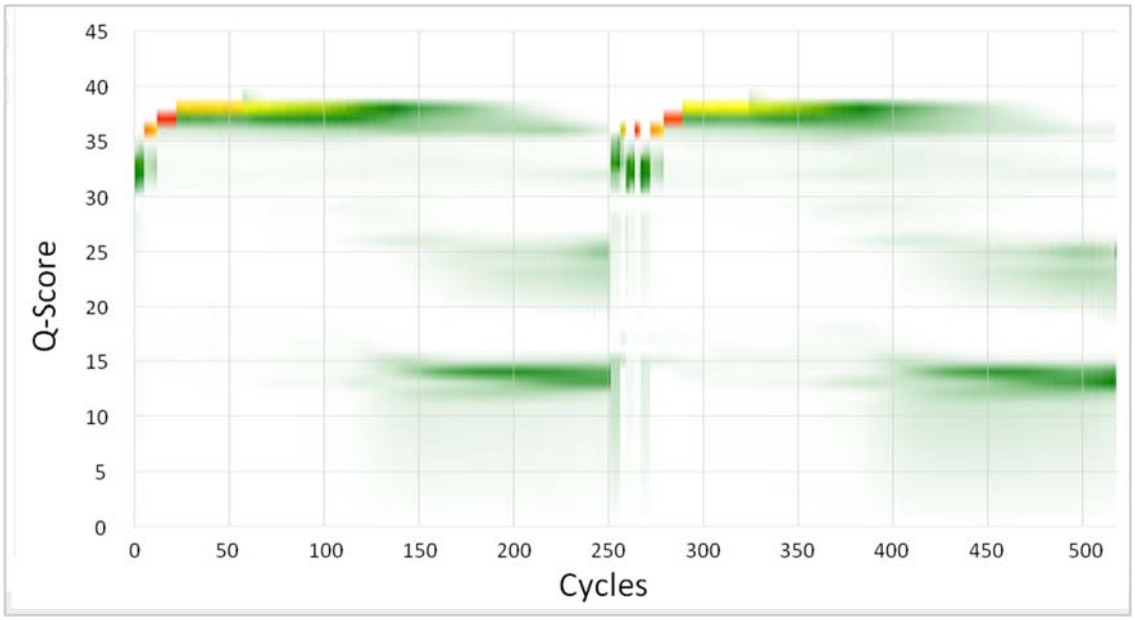

Figure 3.3: The quality score (Q-Score) distribution for each cycle showing a drop in Q-Score towards the later cycles of each read.

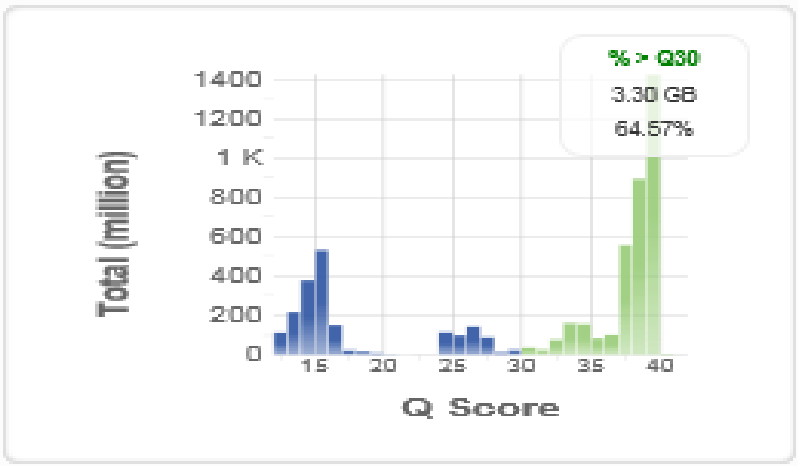

Figure 3.4: The quality score (Q-Score) distribution for the reads generated on the MiSeq. Blue = bases with a Q-Score <30, Green = bases with a Q-Score >30.

3.2.4. Variant calling

To identify variants in next generation sequence data a variant caller was used. Variant callers detect differences between a reference sequence and the sequence of a sample. Using algorithms the variant caller classes these inconsistencies between the reference sequence and a sample as either likely to be real variants or artefacts. Only variant that are deemed to be real according to the algorithms of the variant caller are reported. Algorithms differ for different variant callers which can result in discrepancies in variant calls. For the purpose of this study it was important to be able to analyse both artefacts caused by sequencing and PCR error as well as true indel events. For this reason the use of a standard variant caller that filters out artefacts would be inappropriate. To circumvent the use of a standard variant caller, a simple in-house caller was created. This caller has been named "Concordant Overlapping Paired Reads Caller" (COPReC). COPReC only reports indels in concordant overlapping reads, therefore reducing the amount of sequencing error. Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University) designed COPReC and performed the variant calling for this MiSeq run. The output consisted of the length variants observed for each homopolymer and how many paired end reads corresponded to each length. COPReC also has the advantage to be able to call low frequency indels that may be filtered out by a conventional indel caller. All indels are of interest as it will also be important to understand the distribution of sequencing and PCR error for different homopolymers.

3.2.5. Polymorphic homopolymers

Repeats with length polymorphisms were undesired because they would complicate the detection of MSI. Even with matched normal controls for each of the MSI-H tumours, polymorphisms would make it impossible to measure the frequency of variant reads caused by MSI in an individual with a read length polymorphism of corresponding length. The reason for this is that PCR amplification from poor quality template DNA can be biased so that the allele ratios for heterozygotes are not 50/50. Ascertaining the frequency of variant reads that can be attributed to an MSI event would therefore be difficult in these circumstances.

Despite attempts to only select monomorphic homopolymers it was clear from the data that two out of the twenty-two sequenced repeats, MX1 and C4orf6, contained read

length polymorphisms in at least one of the controls (see Figure 3.5). Polymorphisms for these two repeats have been registered on dbSNP, but no frequency data were available (see section 3.2.2). Based on that information it was possible that these polymorphisms were extremely rare so the repeats were included. However, for the repeats MX1 and C4orf6 my data suggests the polymorphisms are not rare (see Figure 3.5). Because length polymorphisms will make it harder to detect MSI events, these two repeats were excluded from further analysis.

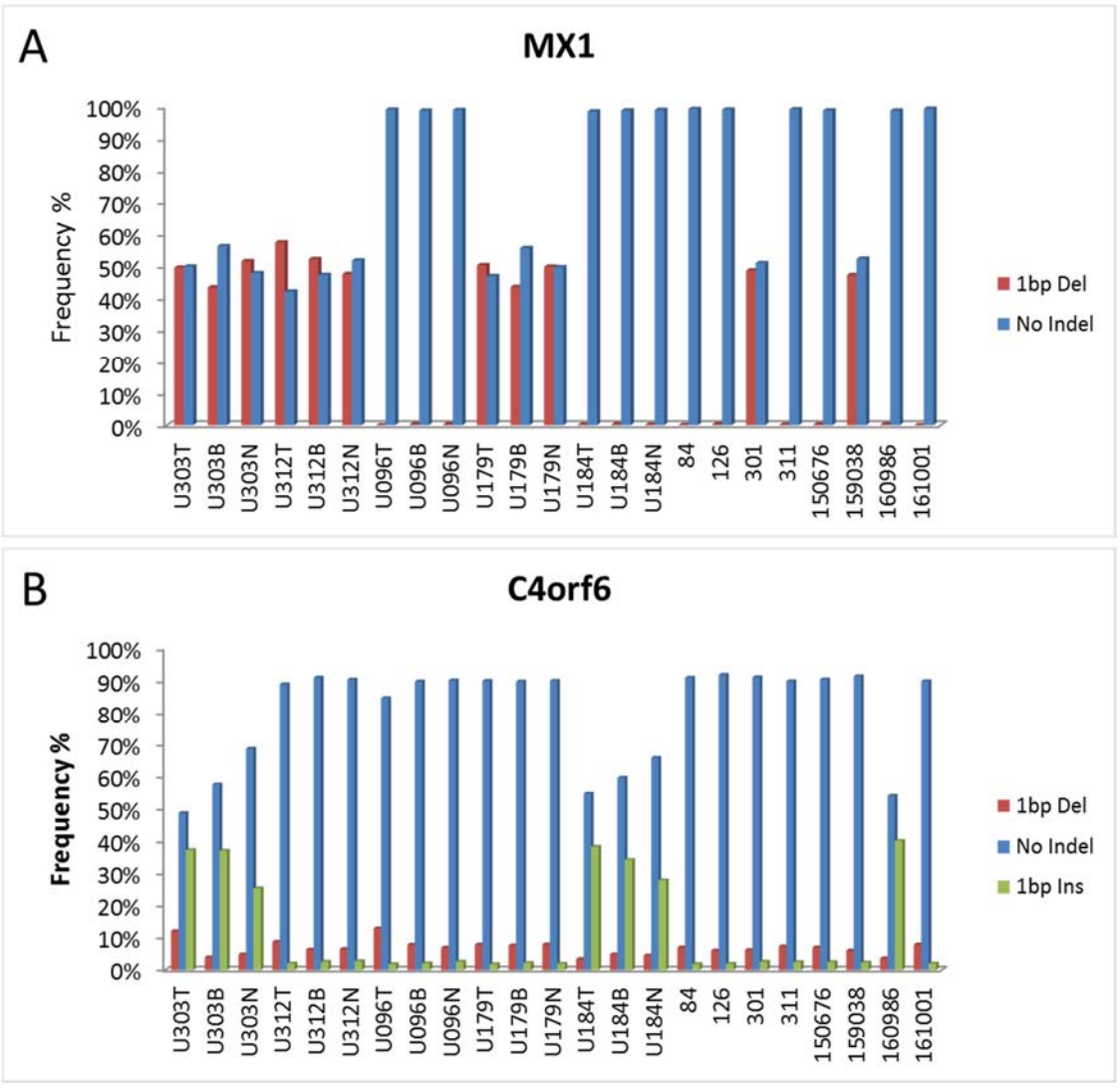

Figure 3.5: Frequency of variant alleles in all samples for markers MX1 and C4orf6. Both homopolymers were known to contain polymorphisms of unknown frequency (see Table 3.2). Panel A: the 8bp repeat MX1. Panel B: The 9bp repeat C4orf6. T= Tumour sample, N= Normal Mucosa, B=Blood

3.2.6. PCR/Sequencing error in short homopolymers

Because of the repetitive nature of homopolymers some sequence and PCR error was expected. To ascertain the levels of errors produced, mean variant read frequencies were calculated for the control samples (see Table 3.4). Variant reads observed in control samples were assumed to be derived from PCR errors during sample preparation. The 7bp repeat has a mean error frequency of 0.2% and 0.3% for all the controls. The 8bp repeats have a mean error frequency of between 2%- 3.7%. The 9bp repeats have a mean error frequency of between 0.4% – 7.7%. The 10bp homopolymers have a mean variant read frequency of between 2.7% -8.5%. The longer repeats had even higher error frequencies (see Table 3.4).

		Mean Error Rates			
	Repeat length (bp)	LS Blood	LS Normal Tissue	Normal Bloods	MSS Tumours
$Axin2	7	0.2%	0.3%	0.2%	0.3%
$AL590078	8	1.4%	1.9%	1.7%	0.5%
$APBB2	8	2.2%	2.2%	2.6%	2.0%
$DEPDC2	8	3.7%	3.6%	3.8%	3.7%
$HPS1	8	2.4%	2.4%	2.4%	2.4%
$IL1R2	8	2.4%	3.7%	2.6%	2.6%
$SLC4A3	8	2.6%	2.8%	2.7%	3.5%
$AC079893	9	0.6%	0.8%	0.4%	0.7%
$AL359238	9	1.5%	1.4%	1.4%	1.5%
$AL390295	9	1.5%	1.5%	1.7%	0.8%
$AL954650	9	7.7%	7.4%	5.7%	6.8%
$AP003532_2	9	1.2%	1.1%	1.3%	1.5%
$TTK	9	3.2%	2.8%	3.2%	3.2%
$AL355154	10	3.0%	2.7%	3.2%	3.8%
$AVIL	10	8.1%	8.5%	7.7%	6.8%
$ASTE1	11	18.4%	21.5%	18.0%	19.4%
$MRPL2	12	75.6%	73.2%	76.5%	89.1%
$EGFR	13	42.5%	49.2%	45.0%	42.1%
$FBXO46	14	47.6%	53.9%	45.4%	44.7%
$FTO	15	77.3%	78.6%	71.9%	87.8%

Table 3.4: Mean error rates consisting of PCR and sequencing error divided into the different control sample groups. The longer repeats have a high error rate (highlighted in orange). LS = Lynch Syndrome.

3.2.7. Levels of instability observed in short homopolymers

For each indel size the percentage of reads with that indel size were calculated and graphs for each homopolymer were plotted. In only one homopolymer, APBB2, there was a higher insertion frequency observed in one of the MSI-H samples compared to the controls (see Figure 3.6).

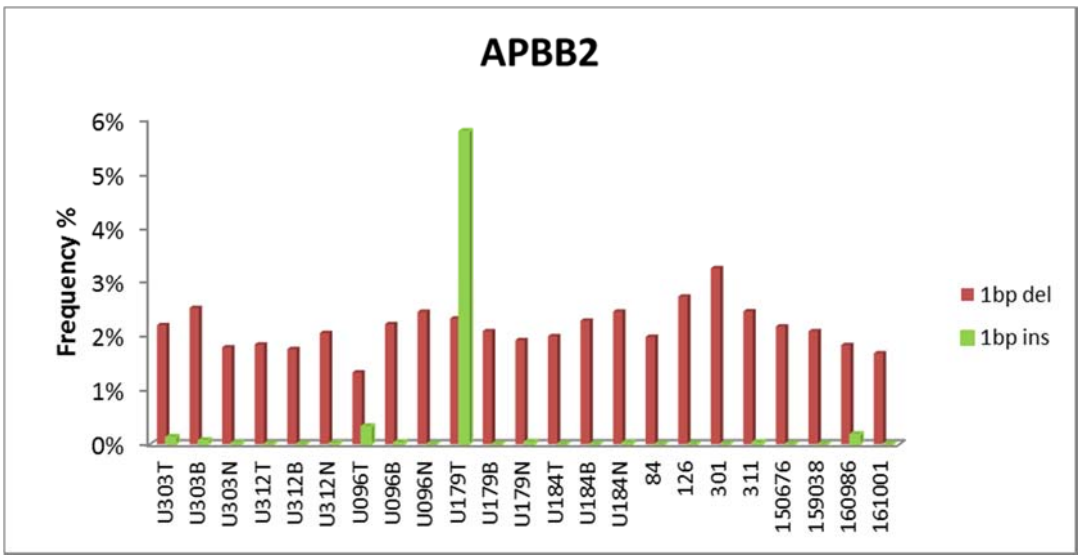

Figure 3.6: Indel rates in the homopolymer APBB2. The U179 tumour sample has an insertion frequency of 5.8%, which is higher than that of any other sample. T= Tumour sample, N= Normal Mucosa, B=Blood

The only 7bp repeat, Axin2, showed no difference in the frequency of variant reads between the control samples and the MSI-H samples (Figure 3.7). For all samples the reference reads made up over 99% of the reads covering the Axin2 homopolymer.

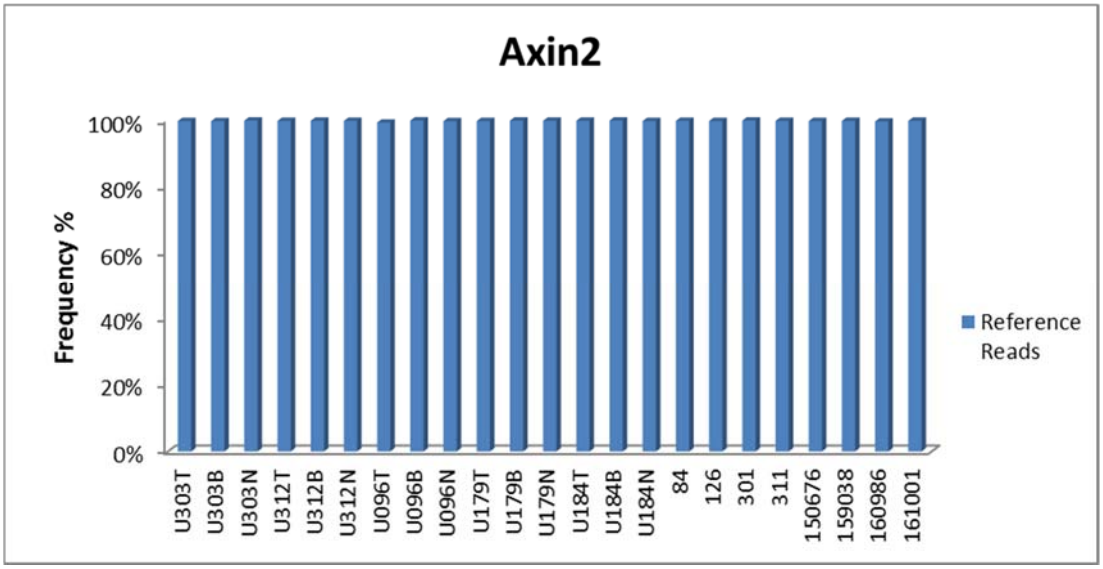

Figure 3.7: The frequency of reference reads for the 7bp homopolymer Axin2. T= Tumour sample, N= Normal Mucosa, B=Blood

Because of varying levels of PCR/sequencing error within repeats of the same length, it is not easy determining a cut off value for distinguishing between background error and MSI events. In this chapter arbitrary thresholds were set for calling repeats unstable. Cut off values were not calculated because a low number of repeats and samples were used. Calculating cut off values will be covered later in this thesis.

For the 8-10bp repeats a deletion was classed as MSI if more than 10% of the reads contained that deletion. For the larger homopolymers there was a lot of PCR/sequencing error (see Table 3.4). For the longer homopolymers MSI events presented as larger deletions compared to the background noise (see Figure 3.8). For the 11-13bp repeats an event was classed as being caused by MSI if there was a 2bp deletion or larger which accounted for ≥ 10% of the reads. The 14-15bp homopolymers were classed as unstable if there was a 3bp deletion or larger which contained ≥ 10% of the reads for that homopolymer.

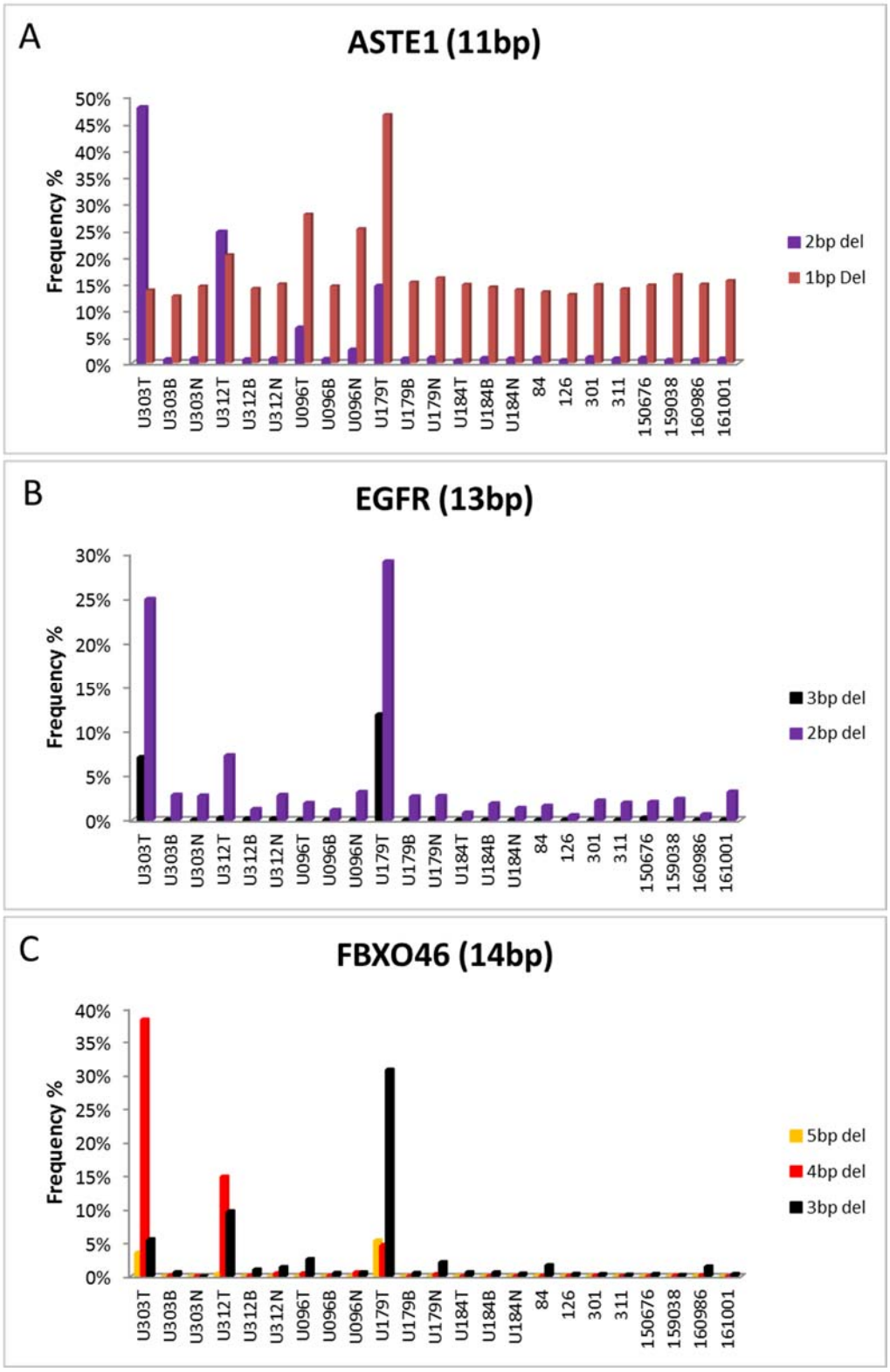

Figure 3.8: In the larger homopolymers MSI was observed as larger deletions. Panel A: The 11bp repeat ASTE1 with a 2bp deletion present in over 10% of the reads for samples U303T, U312T and U179T. Panel B: The 13bp repeat EGFR with a 2bp deletion present in over 10% of the reads for samples U303T and U179T. Panel C: The 14bp repeat FBXO46 with a 3bp or 4bp deletion present in over 10% of the reads for samples U303T, U312T and U179T. T= Tumour sample, N= Normal Mucosa, B=Blood

Variant reads containing deletion frequency levels consistent with instability were observed for 10 out of the 20 homopolymers in at least one tumour. All 20 homopolymers were classed as stable in all of the control samples. The ten homopolymers and deletion sizes which best separated the MSI-H samples from the controls can be found in Figure 3.9. All of the Lynch Syndrome patient tumour samples, with the exception of the tumour from patient U184, had at least three unstable homopolymers. In fact based on the data obtained for the U184 tumour sample, this sample behaves like a MSS sample. The tumour U096 had two unstable 8bp repeats and one unstable 10bp repeat, while there was no evidence of instability in the longer 11bp-14bp repeats. The other Lynch Syndrome tumours, with the exception of the U184 tumour sample, all had at least two unstable 11bp-14bp repeats.

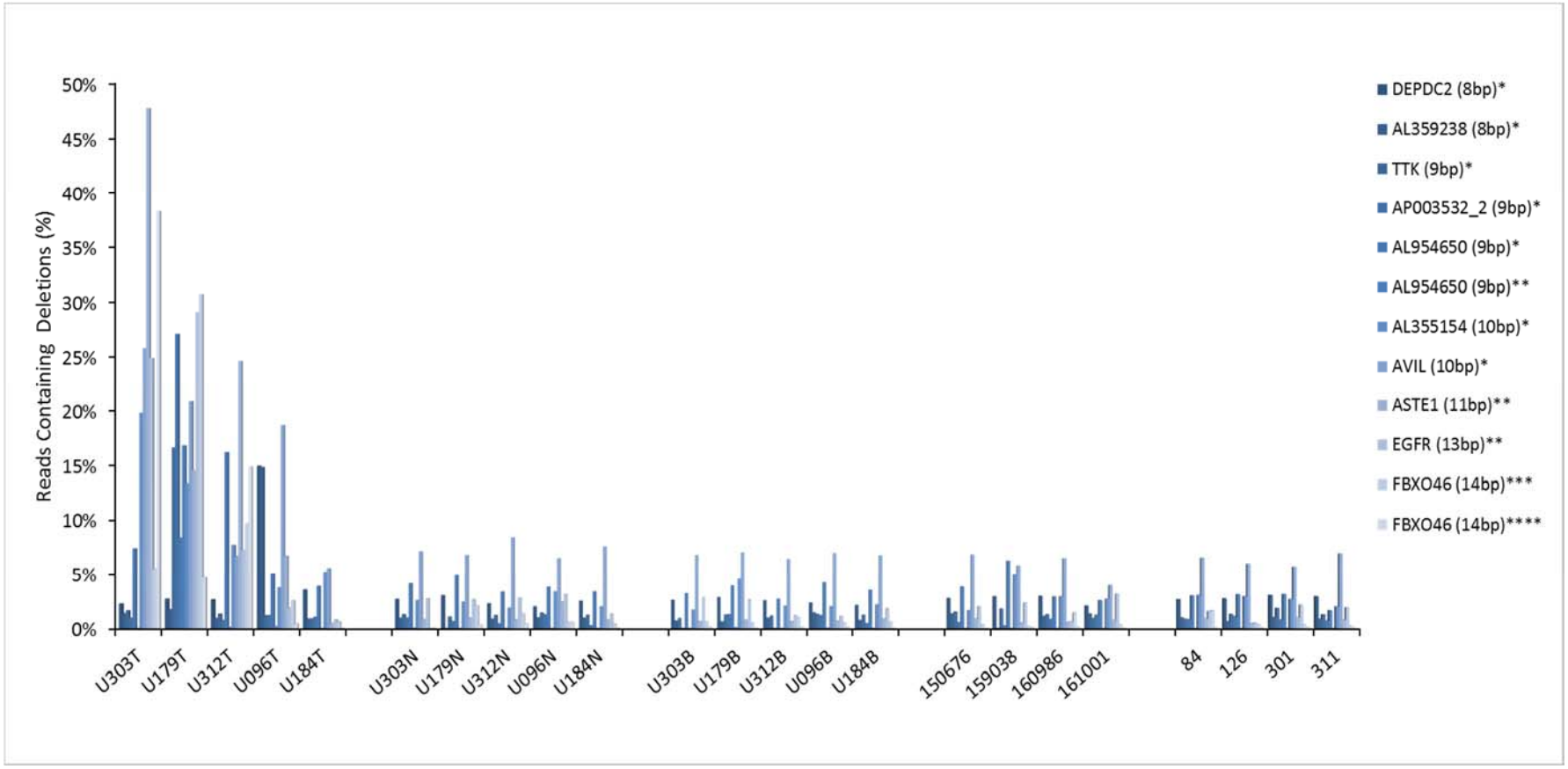

Figure 3.9: Results for the ten markers with elevated deletion frequencies in Lynch Syndrome tumours. The deletion sizes shown are the best for separating the tumours from the Lynch Syndrome patients and the controls. Four out of the five tumours from Lynch Syndrome patients show an increased deletion frequency in at least three homopolymers compared to the controls. No MSI events were observed for the tumour from patient U184. T= Tumour sample, N= Normal Mucosa, B=Blood.

3.2.8. Fragment analysis MSI results

The five tumours obtained from Lynch Syndrome patients were assumed to be MSI-H because of their origin. However the results of the next generation sequencing assay provide no evidence for instability in the tumour derived from patient U184 using the 20 markers analysed here (see Figure 3.9). Patient U184 has a germline Missense mutation in exon 8 of MLH1 (c.677G>T; p.Arg226Leu). It was therefore assumed that a tumour derived from this patient would be mismatch repair deficient and MSI-H. However, it had not been formally tested prior to this analysis.

Conventional MSI tests were therefore performed to confirm the MSI status of all five Lynch Syndrome tumours. The kit used for the MSI test was the Promega MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, United States of America). The results of the MSI assay confirmed that the tumours from patients U096, U179, U303, and U312 are MSI-H (see Table 3.5). The fragment analysis results for the tumour from patient U184 show that this tumour is MSS, indicating that is a sporadic tumour unrelated to the Lynch Syndrome associated predisposition. This is consistent with the results obtained from the sequencing of short homopolymers (see Figure 3.10). My interpretations of the fragment analysis traces was confirmed by Ottie O'Brien (Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust).

	NR-21	BAT26	BAT25	NR-24	MONO-27
U096 Tumour	unstable	unstable	unstable	unstable	unstable
U179 Tumour	unstable	unstable	unstable	unstable	unstable
U184 Tumour	stable	stable	stable	stable	stable
U303 Tumour	unstable	unstable	unstable	unstable	unstable
U312 Tumour	unstable	unstable	unstable	unstable	unstable

Table 3.5: Results from a standard fragment analysis test results for tumours from Lynch Syndrome patients U096, U179, U184, U303, and U312. The test was performed using a Promega MSI Analysis System Version 1.2 kit.

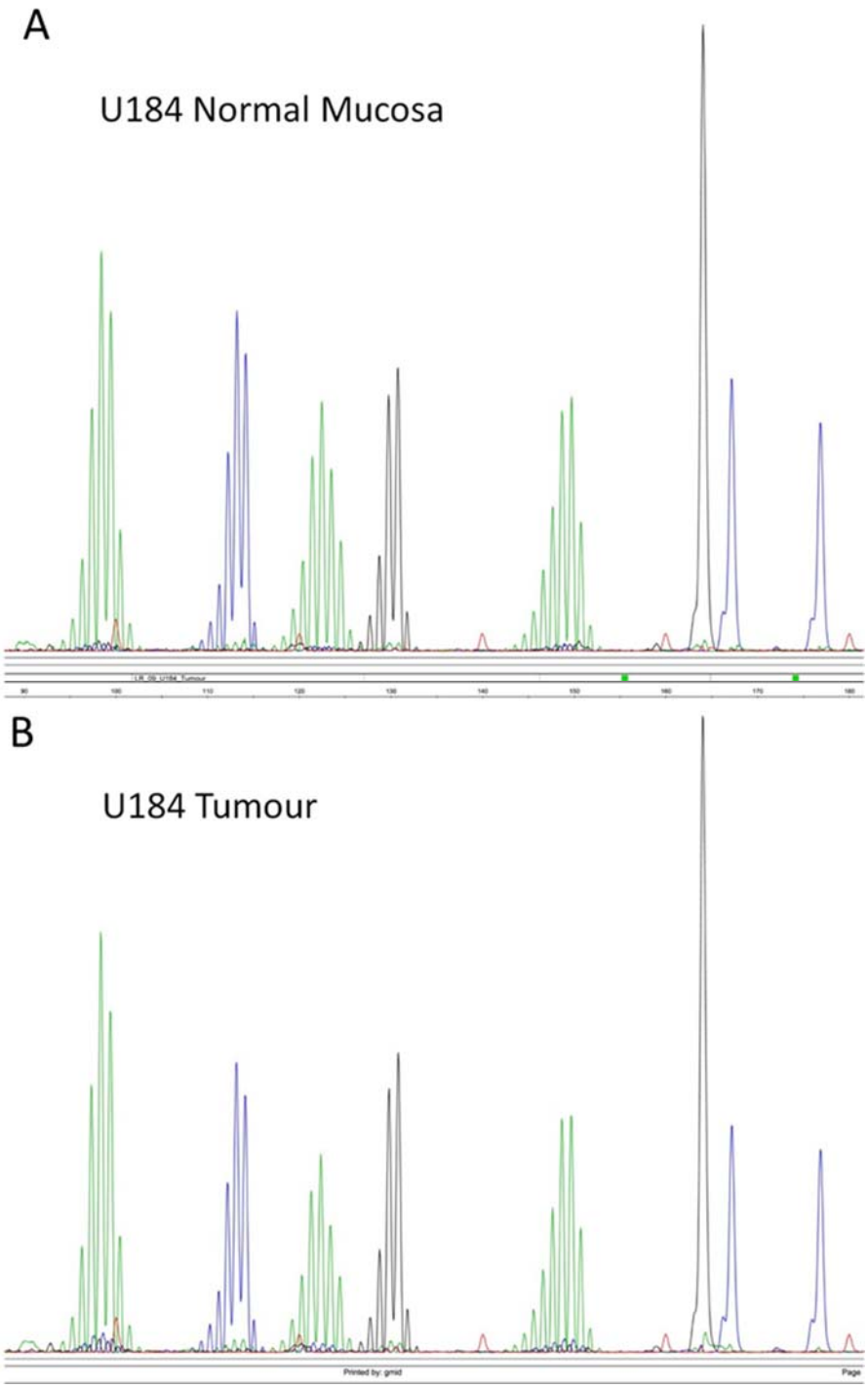

Figure 3.10: Results from a standard fragment analysis test using a Promega MSI Analysis System Version 1.2 kit. Panel A shows the test results for the U184 normal mucosa sample, and panel B shows the test results for the U184 tumour sample. There is no difference in the stutter pattern of the tumour and normal mucosa, both of which show a stutter pattern consistent with the absence of MSI. Therefore according to these fragment analysis results the U184 tumour sample is microsatellite stable (MSS).

3.2.9. Assessing the value of neighbouring SNPs

The panel of homopolymers was selected to include SNPs with a high minor allele frequency in close proximity to each repeat. In heterozygous individuals it is therefore possible to distinguish between alleles for repeats where reads span both the repeat and the SNP. For homopolymers that show microsatellite instability in at least one of the MSI-H samples, neighbouring heterozygous SNPs were used to investigate the distribution of variant reads between the two alleles. The aim was to determine whether MSI is an allele dependent event or if MSI affects both alleles. A SNP was not considered heterozygous if one allele had a read count of less than 10% of the total read count. The criteria that repeats were not analysed if one allele has less than 10% of the total read count was used because such an extreme allele imbalance might indicate sample contamination.

For four of the homopolymers MSI was observed in at least one individual with a heterozygous SNP. In total, there are five instances where MSI is observed in a homopolymers with a neighbouring heterozygous SNP (see Figure 3.11 and Figure 3.12). In all five of these instances there was evidence of bias between indel frequencies for the two alleles. The bias ranged from one allele showing 4.9 times the frequency of reads on one allele compared to the other allele (U179 tumour 1bp deletion, Figure 3.11 panel A) up to 8.4 times the frequency of reads on one allele compared to the other allele (U303 tumour 4bp deletion, Figure 3.12 panel A).

Some allelic imbalance is also present in the control samples. The U179 blood sample and U179 normal tissue have a 1bp deletion with a frequency of 5.63% and 4.17% respectively on one allele, and no reads with a 1bp deletion on the other allele for the homopolymer AL355154 (see Figure 3.11 panel B). But for neither U179 blood sample and U179 normal tissue does the fraction of reads containing an indel exceed 6% of the reads on an allele for the 10bp homopolymer AL355154. There is also a low read count in both instances with the AL355154 1bp deletion only being observed in 4 reads and 2 reads for U179 blood an U179 normal tissue respectively. This amount of variant reads is not higher than the background PCR and sequencing error rate seen in 10bp repeats (see Table 3.4) and the allelic bias is lower than that observed in the MSI-H samples, suggesting that it may be caused by PCR error and stochastic events during PCR.

The data presented so far suggests that MSI may show allelic bias, however more data would be required to confirm if MSI is an allelic event. If MSI is an allele dependent event where as PCR/Sequencing error effects both alleles, then for instances where a

patient is heterozygous for a SNP it should be possible to distinguish variants resulting from MSI from variants resulting from PCR/sequencing error when there is a sufficient read depth. For low read depths, distinguishing between MSI and sequencing/PCR error may be more problematic because a very small number of reads can make a large difference between allele frequencies, such as seen for the U179 control samples for the AL355154 homopolymer.

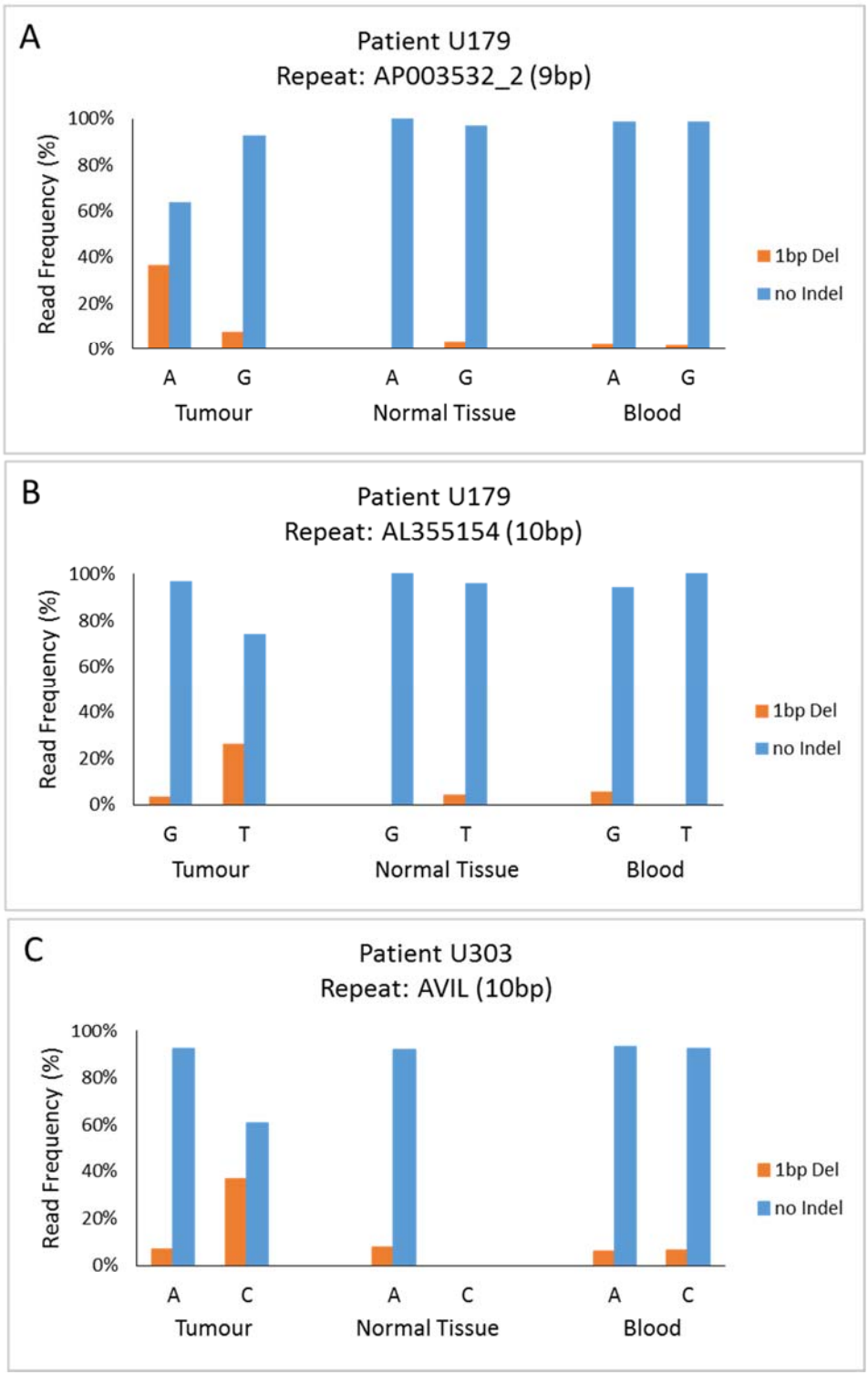

Figure 3.11: The percentage distribution of reads in three short repeats (9bp-10bp) with a neighbouring heterozygous SNPs in MSI-H tumours and matched normal mucosa and blood. These results show that there is an allele bias in the MSI-H tumours with one allele showing a higher frequency of a 1bp deletion compared to the other allele.

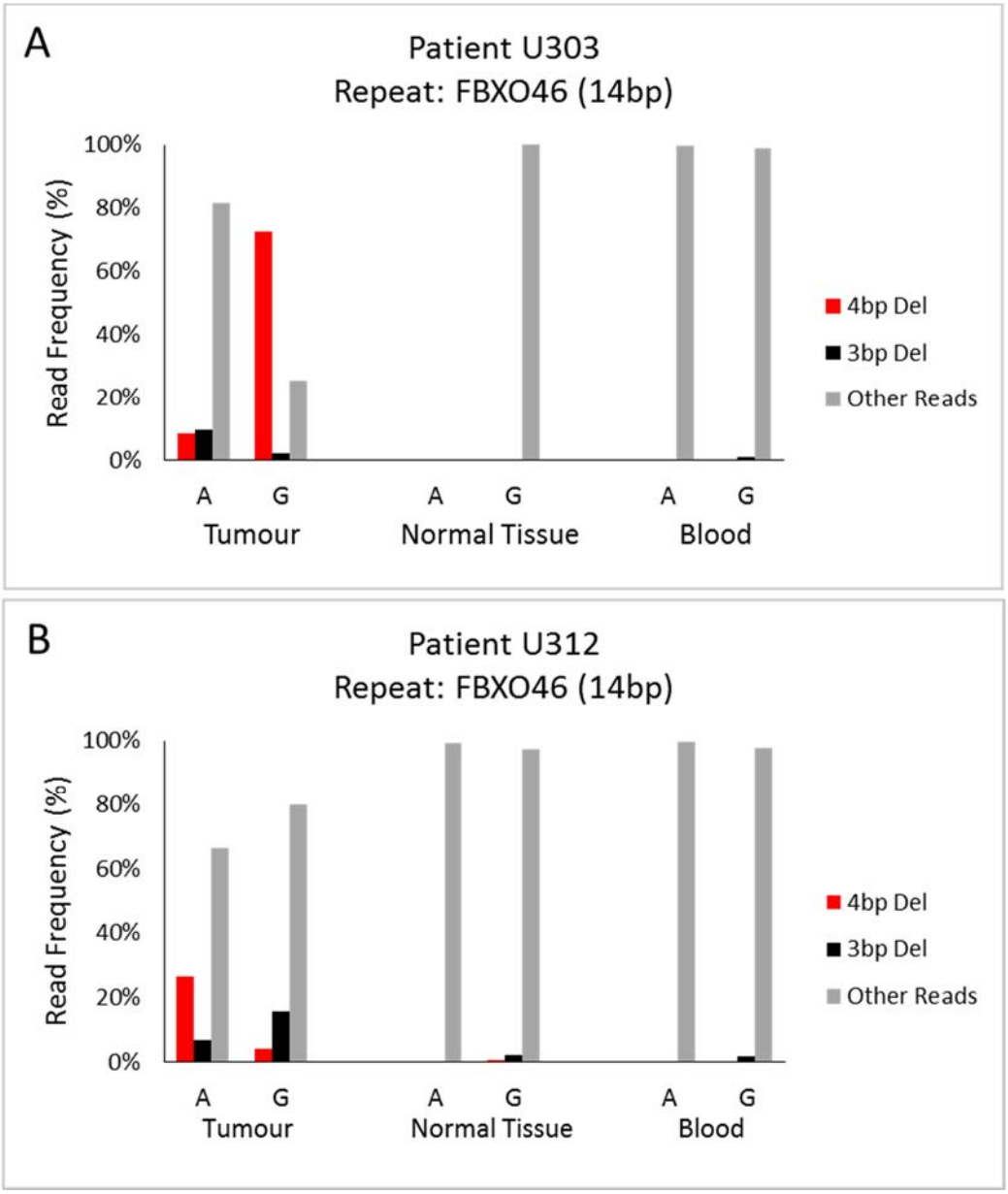

Figure 3.12: The percentage of reads corresponding to a 3bp and 4bp event in the presence of a heterozygous SNP in two Lynch Syndrome patients. These results show that there is an allele bias in the MSI-H tumours with one allele showing a higher frequency of reads corresponding to a 4bp deletion compared to the other allele.

3.2.10. U303 allelic dropout and identity testing

Due to discrepancies at SNP positions between the genotype of the U303 normal mucosa and the other U303 samples (see Figure 3.11 panel C, and Figure 3.12 panel A), an identity test was performed for to confirm that the U303 normal mucosa and U303 blood sample are derived from the same individual. The identity test was performed using the Promega PowerPlex 16 kit (Promega, Madison, WI, United States of America). Due to the quality of the U303 normal mucosa DNA, the Powerplex 16 markers with alleles larger than 330bp did not amplify well. The identity test was therefore carried out using

12 out of the original 16 markers. The results of this test show that the U303 normal mucosa sample belongs to the same individual as the U303 Blood sample (see Figure 3.13). A matched probability calculation estimated that the likelihood of the U303 normal mucosa originating from someone other than, and unrelated to the individual the U303 blood sample was derived from is in the order of 1 in 3×10^{11}. The matched probability calculation was performed using an F_{ST} value of 0.02 and the database size and allele frequencies from the NorthGene database (NorthGene Ltd, Newcastle upon Tyne, UK). The absence of one allele at the SNP positions for the sample U303 normal tissue is therefore most likely caused by allelic dropout caused by low copy number of starting material as a result of degraded DNA obtained from FFPE tissue. PCR amplification from poor quality template DNA can cause allelic biases with allele dropout (Wang et al., 2012). The DNA for U303 normal tissue always produced the lowest amount of PCR product, which could suggest there was very little starting material for generating amplicons around 300bp in size.

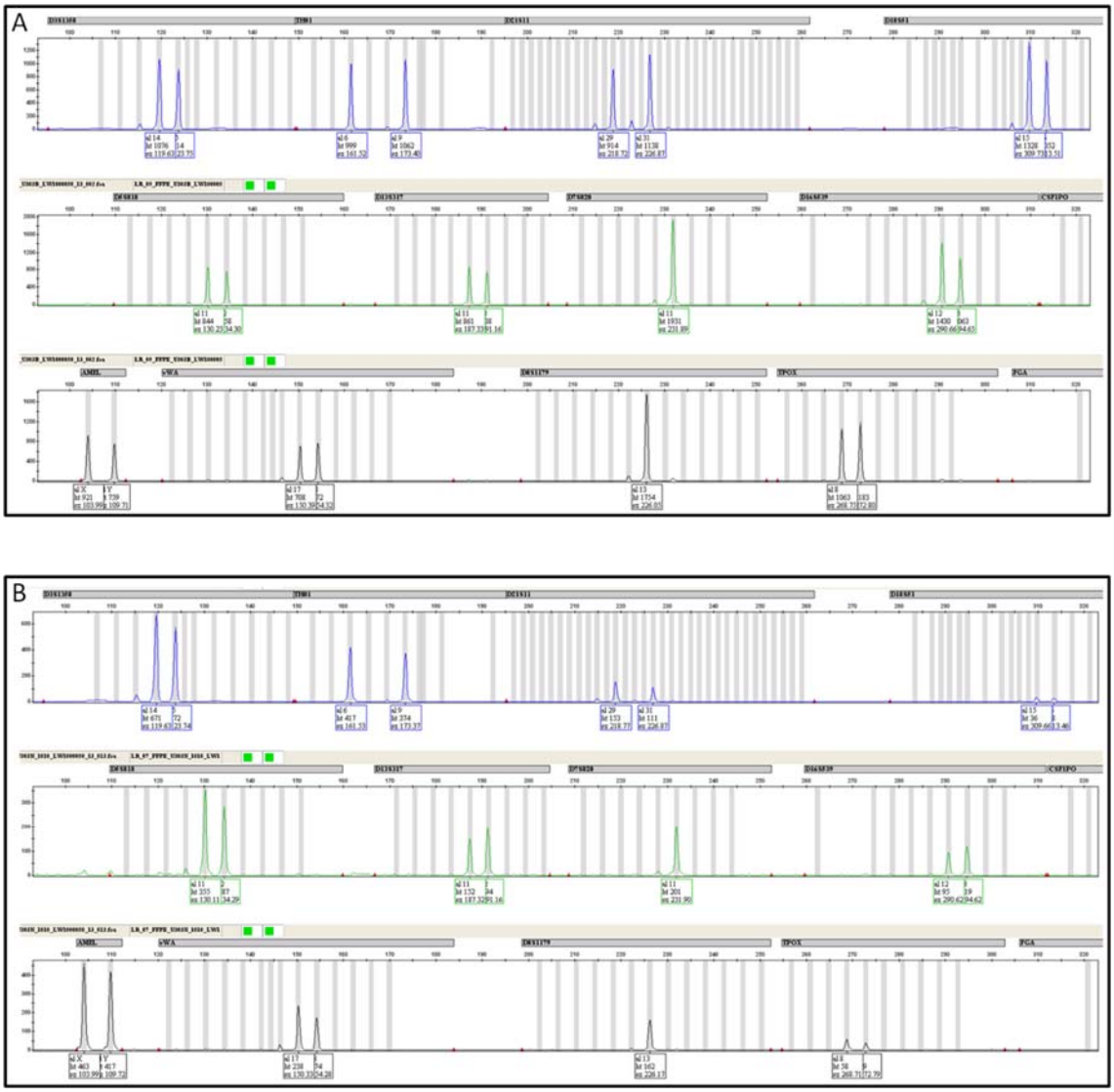

Figure 3.13: The results of a PowerPlex16 identity test for U303 Blood sample and U303 Normal tissue. These results support the hypothesis that both samples belong to the same individual. Panel A: Identity markers for the U303 Blood sample. Panel B: Identity markers for the U303 Normal Mucosa

3.3. Discussion

The aim of this initial work was to use short mononucleotide repeats, previously shown to exhibit MSI, to assess if microsatellite unstable tumours could be typed using a panel of short repeats sequenced using high throughput sequencing. In this chapter, it has been shown that it is possible to detect microsatellite instability using Illumina sequencing. The use of a sequencing based assay for the detection of MSI has been suggested previously. However prior to the start of this PhD project no data has been published to show that a panel of short repeats, sequenced by next generation sequencing, could be used to differentiate between MSI-H and MSS samples. The work conducted by (Cancer Genome Atlas Network, 2012) indicated it may be possible to type MSI using next generation sequencing. However investigating MSI was not the main focus of the (Cancer Genome Atlas Network, 2012) paper, and due to having trouble identifying unstable repeats using a variant caller, no in depth analysis was performed. Using a panel of 20 mononucleotide repeats ranging in size from 7bp to 15bp, and using post-hoc threshold values, it was possible to distinguish the MSI-H samples from the stable samples (see Figure 3.9). The thresholds used for calling a marker unstable were ≥10% of reads containing a deletion for 7-10bp repeats, ≥10% of reads containing a ≥2bp deletion for 11-13bp repeats, and ≥10% of reads containing a ≥3bp deletion for 14-15bp repeats. Furthermore, the single Lynch Syndrome tumour U184, which did not show instability was subsequently found to be MSS using a standard fragment analysis test (see Figure 3.10). Lynch Syndrome patients can develop MSS tumours (Giuffre et al., 2005). It was therefore not wise to assume that all five tumours from Lynch Syndrome patients analysed in this chapter would be MSI-H just because the patients had germline mutations in mismatch repair genes.

Other than one insertion in the repeat APBB2 for the U179 tumour sample, there were no other insertion events where a substantial difference between MSI-H samples and normal controls was observed. There were, however, several different deletion events observed in MSI-H samples. This may suggest that deletions are more indicative of MSI than insertions.

Interpretation of the data obtained so far indicates that shorter repeats are less susceptible to MSI, and susceptibility increases with repeat length. This is consistent with what is reported in the literature (Sammalkorpi et al., 2007, Vilkki et al., 2002, Woerner et al., 2003). For the 8bp repeats only two out of the six repeats were unstable in any of

the samples. Out of the six 9bp repeats, four showed instability in at least one sample. Whereas the unstable 8bp repeats only showed instability in one sample, one of the 9bp repeats showed instability in two samples. All of these results indicate that 9bp repeats are more prone to MSI than 8bp repeats. Both of the sequenced 10bp repeats showed instability. The 10bp repeats Avil and AL355154 were unstable in three and two out of the four MSI-H samples respectively. Each of the 11, 12, 13, and 14bp repeat were unstable in at least one of the MSI-H samples. Adding further evidence that susceptibility to MSI increases with length. However, one of the tumours, U096 tumour, only displayed instability in the 8bp and 10bp repeats.

The results of this analysis suggest that PCR and sequencing error frequencies increase with homopolymer length. This is consistent with results by Fazekas et al. (2010), There are, however, also large differences in error rates between repeats of the same length. This suggest that sequence context also plays a large role when it comes to sequence/PCR error rates. PCR replication of homopolymers is known to be prone to replication errors resulting from polymerase slippage (Clarke et al., 2001, Fazekas et al., 2010, Flores-Renteria and Whipple, 2011). The effect of these errors on Sanger sequence data has been well documented (Fazekas et al., 2010, Flores-Renteria and Whipple, 2011). The fidelity of different Taq polymerases has also been investigated in attempts to reduce replication errors (Fazekas et al., 2010). To investigate the amount of PCR and sequencing error that could be expected from next generation sequencing data the frequency of variant reads for different homopolymer sizes were evaluated in a control exome. Based on this investigation it was concluded that that the levels of PCR and sequencing errors would be almost negligible up to repeat lengths of 10bp and it might still be possible to gain valuable information about MSI from repeats of up to 14-15bp.

However the levels of background noise in my sequencing data is higher than seen in the control exome that was initially analysed for sequencing error (see Figure 3.1 and Table 3.4). The reason for this is likely to be due to the number of PCR cycles used in the initial PCR and during the library prep. Because of the low quality of DNA obtained from the FFPE samples and the relatively large amplicon sizes used (a requirement of the Nextera XT library prep), a large number of PCR cycles were needed to obtain a sufficient quantity of DNA for the Nextera XT library prep after PCR clean up. As a consequence the levels of PCR error are very high in the 11-15bp repeats and any 1bp deletions caused by MSI would be hard to distinguish from this background noise. Interestingly MSI appears to be manifest as larger 2-4bp deletions in these long repeats. At the deletion size

of 2bp where MSI is observed in the 11-12bp repeats, the background PCR error, measured as the frequency of reads in control samples, is low (see Figure 3.8). For the 13-14bp repeats there is even more PCR error, but the MSI events are still identifiable. However as the 11bp and the 12bp repeat both show MSI in three out of the four MSI-H samples while the 13bp and 14bp repeat show instability in two and three of the MSI-H samples respectively, there does not seem to be much value gained by using repeats longer than 11-12bp. For the 15bp repeat there is so much PCR error that only a fraction of the reads correspond to the reference sequence in the control samples. Interestingly the FFPE derived DNA did not cause any higher level of PCR and sequencing error compared to the DNA derived from blood. This suggest that DNA degradation in FFPE tissues does not cause indel events in homopolymers. This means that FFPE tissues are suitable templates for a sequencing based MSI testing.

For next generation sequence data there is the added complication of the use of a variant caller. In this chapter the problems of using a standard variant caller were avoided by developing a simple caller which makes no assumptions of data distributions for calling variants (ie. not looking for homozygotes heterozygotes etc.). The caller simply reports the number of paired-end reads corresponding to each variant. The only filtering which is used is that only matching overlapping paired-end reads are reported. The data were then analysed by eye after creating graphs in Microsoft Excel.

As an additional tool for analysing MSI in short homopolymers, neighbouring SNPs were chosen to allow the analysis of individual alleles. However, there were only four instances where MSI was observed in a homopolymer with a heterozygous SNP. For each of these instances there was a bias between the alleles with one allele having $\geq$5x the frequency of reads compared to the other allele. This is expected as MSI is sometimes seen to affect a single allele, reflecting the origin of the length changes in a homopolymer as a one hit event, consistent with MSI being caused by random errors during DNA replication which are not rectified as opposed to a targeted event towards some microsatellites. These results suggest that SNPs may allow these events to be investigated in more detail. If this is the case then SNPs could be used as another tool to differentiate between sequencing/PCR error and MSI. This would of course be limited to instances where there was a heterozygous SNP.

Analysing the SNP data revealed a complication with using degraded FFPE tissues that were not foreseen. For one of the samples, U303 normal tissue, there was

allelic dropout for some of the amplicons. This is most likely due to there being very little starting material to produce ≥300bp amplicons as a result of the DNA from that sample being very degraded. One of the challenges of genetic cancer diagnostics is obtaining good DNA from tumour specimens preserved by formalin fixation and paraffin embedding. As this process leads to the degradation and fragmentation of DNA. DNA fragmentation limits the PCR amplicon size that can be used. The degree of DNA degradation is dependent on the fixative used and its pH, the duration of fixation and how long the fixed specimen is stored and at what temperature (Gilbert et al., 2007). The failure rate of MSI tests are often not reported in the literature, but DNA degradation in FFPE tumour samples is a known reason for the failure of MSI tests (Snowsill et al., 2014). In this chapter, the quality of DNA limited the number of MSI-H samples used as well as being the suspected cause of allelic dropout in one sample. Because of the library prep method chosen, PCR amplicons of ≥300bp were required. This meant that out of samples from eight Lynch Syndrome patients, only samples from five patients were of sufficient quality to produce PCR amplicons from both tumour and matched normal tissue. As a result the amount of data was limited by the number of samples of sufficient quality that were obtained.

Three repeats containing SNPs with a high minor allele frequency that could cause length polymorphisms of the repeats were included by mistake (see section 3.2.2). None of these repeats were found to be polymorphic in any of the samples sequenced. These three repeats were not found to be unstable in any of the MSI-H samples and were therefore not used in any subsequent analysis.

3.3.1. Conclusions

In Conclusion, it is possible to distinguish between the MSI-H and MSS stable samples using Illumina sequencing. These results show that creating a next generation sequencing MSI assay the using short homopolymers is feasible. Eleven out of the twenty homopolymers analysed were unstable in at least one MSI-H sample. For the 9bp homopolymers four out of six showed MSI for at least one sample, while both 10bp homopolymers showed MSI for at least one sample. These are probably the best lengths for an MSI test. However, the 8bp homopolymers that showed MSI had the least noise (sequencing/PCR error). 7bp repeats may also be of value because of low background error if they have a reasonable susceptibility to MSI. Unfortunately, not enough 7bp

repeats were sequenced to assess their susceptibility to MSI. 11bp and 12bp homopolymers were deemed interesting due to the prevalence of 2bp deletions. The repeat sizes selected for further study were therefore 7-12bp repeats. The next step will be to identify the most unstable 7-12bp repeats for an MSI assay. This will be addressed in the next chapter.

A final MSI test will need robust thresholds for calling instability. In this chapter, arbitrary thresholds were set. For a final panel of homopolymers thresholds for defining markers as unstable will need to be calculated. Different repeat sizes will require different thresholds for the calling of microsatellite instability. It is also possible that individual repeats may require individual thresholds because the levels of PCR error vary for repeats of the same length. As the level of PCR error increases it will be increasingly difficult to differentiate between MSI-H samples and MSS samples. Because both the susceptibility of repeats to MSI and PCR error increases with repeat length a compromise between these factors has to made when choosing a final panel of repeats. Because only one repeat of each size from 11-15bp was sequenced there is very limited data for these repeat sizes. However, the results that were obtained suggest that even the 11bp and 12bp repeats are highly susceptible to MSI and not much is gained by analysing longer repeats than theses. It was therefore concluded that 7-12bp repeats would be the best lengths to use. The SNPs evaluated so far (SNPs in MSI-H samples with an unstable homopolymer) show that MSI occurred in one allele for those samples. This indicates that the second allele could be used as a negative control or a ratio of instability between two alleles could be incorporated into an MSI test. The number of markers for a MSI panel would have to be chosen taking into account the degree of instability of the markers. The number of markers needed for a panel will also depend on how susceptible the markers are to MSI events. To date there have been no attempts examine whole genome sequences to identify the most unstable short repeats for the use in an MSI test. This is the focus for chapter 4.

Chapter 4. Identification and analysis of highly variable homopolymers from next generation sequence

4.1. Introduction and aims

4.1.1. Introduction

Changes in microsatellite lengths occur due to strand slippages during DNA replication that lead to the template strand and nascent strand aligning out of register (Ellegren, 2004). When DNA replication continues with the strands out of register the result is an insertion or deletion. Homopolymers are the most unstable repeat type in mismatch repair deficient cells (Lang et al., 2013, Yoon et al., 2013, Umar et al., 2004). This makes homopolymers the most suitable repeat for a sequencing based MSI test, and also makes homopolymers a good choice for studying MSI. In chapter 3 it was shown that there is a positive correlation between homopolymer length and instability in MSI-H samples. It is also well documented in the literature that in general the length of a repeat is positively correlated with mutation rates in MSI-H cancers (Ellegren, 2004, Lang et al., 2013). However, there are also other factors that also play a role in the mutability of repeats such as repeat unit and sequence context. For example, G/C homopolymers have a higher mutability than A/T homopolymers (Ellegren, 2004, Sammalkorpi et al., 2007). The base composition of the sequence surrounding a microsatellite also plays a large role in the susceptibility of a repeat to MSI (Chung et al., 2010). For example, Harfe and Jinks-Robertson (2000) found that altering the 3 bases on either side of a 10bp homopolymer had up to a fourfold effect on the stability of the homopolymer. Another example of sequence context having an effect on homopolymer stability is that closely situated homopolymers are more mutable than a single homopolymer of the same length (Lang et al., 2013, Ma et al., 2012). In chapter 3 it was established that the MiSeq platform is appropriate for sequencing homopolymers and detecting microsatellite instability, but the frequency of instability was variable among different homopolymers. The optimal homopolymer length for an MSI test is still unclear, and it not clear if the repeats that have been reported in the literature are the best markers for an MSI test, or how many will be needed. In addition, the appropriate thresholds for distinguishing instability and error remain to be defined. Analysis of whole genome sequences may be informative

because there are many factors that increase the susceptibility of a repeat to MSI and it would be advantageous to find the most unstable homopolymers.

4.1.1.1. MSI within genes and intergenic regions

Although whole genome sequence data has been generated from CRCs (Cancer Genome Atlas Network, 2012), the dynamics of MSI in colorectal cancer has not been investigated in detail using whole genome sequence data. Next generation sequencing has allowed the identification of a large set of more informative markers for the identification of MSI in colorectal cancer. The focus of the literature to date had however been on exonic repeats using RNAseq data and exome data (Cancer Genome Atlas Network, 2012, Lu et al., 2013, Salipante et al., 2014, Terdiman et al., 2001). This limits the number of repeats that have been investigated, especially because less than 2% of the genome is coding, and homopolymers are under represented in exonic sequences compared to the rest of the genome (Borstnik and Pumpernik, 2002). There are also very few G/C homopolymers in the genome compared to A/T homopolymers (see Figure 4.1). Therefore, G/C homopolymers have not been studied in any great detail in MSI-H CRC samples because only investigating G/C homopolymers in coding sequences limits the numbers available for study.

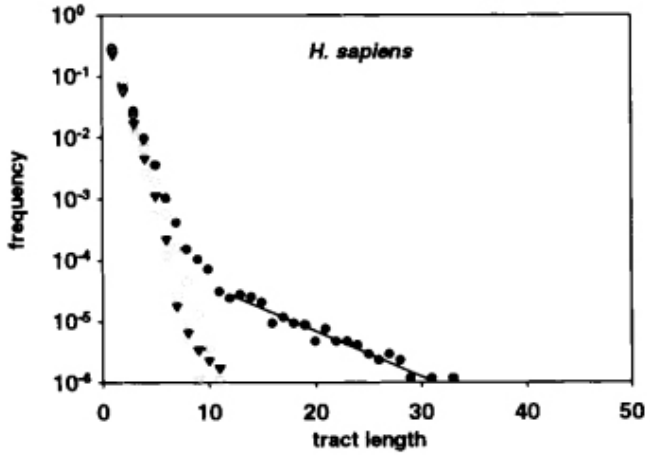

Figure 4.1: The frequency of different lengths of A/T and G/C homopolymers in the human genome. ▼G/C homopolymers. ● A/T homopolymers. This figure is modified from Dechering et al. (1998).

In exons a mutation in a repeat is likely to affect the function of the cell, and many repeats are therefore highly conserved. Due to most exonic homopolymers being highly conserved, a larger number of variable homopolymers are likely to be found in intergenic

regions. A few mutations will be more common because they give a replication advantage to the tumour cells affected. For example two repeats that show high mutation rates in colorectal cancers are a 10bp poly A repeat in the TGFBR2 gene and a 8bp poly A repeat in the ACVR2 gene. These two repeats were not included in the initial analysis due to the lack of a close SNP with a high minor allele frequency, but they have been analysed in other studies. One study looking at colorectal cancer cell lines and xenographs found that 22 out of 24 of the samples had a bi-allelic inactivation of the ACVR2 gene (Hempen et al., 2003). In 21 of the samples with bi-allelic inactivation there was a frameshift causing indel in the A8 repeat of ACVR2 exon 10 in at least one of the alleles. All 24 samples had an indel in at least one of the alleles of the A10 repeat in TGFBR2 (Hempen et al., 2003). These mutations in these repeats will be over represented in MSI-H tumours because they cause frameshift mutations that inactivate tumour suppressor genes leading to increased tumour cell proliferation.

Most mutations in coding homopolymers will affect the cell in a way that is not conducive to tumour cell proliferation and survival, leading to cells with theses mutations being selected against. Most highly unstable repeats in coding sequences are therefore limited to repeats where mutations do not lead to a survival disadvantage for tumour cells. It is therefore likely that there will be many more homopolymers that are highly variable in MSI-H tumours in intergenic regions which are not so highly conserved and where mutations are less likely to result in a negative selection pressure for cells. Furthermore, the sequence context is likely to be much more variable in non-coding regions, and this may affect repeat stability. For example, repeats situated within a few base pairs of each other are likely to be rare in coding sequences, and closely situated repeats that do not lead to a survival disadvantage for cells when mutated are likely to be rarer.

Because there are many factors that increase or decrease the susceptibility of a repeat to MSI it would be advantageous to use whole genome sequences to select the most unstable repeats for MSI testing. Having a larger set of unstable homopolymers would also allow for homopolymers in close proximity to SNPs with a high minor allele frequency to be chosen for further study. This would benefit the analysis of the allelic distribution of MSI. Whole genome sequence data for MSI-H and MSS CRCs including matching normal tissue was available from The Cancer Genome Atlas Network. This data had not previously been mined for unstable homopolymers in CRCs because of problems identifying indels in homopolymers using standard variant callers (Cancer Genome Atlas Network, 2012). The problems which were encountered included PCR and sequencing

error in controls which made it harder to identify differences between controls and MSI-H samples as well as the challenge of using standard variant callers to identify indels in repeats. The low coverage of the whole genome data (~3-4 fold sequence coverage) also makes it harder to distinguish PCR and sequencing errors from real mutations. In this chapter whole genome sequence data generated by The Cancer Genome Atlas Network is utilised to identify differences in homopolymer indel distributions between MSI-H tumours, matched normal tissue and MSS tumours. A list of unstable homopolymers is generated with information on neighbouring SNPs, which will later be used to identify homopolymers which are highly susceptible to MSI for further study. To overcome the challenges of using low coverage sequence data the variant calls from all tumour in each group were pooled prior to analysis.

4.1.1.2. Accuracy of indel identification by variant callers

Despite the availability of low pass whole genome sequence data for repeat identification, at the outset of this work there was no consensus as to the most appropriate variant caller for indel identification in homopolymers. A potential issue with identifying highly unstable homopolymers is that for indels there is still very little consistency between different variant callers (Li, 2014, O'Rawe et al., 2013). O'Rawe et al. (2013) assessed three different variant calling pipelines (SOAPindel, BWA-GATK, SAMtools) and discovered that there was only a 26.8% concordance between the indels being called using those pipelines. 28.5% of the indels were unique to GATK, 22.4% unique to SOAPindel, and 7.8% unique to SAMtools (O'Rawe et al., 2013). Pabinger et al. (2014) compared the number of indel calls made by CRISP, GATK, SAMtools, SNVer and VarScan 2, and they called 259, 1959, 234, 332 and 1896 indels respectively, with GATK and VarScan having the largest number of indels in common (~57%). More recently, Houniet et al. (2015) have evaluated the indel callers Samtools, Dindel and GATK for their ability to identify indels in exome sequences. The results of their analysis showed that Samtools had a sensitivity of less than 0.05 for identifying indels while GATK had a sensitivity of around 0.35 and Dindel had a sensitivity ranging from ~0.17 – ~0.38 depending on which aligner was used.

Sequencing and PCR errors are likely to cause problems for indel calling. There could be a lot of "noise" in the form of sequencing and PCR error around homopolymers, making it important to be able to pick out real indels from the background noise. Most

indel callers have advanced models for dealing with indels in homopolymers caused by PCR errors. However there is concern that PCR errors in homopolymers are still not being modelled well, with different variant callers calling different indels (Li, 2014).

Another challenge when it comes to calling indels is obtaining the correct gapped alignment around indels. This is especially true for low complexity regions where realignment of indels is a large challenge (Li, 2014). As many homopolymers are present in low complexity regions, alignment problems can make it harder to differentiate between MSI and false positive indels. Li (2014) reported that 50-70% of the heterozygous indels detected in their CHM1 cell line sequence data would not exist with better realignment. Equally, true indels may be lost after being filtered out by low-complexity filters. Another alignment problem can be caused because the human genome sequence still has gaps where sequence information is missing. Missing paralogous sequences can lead to incorrect alignments and the generation of errors (Li, 2014).

Furthermore, most variant callers are geared towards bi-allelic genomes not cancer genomes with multiple alleles. This means that the programs are expecting variants in either a heterozygous or a homozygous form. This can lead to allele bias filters removing low frequency variants because they do not meet the criteria for being called heterozygous.

Because of the limitations to indel callers mentioned above it will be important to assess available callers, select the most appropriate for identifying indels in homopolymers from whole genome data, and be aware of any limitations which may affect the selection of appropriate homopolymers for an MSI test. In chapter 3, repeats were analysed with a simple indel caller, COPReC, that addresses some of these issues, however this caller cannot be used on the whole genome sequence data because this caller uses overlapping paired end reads. Therefore, for the work conducted in this chapter another caller was needed.

4.1.2. Aims

At the outset of this work, no genome wide analysis of short homopolymer stability in CRCs had been performed, despite the availability of low pass sequence data. However, such an analysis would be required to identify the most variable and informative markers for use in a sequence based MSI test. The lack of consistency of

available variant callers was, however, a barrier to such an analysis. The aims for this work in this chapter are to:

- Assess the three most commonly used variant callers Dindel, GATK and VarScan to find the most appropriate for indel discovery in homopolymers.
- Assess the impact of size and homopolymer sequence upon PCR/sequencing error and instability in CRCs
- Evaluate the indel distribution in 7bp-12bp homopolymers in MSI-H samples using whole genome sequence data.
- Discover homopolymers that are highly variable in MSI-H samples, but not in control samples, to enable further assessment of these repeats for inclusion in a sequence based MSI test.

4.2. Results

4.2.1. Comparison of variant callers

To identify an appropriate indel caller for genome wide homopolymer analysis, firstly a single control exome sequence was analysed to compare the indel calling of Dindel (Albers et al., 2011), GATK (DePristo et al., 2011), and VarScan (Koboldt et al., 2009). These are 3 of the most commonly used indel callers (Neuman et al., 2013). The Illumina reads generated from 1 normal control exome sequence were provided by Dr Mauro Santibaez-Koref, (Institute of Genetic Medicine, Newcastle University). The aim was to find parameters for indel calling that would allow MSI to be distinguished from microsatellite stability.

For sequence alignment the Burrows–Wheeler Aligner (BWA) (version 0.6.2) (Li and Durbin, 2009) was used. Samtools (version 0.1.8) (Li et al., 2009) was used to create a sorted bam file and the pileup file needed for variant calling with Varscan. Variant calling from the pileup file was achieved using Varscan pilup2indel (version 2.2.2). Variant calling was also performed using Dindel (version 1.01) with the sorted bam file as input. Prior to indel calling using GATK duplicate sequences were removed from the sorted bam file using Picard (version 1.75) (PICARD [http://picard.sourceforge.net]). GATK (version 2.2.9) was then used to realign around indels before variant calling was performed using GATK's UnifiedGenotyper (version 2.2.9) and the HomopolymerRun tool for annotation. See method sections 2.8.6 for further information about the methods used for variant calling.

The differences in indel calling between Varscan, Dindel and GATK were initially assessed by seeing how many indels in homopolymer ≥6bp each program had found between positions 1-3395973 on chromosome 1 in the control exome sequence. This was achieved by looking through the VCF files and counting the number of homopolymers with indels that had been recorded by each program. Using the positional information for each homopolymer it was possible to identify which indels had been identified by more than one caller.

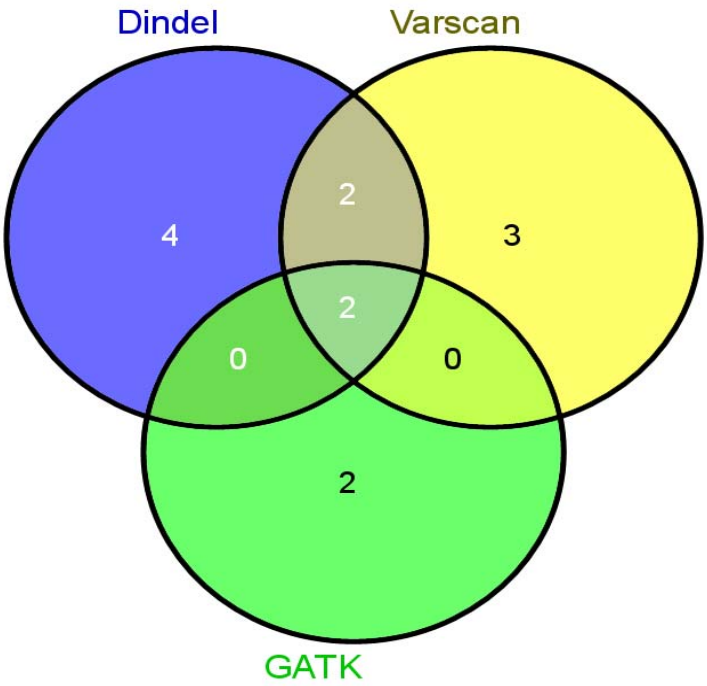

Figure 4.2: Indels called by the variant callers Dindel, VarScan and GATK between positions 1-3395973 on chromosome 1 for one control exome sequence.

Only 2 out of the 13 indels called by these three variant callers were found by all the variant callers (Figure 4.2). The differences in indel calling between Dindel, GATK, and VarScan shows that there is a difference in the algorithms these three programs use to call indels. This is consistent with reports in the literature that there are large differences in indel calling between different callers (Li, 2014, O'Rawe et al., 2013). This also confirms the need to select a caller that is appropriate for calling indels in homopolymers.

Both GATK and Dindel have the option to annotate variants found in homopolymer runs. VarScan on the other hand does not contain this option. This means that the output (the Variant Call Files or VCF files) from VarScan could not easily be filtered for indels in homopolymers. VarScan was therefore not assessed further. To extract all relevant annotations in the VCF files generated by GATK and Dindel a Perl script was generated (Dindel_GATK_compare.pl) that identifies homopolymers >7bp using the homopolymer run annotations in the GATK and Dindel VCF files. This program then and counts and lists the indels that two VCF files have in common as well as counting and listing the indels that are unique to a VCF file using the chromosome and position data to determine if two indels are the same. For both Dindel and GATK variants are left aligned in homopolymers, so variants in the same homopolymer will be given the same position by both programs. This script was then used to compare the Dindel and GATK

indel calls from the control exome being analysed. The results of this comparison are shown in Figure 4.3

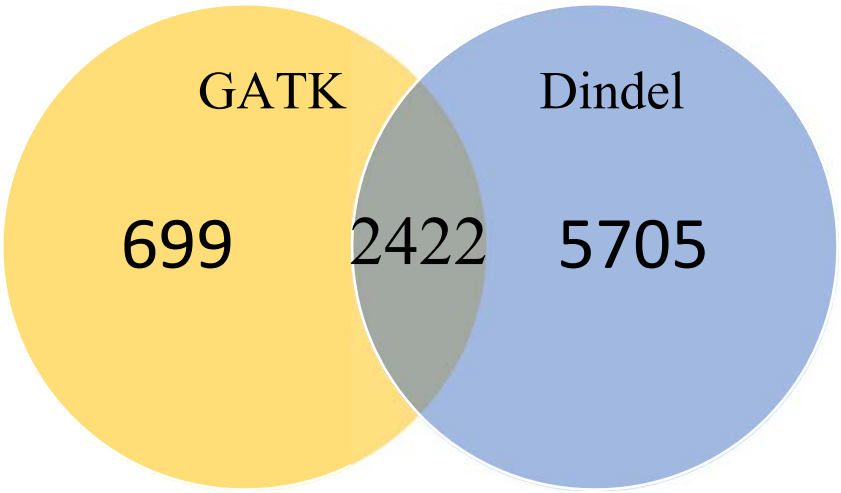

Figure 4.3: The number of indels in homopolymers called by Dindel and GATK across one control exome sequence.

There is a large difference in the number of indels in homopolymers >7bp called by Dindel and GATK, with Dindel identifying 8127 indels and GATK identifying 3121 indels. There are also many indels calls that are unique to each program. The overlap of indels called by both programs consists of 78% of the indel calls made by GATK and 30% of the indel calls made by Dindel. Because the differences in indel calling between programs are large an attempt was made to distinguish between the two methods in terms of their ability to detect indels within homopolymers that are likely to be real. To do this, this a set of criteria were defined to distinguish between indels which were likely to be false calls, and indels which looked real. This was done to enable the quality of indel calls for both programs to be assessed. The criteria used to determine if an indel was likely to be real or if it should be disregarded as a false indel can be found in Figure 4.4. The aim was to identify the variant caller that misses the least real indels and includes the least spurious indels.

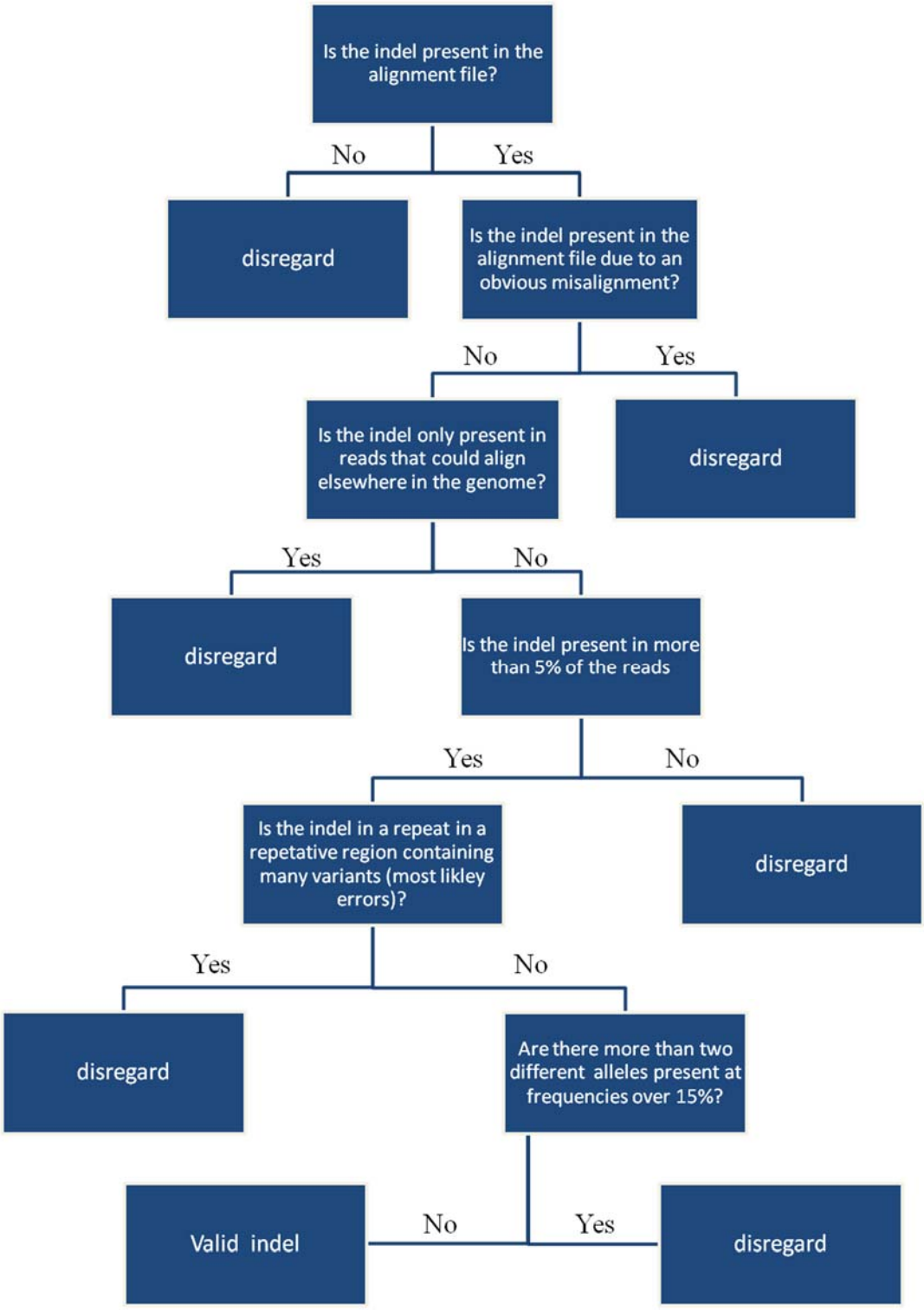

Figure 4.4: Flow chart depicting a method for distinguishing between false and real indel calls.

Using the flow chart shown in Figure 4.4 the first 15 indels were manually classified as "real" or "false positive" in each of the following categories: Indels called by both GATK and Dindel, Indels unique to GATK, and Indels unique to Dindel. The Integrative Genomics Viewer (IGV) (Robinson et al., 2011) was used to visualise the

aligned reads for each indel in order to determine if the indel conformed to my criteria for a real indel. The results of this analysis can be found in Figure 4.5.

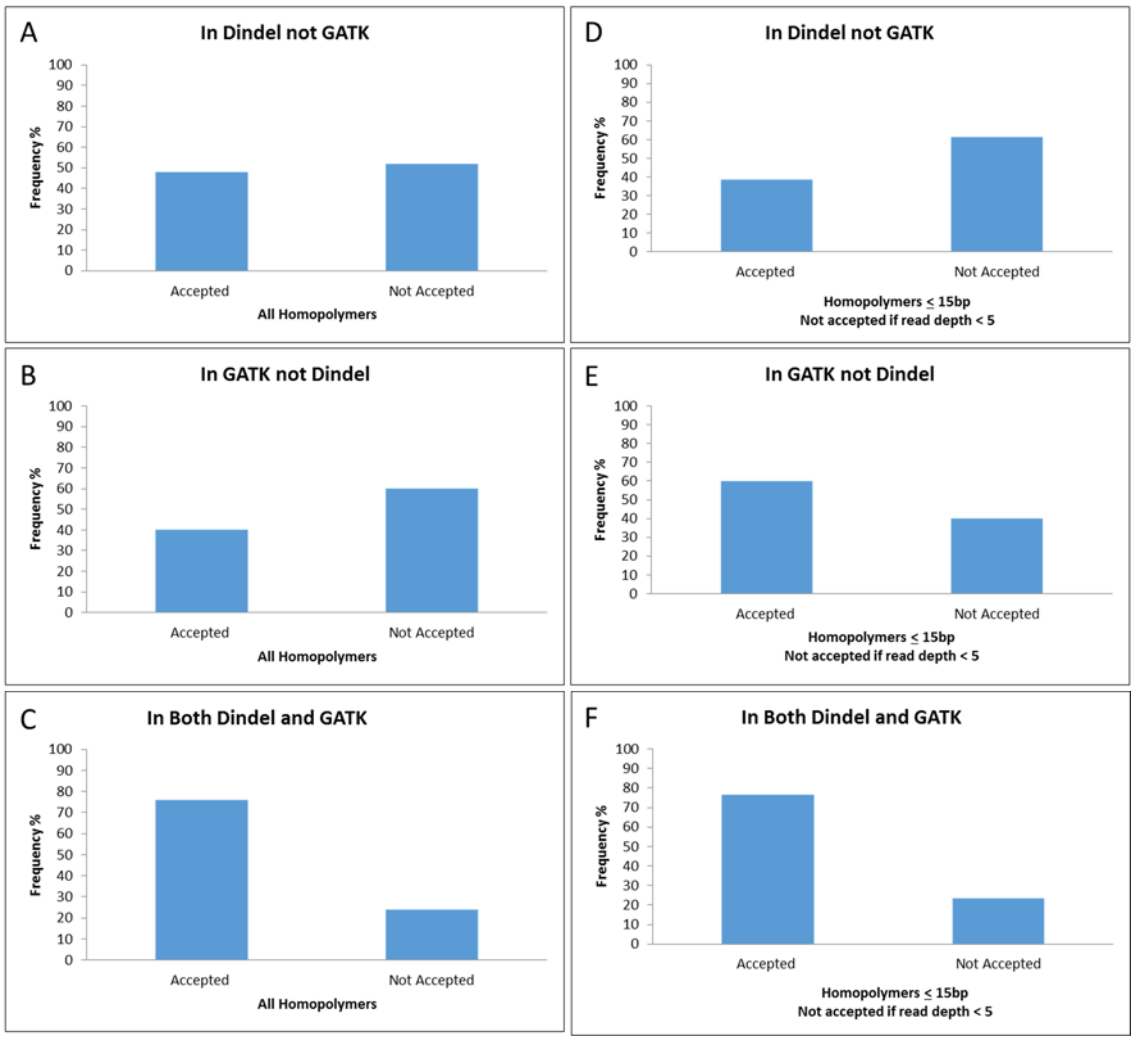

Figure 4.5: A comparison between the variant callers GATK and Dindel. Indels that were deemed to be accepted if they passed the criteria found in the flow chart in Figure 4.4. The first 15 indels in each of the following categories: Indels called by both GATK and Dindel, Indels unique to GATK, and Indels unique to Dindel were analysed manually.

Based on the results shown in Figure 4.5 GATK was deemed to be better than Dindel for analysing indels in homopolymers. This is because GATK had a larger number of indel calls that passed filter for the variants that were unique to the caller compared to Dindel for homopolymers ≤15bp. 60% of the indels unique to GATK pass filter (see Figure 4.5 panel E) as opposed to 38.5% for indels unique to Dindel (see Figure 4.5 panel D), and this only increases to 76.5% when both used (see Figure 4.5 panel F). So based on this limited analysis, GATK is superior for calling indels in homopolymers ≤15bp. Dindel had the largest number of indel calls (see figure Figure 4.3), but a higher rate of calls that did not pass filter for homopolymers ≤15bp compared to GATK would mean

that it would be harder to distinguish between MSI-H and MSS samples as there would be more false positive indel calls in both groups.

Subsequent to the initial analysis of the variant callers and GATK being chosen as the most appropriate variant caller for identifying indels in homopolymers, a problem with GATKs variant calling was discovered. This problem was a fault in the HomopolymerRun annotation, which GATK uses to annotate homopolymer runs. Therefore GATK's TandemRepeatAnnotator was used instead of the HomopolymerRun annotator for all subsequent analyses. This is likely to have made GATK even better at identifying indels in homopolymers than previously because some homopolymers were being missed using the HomopolymerRun annotator.

4.2.2. Homopolymer analysis of CRCs and controls from TCGA whole genome data

In order to study the indel distribution for different homopolymer lengths in MSI-H colorectal cancers and compile a list homopolymers susceptible to MSI, analyses of the MSI-H tumour whole genome sequences was chosen and compared to matched normal and MSS whole genome sequences. The Cancer Genome Atlas Network (TCGA) produced the whole genome sequences used in this chapter. These sequences have an average ~3-4 fold sequence coverage, which means that it would be beneficial to pool the variant calls of all samples of the same type prior to analysing the data instead of analysing each sample separately.

Low depth whole genome sequences consisting of 12 MSI-H tumours, 12 MSS tumours and matched normal tissue for 11 of the MSI-H tumours were mined for indels in all 7-12bp homopolymers using the Burrows–Wheeler Aligner (BWA) (version 0.6.2), Samtools (version 0.1.8), the GATK UnifiedGenotyper (version 2.2.9), and Perl scripts (see methods section 2.8.7). First, the BAM files obtained from TCGA were converted to fastq files using bam2fastq (version 1.1.0) (bam2fastq software [http://gsl.hudsonalpha.org/information/software/bam2fastq]). Then BWA was used to covert the fastq files to SAM files, which were then converted to BAM files and sorted using Samtools. Picard (version 1.75) was used to exclude sequence duplicates. GATK was then used to merge all BAM files creating a multi-sample BAM file and to perform a realignment around indels for the multi-sample BAM file. Realignment of all samples in a multi-sample BAM file was done to ensure consistent alignment of low pass data

from all samples. GATK's UnifiedGenotyper was used to produce a file containing raw variant calls with GATK's TandemRepeatAnnotater to annotate homopolymers. All homopolymers with known polymorphisms as of dbSNP (version 137, hg19) were also annotated. The selection of indel calls in 7-12bp homopolymers from the GATK variant call file was performed using the Perl script Perl_SelectVariants_RPA_RU.pl (see methods section 2.8.7.2). Analysis of read frequencies and indel size distributions were done using Perl scripts (see methods sections 2.8.7.2 and 2.8.7.3). Because of the low pass nature of the sequence data, all reads from each group (MSI-H samples, MSS samples and matched normal for the MSI-H samples) were combined before analysis. Any SNPs with a high minor allele frequency from dbSNP (version 137, hg19) (Sherry et al., 2001) within 30bp of the start of repeats were annotated using the Perl script AnnotateCloseSNPs.pl (see methods section 2.8.7.4). All homopolymers with known polymorphisms were excluded.

218181 variable 7-12bp homopolymers were identified. A/T and G/C homopolymer were analysed separately.

4.2.3. Indel frequencies in A/T homopolymers from whole genome data

To investigate the stability of short A/T homopolymers in tumours with mismatch repair defects at the genome level, the indel profiles of all A/T 7-12bp repeats called by GATK within whole genome sequence data from CRC tumours, after removal of all repeats with common polymorphisms (dbSNP version 173) were analysed. 216495 A/T homopolymers were identified. For the A/T homopolymers the distribution of variant read frequencies in the MSI-H tumours differed from those of the MSS tumours and matched normal samples for all homopolymer lengths investigated (see Figure 4.6). The distribution of variant read frequencies for the MSS tumours and matched normal samples are the similar (see blue and green lines Figure 4.6). Because the distribution of variant read frequencies is the same in both the control sample groups, but different in the MSI-H samples (red line Figure 4.6) it is likely that the difference reflect MSI.

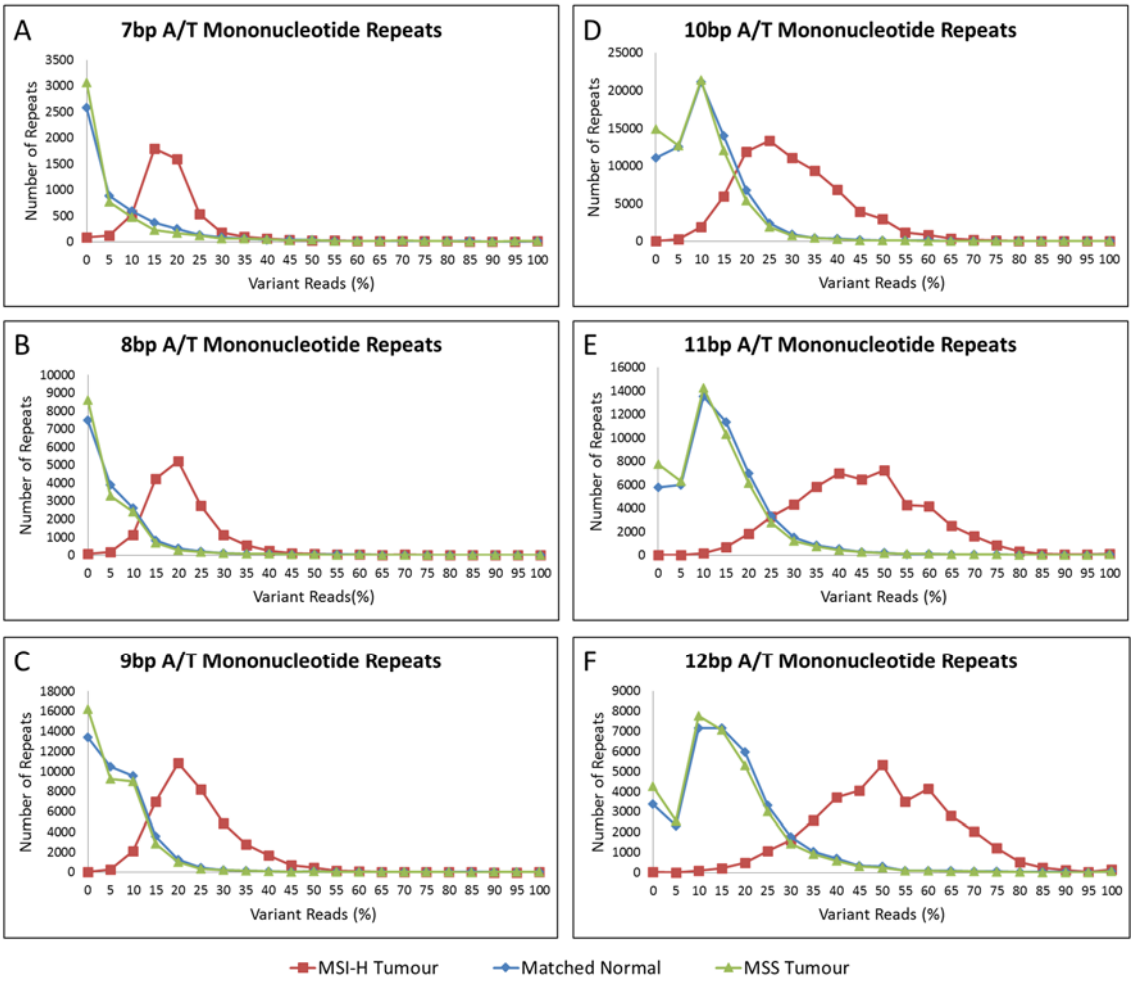

Figure 4.6: Frequencies of variant reads in homopolymers for MSI-H tumours, matched normal tissue, and MSS tumours. Only homopolymers with no known polymorphisms were included in this analysis. Panels A-F: Variant read frequencies in 7bp-12bp A/T homopolymers.

The variant reads in the MSS tumours and matched normal samples in Figure 4.6 are presumed to be caused mainly by sequencing and PCR error. For the 7bp-9bp repeats the peak of the curve for the MSS samples is at a variant read frequency of zero, which means that for a large number of repeats there has been no PCR/sequencing error detected (see Figure 4.6 panels A-C). For these repeats the number of homopolymers decreases with an increased frequency of variant reads until the graphs level out around at a variant read frequency of ~25%. For the 10bp-12bp repeats the peaks of the curve for the MSS tumours and matched normal samples is no longer at zero percent (see Figure 4.6 panels D-F). There has been a shift in the peak of the curve to 10% for the 10bp and 11bp repeats, and to 10%-15% for the 12bp repeats. This shift in the curve is presumed to be caused by an increase in PCR/sequencing error for these longer repeats. An increase in PCR error with repeat length is consistent with results from chapter 3 and results reported in the literature (Clarke et al., 2001, Fazekas et al., 2010, Flores-Renteria and Whipple, 2011).

The MSI-H samples had more variant reads than MSS and normal samples (see Figure 4.6). The distributions of variant reads vary, and the peak frequency increases steadily with homopolymer size. However, the range of the distribution also increases with homopolymer length. The peak of the curve for the MSI-H samples is at a higher variant read frequency compared to the control samples for all repeat lengths. This will be due to the presence of variant reads caused by MSI as well as variant reads caused by PCR/sequencing error. It is to be expected that the levels of PCR and sequencing error in the MSI-H samples would be equivalent to that observed in the controls for the same repeat length.

The frequency of variant reads in the MSI-H samples increases with repeat length (see Figure 4.6). This is consistent with longer repeats being more prone to MSI (Sammalkorpi et al., 2007, Vilkki et al., 2002, Woerner et al., 2003). For the 7bp repeats the majority of microsatellite unstable repeats have a variant read frequency of between 10%-25%. This increases with repeat length up to the 12bp repeats where the majority of microsatellite unstable repeats have a variant read frequency of between 30%-80%.

4.2.4. Indel frequencies in G/C homopolymers from whole genome data

For the G/C homopolymers there are so few homopolymers of each repeat length that it is more difficult to discern patterns in the data than it was for the A/T homopolymers. In total 1686 G/C homopolymers were identified after the removal of all repeats with common sequence length variants using dbSNP version 173. The distribution of variant reads for the MSS tumours and matched normal samples are the same in the 7bp-10bp G/C homopolymers (see Figure 4.7 panels A-D). This suggests that there are enough homopolymers present for these repeat lengths to show the trends of the data. For the 11bp and 12bp homopolymers there are so few repeats that there is not enough data for a proper analysis (see Figure 4.7 panels E and F).

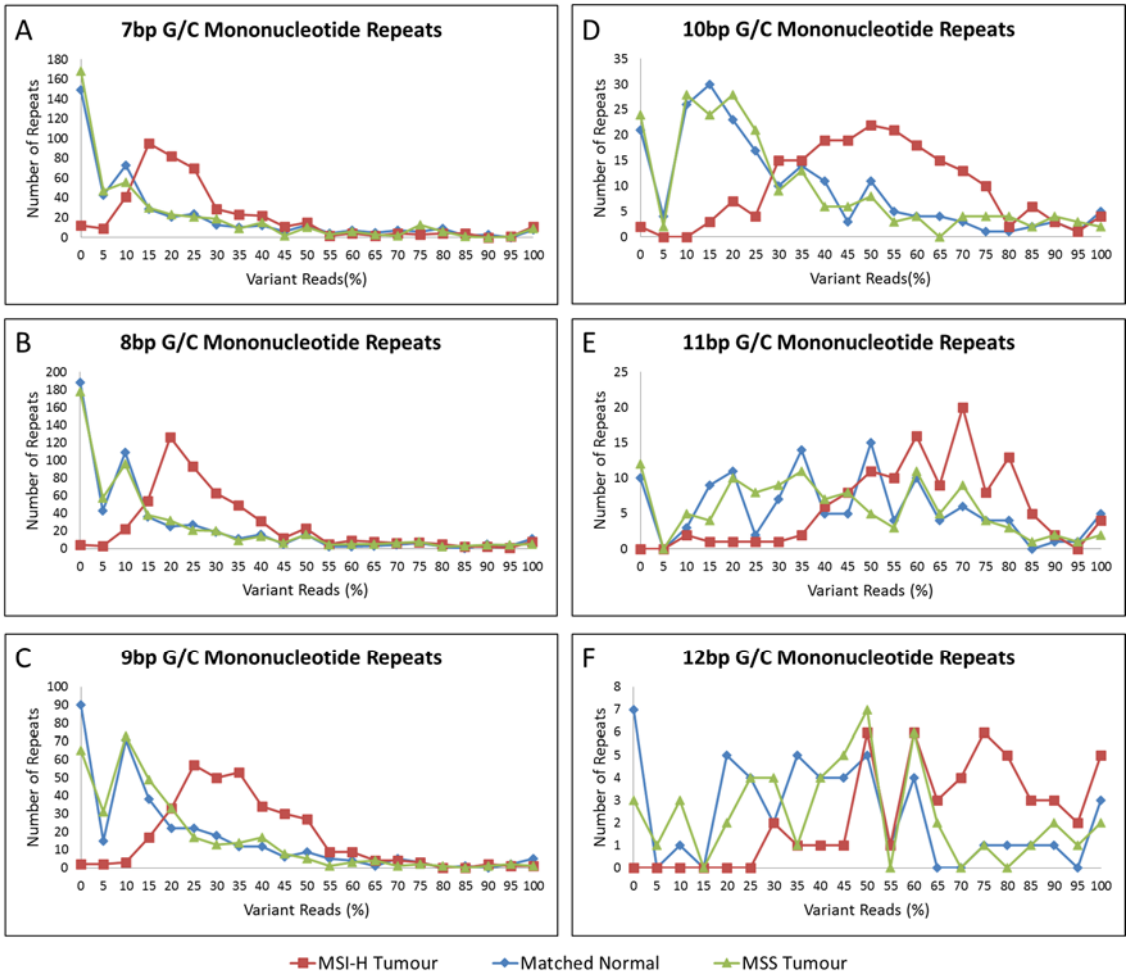

Figure 4.7: Frequencies of variant reads in homopolymers for MSI-H tumours, matched normal tissue, and MSS tumours. Only homopolymers with no known polymorphisms were included in this analysis. Panels A-F: Variant read frequencies in 7bp-12bp G/C homopolymers.

The variant read distribution in the MSI-H samples differs from the distributions in the control samples, with the homopolymers in the MSI-H samples having a higher frequency of variant reads (see Figure 4.7). The results in Figure 4.7 show that, as observed with the A/T homopolymers, MSI increases with repeat length for the G/C repeats. Interestingly the MSI-H samples have a greater variant read frequency for the G/C homopolymers compared to the A/T homopolymers of an equivalent length (see Figure 4.6 and Figure 4.7). For example, most of the 9bp G/C homopolymers in the MSI-H samples have a variant read frequency between 10% and 55%, whereas the 9bp A/T homopolymers in the MSI-H samples have a variant read frequency between 5% and 45%. Also the shape of the curve for the 10bp G/C repeats in the MSI-H samples is more reminiscent of the curve for the 11bp or 12bp A/T repeats than the 10bp A/T repeats. The frequency of variant reads in the control samples also increase with repeat length (see Figure 4.7). These variant reads are presumed to be caused by PCR and sequencing error.

The frequency of variant reads for the control sample C/G homopolymers is also higher than for the equivalent A/T homopolymer repeat lengths (see Figure 4.6 and Figure 4.7).

4.2.5. The distributions of indel sizes

To investigate the size distribution of variant reads in MSI-H tumours, variant read lengths were analysed, and results for 7bp - 12bp A/T repeats are shown in Figure 4.8. The prevalence of deletions is higher than insertions in the MSI-H sample group. This suggests that deletions are more indicative of MSI than insertions, and is consistent with results seen in gastric cancers (Yoon et al., 2013). As the repeat size increases the fraction of the indels that is made up of insertions dwindles until in the 10bp repeats the excess of 1bp insertions in the MSI-H tumours compared to the control samples is marginal (see Figure 4.8 panel D). In the 11bp and 12bp repeats there is no excess in insertions in the MSI-H samples compared to the controls (see Figure 4.8 panel E and F).

For the 7bp, 8bp and 9bp homopolymers most of the deletions were 1bp in length (see Figure 4.8 panels A-C). However, there is an emergence of additional variant homopolymer lengths in the larger repeats (see Figure 4.8 panels D-F). In the 10bp and 11bp repeats there are more 2bp deletions observed in the MSI-H samples compared to the controls. For the longer 12bp homopolymers there was a surplus of 1bp, 2bp and 3bp deletions in the MSI-H group compared to the controls. The deletions present in the controls are at a lower frequency than in the MSI-H samples, and as mentioned before are assumed to be derived from PCR artefacts and sequence error. In the control samples there are more deletions than insertions present for all repeat sizes (see Figure 4.8).

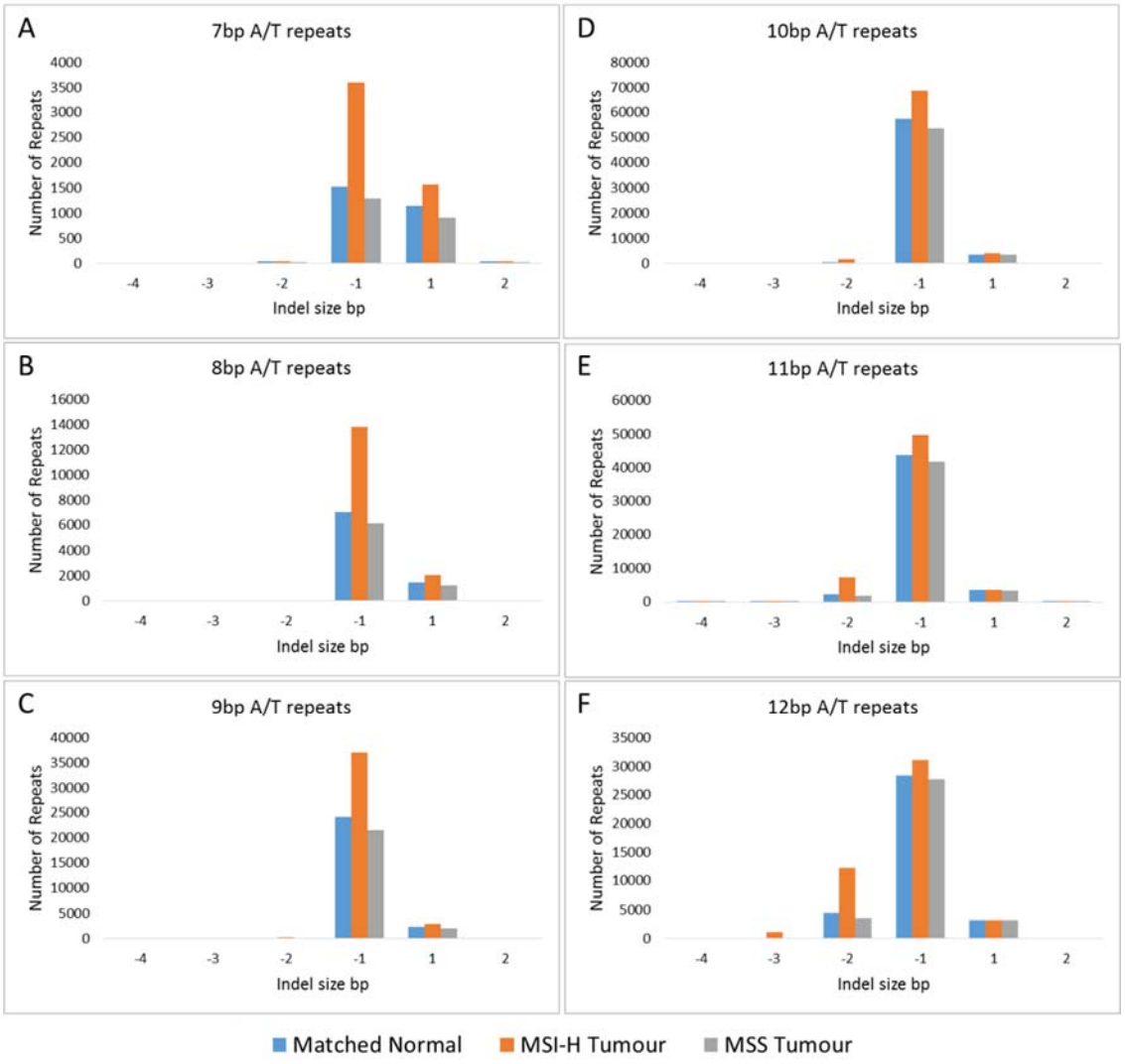

Figure 4.8: The indel distributions observed in the 7bp -12bp A/T homopolymers extracted from whole genome sequence data. Repeats were recorded as including an indel size if the indel was observed one or more reads. Panel A: The distribution of indels in the 7bp repeats. Panel B: The distribution of indels in the 8bp repeats. Panel C: The distribution of indels in the 9bp repeats. Panel D: The distribution of indels in the 10bp repeats Panel E: The distribution of indels in the 11bp repeats. Panel F: The distribution of indels in the 12bp repeats.

In general, the indels seen in the MSI-H samples are at a higher frequency than in the control samples (see Figure 4.9). For Figure 4.9 indel frequencies were deliberately but arbitrarily chosen for each repeat length to highlight the differences between MSI-H samples and controls. For the 7bp and 8bp A/T repeats there is a large excess of both 1bp deletions and 1bp insertions at a frequency ≥10%. As the repeat size increases the fraction of high frequency insertions diminishes, while there is an emergence of high frequency 2bp and 3bp deletions in the MSI-H samples (see Figure 4.9).

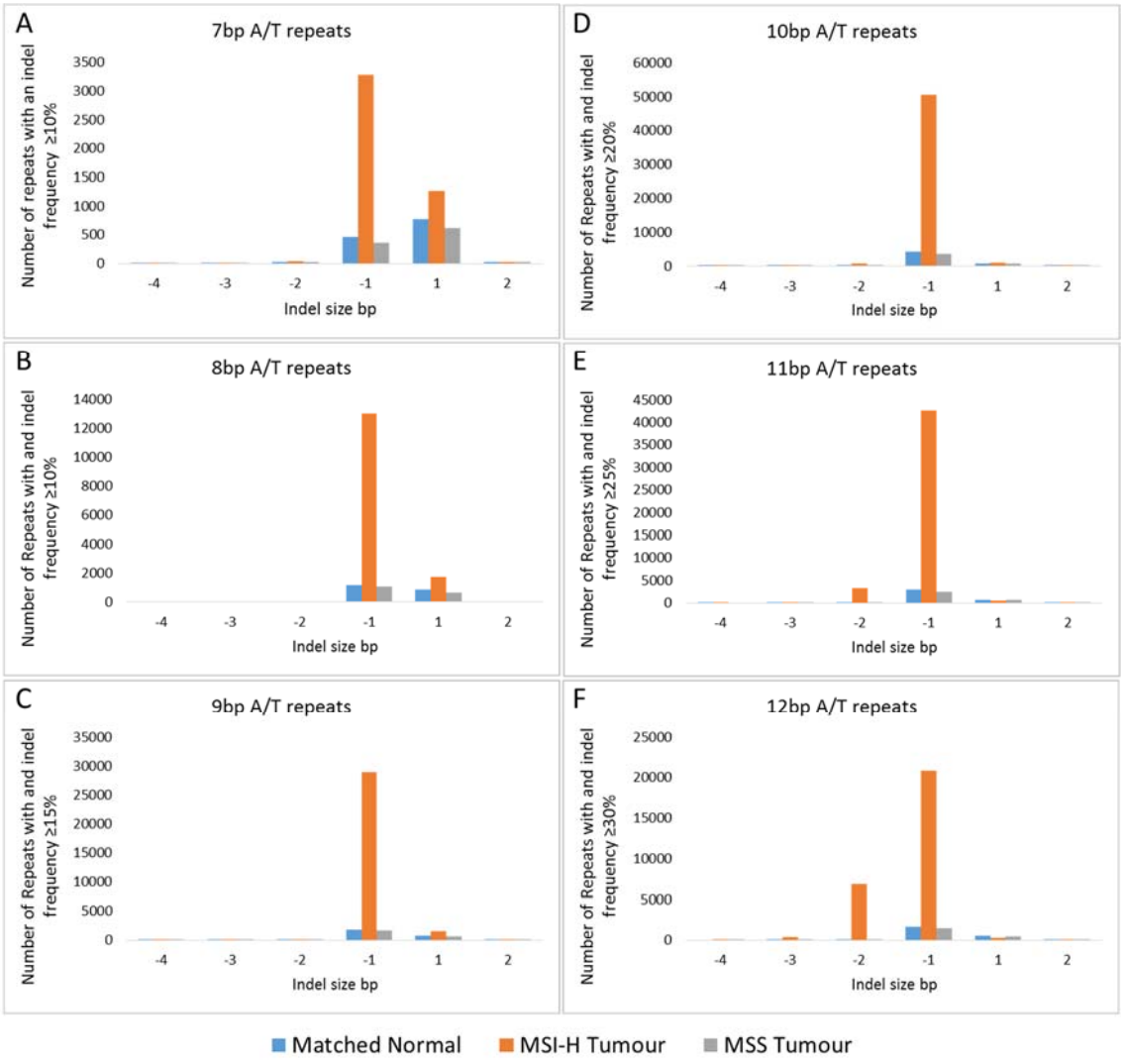

Figure 4.9: The distributions of high frequency indels observed in the 7bp -12bp A/T homopolymers extracted from whole genome sequence data. Cut-offs of different indel frequencies for different repeat lengths were deliberately chosen to highlight differences between the MSI-H samples and controls. Panel A: The distribution of indels with a frequency ≥10% for the 7bp repeats. Panel B: The distribution of indels with a frequency ≥10% for the 8bp repeats. Panel C: The distribution of indels with a frequency ≥15% for the 9bp repeats. Panel D: The distribution of indels with a frequency ≥20% for the 10bp repeats. Panel E: The distribution of indels with a frequency ≥25% for the 11bp repeats. Panel F: The distribution of indels with a frequency ≥30% for the 12bp repeats.

Because of the small number of 11bp and 12bp G/C homopolymers the indel distributions in these have not been analysed. In comparison to the A/T repeats of equal size, a greater fraction of the indels observed in the G/C repeats consist of insertions (see Figure 4.8 and Figure 4.10). In the G/C repeats, the fraction of indels consisting of insertions diminishes with increased repeat length, just as seen in the A/T repeats. The distribution of indels in the 7bp-9bp repeats consists of mainly 1bp deletions and 1bp insertions, with an excess of these indel sizes in the MSI-H samples compared to controls. In the 10bp repeats, there is an excess of 1bp and 2bp deletions in the MSI-H samples compared to the control samples (see Figure 4.10 panel D). The 10bp A/T repeats are also

the shortest repeats where an excess of 2bp deletions is seen in the MSI-H samples compared to controls.

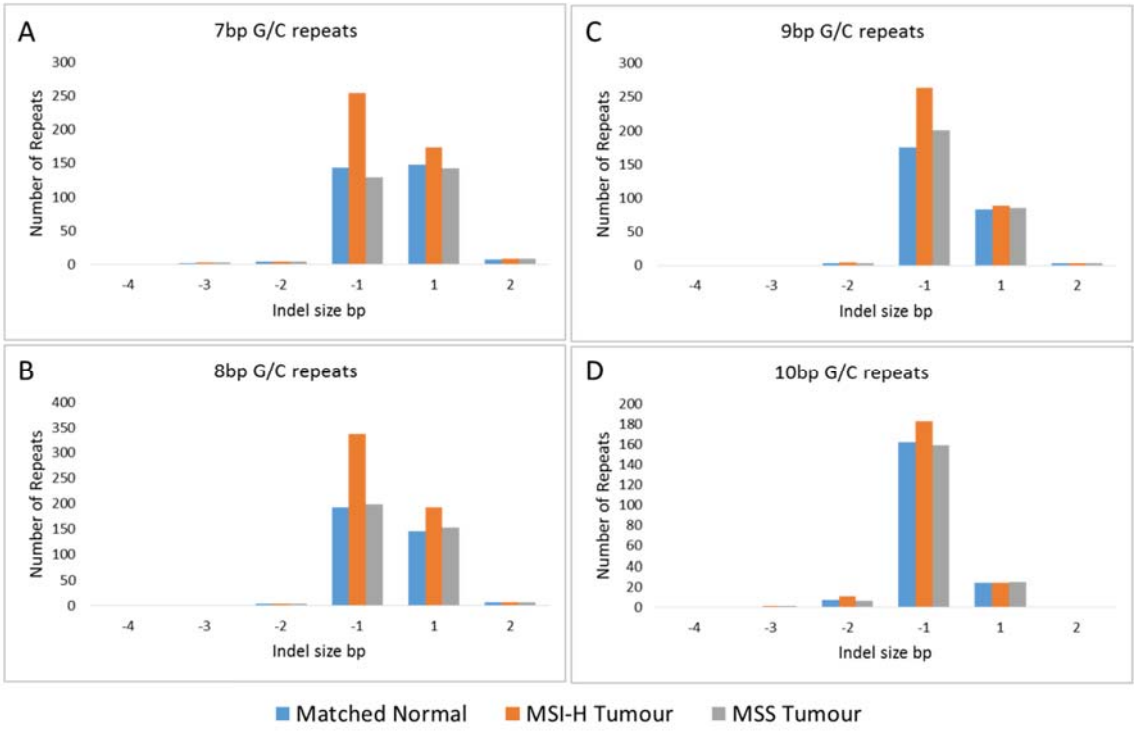

Figure 4.10: The indel distributions observed in the 7bp -10bp G/C homopolymers extracted from whole genome sequence data. Panel A: The distribution of indels in the 7bp repeats. Panel B: The distribution of indels in the 8bp repeats. Panel C: The distribution of indels in the 9bp repeats. Panel D: The distribution of indels in the 10bp repeats.

For the G/C homopolymers there are also a much greater fraction repeats with high frequency indels in the MSI-H samples compared to the controls (see Figure 4.11). In the control samples there are more high frequency insertions than deletions in the 7bp-8bp repeats. There is a larger difference in the number of high frequency deletions between the MSI-H samples and the control samples than the difference in number of high frequency insertions between the two groups. This suggests that, also for the G/C repeats, deletions are more indicative of MSI than insertions.

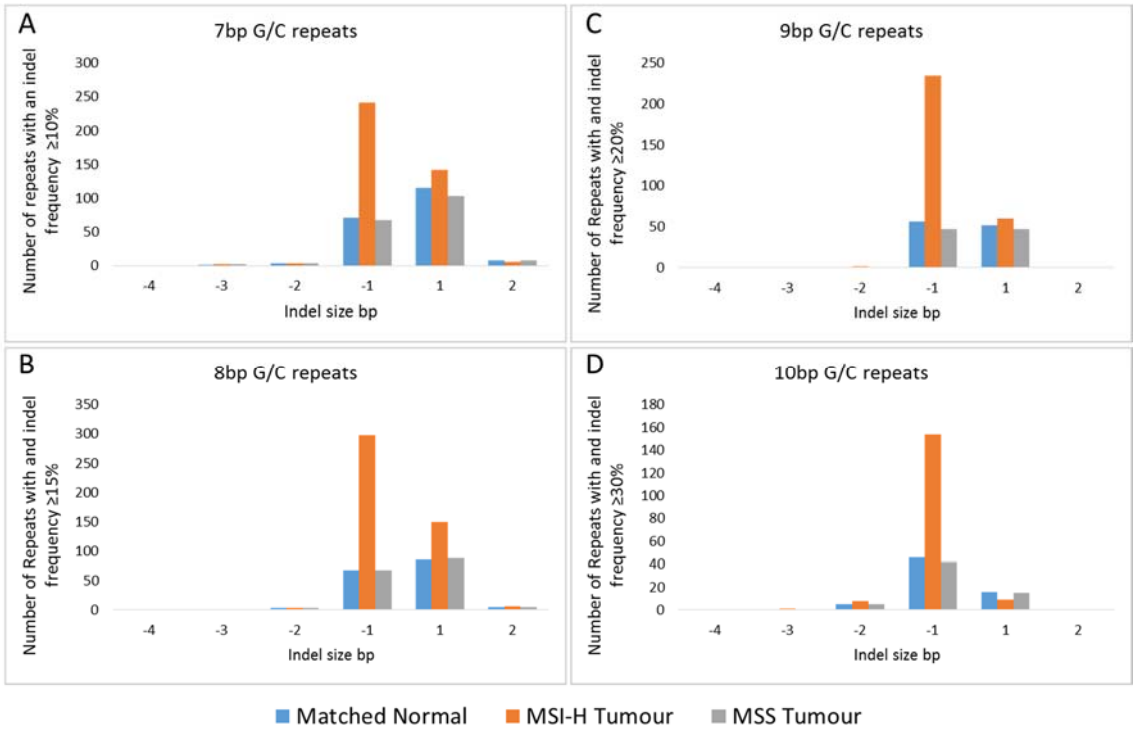

Figure 4.11: The distributions of high frequency indels observed in the 7bp -12bp G/C homopolymers extracted from whole genome sequence data. Panel A: The distribution of indels with a frequency ≥10% for the 7bp repeats. Panel B: The distribution of indels with a frequency ≥15% for the 8bp repeats. Panel C: The distribution of indels with a frequency ≥20% for the 9bp repeats. Panel D: The distribution of indels with a frequency ≥30% for the 10bp repeats.

A file containing the 218181 variable 7-12bp homopolymers that were identified in this chapter can be found in the file Homopolymer_SNP_file_mmr7-12bp on the supplementary CD. This file also contains minor allele frequency annotations for any SNPs within 30bp of the start of each repeat. In the next chapter, this file will be used to identify 7bp-10bp homopolymers with a variant read frequency ≥10% in the MSI-H samples and no variant reads in the controls, and 11bp-12bp homopolymers with a variant read frequency ≥15% in the MSI-H samples and variant allele fraction of ≤5% in the controls. These repeats will be identified for further analysis of and possible inclusion in a sequence based MSI test.

4.3. Discussion

Because no whole genome studies analysing indels in homopolymers had been performed for MSI-H colorectal tumours, in this chapter the indel profiles of 7-12bp homopolymers in 12 MSI-H colorectal tumours and controls were investigated. The aims of this analysis were to analyse the impact of homopolymer length and repeat unit on indel distributions in MSI-H colorectal tumours, and also generate a list of homopolymer that were found to be highly variable in MSI-H samples so that some of these repeats could be assessed in an independent panel of tumours later. In order to achieve these aims, an indel caller, which was appropriate for analysing indels in homopolymers was needed. To find a good indel caller three commonly used indel callers, VarScan, Dindel and GATK, were evaluated.

The comparison between indel callers showed that there are differences in indel calls between the programs VarScan, Dindel and GATK with only 2 out of the 13 indels identified in a small stretch of exome sequence being called by all three programs. Because of this, comparisons were made to determine which program would be the most appropriate for calling indels in homopolymers using a control exome sequence. GATK was chosen as the most appropriate indel caller. This was because GATK had a higher ratio of indel calls that passed filter to indel calls that failed filter for homopolymers ≤15bp compared to Dindel. VarScan was excluded because it had no annotations for homopolymer and it would therefore be difficult to single out indels in homopolymers using this caller. GATKs HomopolymerRun annotator was used in the initial assessment of variant callers, but GATKs TandemRepeatAnnotater was found to be better than the HomopolymerRun annotator because the HomopolymerRun annotator failed to annotate all homopolymers. The TandemRepeatAnnotater was therefore used for all subsequent analyses. Using GATKs UnifiedGenotyper and TandemRepeatAnnotater the distribution of variant read frequencies in the MSI-H tumours differed from those of the controls. The results from the comparisons between MSI-H samples and controls suggests that an appropriate indel caller was chosen.

Using pooled low depth genome sequences the MSI-H samples were easily distinguishable from the controls (see Figure 4.6 and Figure 4.7). The distribution of indel frequencies in the matched normal and MSS samples were the same, and therefore the difference in the distribution in the MSI-H samples can be attributed to mutations accumulated due to failure of the mismatch repair system in the MSI-H tumours. There

were more indels in homopolymers in the MSI-H samples compared to the control samples and most of the indels in the MSI-H samples were found at a higher frequency than the indels in the controls. This suggests that most of the indels caused as a result of MSI occur at a higher frequency compared to indels caused by PCR and sequencing error. As the repeat length increases so does the frequency of variant reads in the MSI-H samples, but so does PCR/sequence error. There is therefore a trade-off between the susceptibility of repeats to MSI events and noise with increased repeat length.

For all repeats, the fraction of high frequency indels consisting of deletions was higher than insertions (see Figure 4.9 and Figure 4.11). For both the A/T and the G/C homopolymers the fraction of indels consisting of insertions was highest in the 7bp homopolymers (see Figure 4.8 and Figure 4.10). This fraction decreases until in the 11bp repeats A/T repeats and the 10bp G/C repeats there is no longer more insertions in the MSI-H samples compared to the controls. These results suggest that deletions are more indicative of MSI than insertions. The distribution of deletion sizes changes with repeat length for 7bp-12bp homopolymers. For 7bp-9bp homopolymer MSI presents mostly as 1bp deletions. In the 10bp and 11bp repeats there were some 2bp deletions as well as the 1bp deletions. For 12bp homopolymers MSI is present in the form of 1bp, 2bp, and 3bp deletions. For the 10bp-11bp there were very few repeats in the control samples that contained high frequency 2bp and 3bp deletions (see Figure 4.9 and Figure 4.11). This suggests that these repeat sizes may be valuable in an MSI test because high frequency deletions ≥2bp in length are very indicative of MSI.

There were a lot less unstable G/C homopolymers discovered compared to A/T homopolymers (216495 A/T homopolymers versus 1686 G/C homopolymers). This is consistent with the data reported by Yoon et al. (2013) in gastric cancers. This is expected because there are fewer G/C homopolymers in the human genome than A/T homopolymers (Dechering et al., 1998). A contributing factor could conceivably be that G/C homopolymers are also less susceptible to MSI than A/T homopolymers. However the literature gives evidence to the contrary (Ellegren, 2004, Sammalkorpi et al., 2007). In the paper Sammalkorpi et al. (2007) the G/C homopolymers investigated showed a higher rate of susceptibility to MSI compared to A/T homopolymers. Because of the extra hydrogen bond between guanine and cytosine slippage events for these bases should be less common during DNA replication. However A/T homopolymers in genomic DNA pack in such a way to allow bifurcated hydrogen bonds to form between bases as well as the usual hydrogen bonds (Nelson et al., 1987). This gives extra rigidity to the DNA

structure of A/T repeats and increases the kinetic energy needed to cause a slippage during DNA replication. This could be one of the explanations for why A/T repeats are less susceptible to MSI than G/C repeats.

In concordance with the literature reporting that G/C homopolymers are more unstable than A/T homopolymers, my results show that 7bp-10bp G/C homopolymers are more unstable than the A/T homopolymers of the same lengths. The MSI-H samples have a greater frequency of variant reads in the G/C homopolymers compared to the same read length A/T homopolymers (see Figure 4.6 and Figure 4.7). Despite the G/C homopolymers having a comparatively higher indel frequency compared to the A/T repeats of the same size, the size distribution of deletions is similar for both repeat types. For example, the excess in 2bp deletions in the MSI-H samples compared to the controls is first seen in the 10bp repeats for both A/T and G/C repeats (see figures Figure 4.8 and Figure 4.10).

4.3.1. Conclusions

In conclusion, the frequency of variant reads in MSI-H colorectal tumours increased with homopolymer length for both A/T and G/C homopolymers. In the MSI-H tumours the number of homopolymers containing deletions was higher than the number of homopolymers containing insertions for all repeat lengths analysed. The deletions caused as a result of MSI occur at a higher frequency compared to the deletions found in normal controls, and in the larger repeat sizes deletions of 2bp and 3bp were observed in the MSI-H samples. Also 218181 of highly variable homopolymers have been identified for further analysis. Many of these are potentially suitable for a sequencing based MSI test.

Chapter 5. Assessing next generation sequencing of short homopolymers identified from whole genome sequence data in microsatellite unstable tumours

5.1. Introduction and aims

5.1.1. Next generation sequencing of MSI-H tumours in 2013- 2014

In chapter 4 low depth whole genome sequences consisting of 12 MSI-H tumours, 12 MSS tumours and matched normal tissue for 11 of the MSI-H tumours were analysed. Because of the low coverage of the whole genome data (~3-4 fold sequence coverage) variant calls from all tumour in each group were pooled prior to analysis. A list of indels in 7-12bp homopolymers was generated containing the frequencies of variant reads in each sample group for all repeats. The list of indels was annotated with all SNPs as of dbSNP (version 137, hg19) (Sherry et al., 2001) within 30bp of the start of the repeats. A total of 218,181 variable 7-12bp homopolymers were identified. The most variable of these are potentially suitable for a next generation sequencing based MSI test. It is not clear if the short homopolymers that have been reported in the literature are the best markers for an MSI test. There could potentially be more unstable repeats among the 218,181 repeats identified in chapter 4. In this chapter, the aim is to sequence some of the repeat identified in the whole genome analysis in a small panel of tumours enabling the most unstable repeats to be selected for a sequencing based MSI test.

The first paper to illustrate the potential use of next generation sequencing for detecting MSI was The Cancer Genome Atlas Network (2012). More papers expanding on this work were published after the whole genome analysis in chapter 4 was conducted (Lu et al., 2013, Niu et al., 2014, Salipante et al., 2014, Yoon et al., 2013). Yoon et al. (2013) analysed MSI in gastric cancers and gastric cancer cell lines using RNA sequencing. They also performed whole genome sequencing of 3 MSI-H and 3 MSS gastric cancer cell lines. In concordance with results in chapter 4, Yoon et al. (2013) found that mutations in mononucleotide repeats that are caused by MSI consist mainly of deletions. Yoon et al. (2013) also reported that the susceptibility of mononucleotides to MSI is dependent on repeat length with longer repeats being more prone to MSI, which is consistent with the results from my whole genome analysis.

In their study Yoon et al. (2013) discovered mononucleotide repeats with deletions in MSS cell lines and concluded that mismatch repair deficiency may not be the only factor that affects the frequency of deletions in mononucleotide repeats in gastric cancers (Yoon et al., 2013). 27.2% of the deletions identified in all 3 MSI-H gastric cancer cell line whole genome sequences were also seen in MSS gastric cancer cell lines (Yoon et al., 2013). A few markers with deletions in stable tumours will therefore most likely have to be taken into account when developing a test for colorectal cancers.

Methods other than using a set panel of short repeats for an MSI test have also been suggested (Lu et al., 2013, Niu et al., 2014, Salipante et al., 2014). Using RNA-Seq data Lu et al. (2013) found that the proportion of microsatellite insertions over all insertions divided by the proportion of microsatellite deletions over all deletions could be used to reliably predict the MSI status of MSI-H and MSS tumours. Niu et al. (2014) have developed a software tool (MSIsensor) for differentiating between MSI-H and MSS tumour samples when given tumour and matched normal whole genome sequences. A similar approach has been developed by Salipante et al. (2014). They have developed a pipeline (mSINGS) that can determine a sample's MSI status using exome sequence data or sequence data from the capture panels ColoSeq and UW-OncoPlex. The mSINGS software analysis uses a panel of mononucleotide repeats specific to each of the three sequencing approaches and calls MSI based on the fraction of mononucleotide repeats that show the emergence of new variant read lengths. Repeats are classed as unstable if the number of variant read lengths exceed the mean number of read lengths + 3x the standard deviations measured in control samples. Approaches like this may be used in the future if performing whole genome sequencing, exome sequencing or gene panel sequencing of all tumours becomes an economically viable and routine method for analysing colorectal tumours. However in the near future sequencing a small a panel of short repeats will be a more cost effective way of determining the MSI status of cancers. The number of markers in such a repeat panel will have to reflect how susceptible to microsatellite instability the markers are. One thing to consider will be that a shorter microsatellite will produce less PCR error whereas larger microsatellites are more susceptible to MSI.

5.1.2. Receiver operator characteristics as a method for assessing the ability short homopolymers for differentiating between MSI-H and MSS tumours

Receiver Operator Characteristics (ROC) curves are curves where the true positive rate (sensitivity) is plotted against the false positive rate (1-specificity) for a range of different threshold values. ROC are used for data that can be classified using binary classification as either positive or negative for a trait. In this chapter the trait of interest is microsatellite instability. ROC curves can be an important tool for evaluating thresholds of diagnostic tests. These curves are used as a visual aid for analysing the sensitivity and specificity for different thresholds. Setting thresholds is often a trade off between sensitivity and specificity. If a high threshold is chosen there is a risk that some individuals with the disease but low test values will be missed. In this case the specificity of the test will be high at the expense of sensitivity. On the other hand if thresholds are set low there could be a risk of individuals without the disease receiving a positive test result. In this case the specificity of the test would be low to facilitate a higher sensitivity.

For a ROC curve the true positive rate is plotted along the y axis and the false positive rate is plotted along the x-axis. Each point along the curve represents the true positive and false positive rate at a given threshold. For the work in this chapter the frequency of reads containing deletions will be used for the thresholds. Therefore each the ROC curves would show the true positive and false positive rate across a range of deletion frequency thresholds. The samples would be plotted in order of decreasing deletion frequencies. Each MSI-H sample would be plotted as an increase in the true positive rate and each MSS sample would be plotted as an increase in the false positive rate. An example of a ROC curve can be found in Figure 5.1.

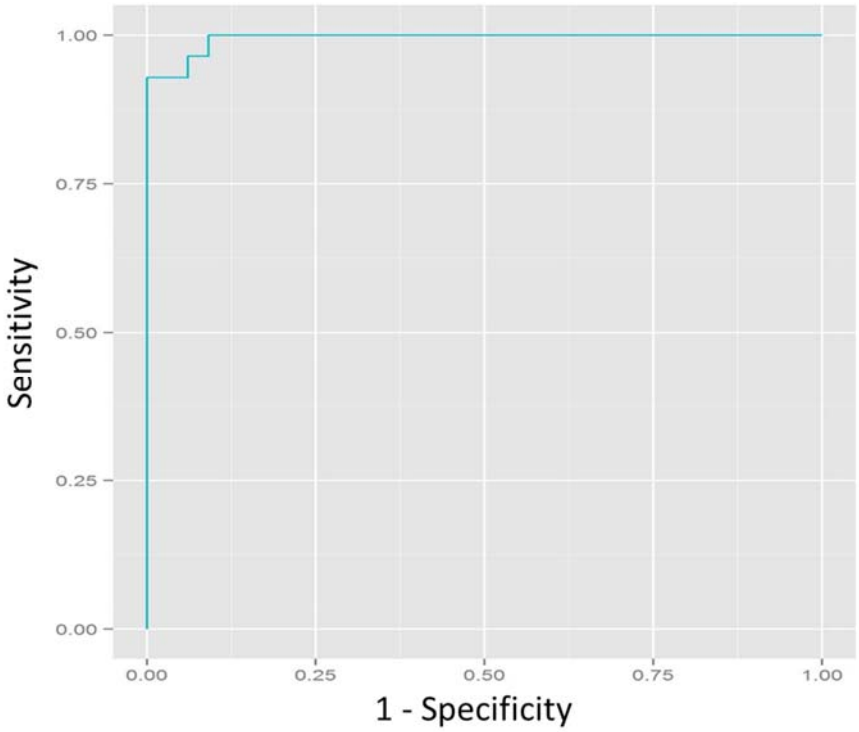

Figure 5.1: ROC curve

The area under the receiver operating characteristic curve (AUC) is a measure of how well a test can differentiate between individuals that have a disease and individuals that do not have the disease. In this chapter AUC is employed to evaluate the ability of individual homopolymers for differentiating between MSI-H and MSS tumours. An AUC of 1 would indicate that all MSI-H samples have a higher deletion frequency than the MSS samples. On the other hand an AUC value of 0.5 would mean that a repeat has no discrimination power because there would be 50-50 chance that any randomly chosen MSI-H sample would have a higher deletion frequency than any randomly chosen MSS sample. Possible AUC values range between 0 – 1, and any AUC vale of x indicates that there is an x ×100 percent chance that any randomly chosen MSI-H sample would have a higher deletion frequency than any randomly chosen MSS sample.

5.1.3. Aims

In chapter 3 it was established that short mononucleotide repeats (7bp-14bp) are susceptible to microsatellite instability and that these short repeats might be used to differentiate between MSI-H and MSS samples using next generation sequencing on the Illumina MiSeq platform. It was also shown that by using neighbouring SNPs it is possible to distinguish between the two alleles for repeats with a neighbouring

heterozygous SNP to determine which allele contains the variant reads. Using this technique it will be shown that there may be an allelic bias for variant reads of repeats affected by MSI. However, due to the low number of unstable repeats in the chosen panel of MSI-H tumours, only 5 examples of repeats with a heterozygous SNP were available for study. In chapter 4 whole genome sequence data from MSI-H colorectal cancers were mined to identify new homopolymers that are highly variable in MSI-H tumours and have closely situated neighbouring SNPs with high minor allele frequencies. In this chapter, approximately 100 of the newly discovered repeats are tested to assess their variability in a small panel of tumours. This was done to find out if these repeats are unstable in a different panel of MSI-H tumours. In this chapter the targets are:

- Using the list of unstable repeats generated from the whole genome analysis in chapter 4, select homopolymers that show a high frequency of variant reads in MSI-H tumours for sequencing in a small panel of MSI-H tumours and controls.

- Assess the level of instability in the chosen repeats in the new panel of tumours to enable the selection of the most variable repeat for use in a future sequencing based MSI test.

- Evaluate the use of SNPs for analysing allelic distribution of MSI in a larger panel of repeats.

5.2. Results

5.2.1. Choosing repeats identified in the whole genome analysis for investigation in a new panel of MSI-H tumours

A total of 218,181 variable 7-12bp homopolymers were identified from the whole genome analysis in chapter 4. To validate specific repeats for MSI detection, some of the most unstable homopolymers identified in the whole genome analysis were selected for further analysis. The list of 218,181 variable 7-12bp homopolymers was narrowed down by filtering for repeats with a read depth ≥20x in each group (MSI-H, matched normal for the MSI-H samples, and MSS samples). Repeats with common polymorphisms (dbSNP version 173, hg19) (Sherry et al., 2001) were excluded. 7-10bp repeats were selected if they had a variant read fraction of 10% or higher in the MSI high sample group and no variant reads in the controls. For the 11-12bp repeats were selected if they had a variant read fraction of 15% or higher in the MSI-H samples and a variant read fraction of ≤5% in the controls. A variant read fraction of ≤5% in 11-12bp repeats was presumed to be caused by sequencing and PCR error. Homopolymers with low indel frequencies in the control samples were desired because it would be easier to cope with repeats with a low background error rate. It is presumed that variation in background errors could to some extent be attributed to sequence context.

The Perl script AnnotateCloseSNPs.pl was used to annotates SNPs within 30bp of the start of repeats (see methods section 2.8.7.4 for further detail). Homopolymers were selected to insure the inclusion of SNPs with a high minor allele frequency within 30bp were selected. If there were more than one SNP detected within 30bp of a repeat, the minor allele frequencies were added together as a quick method to assess the value of the SNPs. Repeats were only selected if there were SNPs within 30bp of the repeat with minor allele frequencies, which summed up to least a frequency of 0.2. In total 529 A/T homopolymers fitted these criteria. Because there were few G/C homopolymers in the data set the criteria for including SNPs within 30bp of the repeat was omitted and the requirement for a read depth ≥20x in each group was relaxed. This resulted in a data set of 33 G/C homopolymers. A list of all these repeats can be found on the supplementary CD (File names: “GC_SNPfile_sorted.xlsx” and “AT_SNPfile_sorted.xlsx”).

The UCSC Genome browser (Kent et al., 2002) was used to assess the possibility of creating primers for the homopolymers that passed the above criteria. Many of the 529

A/T homopolymers and 33 G/C homopolymers that met the selection criteria above were situated in regions of low complexity such as LINES and SINES, which limited the number of repeats where primers could be produced without the risk of miss-priming. The 120 most variable repeats for which suitable primers could be produced were selected to assess the utility of these specific mononucleotides for sequence based detection of MSI repeat length variation.

The selected 120 unstable mononucleotide repeats (7-12bp) were amplified from FFPE tissue and sequenced using the Illumina MiSeq. The FFPE tissues consisted of a selection of 6 Lynch Syndrome tumours, matching normal mucosa for 5 of these tumours, and 6 MSS tumours (see Table 5.1). Up to 120 repeats were amplified for each sample. For the matched normal tissue there was too little material to enable the sequencing of all 120 repeats so this material was only used for a selection of repeats. For the other samples the amount of available DNA was also in a limited supply. ~300bp amplicons were produced using the high fidelity Pfu-based Herculase II Fusion DNA polymerase and 35 PCR cycles. Amplicons were quantified using Qiagen QIAxcel, then pooled at a roughly equimolar concentration. Agencourt AMPure XP beads were used for PCR clean-up. After PCR clean-up the amplicon pools were diluted to a concentration of 0.2ng/µl before Library Prep using the Illumina Nextera XT kit (Illumina, San Diego, CA, United States of America).

Samples	Sample Type	Lynch Syndrome Patients Number
U029 Tumour	Lynch Syndrome Tumour	U029
U096 Tumour	Lynch Syndrome Tumour	U096
U179_H03 Tumour	Lynch Syndrome Tumour	U179
U179_H12 Tumour	Lynch Syndrome Tumour	U179
U303 Tumour	Lynch Syndrome Tumour	U303
U312 Tumour	Lynch Syndrome Tumour	U312
U029 Normal	Normal Mucosa	U029
U096 Normal	Normal Mucosa	U096
U179 Normal	Normal Mucosa	U179
U312 Normal	Normal Mucosa	U312
169259	MSS tumour	n/a
169736	MSS tumour	n/a
169836	MSS tumour	n/a
170146	MSS tumour	n/a
170402	MSS tumour	n/a
171223	MSS tumour	n/a

Table 5.1: Tissue samples consisting of Lynch Syndrome tumours, matching normal tissue for the Lynch Syndrome tumours and MSS tumours.

A list containing the 120 mononucleotide repeats can be found in table Table 5.2. Primer design, PCR amplification and the QIAxcel quantification for 75 of the homopolymers was carried out by the students Ghanim Alhilal (Institute of Genetic Medicine, Newcastle University) and Iona Middleton (Institute of Genetic Medicine, Newcastle University) under my supervision. I did primer design, PCR amplification and the QIAxcel quantification for the remaining 45 homopolymers. Many of the failed PCRs for both students were repeated by me.

Repeat Name	Repeat Size	Repeat Position	SNP1	SNP2	SNP3
GM04	7bp	chr13:92677561	rs9560900		
GM19	7bp	chr11:114704378	rs142833335	rs190597109	rs10502196
GM24	7bp	chr10:117432196	rs2532728		
GM25	7bp	chr3:110871917	rs74593281	rs6437953	rs188039266
GM27	7bp	chr11:85762247	rs669813	rs181565251	rs146406522
GM30	7bp	chr14:53111542	rs12880534		
IM13	7bp	chr2:235497098	rs6721256	rs183025093	rs187312036
IM14	7bp	chr7:80104530	rs11760281		
IM19	7bp	chr9:82475000	rs72736428	rs186539440	rs4877153
IM20	7bp	chr13:57644695	rs6561918		
IM22	7bp	chr7:90135495	rs10487118	rs10487117	rs139214151
IM23	7bp	chr6:72729530	rs557365		
IM26	7bp	chr3:166053586	rs2863375		
IM27	7bp	chr7:35079238	rs4723393	rs112516918	
IM43	7bp	chr21:32873760	rs9981507		
IM55	7bp	chr3:143253844	rs13099818		
IM61	7bp	chr12:73576422	rs34696106		
IM66	7bp	chr17:48433966	rs147847688	rs141474571	rs4794136
IM67	7bp	chr7:22290894	rs67082587	rs57484333	
IM69	7bp	chr9:92765722	rs1036699		
LR04	7bp	chr1:4677109	rs113646106	rs2411887	
LR06	7bp	chr18:20089449	rs501714		
LR08	7bp	chr11:56546205	rs181578273	rs7117269	
LR13	7bp	chr8:21786971	rs2127206		
LR15	7bp	chr8:92077209	rs56084507		
LR25	7bp	chr16:63209545	rs76192782	rs79880398	rs4949112
LR45	7bp	chr2:226938121	rs180896305	rs1522818	rs144175764
LR47	7bp	chr10:20506728	rs11597326	rs12256106	
LR49	7bp	chr15:93619047	rs80323298	rs201097746	rs12903384
LR50	7bp	chr2:76556320	rs925991	rs144630203	
LR51	7bp	chr10:51026724	rs8474		
GM03	8bp	chr4:120206446	rs17050454	rs10032299	
GM08	8bp	chr21:36575085	rs2834837	rs115025058	
GM09	8bp	chr20:6836976	rs6038623		
GM16	8bp	chr6:100743595	rs7765823		
GM20	8bp	chr7:142597494	rs6961869	rs6961877	
IM15	8bp	chr6:91455181	rs1231482		
IM21	8bp	chr1:215136389	rs181787229	rs1901621	rs1901620
IM25	8bp	chr12:24568356	rs10771087		
IM39	8bp	chr2:103233866	rs76771828	rs190979688	rs187315716
IM40	8bp	chr4:84074813	rs10516683		
IM41	8bp	chr6:147948940	rs1944640	rs112075239	
IM57	8bp	chr3:81210016	rs35085583		
IM59	8bp	chr8:108359000	rs10156232		
IM63	8bp	chr3:115816065	rs34764455		
IM68	8bp	chr12:129289692	rs10847692		
LR02	8bp	chr4:134947775	rs189671825	rs192703656	rs1494978
LR18	8bp	chr1:220493934	rs191265856	rs199830128	rs74940412
LR19	8bp	chr12:29508668	rs10843391	rs186762840	
LR20	8bp	chr1:64029633	rs146973215	rs191572633	rs217474
LR27	8bp	chr4:72877514	rs55894427	rs74733006	
LR31	8bp	chr3:62995577	rs183248146	rs2367592	
LR46	8bp	chr20:10660084	rs143884078	rs182346625	rs6040079
GM05	9bp	chr2:216770762	rs6704859		
GM06	9bp	chr16:77496517	rs6564444	rs143453795	rs145573459
GM10	9bp	chr1:59891623	rs946576	rs182557762	
GM11	9bp	chr5:166099890	rs347435		
GM15	9bp	chr7:97963736	rs6465672		
GM17	9bp	chr11:95551110	rs666398		
GM21	9bp	chr3:142695338	rs185182		
GM23	9bp	chr5:11345920	rs184237728	rs32123	
GM28	9bp	chr5:29209380	rs4130799		
IM16	9bp	chr18:1108766	rs114923415	rs73367791	rs59912715
IM17	9bp	chr13:31831504	rs932749		
IM42	9bp	chrX:96502620	rs1409192		
IM44	9bp	chr12:9797065	rs201750704	rs4763716	
LR05	9bp	chr2:10526616	rs111286197	rs13431202	
LR10	9bp	chr1:81591387	rs111814302	rs1768398	rs1768397

LR14	9bp	chr17:69328494	rs9895642		
LR21	9bp	chr15:50189464	rs182900605	rs80237898	rs2413976
LR24	9bp	chr1:153779428	rs192329538	rs1127091	
LR28	9bp	chr12:81229785	rs185642078	rs28576612	rs10862196
LR34	9bp	chr3:115377097	rs187521190	rs192106258	rs9883515
LR40	9bp	chr2:13447469	rs6432372		
GM01	10bp	chr11:28894428	rs7951012		
GM22	10bp	chr14:43401009	rs58274313		
GM26	10bp	chr14:49584750	rs187027795	rs11628435	
GM29	10bp	chr3:70905559	rs2687195		
IM07	10bp	chr6:100701947	rs189035042	rs6915780	
IM12	10bp	chr8:23602937	rs389212		
IM33	10bp	chr8:25731926	rs202225742	rs35644463	rs113180202
IM34	10bp	chr7:83714718	rs1524881		
IM35	10bp	chr11:84425221	rs67283158	rs10792775	rs116387070
IM37	10bp	chr17:50813569	rs2331498		
LR26	10bp	chr16:80050257	rs4889066	rs187883346	
LR29	10bp	chr6:78198348	rs1778257		
LR30	10bp	chr11:105445091	rs7933640		
LR32	10bp	chr19 :37967219	rs7253091		
LR35	10bp	chr8:130384501	rs4733547		
LR39	10bp	chr17:66449341	rs2302784		
GM02	11bp	chr1:116246109	rs10802173	rs148789685	
GM07	11bp	chr7:93085747	rs2283006		
GM13	11bp	chr12:107492626	rs34040859	rs77265275	rs201488736
GM14	11bp	chr3:177328817	rs6804861		
IM28	11bp	chr9:5122910	rs10815163		
IM32	11bp	chr18:42045500	rs8087346		
IM45	11bp	chr4:99545419	rs189419054	rs2178216	
IM52	11bp	chr21:22846823	rs74462385	rs9982933	rs2155801
IM53	11bp	chr9:20662629	rs182630429	rs140426089	rs12352933
IM54	11bp	chr21:33710014	rs13046776		
IM65	11bp	chr13:25000863	rs7324645	rs9511253	
LR01	11bp	chr13:97387479	rs1924584	rs4771258	
LR11	11bp	chr2:217217870	rs13011054	rs147392736	rs139675841
LR12	11bp	chr14:47404235	rs187434561	rs144159314	
LR16	11bp	chr3:8522416	rs148171413	rs6770049	
LR17	11bp	chr14:55603030	rs79618905	rs77482253	rs1009977
LR23	11bp	chr2:142013941	rs434276	rs146141768	
LR33	11bp	chr4:138498649	rs200714826	rs4637454	rs111688169
LR48	11bp	chr12:77988096	rs11105832		
GM18	12bp	chr10:8269565	rs113251670	rs189036006	rs533236
IM47	12bp	chr21:22734436	rs2588655	rs149325240	rs232496
IM49	12bp	chr3:56682065	rs7642389		
IM50	12bp	chr20:37048155	rs1739651	rs145870165	
IM51	12bp	chr5:128096988	rs4836397		
IM64	12bp	chr16:14216095	rs201451896	rs112858435	rs75477279
LR36	12bp	chr4:98999722	rs182020262	rs17550217	
LR41	12bp	chr4:34074106	rs190518698	rs6852667	
LR43	12bp	chr5:86199060	rs201282399	rs10051666	rs6881561
LR44	12bp	chr10:99898285	rs78876983	rs7905388	rs7905384
LR52	12bp	chr16:63861440	rs2434849		

Table 5.2: A list of the 120 mononucleotide repeats sequenced. This list contains the designated repeat names, the length and location (genome build hg19) of each mononucleotide repeat, and the rs numbers of neighbouring SNPs. The repeat name indicates who performed the primer design and PCR amplification (GM= Ghanim Alhilal, IM= Iona Middleton, and LR = Lisa Redford).

For the MiSeq run a MiSeq Reagent Kit (v3 600-cycles) (Illumina, San Diego, CA, United States of America) was used. Sequencing was done on the Illumina MiSeq to an average read depth of >10000 paired end reads per amplicon. A cluster density of 1604 K/mm^2 was achieved on the Illumina flow cell and a Q-Score of over 30 was obtained for 60.6% of the bases sequenced (see Figure 5.2). There was a drop in Q-Score towards the

latter cycles (see Figure 5.3). This is believed to be due to reaching the end of some of the amplicons being sequenced. A total read depth of 30,107,152 was obtained across all samples for this MiSeq run. The sample U179_H12 tumour had the lowest read depth with 468,565 reads and the sample 169259 had the highest total read depth of 4,047,041 reads.

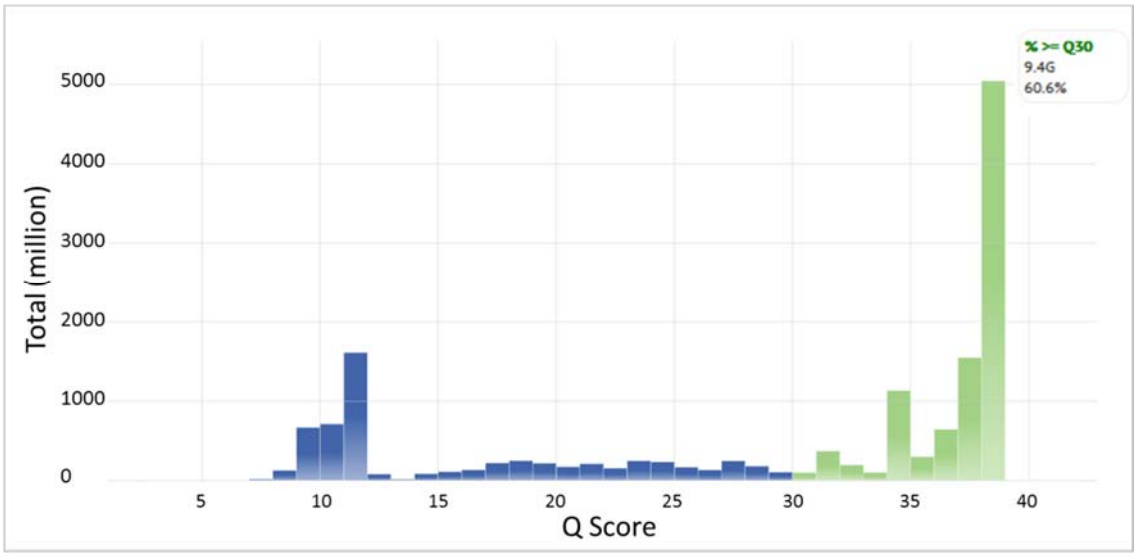

Figure 5.2: The quality score (Q-Score) distribution for the reads generated on the MiSeq. Blue = bases with a Q-Score <30, Green = bases with a Q-Score >30.

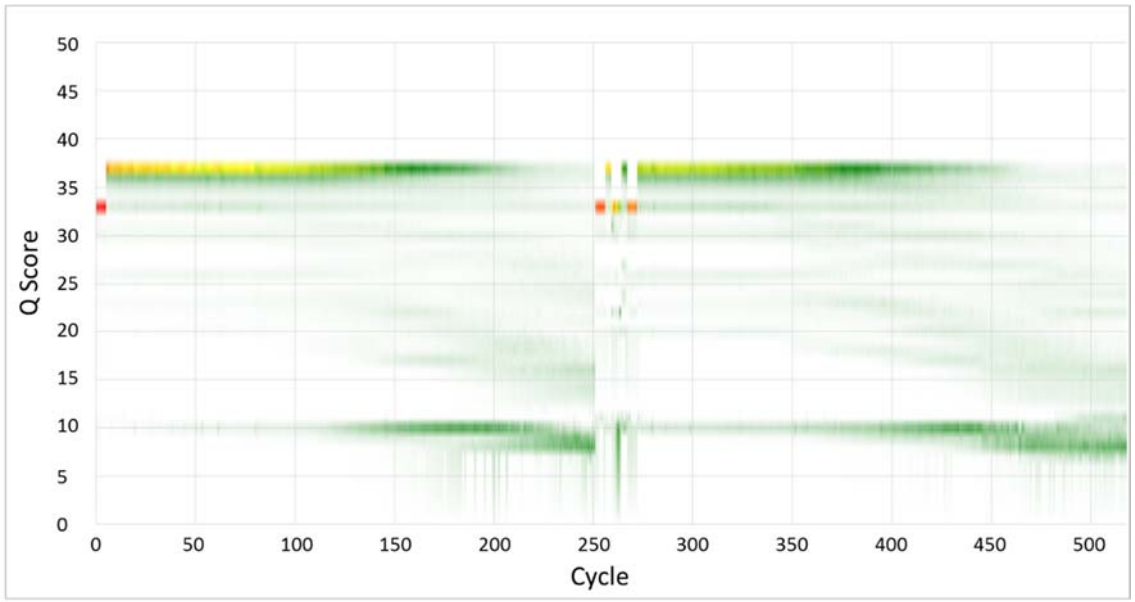

Figure 5.3: The quality score (Q-Score) distribution for each cycle showing a drop in Q-Score towards the later cycles of each read.

Variant calling was performed using the variant caller COPReC, run by Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University). This is the same variant caller that was used in chapter 3 (see methods section 2.8.6.2). Graphs were only created for repeats with a minimum of 100 paired end reads spanning the repeat. The criteria of a minimum of 100 paired end reads was used to prevent a misrepresentation of variant frequencies caused by PCR duplicates which may happen at low read depths.

5.2.2. Fragment analysis to determine the MSI status of Lynch Syndrome tumours

To confirm the MSI status of the Lynch Syndrome samples used in this chapter a standard fragment analysis was carried out on all of these samples using the Promega MSI Analysis System Version 1.2 kit (Promega, Madison, WI, United States of America).

The fragment analysis traces for tumours and matching normal tissue from patients U029 and U179 can be found below in Figure 5.4. Two tumours were analysed for patient U179. These are two separate tumours, one of which was removed from the patient in 2003 and the other in 2012. The U029 tumour and both U179 tumours had instability at all five markers confirming the diagnosis as MSI-H (see Figure 5.4). My interpretations of the fragment analysis traces was confirmed by Ottie O'Brien (Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust).

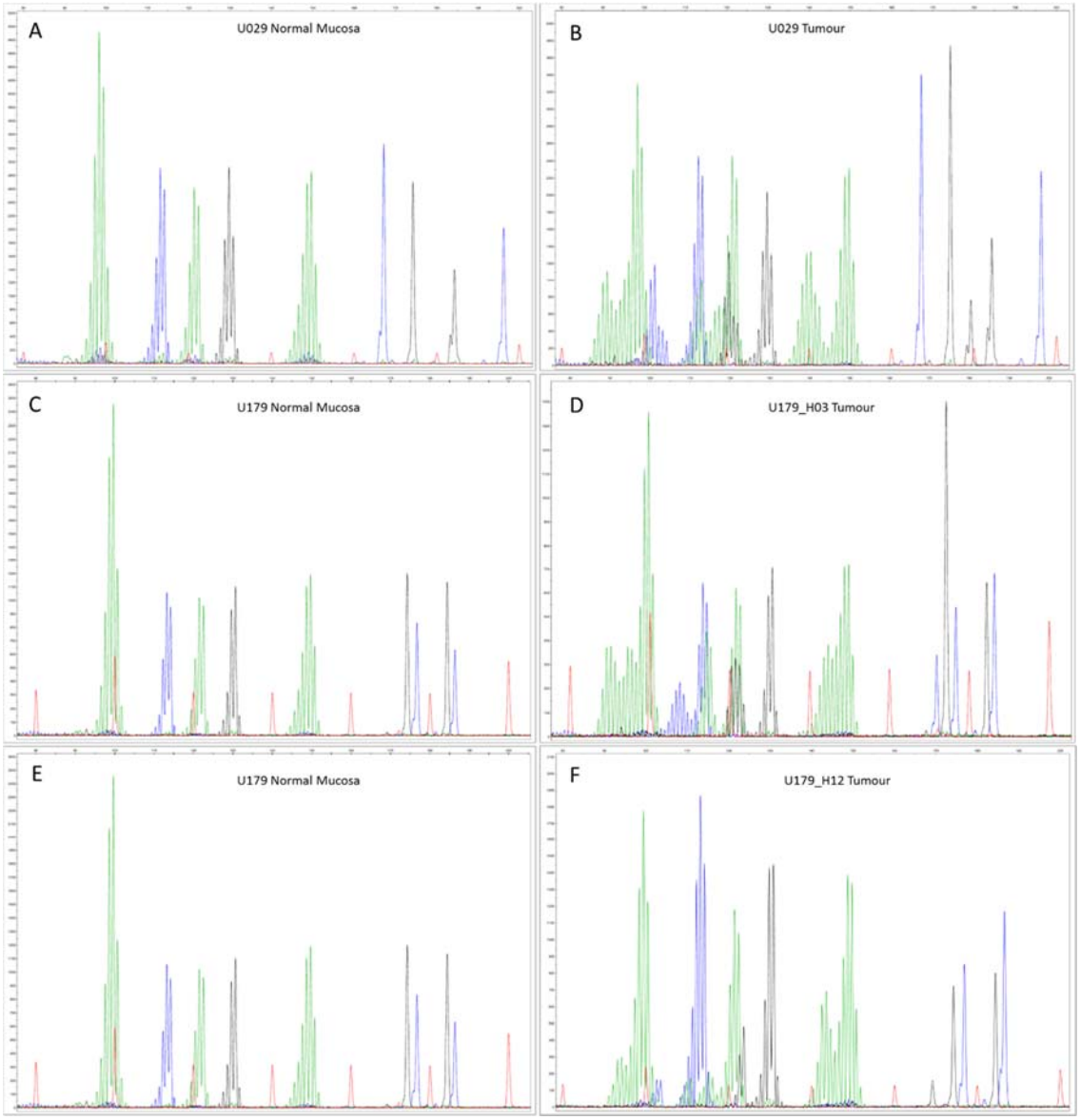

Figure 5.4: Results for the U029 and U179 tumours using a standard fragment analysis test. Panels A and B show results for the U029 normal mucosa sample and tumour sample respectively, Panels C and D show results for the U179 normal mucosa sample and the U179_H03 tumour sample respectively, Panels E and F show results for the U179 normal mucosa sample and the U179_H12 tumour sample respectively.

The fragment analysis traces for tumours and matching normal tissue from patients U312, U303 and U096 can be found in Figure 5.5. The U312 Tumour was confirmed as being MSI-H with instability detected at three markers (BAT26, BAT25 and NR-24) (see Figure 5.5 panels A and B). The U303 tumour was also confirmed as being MSI-H with instability detected at all five markers (see Figure 5.5 panels C and D). The U096 tumour on the other hand did not show any instability at any of the marker (see Figure 5.5 panels E and F). The classification of the three tumours from patients U312, U303 and U096 was confirmed by Ottie O'Brien (Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust).

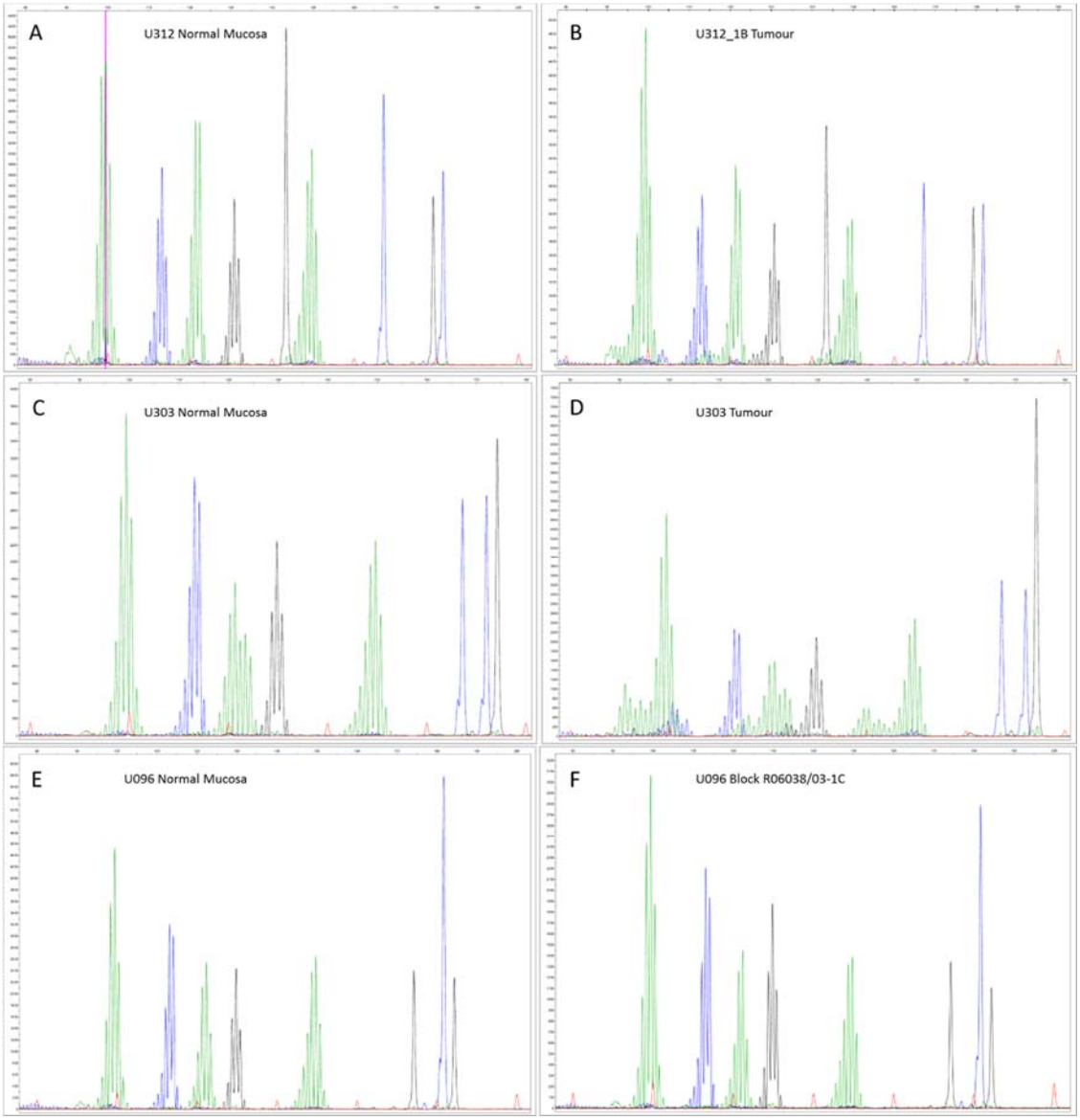

Figure 5.5: Results for the U312, U303 and U096 tumours using a standard fragment analysis test. Panels A and B show results for the U312 normal mucosa sample and tumour sample respectively, Panels C and D show results for the U303 normal mucosa sample and tumour sample respectively, Panels E and F show results for the U096 normal mucosa sample and tumour sample respectively.

The documentation for the wax block from which the U096 tumour sample was derived (block R06038/03-1C) was rechecked. This revealed that wax block R06038/03-1C was not a part of the tumour from patient U096, but a piece of the distal resection margin. The wrong wax block had been cut and I was provided with the wrong sample. This means that sequenced samples consisted of 5 MSI-H tumours, 6 MSS tumours, matched normal mucosa for 4 of the MSI-H tumours and normal mucosa from patient U096.

5.2.3. Read length variation in 7bp - 12bp repeats

To assess the instability rates of 120 7-12bp repeats these were sequenced in up to 5 MSI-H cancers and controls. For the 7bp repeats, variant reads with 1bp deletions were observed in the MSI-H samples at a frequency that notably differed from what was observed in the control samples (see Figure 5.6).

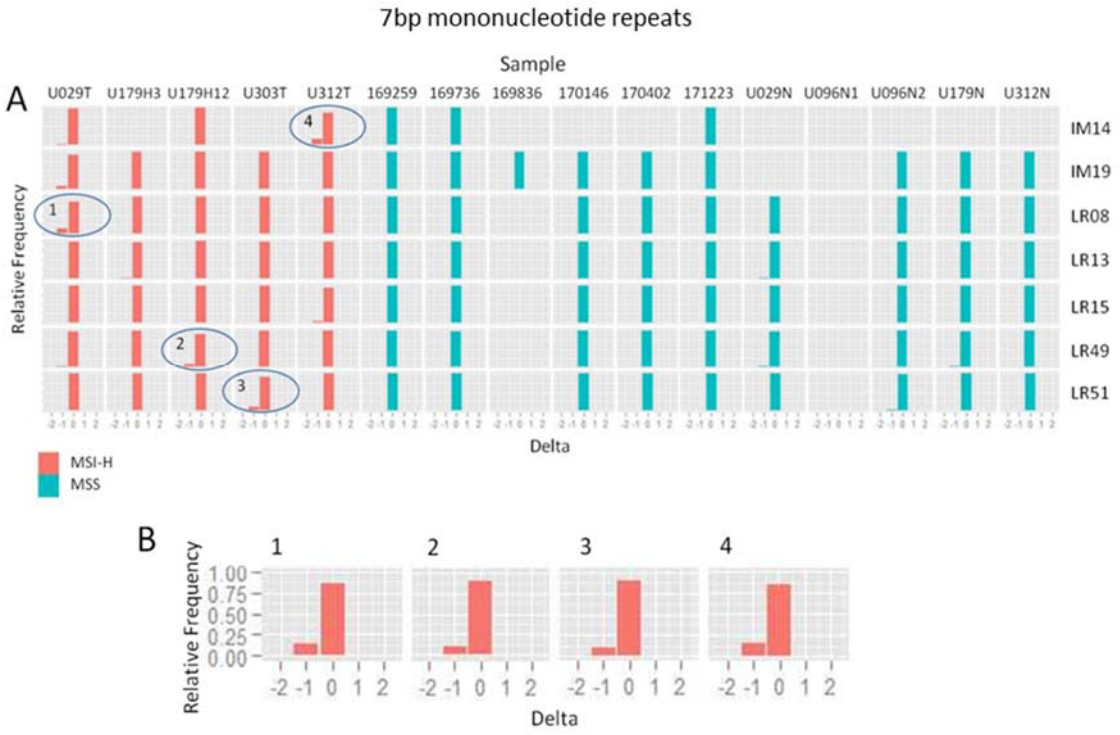

Figure 5.6: Read length variation in 7bp repeats. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the twenty-seven 7bp mononucleotide repeats sequenced. Panel B: Four repeats with 1bp deletions in MSI-H samples. Delta = change in homopolymer length.

For the 8bp repeats, variant reads in the MSI-H samples also presented as 1bp deletions and these were found at a higher frequency compared to the 1bp deletions observed in the 7bp repeats (see Figure 5.7). Interestingly the repeat LR46 showed instability in all five of the MSI-H tumours. This might suggest that this 8bp repeat is highly susceptible to MSI (see Figure 5.7).

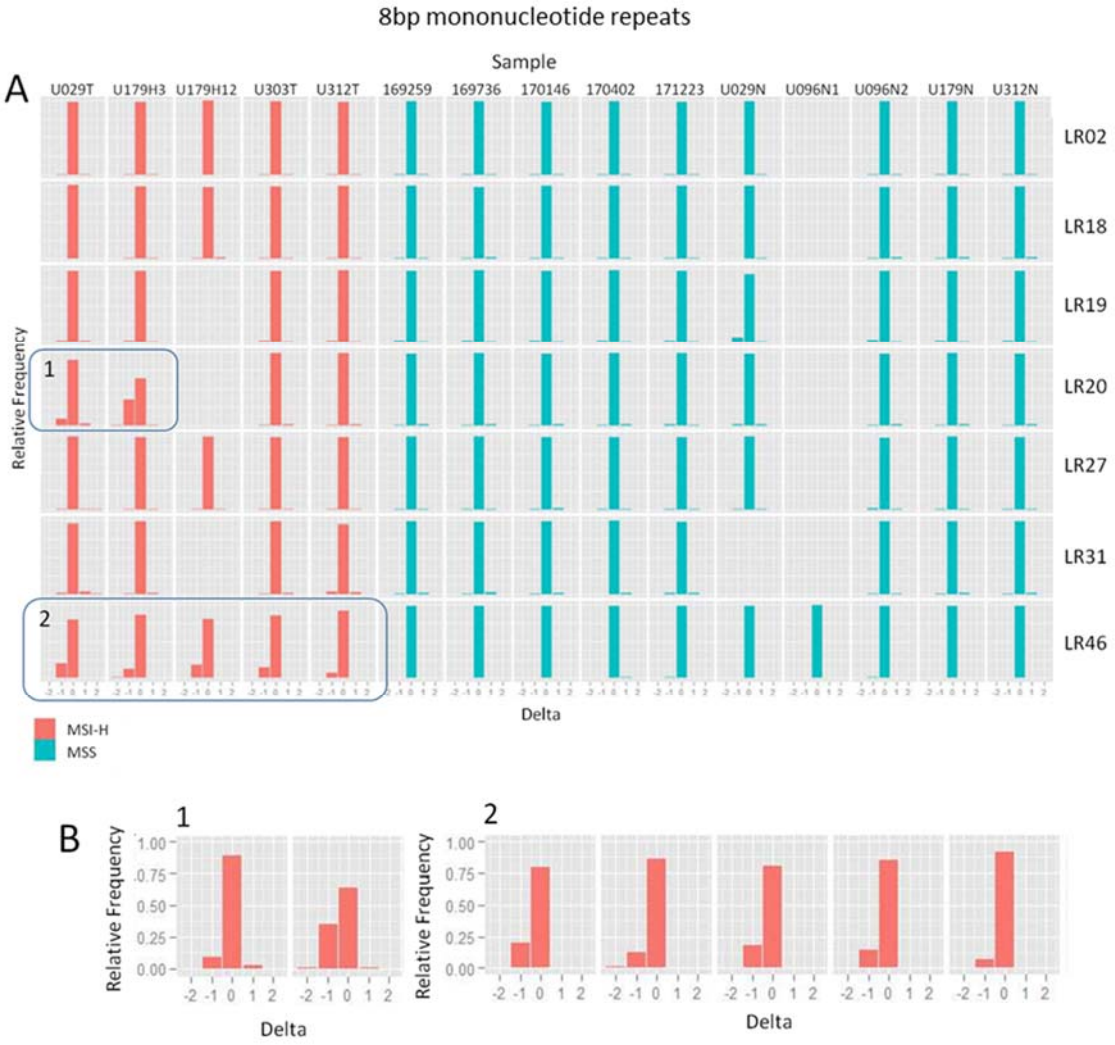

Figure 5.7: Read length variation in 8bp repeats. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the twenty-two 8bp mononucleotide repeats sequenced. Panel B: Read length variation in MSI-H samples for repeats LR20 and LR46. Delta = change in homopolymer length

For the 9bp mononucleotide repeats most of the variant reads present in the MSI-H samples are 1bp deletions but there are some 2bp present in the MSI-H samples for example see repeats LR10 sample U029 tumour and IM16 sample U179H03 tumour (see Figure 5.8). The repeat LR05 has a high 1bp deletion frequency in both the MSI-H samples and controls. This could be a result of LR05 being a G/C mononucleotide repeat while all the other 9bp repeats sequenced are A/T repeats. A/T homopolymers in genomic DNA pack in such a way to allow bifurcated hydrogen bonds to form between bases as well as the usual hydrogen bonds (Nelson et al., 1987). This increases the energy needed to cause slippage during DNA replication for A/T repeats and could be a reason for G/C repeats having more PCR error compared to A/T repeats of the same length.

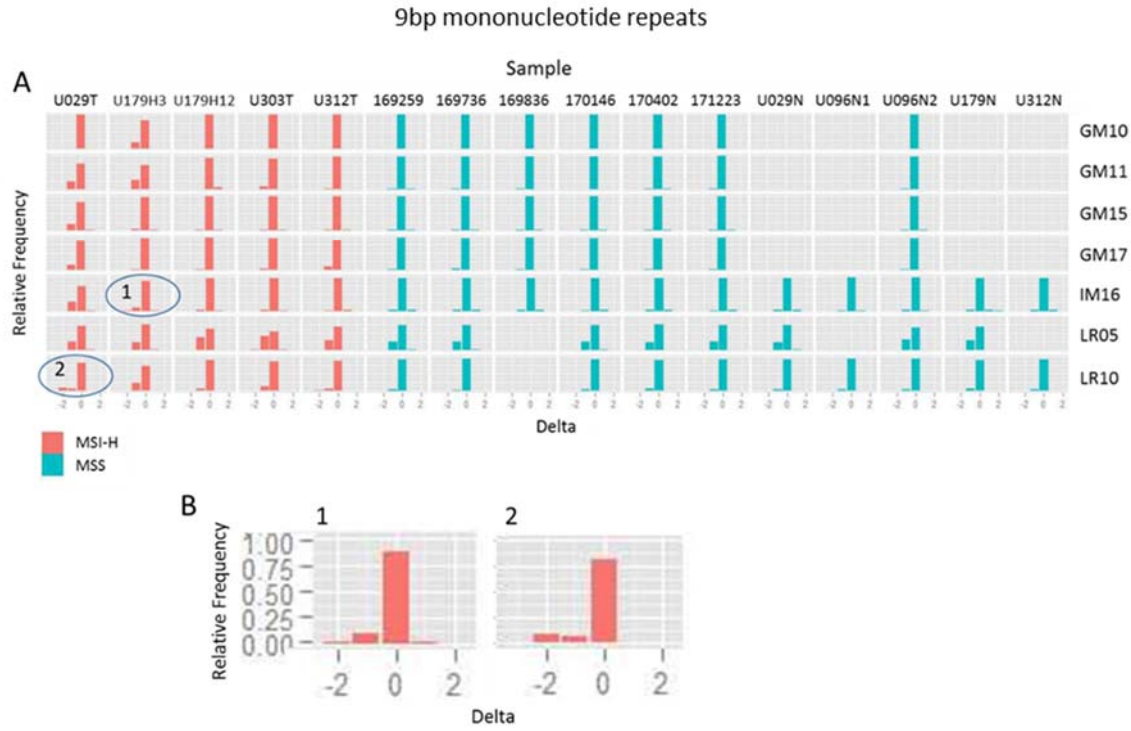

Figure 5.8: Read length variation in 9bp repeats. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the twenty-one 9bp mononucleotide repeats sequenced. Panel B: Two repeats with 1bp and 2bp deletions in MSI-H samples. Delta = change in homopolymer length

For the 10bp repeats both 1bp and 2bp deletions are present in the MSI-H samples (see Figure 5.9). Variant reads in the form of 1bp deletions are also present in the control samples, but at a lower frequency than seen in the MSI-H samples. The repeats IM37 and IM34 also have low levels of 2bp deletions in the controls.

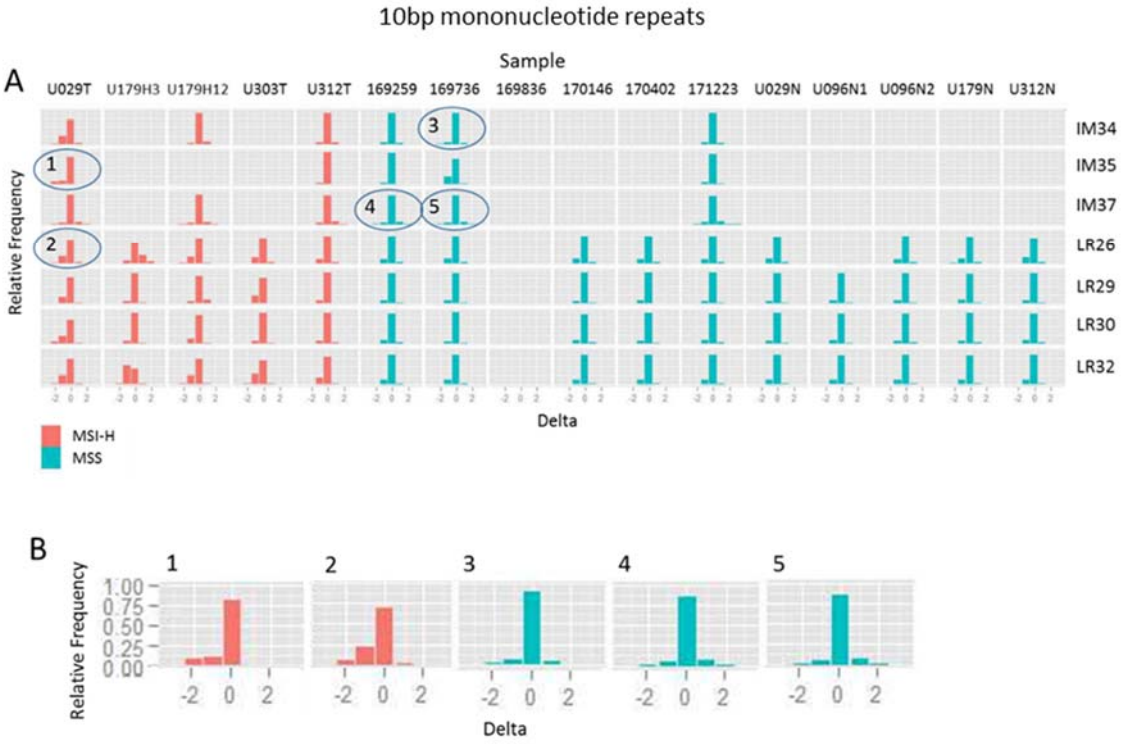

Figure 5.9: Read length variation in 10bp repeats. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the sixteen 10bp mononucleotide repeats sequenced. Panel B: Repeats with 1bp and 2bp deletions. Delta = change in homopolymer length

For the 11bp mononucleotide repeats there is a change in the shape of the graphs for many of the MSI-H samples compared to the controls with an emergence of new variant repeat lengths of 9-10bp (see Figure 5.10). The repeat LR16 is likely to contain a polymorphism because there is a roughly equal amount of reference reads and reads with 1bp deletions for the two tissue biopsies from patient U096. Both tissue biopsies from U096 are normal tissue and the biopsies were taken 1 year apart so the reads containing 1bp deletions are highly unlikely to be mutations. The repeats LR01 and LR32 also contain potential polymorphisms (see Figure 5.10 sample 169259 for repeat LR01 and sample 169736 for repeat LR23). On the other hand, it is also possible that some of the variant reads seen in the MSI-H samples are due to polymorphisms. It can be hard to tell the difference between polymorphisms and genuine mutations in cases such as samples U303 tumour and U179_H12 tumour for repeats GM14 and IM28 respectively, where no normal tissue was available for comparison.

Figure 5.10: Read length variation in 11bp repeats. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the nineteen 11bp mononucleotide repeats sequenced. Panel B: Repeats with potential polymorphisms. Delta = change in homopolymer length

For the 12bp repeats there were a lot of variant read lengths observed in the MSI-H samples which were not present in any of the controls (see Figure 5.11). There were more variant reads in the MSI-H samples than reference reads for some of the markers. This was never seen in the control samples. Deletions as large as 4bp and even 5bp were observed in the MSI-H samples for some of the repeats (see Figure 5.11 repeats LR41 and LR52).

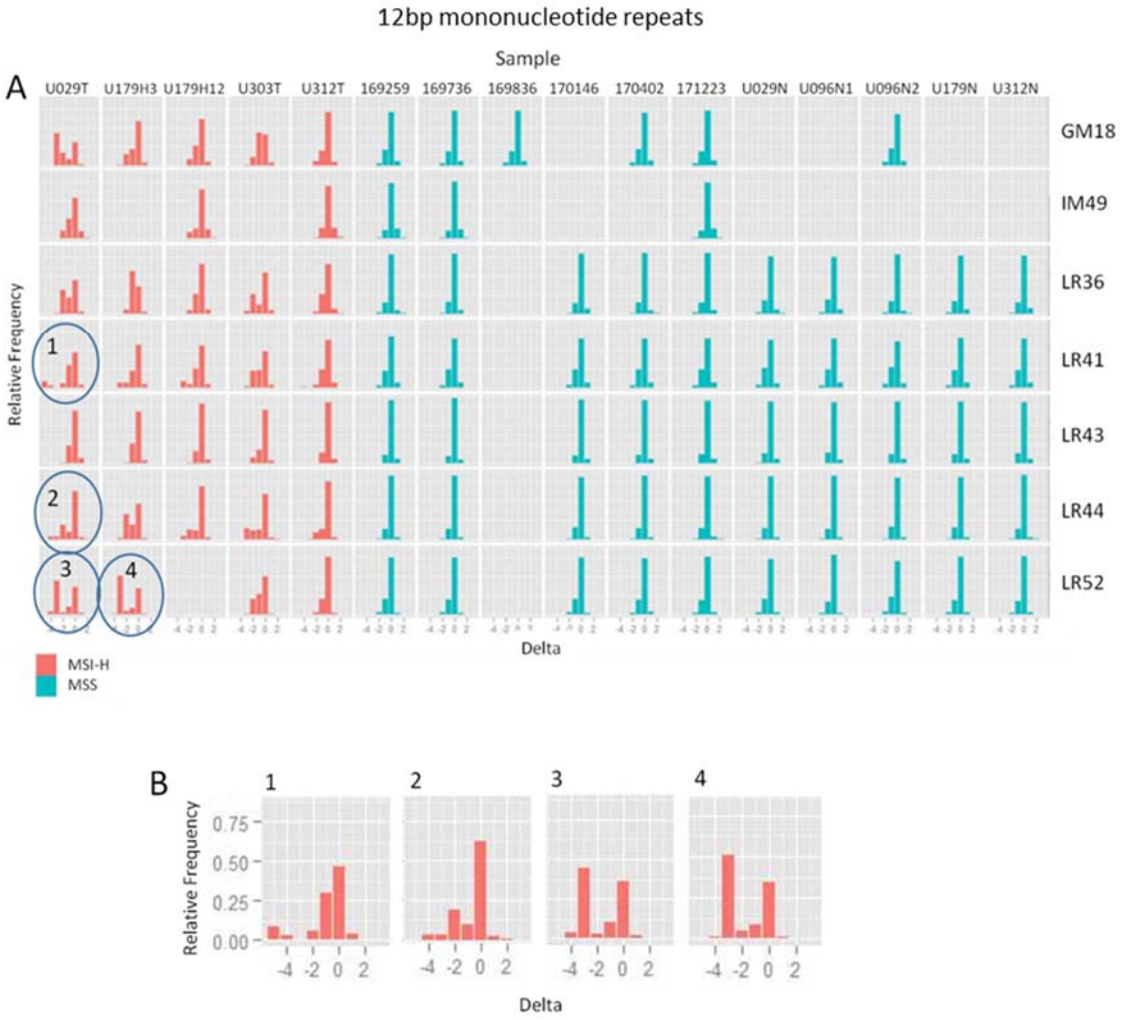

Figure 5.11: Read length variation in 12bp repeats.This figure shows seven of the 12bp mononucleotide repeats sequenced. Panel A: Variant read frequencies in MSI-H tumours and controls for seven out of the eleven 12bp mononucleotide repeats sequenced. Panel B: Repeats with 4bp and 5bp deletions in MSI-H samples. Delta = change in homopolymer length

5.2.4. Deletion frequencies in repeats identified by whole genome sequencing

The data can also be analysed in a different way, which highlights the trends mentioned before. For the patient U096 two separate sets of normal mucosa were analysed. For the data analysis in this section results obtained from wax block R06038/03-1C were used with the exception of repeats IM64 (12bp A/T repeat) and IM67 (7bp G/C mononucleotide repeat) which failed to be sequenced from the R06038/03-1C sample. Only one of the two U096 samples was used because they were duplicates with little difference between the two. A/T and G/C repeats were plotted separately.

For most 7bp mononucleotide repeats there was a deletion frequency of less than 1% in the control samples (see Figure 5.12). In only three cases was there a deletion frequency of over 2% in the control samples. These cases consisted of the repeat LR49 in samples U029 normal mucosa and U179 normal mucosa with deletion frequencies of 3% and 2.9% respectively, and the repeat LR51 with a deletion frequency of 3.3% in the normal mucosa from patient U096. In the MSI-H samples there were a few repeats which had a larger deletion frequency than the rest. Seven repeats had a deletion frequency above 4% in the MSI-H samples (see Table 5.3). These consisted of three repeats for the U029 tumour sample, one repeat for the U179_H12 tumour sample, one repeat for the U303 tumour sample, and two repeats for the U312 tumour sample.

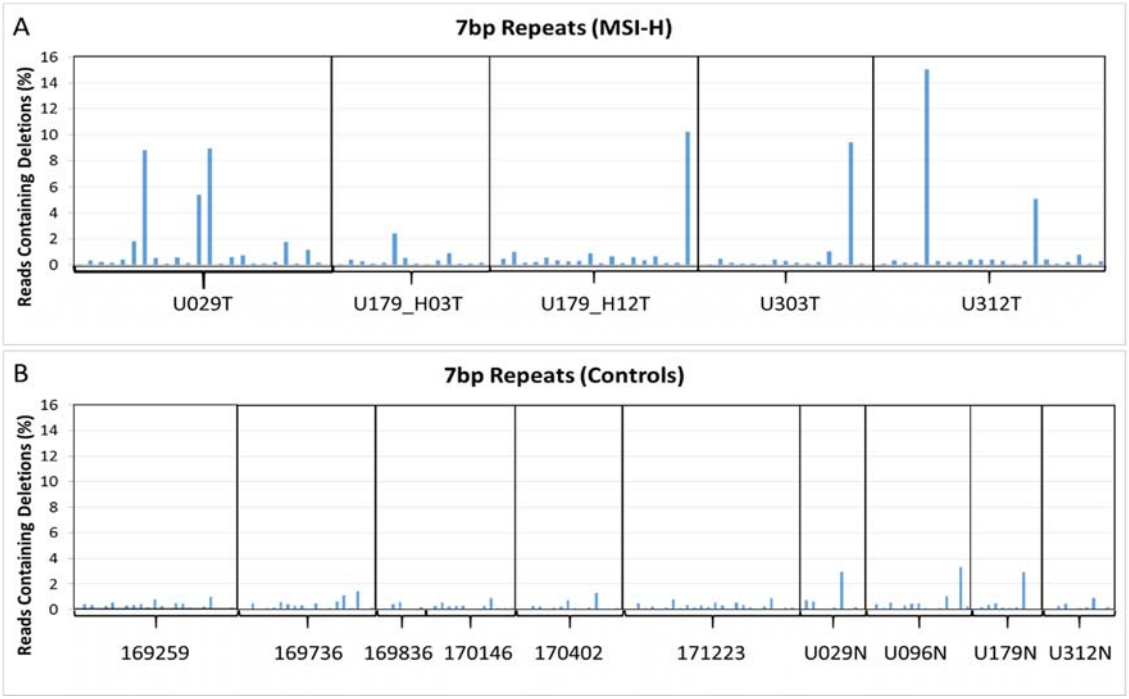

Figure 5.12: Deletion frequencies in all the 7bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H12		U303T		U312T	
Repeat	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**
IM19	8.3%	**LR49**	10.2%	**LR51**	9.4%	**IM14**	15.0%
IM43	5.4%					**LR15**	5.1%
IM55	8.9%						

Table 5.3: 7bp repeats with a deletion frequency ≥4% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

In the 8bp mononucleotide repeats the deletion frequencies were consistently below 3% in the controls with the exception of repeat LR19 from the U029 normal mucosa where the deletion frequency is 5.8%. In the MSI-H samples there were 12 repeats with a deletion frequency ≥5% (see Figure 5.13). Sample U029 tumour had five repeats with a deletion frequency ≥5%, U179_H03 tumour had two, U179_H12 tumour had three, U303 tumour had one, and U312 tumour had one (see Table 5.4).

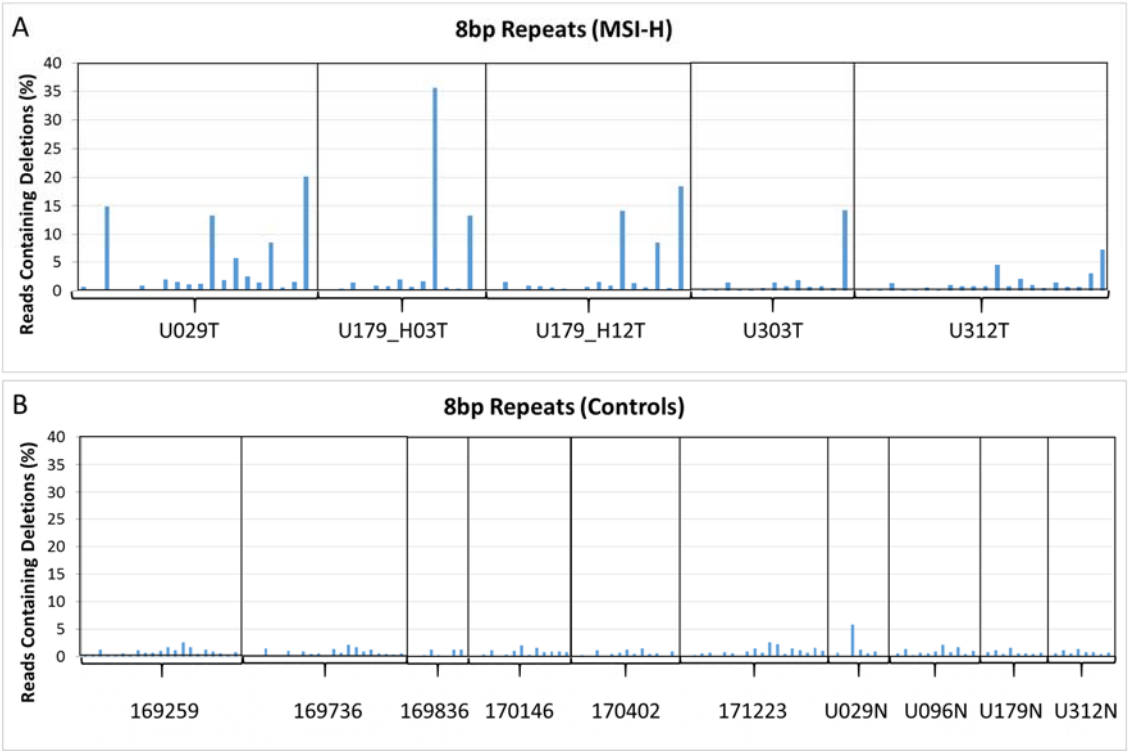

Figure 5.13: Deletion frequencies in all the 8bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H03		U179_H12		U303T		U312T	
Repeat	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**
GM09	14.9%	**LR20**	35.6%	**IM59**	14.2%	**LR46**	14.3%	**LR46**	7.2%
IM41	13.3%	**LR46**	13.2%	**LR19**	8.6%				
IM59	5.7%			**LR46**	18.5%				
LR20	8.6%								
LR46	20.1%								

Table 5.4: 8bp repeats with a deletion frequency ≥5% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

In the 9bp mononucleotide repeats there were no repeats in the control samples with a deletion frequency ≥10%, while four of the five MSI-H samples had repeats with a deletion frequency ≥10% (see Figure 5.14). These consisted of samples U029 tumour with 7 repeats, U179_H03 tumour with 4 repeats, U303 tumour with 2 repeats, and U312 tumour with 2 repeats (see Table 5.5).

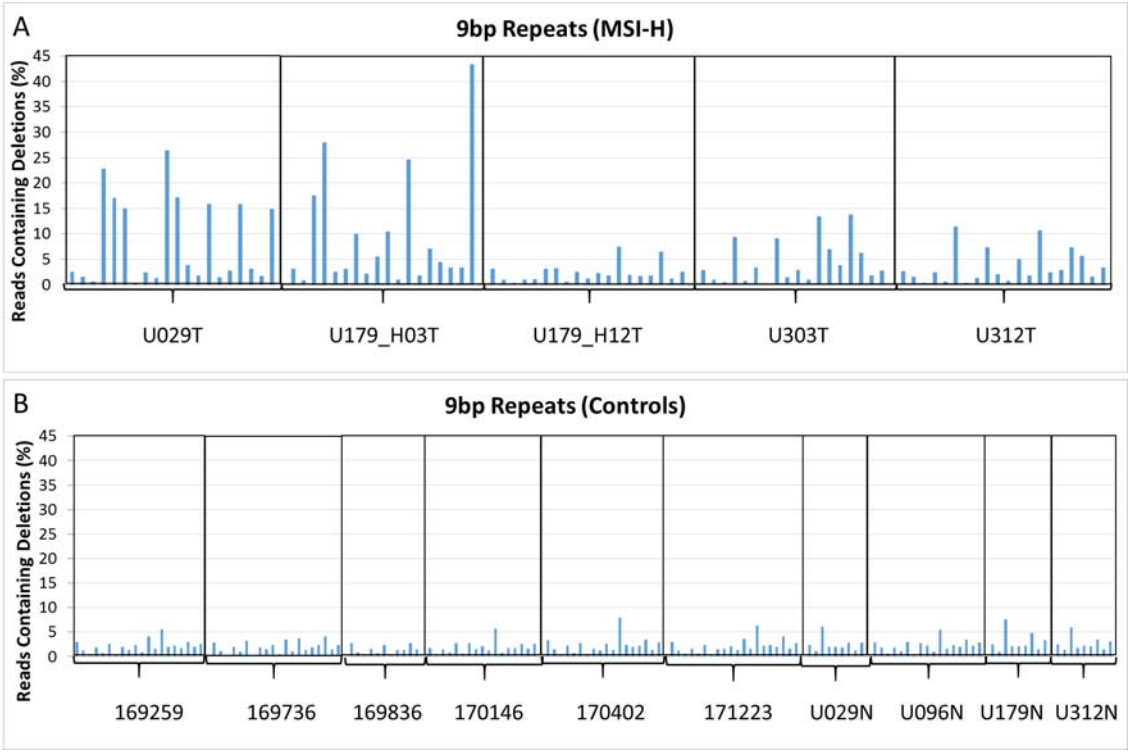

Figure 5.14: Deletion frequencies in all the 9bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H03T		U303T		U312T	
Repeat	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**
GM11	22.9%	**GM10**	17.6%	**LR10**	13.4%	**GM17**	11.4%
GM15	17.1%	**GM11**	28.0%	**LR24**	13.9%	**LR10**	10.7%
GM17	15.1%	**IM16**	10.5%				
IM16	26.5%	**LR10**	24.7%				
LR10	15.9%	**LR40**	43.4%				
LR24	15.8%						
LR40	14.9%						

Table 5.5: 9bp repeats with a deletion frequency ≥10% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

For the 10bp repeats there was one repeat with a deletion frequency ≥20% in the control samples (see Figure 5.15). This repeat was IM35 in sample 169736 which had a deletion frequency of 23.6%. In the MSI-H samples there were several repeats with a deletion frequency ≥20% (see Figure 5.15). These consisted of 6 repeats for tumour U029, 1 repeat for tumour U179_H03, 3 repeats for tumour U179_H12, 2 repeats for tumour U303, and 1 repeat for tumour U312 (see Table 5.6).

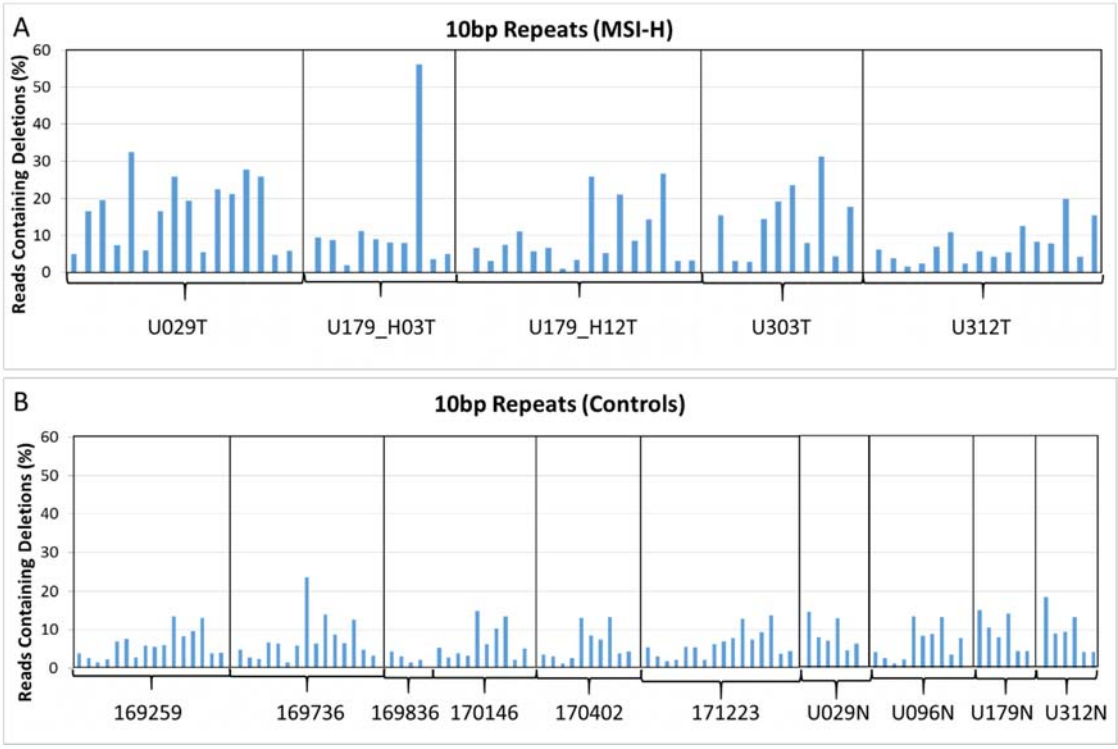

Figure 5.15: Deletion frequencies in all the 10bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H03T		U179_H12T		U303T		U312T	
Repeat	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**	**Repeat**	**DF**
IM07	32.4%	**LR32**	56.2%	**IM35**	25.7%	**LR29**	23.4%	**LR32**	20%
IM34	25.7%			**LR26**	21.1%	**LR32**	31.2%		
LR26	22.4%			**LR32**	26.6%				
LR29	21.2%								
LR30	27.8%								
LR32	25.7%								

Table 5.6: 10bp repeats with a deletion frequency ≥20% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

Most 11bp homopolymers sequenced in the control samples have a deletion frequency below 30% (see Figure 5.16). There were four exceptions. The repeat LR01 had a deletion frequency of 62.8% and 35.4% in the MSS tumour 169259 and the U179 normal mucosa. The repeat LR23 had a deletion frequency of 46% in the MSS tumour 169736 and the repeat LR16 had a deletion frequency of 57.3% in the sample from patient U096. In the MSI-H samples there were many repeats that that had a deletion frequency ≥ 30% (see Figure 5.16). These consisted of 13 repeats for the U029 tumour sample, 5 repeats for the U179_H03 tumour sample, 5 repeats for the U303 tumour sample, two repeats the U179_H12 tumour sample, and only one repeat for the U312 tumour sample (see Table 5.7).

LR01, LR16, LR23 had high deletions frequencies in the control samples and as mentioned before this could potentially be due to polymorphisms. The mononucleotide repeats used in this study have all been screened for polymorphisms using the dbSNP (version 137), but there may still be polymorphism present for some of the repeats which have not been registered in this version of dbSNP. It is also possible that the variant reads for some of the Lynch Syndrome tumours may be a result of polymorphisms.

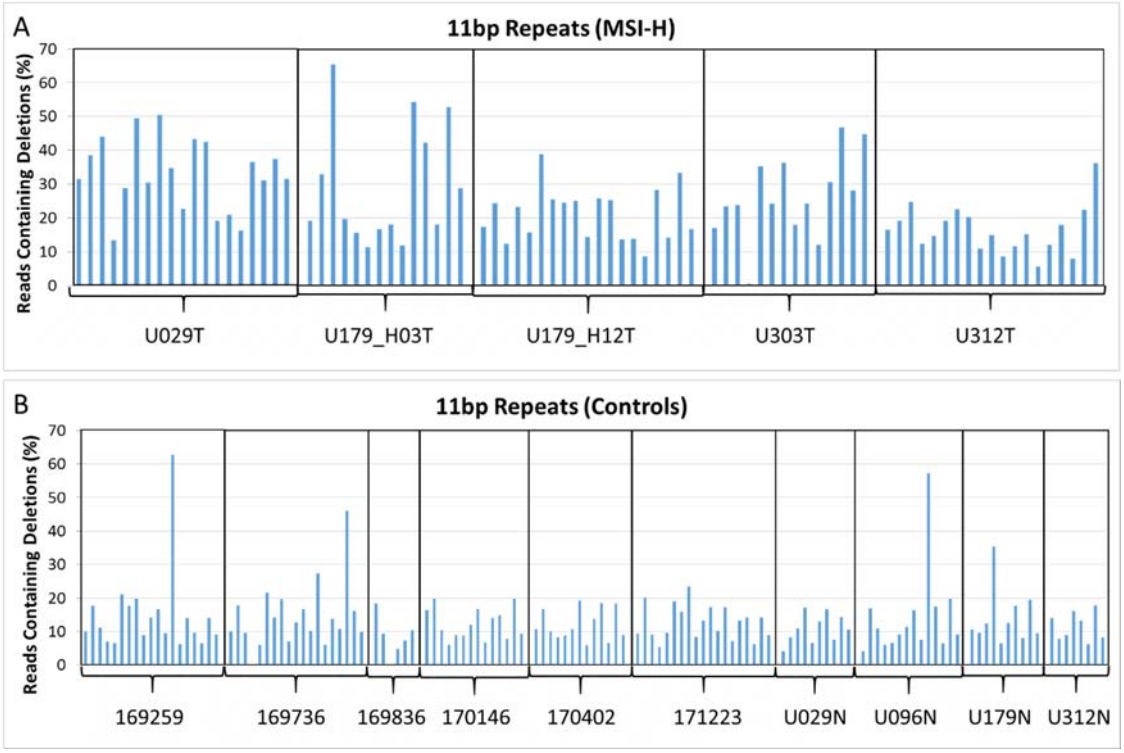

Figure 5.16: Deletion frequencies in all the 11bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H03T		U179_H12T		U303T		U312T	
Repeat	DF	Repeat	DF	Repeat	DF	Repeat	DF	Repeat	DF
GM02	38.4%	GM02	32.9%	IM28	38.8%	GM14	35.3%	LR48	36.1%
GM07	44.0%	GM07	65.3%	LR33	33.3%	IM65	36.3%		
IM28	49.3%	LR16	54.3%			LR17	30.5%		
IM32	30.4%	LR17	42.2%			LR23	46.8%		
IM45	50.4%	LR33	52.8%			LR48	44.8%		
IM52	34.7%								
IM54	43.2%								
IM65	42.4%								
LR12	31.4%								
LR17	36.5%								
LR32	31.2%								
LR33	37.4%								
LR48	31.5%								

Table 5.7: 11bp repeats with a deletion frequency ≥30% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

The 12bp mononucleotide repeat IM51 had a deletion frequency above 40% in the two of the MSS tumours 169259 and 171223. It is possible that the repeat IM51 contains a polymorphism for these two samples. Other than this repeat there were no more repeats with a deletion frequency ≥ 40% in the control samples (see Figure 5.17). Out of the MSI-H tumours there were three which contained repeats with a deletion frequency ≥ 40% (Figure 5.17). The three samples were the U029 tumour with 5 repeats, the U179_H03 tumour with 3 repeats, and the U303 tumour with 4 repeats (see Table 5.8).

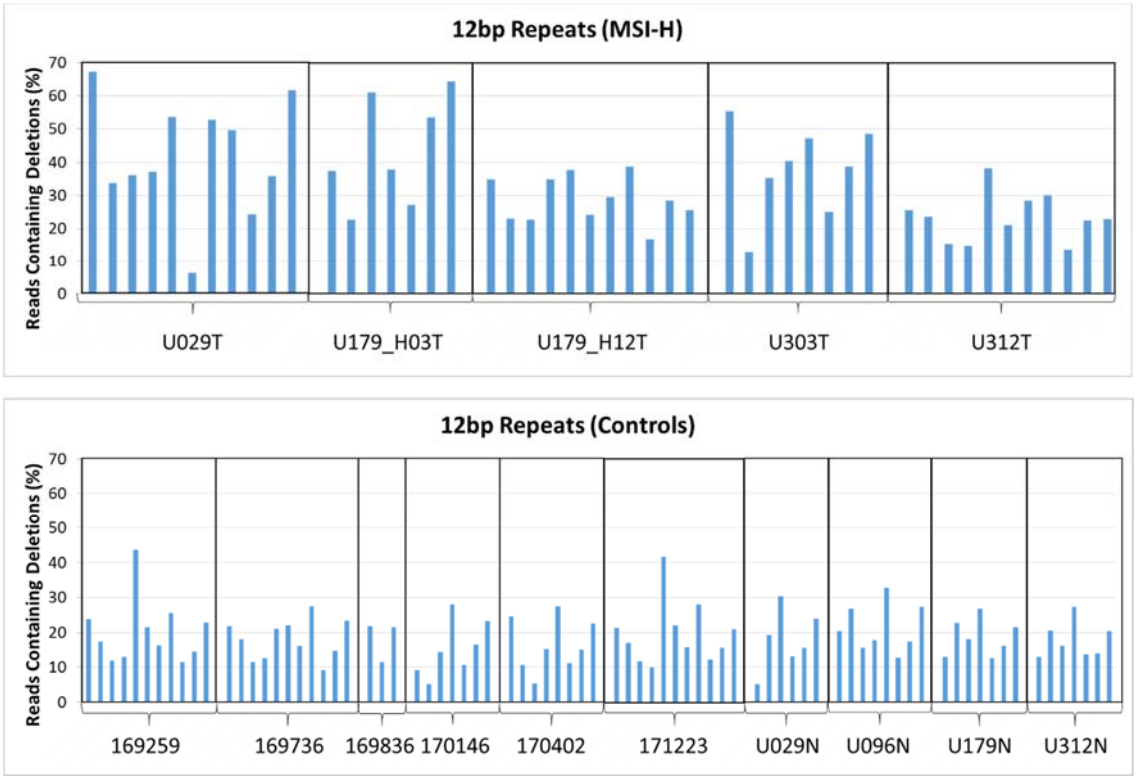

Figure 5.17: Deletion frequencies in all the 12bp mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

U029T		U179_H03T		U303T	
Repeat	DF	Repeat	DF	Repeat	DF
GM18	67.4%	LR36	61.0%	GM18	55.4%
IM51	53.7%	LR44	53.4%	LR36	40.6%
LR36	52.7%	LR52	64.4%	LR41	47.4%
LR41	49.5%			LR52	48.5%
LR52	61.8%				

Table 5.8: 11bp repeats with a deletion frequency ≥40% in the MSI-H samples. T= tumour sample, DF= deletion frequency.

With an increase in repeat length there was a higher deletion frequency observed in the unstable repeats of the MSI-H samples. There was however also an increase in the deletion frequencies of the controls and an increase in the variation of deletion frequencies between different repeats of the same size. An increase in repeat length meant an increase in the average deletion frequency for both MSI-H tumours and controls (see Figure 5.18). There was a larger increase in the average deletion frequency of the MSI-H tumours with increased repeat length compared to the controls.

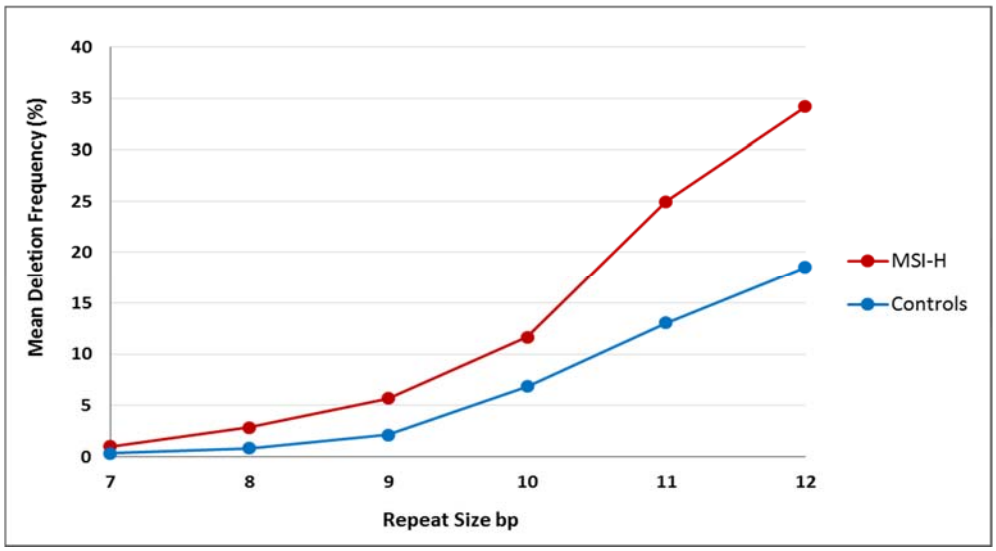

Figure 5.18: Mean deletion frequencies for the A/T mononucleotide repeats.

For the 7bp-8bp G/C mononucleotide repeats all of the control samples had a low background deletion frequency of less than 1% (see Figure 5.19). Both 8bp G/C repeats also had a deletion frequency of less than 1% in all the MSI-H samples. The three 7bp G/C repeats IM66, IM67 and LR08 had a higher deletion frequency than observed in any of the controls for at least one of the MSI-H samples (see Figure 5.19). The repeat IM66 had a deletion frequency of 16.9% and 14.3% respectively in samples U029 tumour and U179_H12 tumour. The repeat IM67 had a deletion frequency of 9% and 2.8% for the U179_H12 tumour sample and the U303 tumour sample respectively. The repeat LR08 had a deletion frequency of 14.7% in the U029 tumour.

For the 9bp G/C mononucleotide repeat LR05, the deletion frequency in all of the control samples was between 23-33% (see Figure 5.19). This is higher than any of the deletion frequencies seen in the controls for the 9bp A/T repeats (see Figure 5.14). Out of the five MSI-H samples, two samples had a higher deletion frequency for the repeat LR05 than was seen in the control samples (see Figure 5.19). These two samples were U179_H12 tumour where LR05 had a deletion frequency of 38.1% and sample U303 tumour where LR05 had a deletion frequency of 44.3%.

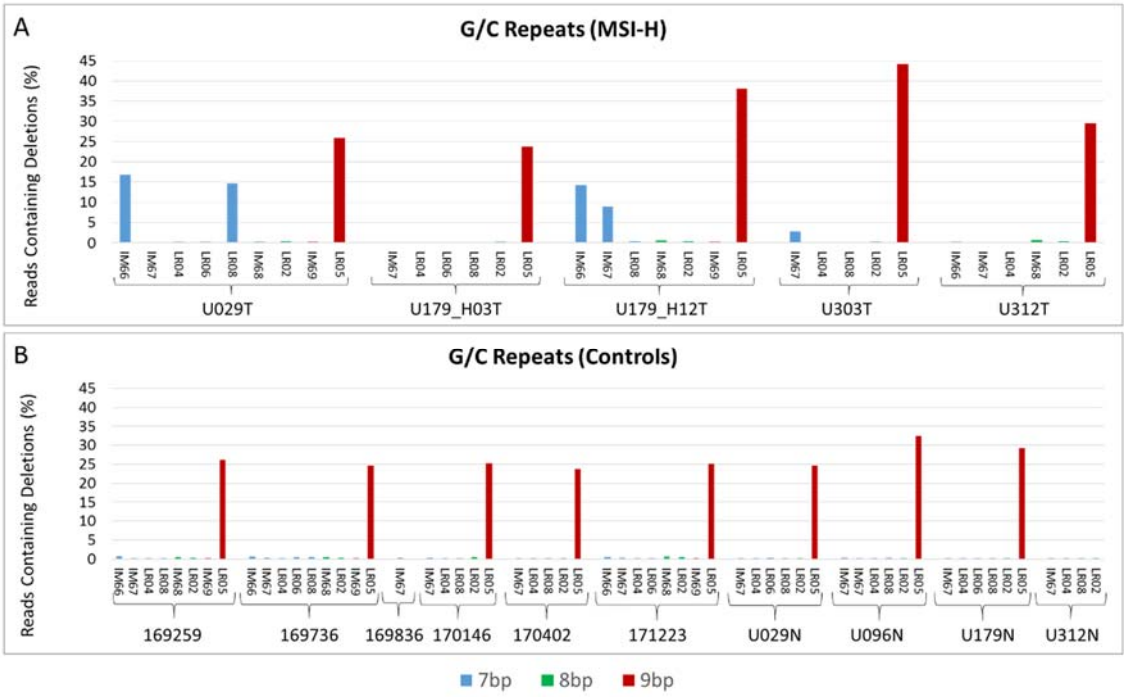

Figure 5.19: Deletion frequencies in all the G/C mononucleotide repeats. Panel A shows the deletion frequencies in the MSI-H tumours. Panel B shows the deletion frequencies in the controls.

Despite the existence of variant reads from PCR based error, Figure 5.20 shows that use of multiple short repeats can readily identify MSI-H tumours that exhibit limited instability as assessed by fragment analysis.

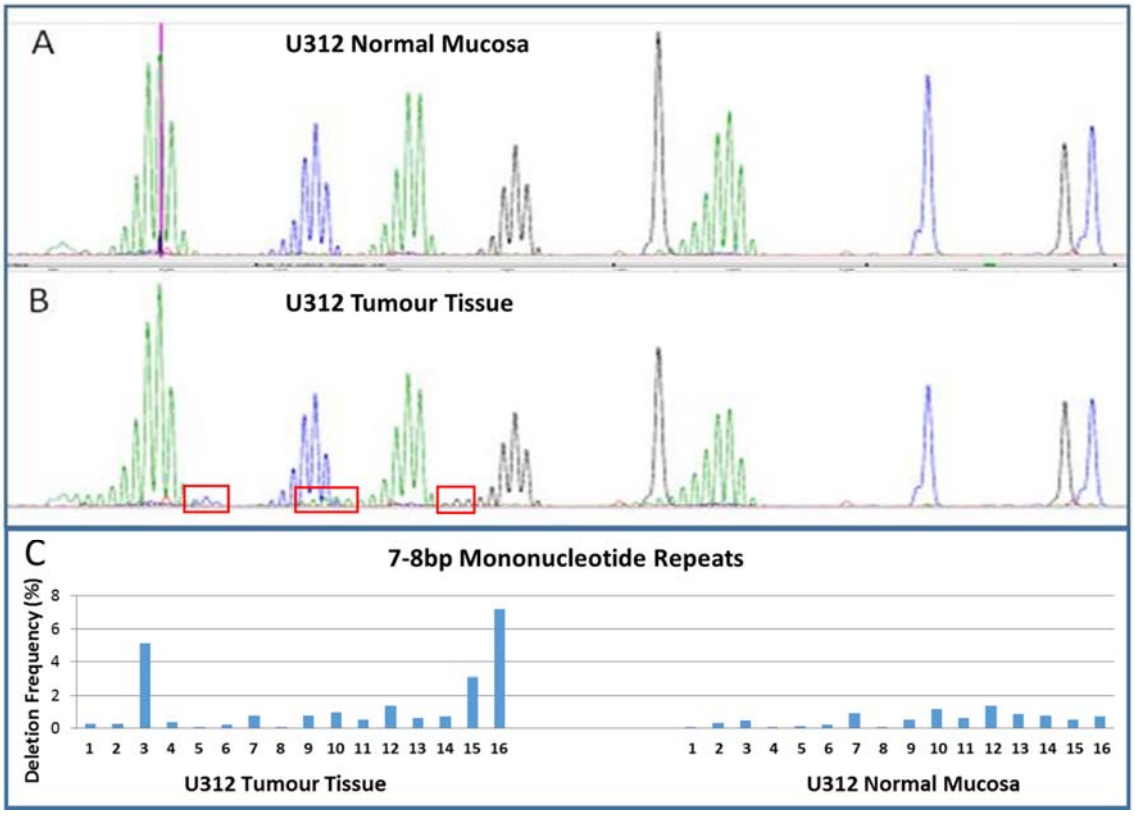

Figure 5.20: Comparison between using a standard fragment analysis test and using the short 7-8bp markers that were sequenced in both tumour and normal tissue. A: Fragment analysis results for the U312 normal mucosa. B: Fragment analysis results for the U312 tumour. C: Sixteen 7-8bp mononucleotide repeats that were sequenced in both the U312 tumour and normal mucosa.

5.2.5. The allelic distribution of MSI in variable repeats identified from whole genome sequences data

All A/T repeats and most of the G/C repeats sequenced had neighbouring SNPs with a high minor allele frequency. Homopolymers with these neighbouring SNPs with a high minor allele frequency were chosen to enable the study of allelic bias for these homopolymers.

In Figure 5.21 there are some examples of allelic bias in MSI-H tumours. For the 7bp and 8bp repeats, the reads containing a 1bp deletion are mostly present on one allele (see Figure 5.21 panels A-B). For the 11bp repeat IM65 in the U029 tumour sample there is an imbalance between the two alleles both for the 1bp deletion (Fisher's exact test: p-value $<10^{-100}$) and for the 3bp deletion (Fisher's exact test: p-value 3.1×10^{-72}) (see Figure 5.21 panel D). This suggests this repeat has had two separate replication mistakes, which have not been rectified by the compromised mismatch repair system. For the 12bp repeat LR36 in the U303 tumour sample there are significantly more reads containing a 2bp

deletion on the allele with an A at the SNP site than the allele with a T (Fisher's exact test: p-value 4.22×10^{-36}).

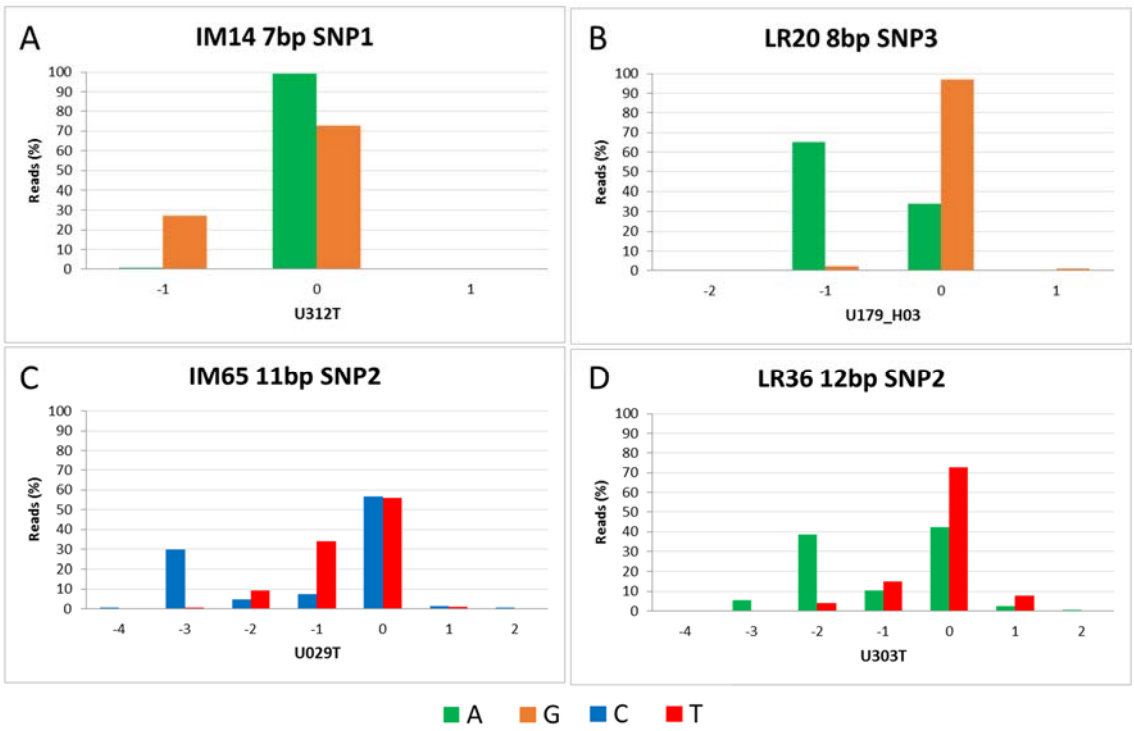

Figure 5.21: Examples of allelic imbalance in different lengths of mononucleotide repeat. Panel A the repeat IM14 in tumour U312, Panel B the repeat LR20 in tumour U179_H03, Panel C the repeat IM65 in tumour U029, Panel D the repeat LR36 in tumour U303.

To investigate allelic bias across all samples and all heterozygous repeats the Perl scripts FisherTest_AllDeletions.pl and FisherTest_IndividualIndels.pl were written. Both scripts use the output files of our in house variant caller COPReC as input files. The Perl scrips identify repeats that are heterozygous for a neighbouring SNP and perform a Fisher's exact test to determine if the fraction of variant reads is significantly different between the two alleles. Repeats were defined as heterozygous if there were 100 paired end reads spanning both SNP and repeat for each allele and one allele did not have less than 10% of the total read count. The criteria of a minimum of 100 paired end reads per allele was used to prevent a misrepresentation of variant frequencies caused by PCR duplicates. The criteria that repeats were not analysed if one allele has less than 10% of the total read count was used because such an extreme allele imbalance might indicate sample contamination. The Fisher's exact test calculations were performed using an external module integrated into my Perl program. The module was written by Pedersen T. (https://metacpan.org/pod/Text::NSP::Measures::2D::Fisher::twotailed). The script FisherTest_AllDeletions.pl calculates the fraction of reads that contain a deletion and the fraction of reads that do not contain a deletion for each allele and performs a Fisher's

exact test to see if there is a significant difference in deletion distribution between the two alleles. The script FisherTest_IndividualIndels.pl calculates the fraction of reads that correspond to each individual insertion and deletion size, then calculates if there is a significant difference between the two alleles for each separate indel size.

Figure 5.22 shows the results for the Fisher's exact test where the significance of differences in total deletion frequencies between the two alleles of repeats were calculated. The repeats plotted in Figure 5.22 include only repeats where the neighbouring SNP was classified as heterozygous. In some cases, a repeat had more than one neighbouring heterozygous SNP and in these cases, all heterozygous SNP repeat combinations were plotted. This method was chosen because different SNPs would have a different number of reads spanned both SNP and repeat. Therefore, different repeat and SNP combinations could provide different levels of significance for allelic bias. The results of the two-tailed Fisher's exact test indicate that there is more allelic bias in the MSI-H samples compared to the MSS samples (see Figure 5.22). To Bonferroni correct a p-value of 0.01, this p-value was divided by the number of heterozygous SNP repeat combinations ($0.01/519 = 1.9\times10^{-5}$). A table containing the number of repeats with a statistically significant p-value can be found in Table 5.9. There were 52 repeats with a statistically significant p-value in the MSI-H samples compared to 12 in the controls. There are three mononucleotide repeats in control samples that have an allelic bias with a p-value below 10^{-20} (see Figure 5.22). These include both U096 samples where there is a large bias between the alleles for the repeat LR16. As mentioned before the LR16 repeat is almost certainly polymorphic in patient U096 and this would explain the level of bias in deletion frequency seen between the two alleles of this repeat. The third repeat with a p-value below 10^{-20} is LR23 in the MSS tumour 169736. This is also a potential polymorphism.

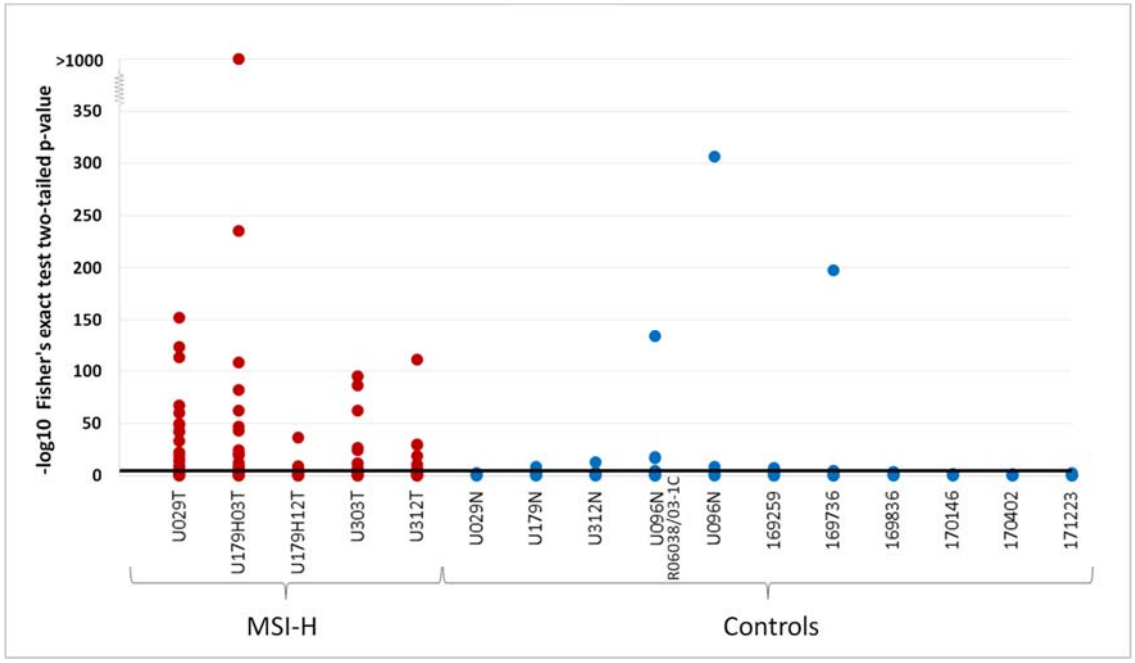

Figure 5.22: Allelic bias in deletion frequency for MSI-H samples and MSS samples measured using the p-value of a two tailed Fisher's exact test. Red = MSI-H samples, Blue = MSS samples. The line corresponds to a Bonferroni corrected p-value of 0.01.

Tumour sample	Sample Type	Number of repeats with a significant allelic bias (p-value ≤ 1.9×10-5)
U029T	MSI-H Tumour	16
U179H03T	MSI-H Tumour	16
U179H12T	MSI-H Tumour	4
U303T	MSI-H Tumour	8
U312T	MSI-H Tumour	8
U029N	Normal Mucosa	0
U179N	Normal Mucosa	1
U312N	Normal Mucosa	1
U096N R06038/03-1C	Normal Mucosa	3
U096N	Normal Mucosa	2
169259	MSS Tumour	4
169736	MSS Tumour	1
169836	MSS Tumour	0
170146	MSS Tumour	0
170402	MSS Tumour	0
171223	MSS Tumour	0

Table 5.9: The number of repeat with a Bonferroni corrected p-value of 0.01 ($0.01/519 = 1.9\times10^{-5}$) for each tumour sample.

Repeats with a neighbouring heterozygous SNP were also analysed to determine the significance of bias between the two alleles for individual indel sizes using the script FisherTest_IndividualIndels.pl. This was done using a two-tailed Fisher's exact test where the frequency of each individual indel size was interrogated (see methods section 2.9.1). For each allele the reads were classed as containing the indel size under investigation or does not contain the indel size under investigation. For each repeat, the

indel with the lowest p-value was recorded in Table 5.10. If there were multiple heterozygous SNPs neighbouring a repeat then the SNP where the lowest p-value was obtained was used.

The MSI-H samples have the highest number of heterozygous repeats with an indel event which is significantly biased between the two alleles. Up to a significance level of p-value $<10^{-10}$ there are a higher number of repeats in the MSI-H samples (see Table 5.10). However, the number of repeats sequenced differs between samples and the number of heterozygous repeats also differ between samples. For the MSI-H samples the fraction of the heterozygous repeats that contain allelic imbalance for individual indel sizes is generally higher than seen in the controls. The U179_H03 tumour sample has an allelic imbalance at a significance level of p-value $<10^{-10}$ for 46% of the heterozygous repeats, U029 tumour for 45% of the heterozygous repeats, U303 tumour for 21% of the heterozygous repeats, U179_H12 tumour for 10% of the heterozygous repeats, and the U312 tumour for 11% of the heterozygous repeats. The fraction of the heterozygous repeats that contain allelic imbalance for individual indel sizes is also high in the U096 controls. For the U096 sample from block R06038/03-1C there is an allelic imbalance at a significance level of p-value $<10^{-10}$ for 10% of the heterozygous repeats and for the other U096 sample (CAPP2 wax block label: U096 normal 23.12.02) an allelic imbalance in 17% of the repeats.

The U096 patient sample from block R06038/03-1C had three repeats with an allelic bias for 1bp deletions of a significance level of p-value $<10^{-10}$. These three repeats were LR16 (p-value $<10^{-100}$), LR27 (p-value $2.9\text{x}10^{-17}$), and LR51 (p-value 2.1×10^{-18}). LR16 is suspected to be polymorphic in patient U096. The U096 sample (U096 normal 23.12.02) shows allelic bias for a 1bp deletion in the repeat LR16 which is believed to be a polymorphism.

Status	Sample	Repeats with 2 alleles	p-value <1E-10	p-value <1E-7	p-value <1E-5	p-value <1E-3	p-value <0.05
Lynch Tumour	U029T	42	19	19	19	20	25
Lynch Tumour	U179T H03	37	17	19	19	20	24
Lynch Tumour	U179T H12	41	4	6	6	9	13
Lynch Tumour	U303T	38	8	8	9	10	17
Lynch Tumour	U312T	45	5	7	9	10	17
Normal Mucosa	U029N	17	0	0	0	0	3
Normal Mucosa	U179N	20	0	1	2	4	9
Normal Mucosa	U312N	18	1	1	1	1	4
Normal Mucosa	U096N R06038/03-1C	29	3	3	3	4	10
Normal Mucosa	U096N (23.12.02)	6	1	2	2	2	3
MSS Tumour	169259	49	0	0	1	6	10
MSS Tumour	169736	39	1	1	1	3	9
MSS Tumour	169836	16	0	0	0	1	3
MSS Tumour	170146	19	0	0	0	0	2
MSS Tumour	170402	33	0	0	0	0	0
MSS Tumour	171223	37	0	0	0	0	5

Table 5.10: The number of repeats with allelic bias for individual indels sizes measured using the p-value of a two tailed Fisher's exact test.

5.2.6. Identifying the most informative homopolymers

The ability of each repeat to discriminate between the MSI-H samples and the MSS samples was assessed using the area under the receiver operating characteristic curve (AUC). Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University) performed the AUC calculations. Receiver operating characteristic curves are a method of measuring true positive and false positive rates. In this case the AUC is a measure of how well a given homopolymer can differentiate between the MSI-H and MSS samples.

Using AUC as a measure of a repeat's ability to discriminate between MSI-H samples and controls, it was concluded that the discrimination power increased with the length of repeat for the A/T mononucleotide repeats (see Figure 5.23). The 11bp and 12bp mononucleotide repeats achieved the best discrimination between MSI-H samples and controls with a median AUC of above 0.95 for separating the MSI-H and control samples (see Figure 5.23). Because the longer repeats are better able to separate the MSI-H and control samples it was decided to include mainly long repeats in the final panel of repeats. However some MSI-H samples are easier to identify using the shorter repeats, for example the U312 tumour sample analysed in this chapter and the U096 tumour sample analysed in chapter 3. The U096 tumour sample in chapter 3 did not show any deletion frequencies above what was observed in the controls for the 12bp-15bp mononucleotide repeats sequenced, but did have a 1bp deletion frequency above 14% for two 8bp repeats

(DEPDC2 and AL359238) and a 1bp deletion consisting of 18.6% of the reads for one 10bp repeat (AVIL). The U312 tumour sample analysed in this chapter only showed a deletion frequency above 1.5x the deletion frequency seen in any of the control samples for the 11bp repeats GM07, GM14 and LR48. No 12bp repeats had a deletion frequency above 1.5x the deletion frequency seen in any of the control samples for the U312 tumour sample. The U312 tumour sample was easier to identify as MSI-H using the shorter 7bp-10bp repeats (see Figure 5.12 - Figure 5.15). This suggests that it may be beneficial to include some of the shorter repeats in a final panel.

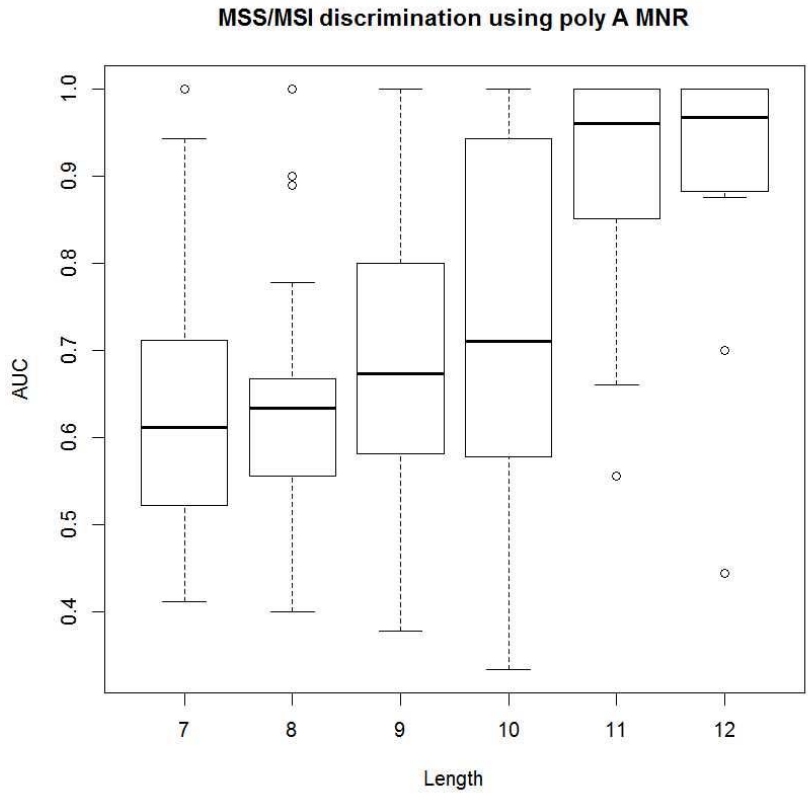

Figure 5.23: Box plot showing the ability of different mononucleotide lengths to separate between MSI-H samples and MSS samples using the area under the receiver operating characteristic curve (AUC).

The aim was to select markers with a clear difference between unstable repeats and background PCR and sequencing error. The aim was also to select repeats which showed instability in as many of the MSI-H samples as possible. To achieve this, markers were classed as unstable in the MSI-H samples if a marker had a deletion frequency of >5% and at least two times the deletion frequency of any of the control samples for the 7-9bp repeats. For the 10bp-12bp repeats a marker was classed as unstable in a MSI-H sample if it had a deletion frequency of >5% and at least 1.5 times higher than seen in any control sample for the same repeat. These thresholds were chosen arbitrarily. The

markers that were chosen for further investigation where markers that were classed as unstable in at least 60% of the MSI-H samples and also had an AUC of at least 0.9. Marker LR11was also chosen, despite having a lower AUC (AUC: 0.82), because the AUC for the SNP was higher than 0.9, making this an interesting marker for studying allelic bias. The markers that were chosen are highlighted in grey in Table 5.11 and Table 5.12. Most of these repeats are long repeats of 10bp or longer, with the exception of the 8bp repeat LR46. The repeat LR46 was chosen because it had a deletion frequency in all five MSI-H tumours of more than 2 times the deletion frequency of any of the control samples. None of the G/C repeats analysed in this chapter were chosen as part of the final panel because repeats with a higher instability rate and higher AUC for the MSI-H samples were available from the A/T repeats sequenced.

MS size	MS	MSI-H (>100 reads)	MSI >2x largest MSS (deletion frequency)	Fraction of MSI-H detected (>100 reads)	AUC	Bias.AUC SNP	Info
7bp	IM14	3	1	0.33	0.89	NA	
7bp	IM43	3	1	0.33	0.56	NA	
7bp	IM55	3	1	0.33	0.67	NA	
7bp	LR51	3	1	0.33	0.56	NA	
7bp	IM19	5	1	0.2	0.76	NA	
7bp	LR15	5	1	0.2	0.62	NA	
7bp	LR49	5	1	0.2	0.51	NA	
7bp	GM04	4	0	0	0.53	NA	
7bp	GM19	5	0	0	0.51	NA	
7bp	GM25	4	0	0	0.49	NA	
7bp	GM27	5	0	0	0.66	NA	
7bp	GM30	5	0	0	0.94	NA	
7bp	IM13	2	0	0	0.56	NA	
7bp	IM22	3	0	0	0.67	NA	
7bp	IM23	3	0	0	0.67	NA	
7bp	IM26	3	0	0	0.89	NA	
7bp	IM27	3	0	0	0.44	NA	
7bp	IM61	3	0	0	0.56	NA	
7bp	LR13	5	0	0	0.47	NA	
7bp	LR25	4	0	0	0.41	NA	
7bp	LR45	5	0	0	0.8	NA	
7bp	LR47	5	0	0	0.6	0.73	
7bp	LR50	4	0	0	0.66	0.67	
7bp	GM24	1	0	0	0	NA	
8bp	LR46	5	5	1	1	0.83	chosen
8bp	IM59	3	2	0.67	0.67	NA	
8bp	LR20	4	2	0.5	0.67	1	
8bp	IM41	3	1	0.33	0.89	NA	
8bp	GM09	5	1	0.2	0.9	NA	
8bp	GM03	4	0	0	0.66	NA	
8bp	GM08	5	0	0	0.63	NA	
8bp	GM16	5	0	0	0.46	NA	
8bp	GM20	5	0	0	0.4	NA	
8bp	IM15	3	0	0	0.44	NA	
8bp	IM20	2	0	0	0.56	NA	
8bp	IM21	3	0	0	0.56	NA	
8bp	IM25	3	0	0	0.67	NA	
8bp	IM39	3	0	0	0.78	NA	
8bp	IM40	3	0	0	0.44	NA	
8bp	IM57	5	0	0	0.6	NA	
8bp	IM63	5	0	0	0.64	NA	
8bp	LR18	4	0	0	0.53	NA	
8bp	LR19	5	0	0	0.56	NA	
8bp	LR27	5	0	0	0.56	NA	
8bp	LR31	4	0	0	0.66	NA	
9bp	GM11	5	3	0.6	0.8	NA	
9bp	LR24	5	3	0.6	0.84	NA	
9bp	GM17	5	2	0.4	0.94	NA	
9bp	GM28	5	2	0.4	0.57	NA	
9bp	IM16	5	2	0.4	0.78	NA	
9bp	LR40	5	2	0.4	0.7	NA	
9bp	GM21	4	1	0.25	0.68	NA	
9bp	GM10	5	1	0.2	0.54	NA	
9bp	GM15	5	1	0.2	0.6	1	
9bp	GM23	5	1	0.2	0.8	NA	
9bp	IM17	5	1	0.2	0.59	NA	
9bp	LR10	5	1	0.2	0.96	NA	
9bp	LR14	5	1	0.2	0.38	NA	
9bp	LR21	5	1	0.2	0.8	NA	
9bp	GM05	5	0	0	0.57	0.5	
9bp	GM06	5	0	0	0.66	NA	
9bp	IM42	3	0	0	0.56	NA	
9bp	IM44	3	0	0	1	NA	
9bp	LR28	5	0	0	0.67	1	
9bp	LR34	5	0	0	0.62	NA	

Table 5.11: Table containing information for all the 7bp-9bp A/T repeats sequenced. Repeats were classed as unstable if they had a deletion frequency above 5% and >2x higher than any of the control samples.

MS size	MS	MSI-H (>100 reads)	MSI >1.5x largest MSS (deletion frequency)	Fraction of MSI-H detected (>100 reads)	AUC	Bias.AUC SNP	Info
10bp	GM29	5	4	0.8	0.94	NA	chosen
10bp	LR32	5	4	0.8	1	NA	chosen
10bp	GM01	5	2	0.4	0.94	NA	
10bp	GM22	5	2	0.4	0.97	0.73	
10bp	GM26	5	2	0.4	0.87	0.89	
10bp	LR29	5	2	0.4	0.68	NA	
10bp	LR39	5	2	0.4	0.74	NA	
10bp	IM07	3	1	0.33	0.78	NA	
10bp	IM33	3	1	0.33	0.56	NA	
10bp	IM34	3	1	0.33	0.67	NA	
10bp	LR30	5	1	0.2	0.42	NA	
10bp	IM12	3	0	0	0.67	NA	
10bp	IM35	3	0	0	0.33	NA	
10bp	IM37	3	0	0	1	NA	
10bp	LR26	5	0	0	0.6	0.65	
10bp	LR35	5	0	0	0.51	NA	
11bp	GM14	5	5	1	1	NA	chosen
11bp	LR48	5	5	1	1	NA	chosen
11bp	GM07	5	4	0.8	1	1	chosen
11bp	GM13	5	4	0.8	0.89	NA	
11bp	LR11	5	4	0.8	0.82	0.92	chosen
11bp	LR17	5	4	0.8	0.98	NA	
11bp	IM28	3	2	0.67	0.78	NA	
11bp	IM54	3	2	0.67	0.67	NA	
11bp	IM52	5	3	0.6	1	NA	
11bp	IM65	5	3	0.6	0.96	0.72	
11bp	LR33	5	3	0.6	1	NA	
11bp	GM02	5	2	0.4	0.94	NA	
11bp	IM32	3	1	0.33	1	NA	
11bp	IM45	3	1	0.33	0.89	NA	
11bp	IM53	3	1	0.33	1	NA	
11bp	LR12	5	1	0.2	1	NA	
11bp	LR01	1	0	0	0.56	NA	Potential polymorphism
11bp	LR16	5	0	0	0.66	0.67	Potential polymorphism
11bp	LR23	5	0	0	0.88	NA	Potential polymorphism
12bp	LR44	5	4	0.8	1	0.73	chosen
12bp	IM49	3	2	0.67	1	NA	chosen
12bp	LR36	5	3	0.6	1	NA	chosen
12bp	LR43	5	3	0.6	0.98	NA	
12bp	LR52	5	3	0.6	0.88	NA	
12bp	GM18	5	2	0.4	0.97	0.89	
12bp	IM50	5	2	0.4	0.89	NA	
12bp	IM47	3	1	0.33	1	NA	
12bp	IM64	4	1	0.25	0.7	NA	
12bp	LR41	5	1	0.2	0.96	NA	
12bp	IM51	3	0	0	0.44	NA	

Table 5.12: Table containing information for all the 10bp-12bp repeats sequenced. Repeats were classed as unstable if they had a deletion frequency above 5% and >1.5x higher than any of the control samples. Only samples with > 100 reads were analysed. Repeats that were chosen for further analysis are highlighted in grey.

MS size	MS	MSI-H (>100 reads)	MSI >2x largest MSS (deletion frequency)	Fraction of MSI-H detected (>100 reads)	AUC	Bias.AUC SNP
7bp	IM66	3	2	0.67	0.67	NA
7bp	LR08	4	1	0.25	0.43	NA
7bp	IM67	5	1	0.2	0.55	NA
7bp	LR04	4	0	0	0.62	NA
7bp	LR06	2	0	0	0.75	NA
8bp	IM68	3	0	0	0.56	NA
8bp	LR02	5	0	0	0.73	NA
9bp	IM69	2	0	0	0.56	NA
9bp	LR05	5	0	0	0.73	NA

Table 5.13: Table containing information for all the G/C mononucleotide repeats sequenced. Repeats were classed as unstable if they had a deletion frequency above 5% and >2x higher than any of the control samples. Only samples with > 100 reads were analysed.

The repeats in Table 5.14, which were taken from the literature and analysed in chapter 3, were also selected as part of the final panel of repeats. Some shorter 8bp and 9bp repeats were included because shorter repeats may be more unstable in some tumours. The U096 tumour in chapter 3 displayed more instability in the shorter repeats than >10bp repeats. There was also a tumour in this chapter, U312 tumour, which displayed more instability in the shorter 7bp-10bp repeats compared to the 11bp and 12bp repeats (see Figure 5.12 - Figure 5.17).

Repeat Name	Size (bp)	Repeat Base	Number of MSI-H samples with instability in this repeat in chapter 3
DEPDC2	8	C	1 out of 4
AL359238	8	A	1 out of 4
AL954650	9	C	1 out of 4
AP003532_2	9	A	1 out of 4
TTK	9	A	1 out of 4
AL355154	10	A	2 out of 4
AVIL	10	A	3 out of 4
ASTE1	11	A	3 out of 4
EGFR	13	A	2 out of 4
FBXO46	14	A	3 out of 4

Table 5.14: Repeats taken from the literature and analysed using a panel of 4 MSI-H tumours and controls in chapter 3.

5.3. Discussion

To validate specific repeats for MSI detection, 120 of the repeats with the highest variant read frequency identified from whole genome sequence data were analysed in a small panel of tumours and control tissues using Illumina sequencing. Repeats were selected to ensure representation of all repeat lengths (7bp-12bp). In addition, repeats linked to a SNP with a high minor allele frequency were chosen to faciliate allele specific variant read identification. This chapter has concentrated on deletion frequencies to gauge MSI because in previous chapters the conclusion has been that deletions are more indicative of MSI than insertions. This is also consistent with results obtained by Yoon et al. (2013) who showed that mutations in mononucleotide repeats that are caused by MSI in gastric cancers are mainly deletions. One of my aims was to select the most variable of the 120 repeats so these could become part of a final panel of repeats for the use as an MSI test. Ten repeats from the literature analysed in chapter 3, have already been chosen to as part of this final panel of repeats.

Of the 120 repeats sequenced, MSI was observed as an increase in deletion frequency in at least one MSI-H cancer for 63 of the A/T mononucleotide repeats. 40% of the short A/T repeats (7bp-9bp) showed MSI, compared to 80% of longer (10bp-11bp) A/T repeats (see Table 5.11 and Table 5.12). However, longer repeats showed more PCR/Sequencing error derived variability in control tissues. The 7bp and 8bp repeats had the lowest instability rates. 7bp -9bp repeats were classed as unstable in a MSI-H tumour if there was a deletion frequency above 5% and >2x the deletion frequency observed in any of the control samples for that repeat. For the 7bp repeats seven out of the 24 repeats sequenced showed instability in at least one MSI-H tumour. For the 8bp repeats only 5 out of the 21 repeats showed instability in at least one MSI-H tumour. However one of the 8bp repeats LR46 was unstable in all five tumours sequenced. This could suggest that there is some attribute of the location of this repeat that makes it more susceptible to MSI. For this reason LR46 was chosen for further study as part of the final panel of repeats to be used on a larger panel of tumours. Fourteen out of the twenty 9bp repeats sequenced showed instability in at least one tumour. Comparatively, five out of the six 9bp repeats from the literature, and analysed in chapter 3, showed instability in at least one of the 5 tumours sequenced. This means that a greater fraction of the 9bp repeats in chapter 3 showed instability in at least one tumour. However, the method used for classifying markers as unstable is different in this chapter to the one used in chapter 3. In this chapter analysis of the total deletion frequency of repeats in the MSI-H samples has been

compared to the total deletion frequency of the same repeat in control samples, while in chapter 3 arbitrary cut-offs were defined for individual deletion sizes.

The longer repeats were better able to discriminate between the MSI-H samples and controls (see Figure 5.23). More of these repeats were therefore chosen as part of the final panel of repeats to be tested on a larger panel of tumours. The 10bp -12bp repeats were classed as unstable in a MSI-H tumour if there was a deletion frequency >1.5x the deletion frequency observed in any of the control samples for that repeat. For the 10bp repeats, instability was observed in 11 out of the 16 repeats for at least one MSI-H tumour. Two of the repeats even showed instability in 4 out of the 5 MSI-H samples analysed. These two repeats were therefore chosen to be part of the final panel or repeats. For the 11bp and 12bp repeats there was instability in 16 out of the 19 repeats analysed and 10 out of the 11repeats analysed respectively. Many of the 11bp and 12bp repeats showed instability in 3-5 of the tumours tested. Four of the 11bp repeats and 3 of the 12bp repeats which showed instability in the largest number of tumours were chosen to become part of the final panel of repeats.

Of the 9 G/C mononucleotide repeats sequenced, only 3 (33%) showed instability in at least one of the MSI-H tumours sequenced. Because more unstable repeats were available from the A/T mononucleotide repeats none of the G/C repeats were chosen for the final panel of repeats to be sequenced in a larger panel of tumours.

Consistently more repeats had higher deletion frequencies in the MSI-H samples compared to the controls. However, there were some exceptions with repeats in control samples showing a high deletion frequency. In some cases, such as for the 11bp repeats LR01, LR16, and LR23 (see Figure 5.16), the reason for a high deletion frequency may be due to polymorphisms. For some of the repeats on the other hand the high deletion frequencies are unlikely to be caused by polymorphisms. This is the case for the 8bp repeat LR19 which has a deletion frequency of 5.8% in the U029 normal mucosa and the 10bp repeat IM35 which has a deletion frequency of 23.6% in the MSS tumour sample 169736. Yoon et al. (2013) analysed MSI in gene regions of gastric cancers and gastric cancer cell lines and mononucleotide repeats with deletions were also discovered in MSS cell lines suggesting that mismatch repair deficiency is not the only cause of deletions in mononucleotide repeats. It is likely a few markers with deletions in stable tumours will therefore have to be taken into account when developing a test for colorectal cancers. This is expected and is also the case for the fragment analysis tests such as the Promega MSI

test where instability in one of the five markers does not mean a sample is classed as MSI-H.

For patient U179 two separate tumours were analysed, one that was resected in 2003 (U179_H03) and a second tumour that was resected in 2012 (U179_H12). The U179_H03 tumour has instability in many repeats which are stable in the U179_H12 tumour. For example for the 9bp repeats (see Figure 5.14) there were five repeats with a deletion frequency above 10% for the tumour from 2003 (GM10: 17.6%, GM11: 28%, IM16: 10.5%, LR10: 24.7%, LR40: 43.4%). All five of these markers were also sequenced in the 2012 tumour but none of them had a deletion frequency above 10% (GM10: 0.4%, GM11 1.0%, IM16: 2.6%, LR10: 7.4%, LR40: 2.5%). The presence of many stable repeats in the U179_H12 tumour which showed instability in the U179_H03 tumour indicates that the 2012 tumour is likely to be a new primary tumour, or possibly a recurrence of the U179_H03 tumour from an earlier clone before the emergence of instability in those repeats.

Mining the whole genome sequences of MSI-H tumours allowed for the selection of many repeats that had neighbouring SNPs with a higher minor allele frequency than were available for study in chapter 3. This means that there were many more repeats with heterozygous SNPs available in the sequenced samples than had been available for the samples in chapter 3. An average of 32 heterozygous SNPs were present in each of the sequenced samples (see Table 5.10). This allowed a more comprehensive study of the allelic bias of deletion frequencies in the MSI-H samples. There were more repeats with an allelic bias in the MSI-H samples compared to the controls (Figure 5.22). There were two repeats (LR16 and LR23) with a two-tailed Fisher's exact test p-value below 10^{-20} in three of the control samples (both of the U096 normal mucosa samples and the MSS tumour 169736). However the repeats LR16 and LR23 look like they are polymorphic in these samples, which would explain the levels of allelic bias. The results for allelic bias of deletion frequencies therefore suggest that if no repeats with polymorphisms are used then allelic bias can be used to confirm some of the deletions as real MSI events as opposed to sequencing and PCR error. In these cases the second allele could be used to determine background PCR and sequencing error rate as an internal control which could be compared to the allele with a high deletion frequency. For the longer 11bp and 12bp repeats it might be better to analyse allelic bias for individual deletion sizes because these repeats sometimes accumulate more than one deletion in MSI-H samples. If two deletions of different sizes occur on different alleles it would still be possible to detect the allelic

bias by looking at individual deletion sizes, while no allelic bias might be detected if total deletion frequency was used to measure allelic bias.

5.3.1. Conclusions

In conclusion, the selected 120 repeats that were highly variable in MSI-H whole genome sequences also showed a high level of instability in the five MSI-H tumours sequenced in this chapter. 40% of the short 7bp-9bp A/T repeats, 80% of the longer 10bp-12bp A/T repeats and 33% of the G/C repeats were unstable in at least one of the MSI-H tumours. Many of the sequenced repeats had neighbouring heterozygous SNPs and there was an excess of repeats showing an allelic bias of reads with deletions in the MSI-H samples compared to the controls.

Chapter 6. Clonality in MSI-H tumours

6.1. Introduction

6.1.1. Clonality within tumours

The majority of colorectal tumours are believed to develop from dysplastic crypts, progressing to adenomas and then to carcinomas before becoming metastatic diseases (Fearon, 2011, Fearon and Vogelstein, 1990). The initial driver mutation that starts a cell on the pathway to become a colorectal tumour is believed to occur in an intestinal crypt. The number of stem cells in an intestinal crypt are small (Barker et al., 2008). Because of the small population size of stem cells and the stem cells in each crypt being separate populations, crypts become monoclonal through genetic drift (Simons and Clevers, 2011). New mutations that arise might therefore drift to fixation in stem cell populations within individual crypts. A cell with a pathogenic mutation could start the process towards the development of a tumour by creating the first dysplastic crypt through this mechanism. Mutant cells then start to expand to neighbouring crypts. Theories on the mechanism of mutant crypt proliferation include crypt fission, epithelial restitution to heal a damaged area, migration of malignant cells across the epithelium down into neighbouring crypts, and dispersal of cells through the basement membrane to neighbouring crypts (Merlo et al., 2006). Kloor et al. (2012) reported cases of crypt fission with MMR deficient crypts showing irregular branching and duplication adding evidence to the theory that dysplastic crypts can spread via crypt fission. Thirlwell et al. (2010) identified partially dysplastic crypts showing a top down growth pattern and histologically normal cells in the base of the crypt. This suggests a spreading of dysplastic cells across the epithelium down into previously unaffected crypts.

There are different hypothesises as to whether the loss of mismatch repair (MMR) function is the first step towards the development of an MSI-H tumour in Lynch Syndrome patients, or if loss of MMR usually occurs at the adenoma stage (Boland, 2012). There is evidence to suggest that a knockout of the MMR genes is not the first mutation in tumour development with the discovery of adenomas that are MSS and adenomas with both MSI-H and MSS regions in Lynch Syndrome patients. A study by Giuffre and colleges used laser dissection to analyse different regions of 18 adenomas

from Lynch Syndrome patients. These adenomas were tested using both immunohistochemistry and MSI fragment analysis and two tumours showed no loss of MMR proteins or microsatellite instability while the rest were MSI-H (Giuffre et al., 2005). For many of the MSI-H tumours, differences in instability in different biopsies were observed. Eight of the tumours even had biopsies with MSI-L or MSS results as well as biopsies that were MSI-H (Giuffre et al., 2005). These finding would be consistent with initially MSS adenomas acquiring a 'second hit mutation' of a MMR gene resulting in microsatellite instability. If this were the case then the loss of MMR function would not be the initial driver mutation which initiated tumorigenesis.

On the other hand, there is also evidence to suggest that a loss of MMR function could be the first mutation which initiates the development of cancers in Lynch Syndrome patients. Kloor et al. (2012) analysed crypts in normal mucosa from Lynch Syndrome patients and discovered crypts with an absence of *MLH1* expression in *MLH1* mutation carriers, and crypts with an absence of *MSH2* expression in *MSH2* mutation carriers. In total, 27 MMR deficient crypt foci were identified (~1 per cm^2) in the normal mucosa of Lynch Syndrome carriers, while none were identified in the normal mucosa of controls. In the Lynch Syndrome patients, each deficient crypt foci consisted of between 1 to 19 crypts. Seven out of the 27 MMR deficient crypt foci were MSI tested and all seven were found to contain microsatellite instability further confirming that crypt foci were MMR deficient. These findings suggest that MMR deficient Lynch Syndrome tumours could arise from crypts which have lost MMR function. If this is the case then loss of MMR gene function may be the first mutation which initiates the transformation from normal mucosa to MSI-H tumour in Lynch Syndrome patents.

For many of the sporadic MSI-H tumours the pathway to a loss of mismatch repair function may be different to what happens in Lynch Syndrome tumours. Many of these tumours may originate as adenomas with CIMP hypermethlation, then during later tumour development, loss of *MLH1* gene expression due to *MLH1* hypermethylation causes the development of MSI (Fearon, 2011). Loukola et al. (1999) analysed adenomas from 378 patients and discovered only six patients with MSI-H adenomas, only one of which turned out to not have a germline mutation in one of the MMR genes. This adds further evidence to the theory that in sporadic MSI-H tumours the loss of MMR function occurs during the development of the tumour and is not the initial driver mutation.

Tumour initiation and progression differs for different types of tumour. For example lung cancers often develop through a field effect, where the development of several separate lung preneoplastic lesions occurs in different locations which can then develop into separate tumours (Wistuba, 2007). In other cases, tumours have a single cell as their point of origin. As the tumours grow, they accumulate mutations and some mutations (driver mutations) give cells an advantage compared to other cells. Different cells with different sets of driver and passenger mutations may arise and become prolific within the same tumour giving rise to different clones and creating heterogeneous tumours. Clones may arise which have a selective advantage allowing them to outcompete the other clones in a tumour. This can lead to selective sweeps which will return a tumour to a monoclonal state (Greaves and Maley, 2012, Merlo et al., 2006). Another possibility is that the number of clones continues to expand as a tumour develops, creating tumours that are a mosaic of different clones. Tumours can be thought of as an ecosystem consisting of evolving clones competing for the available resources such as space and nutrients (Merlo et al., 2006). For some tumours the number of clones detected can be an indicator of tumour progression. One example of a type of tumour where this is true is the Barrett's oesophagus tumour. For this tumour one study found that for every clone identified in a pre-malignant lesion the relative risk of the lesion developing into a adenocarcinoma increased by a factor of 1.43 (Maley et al., 2006). The number of clones is also associated with the chance of drug resistance. With more clones, there is a higher chance that the tumour will survive chemotherapy.

Clonality within tumours can be studied by analysing a set of markers over different regions of a tumour. Studying the clonality of tumours can give new insight into how tumours develop. Theoretically, the order in which different clones in a tumour arose can be deduced by analysing the patterns of mutations for different clones (Merlo et al., 2006). For example, if one clone has a mutation in one marker A and another clone has a mutation in both marker A and marker B, then it is likely that the clone which only has a mutation in marker A arose first. Microsatellites have previously been used to assess the clonality of tumours. One example is the analysis of tumours in the Tasmanian devil (*Sarcophilus harrisii*), population on Tasmania. Siddle et al. (2007) showed that the tumours spreading through the Tasmanian devil population are of the same clonal origin by analysing the length of different microsatellites. This use of microsatellites helped prove that the tumours are being spread as allografts from devil to devil.

6.1.2. Aims

In this chapter, the aim is to use the sequencing of short mononucleotide repeats to investigate the clonal composition of MSI-H tumours. By taking different biopsies from different regions of the same tumour, it should be possible to detect whether there are differences in the instability of short mononucleotide repeats throughout the tumour. Differences in the instability of repeats across a tumour, such as differences in variant repeat lengths, would indicate the presence of different sub-clones within a tumour. Heterozygous SNPs will also be used to determine the allelic origin of variants. This will allow the instability of repeat to be investigated in more detail because variants on different alleles can be identified. For variants located in multiple biopsies it will be possible to deduce whether a variant is present on the same allele and therefore likely to be the result of one mutation, or if a variant is located on different alleles in different biopsies and therefore the result of independent mutations. In this chapter, the aim is to:

- Determine whether there is evidence of clonal evolution in MSI-H colorectal tumours.

- Determine whether the use of heterozygous SNPs to identify on which allele a variant is present provides extra information about the clonality of MSI-H tumours.

6.2. Results

6.2.1. The curation of fresh tissue biopsies for the clonality study

Tumour and tissue samples for the clonality analysis were obtained from the Newcastle Hospitals NHS Foundation Trust after ethical review (REC reference 13/LO/1514). Biopsies were taken from fresh colorectal tumours shortly after resection using the hours of a clock face as a reference point. The side of the tumour closest to the antimesenteric border was defined as 12 o'clock. In some of the tumours it was impractical to use the antimesenteric border as 12 o'clock, for example because the tumours had grown across the antimesenteric border. In these cases the proximal orientation of the tumour was defined as 12 o'clock. Where possible four scalpel biopsies of external tumour tissue were taken from the 3, 6, 9 and 12 o'clock positions round the tumour followed by four fine needle aspiration biopsies taken from the 3, 6, 9 and 12 o'clock positions from deeper within the tumours. If the tumour was too small for this sampling technique then not all 8 biopsies were collected. Normal mucosa was sampled using a scalpel 7-10cm away from the tumour to ensure the normal mucosa biopsies were not contaminated by any tumour tissue (see methods section 2.1.2.4 for more details). A total of 13 tumours were biopsied by Dr Stephanie Needham (Pathology department, Newcastle Hospitals NHS Foundation Trust).

6.2.2. MSI fragment analysis testing of tumours to identify MSI-H tumours

To identify MSI-H tumours, one biopsy from each tumour as well as the matched normal mucosa was tested using the Promega MSI Analysis System, Version 1.2 kit (Promega, Madison, WI, United States of America). The fragment analysis tests showed that for 10 out of the 13 tumours there was no difference in instability between the normal mucosa biopsy and the tumour biopsy. The remaining three tumours were unstable at all 5 markers (see fragment analysis traces Figure 6.1). These three MSI-H tumours were used for the subsequent work described in this chapter.

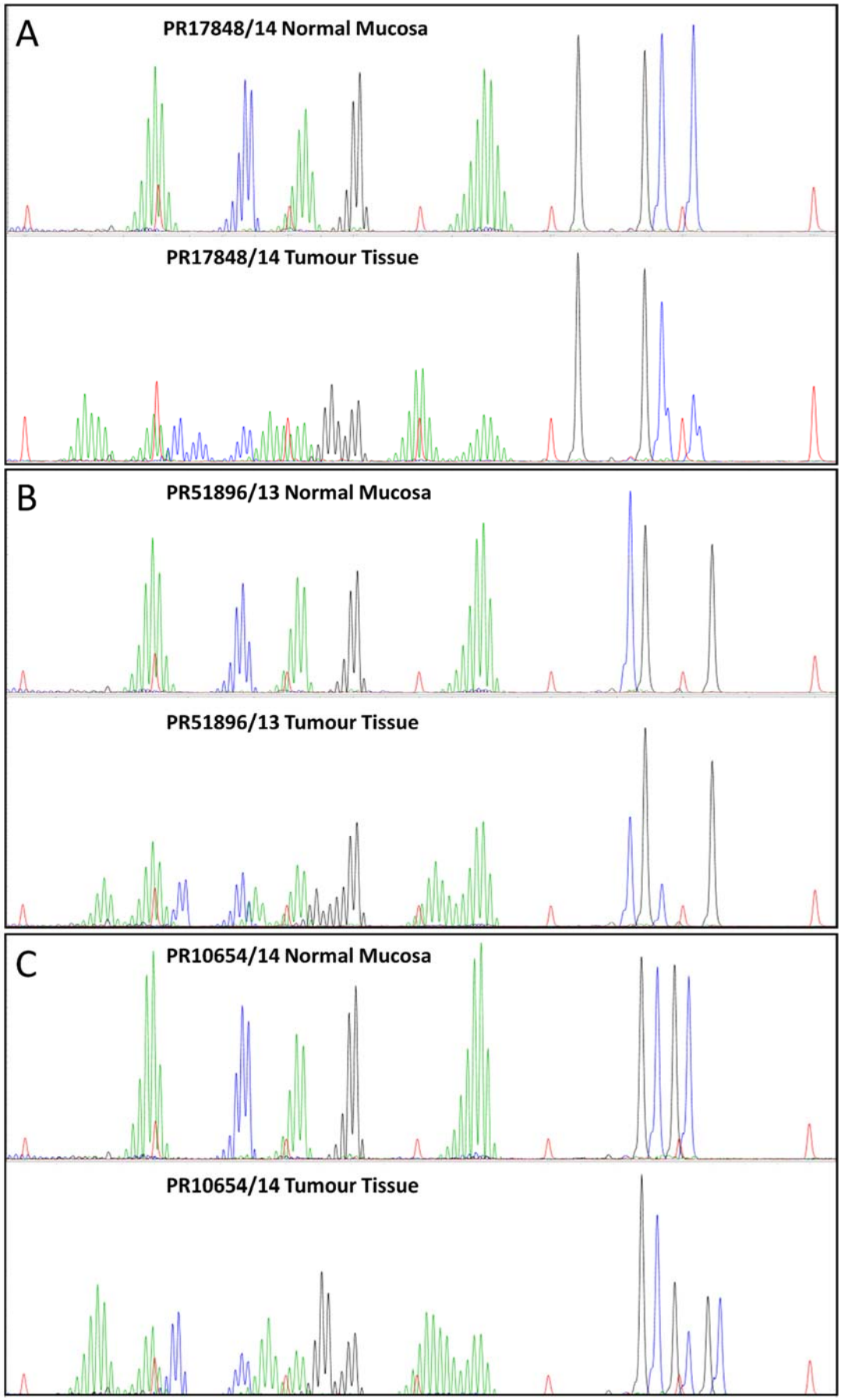

Figure 6.1: Fragment analysis traces for the three MSI-H tumours. Panel A tumour PR17848/14, Panel B tumour PR51896/13, Panel C tumour PR10654/14.

6.2.3. Three MSI-H tumours PR10654/14, PR17848/14 and PR51869/13

For each of the three MSI-H tumours 8 biopsies were taken using the clock face as a reference point. These 8 biopsies consisted of four scalpel biopsies sampling the surface of the tumour at each quadrant of the clock face, and 4 fine needle aspiration biopsies from deeper within the tumours at each quadrant of the clock face.

The tumour PR17848/14 was a large carcinoma. This tumour had grown across the antimesenteric border and it was therefore easier to orientate biopsies with the proximal side of the tumour as 12 o'clock. Normal mucosa was sampled from 10cm away from the tumour. A photo of the tumour prior to processing can be found in Figure 6.2.

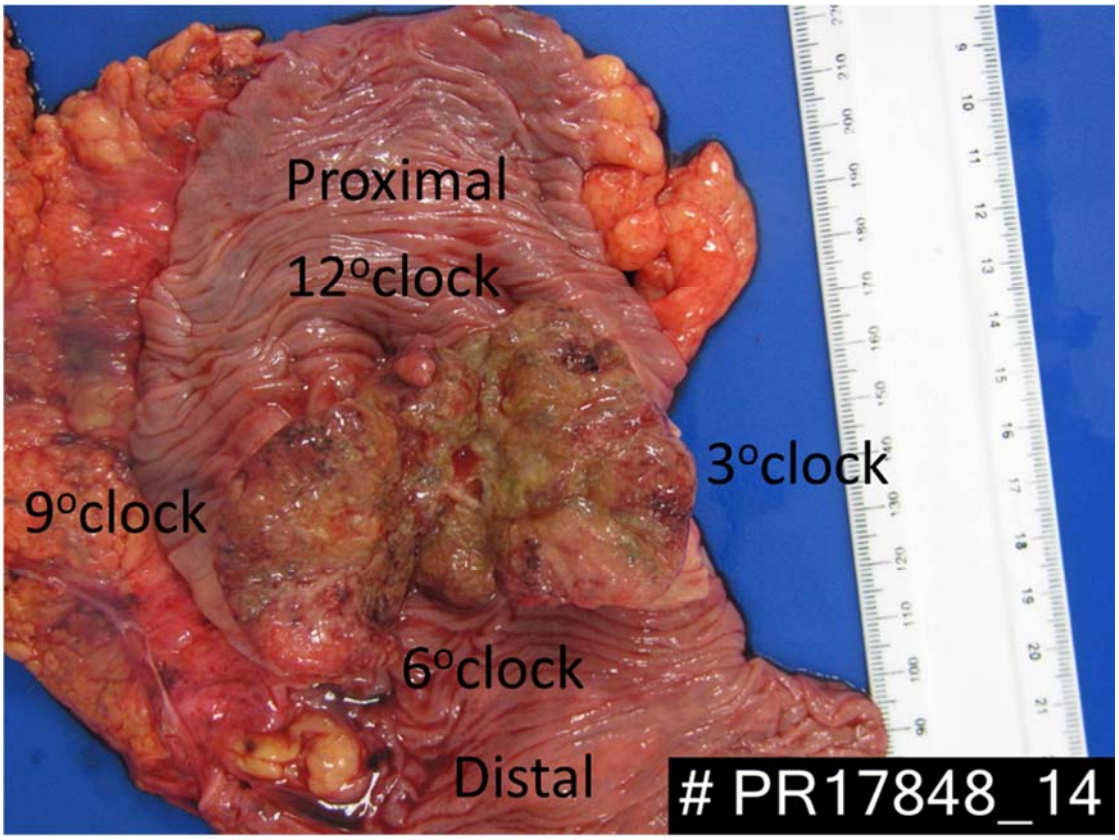

Figure 6.2: The tumour PR17848/14. Biopsies were taken from this tumour using a clock face as reference. The proximal side of the tumour was designated as 12 o'clock.

The tumour PR51896/13 was located close to the ileocaecal valve. This tumour was orientated with the proximal side of the tumour as 12 o'clock for the purpose of obtaining biopsies. Normal mucosa was biopsied 10cm from the tumour. For a photo of the tumour PR51896/13 see Figure 6.3. This tumour had already been processed in formalin fixative prior to being photographed, but all biopsies were taken before the processing of the tumour began.

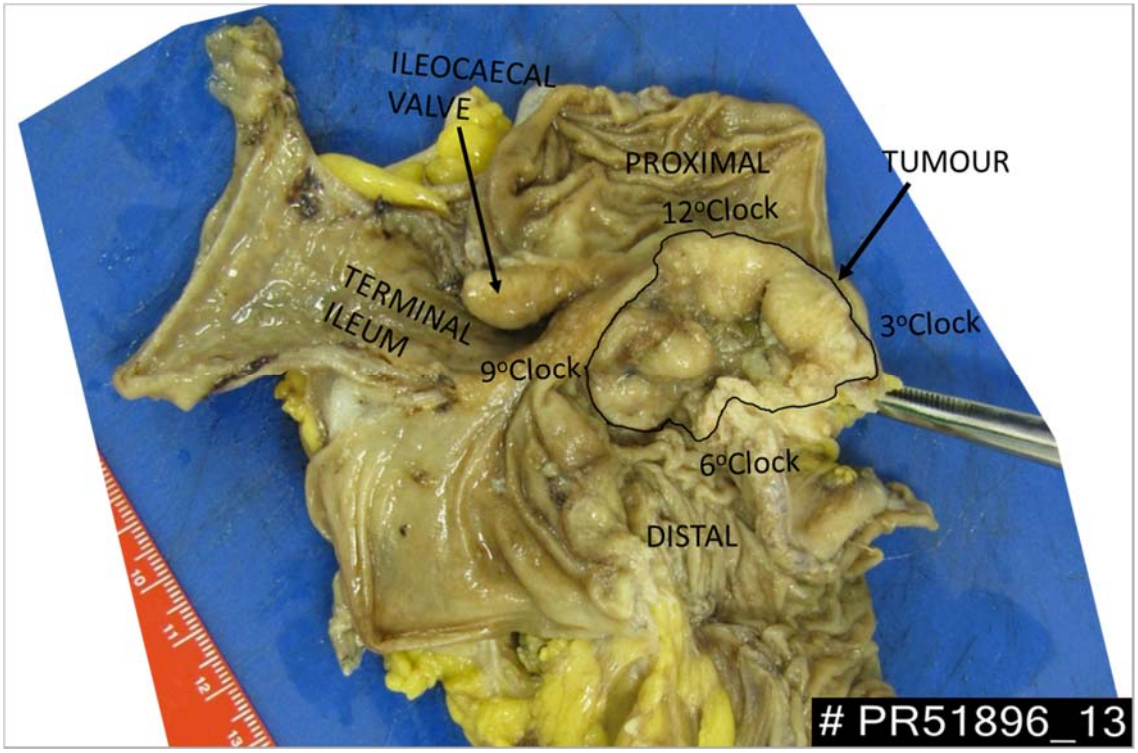

Figure 6.3: The tumour PR51895/13. Biopsies were taken from this tumour using a clock face as reference. The proximal side of the tumour was designated as 12 o'clock.

The tumour PR10654/14 was located where the terminal ileum meets the caecum. The tumour was removed via a limited right hemicolectomy. This tumour was biopsied using the side of the tumour closest to the antimesenteric border as 12 o'clock. Normal mucosa was sampled 7.5cm from the tumour. The tumour PR10654/14 was a necrotic tumour, and there was a risk that the 9 o'clock needle biopsy contained only necrotic tissue. For a photo of the tumour PR10654/14 see Figure 6.4.

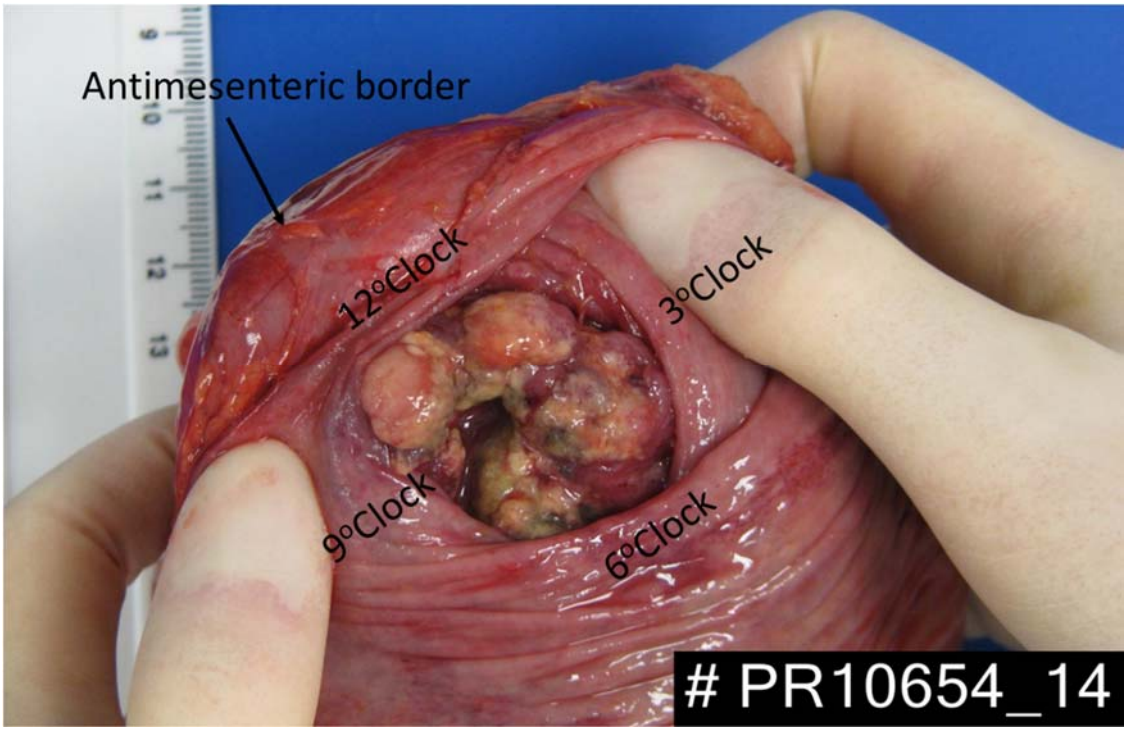

Figure 6.4: The tumour PR10654/14. Biopsies were taken from this tumour using a clock face as reference. The antimesenteric border was designated as 12 o'clock.

6.2.4. Mutation detection in multiple biopsies from MSI-H tumours

DNA concentrations were adjusted to ~10ng/µl for the production of amplicons. Amplicons were produced using Herculase II Fusion polymerase (Agilent, Santa Clara, CA, United States of America) and 28 PCR cycles. The reduction in the number of PCR cycles compared to that used previously was possible because the quality of DNA obtained from the fresh or frozen tumour tissue was better than DNA obtained from FFPE tissue. For each biopsy, 20 mononucleotide repeats were amplified (see Table 6.1 for a list of mononucleotide repeats). Amplicons were sequencing on an Illumina MiSeq using a MiSeq Reagent Kit v3 (600-cycles) (Illumina, San Diego, CA, United States of America) (see methods sections 2.3.2.2 and 2.7.2.2 for more details). The biopsies from tumours PR10654/14, PR17848/14 and PR51869/13 were sequenced on the same MiSeq run as the samples analysed in chapter 7 of this thesis.

Repeat Name	Repeat Size	Repeat Position	SNP1	SNP2	SNP3
LR46	8bp	chr20:10660084	rs143884078	rs182346625	rs6040079
LR24	9bp	chr1:153779428	rs192329538	rs1127091	
GM01	10bp	chr11:28894428	rs7951012		
GM22	10bp	chr14:43401009	rs58274313		
GM26	10bp	chr14:49584750	rs187027795	rs11628435	
GM29	10bp	chr3:70905559	rs2687195		
LR32	10bp	chr19 :37967219	rs7253091		
AVIL	10bp	chr12:58202497	rs2277326		
GM07	11bp	chr7:93085747	rs2283006		
GM14	11bp	chr3:177328817	rs6804861		
LR11	11bp	chr2:217217870	rs13011054	rs147392736	rs139675841
LR17	11bp	chr14:55603030	rs79618905	rs77482253	rs1009977
LR48	11bp	chr12:77988096	rs11105832		
ASTE1	11bp	chr3:130733047			
IM49	12bp	chr3:56682065	rs7642389		
LR36	12bp	chr4:98999722	rs182020262	rs17550217	
LR43	12bp	chr5:86199060	rs201282399	rs10051666	rs6881561
LR44	12bp	chr10:99898285	rs78876983	rs7905388	rs7905384
LR52	12bp	chr16:63861440	rs2434849		
FBXO46	14bp	chr19:46214701	rs34505186		

Table 6.1: A list of the 20 mononucleotide repeats sequenced for the multiple biopsies of tumours PR10654/14, PR17848/14 and PR51869/13. This list contains the designated repeat names, the length and location of each mononucleotide repeat, and the rs numbers of neighbouring SNPs.

Variant calling was performed using COPReC (see methods section 2.8.6.2). The percentage of reads corresponded to each variant repeat length and the percentage of reference reads was calculated to enable the analysis of different indel sizes in the multiple biopsies for tumours PR10654/14, PR17848/14 and PR51869/13. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat. The criteria of a minimum read depth was used to prevent a misrepresentation of variant frequencies caused by PCR duplicates. Table 6.2 contains the mean paired end read depth per mononucleotide repeat for all tumour biopsies. The 6 o'clock needle biopsy for tumour PR10654/14 was underrepresented among the reads generated for this MiSeq run.

	PR51896/13	PR17848/14	PR10654/14
Normal Mucosa	1135	1464	1707
3 o'clock Scalpel Biopsy	1344	1928	1713
6 o'clock Scalpel Biopsy	1100	2570	992
9 o'clock Scalpel Biopsy	1408	1591	1924
12 o'clock Scalpel Biopsy	1247	2383	1855
3 o'clock Needle Biopsy	4733	2347	2363
6 o'clock Needle Biopsy	986	1024	163
9 o'clock Needle Biopsy	1362	1343	1719
12 o'clock Needle Biopsy	1454	1447	1868

Table 6.2: The mean paired end read depth per mononucleotide repeat for the biopsies of tumours PR10654/14, PR17848/14 and PR51869/13.

To analyse the deletions on individual alleles the perl script AlleleicBias_IndividualIndels.pl was written. This script was used to identify repeats that are heterozygous for a neighbouring SNP and calculate the percentage of reads corresponding to each variant repeat length and reference repeat length for both alleles. (see methods section 2.8.7.5 for more details).

6.2.5. The clonal composition of tumour PR17848/14

For 13 out of the 20 homopolymers tested there was a higher deletion frequency within the tumour compared to the normal mucosa. A representative selection of 6 of these repeats can be found in Figure 6.5, the rest can be found in the appendix Figure 9.1. All the unstable repeats with neighbouring heterozygous SNPs showed allelic bias for the tumour 17848/14 with the deletion being present mainly on one allele (see Figure 6.5).

There is instability in all eight tumour biopsies for tumour 17848/14 (see Figure 6.5). There is a lower deletion frequency in the 6 o'clock and 9 o'clock scalpel biopsies compared to the other tumour biopsies. These two biopsies do however show a significantly higher 2bp deletion frequency compared to the normal mucosa for homopolymer LR11 allele 1 (see Figure 6.5 panels A). LR11 allele 1 has a 2bp deletion frequency of 6.3% and 11.4% in the 6 o'clock and 9 o'clock respectively compared to the 2bp deletion frequency of 0.29% in the normal mucosa (two-tailed Fisher's exact test p-values: <0.0000001). This suggests there is some instability within these two biopsies. The low levels of instability in the 6 o'clock and 9 o'clock scalpel biopsies may be due to contamination by normal tissue.

The 9 o'clock needle biopsy stands out as different from the other biopsies for the homopolymers ASTE1, GM14, LR17 and IM49. For the repeat ASTE1 there is a lack of reads with 1bp deletions compared to other biopsies, and reads with a 2bp deletion make up 89% of the reads (see Figure 6.5 panels C). This shows that the 9 o'clock needle biopsy has a high tumour cell content and that the 2bp deletion is likely to be biallelic for this biopsy. The lack of reads with a 1bp mutation suggest that tumour cells with this mutation are underrepresented in the 9 o'clock needle biopsy region. For the repeat GM14 the 9 o'clock biopsy is the only biopsy with a notably different 1bp deletion frequency to the normal control (see Figure 6.5 panels D). This suggests that this is a relatively new mutation which has developed in this location. The absence of a 1bp repeat in ASTE1

and an overrepresentation of a 1bp deletion in GM14 for the 9 o'clock needle biopsy suggests that the 1bp deletions in these repeats are located in different groups of cells. A difference in levels of contamination by normal tissue cannot account for the low 1bp deletion frequency in the other biopsies for GM14, because the level of instability in all of these biopsies, with the exception of the 6 o'clock and 9 o'clock scalpel biopsies, are high in other markers such as LR52. The repeat LR52 has a combined 2bp and 3bp deletion frequency of between 36% and 72% for these biopsies indicating that all of these biopsies have a tumour cell content of ~35% or higher (see Figure 6.5 panels G).

For the repeat LR17 allele 1 there is a notably different level of 2bp deletions in the 9 o'clock needle biopsy compared to the normal mucosa (see Figure 6.5 panels E). The level of 2bp deletions in the other biopsies could be explained by PCR error. The 2bp deletion could be present in the same group of cells with the 1bp deletion in GM14, because both of these deletions are only present in the 9 o'clock needle biopsy. Further evidence that there may be a difference between the 9 o'clock needle biopsy and other biopsies can be seen for the repeat IM49 allele 2. For there is a low frequency of 2bp deletions compared to most of the other tumour biopsies despite evidence which suggests this biopsy has a very high tumour cell content (see Figure 6.5 panel I). This suggests that there is a difference in the composition of tumour cells in this region compared to the rest of the tumour. Further evidence of this is the high 1bp deletion frequency in the 9 o'clock needle biopsy for IM49.

The 9 o'clock scalpel biopsy has a deletion distribution that is reminiscent of 9 o'clock needle biopsy for marker IM49, with no significant difference in 2bp deletion frequency compared to the normal mucosa for allele 2 (two-tailed Fisher's exact test p-value: 0.67) and a 1bp deletion frequency that is significantly higher than what is observed in the normal mucosa (two-tailed Fisher's exact test p-value: 0.0000098) (see Figure 6.5 panels I). This is the only instance where the mutation profile of the 9 o'clock scalpel biopsy differs from the other scalpel biopsies and the 3 and 6 o'clock needle biopsies.

The results discussed above indicate that the 9 o'clock needle biopsy contains a group of cells with a different mutation profile compared to what is present in the other tumour biopsies. This could indicate that there is a distinct sub-clone located in the 9 o'clock needle biopsy region of this tumour. There may also be a difference in the composition tumour cells in the 9 o'clock scalpel biopsy, which shows a lack of the 2bp deletion seen in other biopsies for the homopolymer IM49.

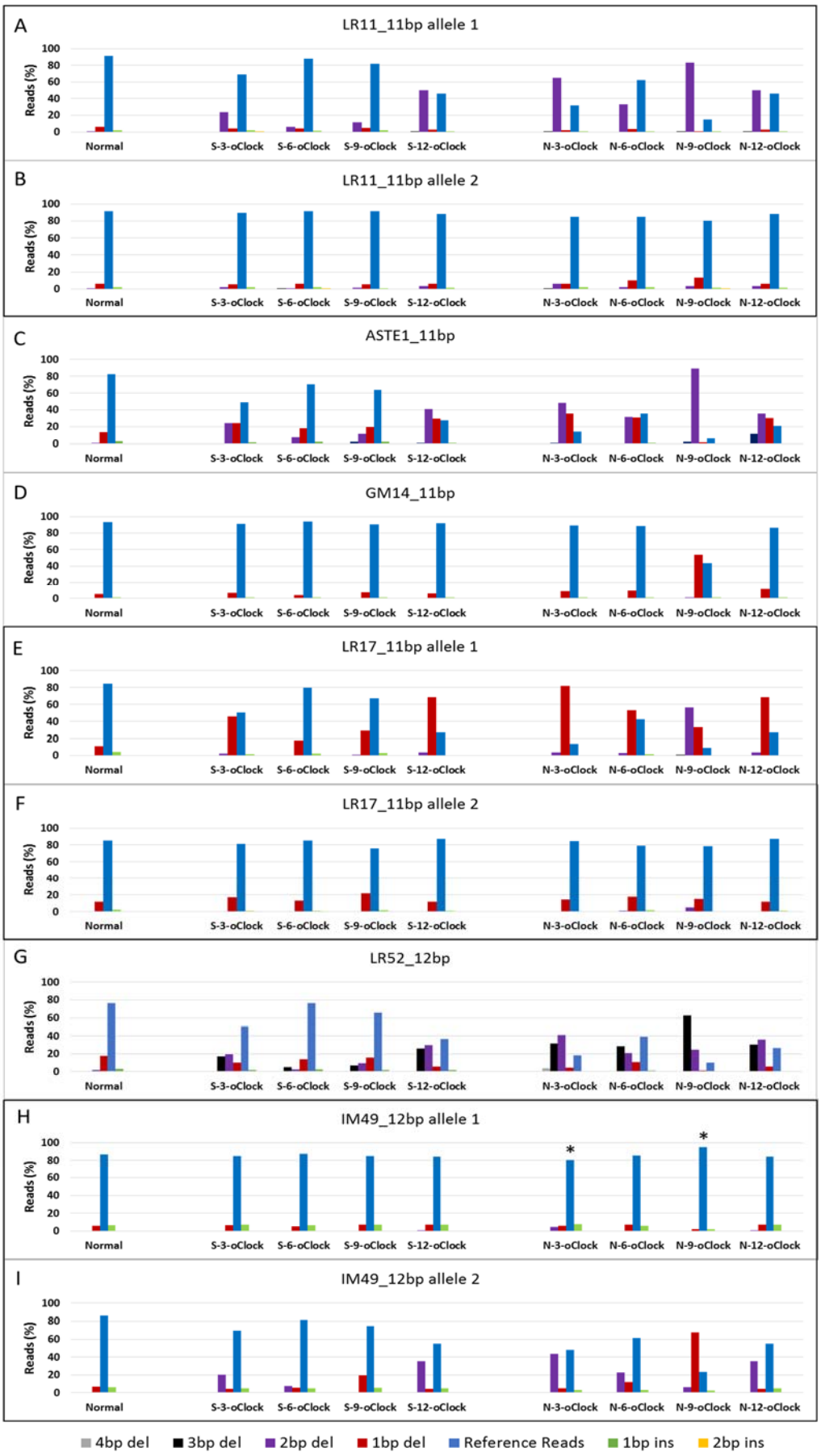

Figure 6.5: Frequencies of variant reads in 6 repeats showing instability for the tumour PR17848/14. Each panel shows the indel frequencies in 8 tumour biopsies and normal mucosa from the same patient. Tumour biopsies were taken from the four quadrants of the tumour according to the clock face. Normal = Normal Mucosa, S = Scalpel Biopsy (a biopsy from the surface of the tumour), N = Needle biopsy (a biopsy from deeper into the tumour tissue). *A total of ≥100 paired end reads for the marker, but less than 100 paired end reads for the allele.

6.2.6. The clonal composition of tumour PR51896/13

For the tumour PR51896/13 14 of the 20 homopolymers analysed showed signs of instability. A representative selection of 7 of these repeats can be found in Figure 6.6, the rest of the repeats can be found in the appendix Figure 9.2 and Figure 9.3. The 9 o'clock scalpel biopsy has the highest frequency of variant reads for most of the repeats sequenced (see Figure 6.6). The 6 o'clock scalpel biopsy on the other hand does not notably differ in variant read frequencies from what is present in the normal mucosa. This biopsy could therefore be classed as MSS. Whether this is a result of contamination by normal tissue or a result of sampling cells that belong to a clone that that arose before the knock out of MMR function is unknown.

For the repeat AVIL the 1bp and 2bp deletions present for this marker are found at a roughly equal ratio in all the tumour biopsies except the 6 o'clock scalpel biopsy which does not show any signs of instability (see Figure 6.6 panel B). This suggests that the 1bp and 2bp deletions are present in the same cells and these cells only became prolific after appearance of both variants.

For the repeat LR46 only the 3 and 9 o'clock scalpel biopsies have a notably higher 1bp deletion frequency compared to the normal mucosa (see Figure 6.6 panel A). Repeats such as ASTE1, LR52, and FBXO46 have deletion levels in the needle biopsies and 12 o'clock scalpel biopsy on an equivalent level to what in seen in the 3 o'clock scalpel biopsy for the same repeats (see Figure 6.6 panels C, G and H). This suggests that lack of a deletion frequency that differs from the normal mucosa in LR46 is not due to normal contamination in the needle biopsies and 12 o'clock scalpel biopsy, but an absence of that mutation in the tumour cells in these biopsies. This could indicate that the 1bp mutation in LR46 occurred late in tumour development and that the 3 and 9 o'clock scalpel biopsies share a clonally distinct population of cells containing this 1bp deletion.

Another difference between biopsies can be seen for GM14 allele 2 were there is a 2bp deletion frequency of 19% in the 12 o'clock needle biopsy (see Figure 6.6 panel E). Many of the other biopsies have a low level of 2bp deletions on allele 1, but the 12 o'clock needle biopsy is the only biopsy with a 2bp deletion frequency that notably differs from the normal mucosa on allele 2. This could suggest that there is a clonally distinct population of cell located in the 12 o'clock needle biopsy region of the tumour.

The repeat FBXO46 also gives evidence to support the hypothesis that there are different sub-clonal populations of tumour cells in tumour PR51896 (see Figure 6.6 panel H). The 12 o'clock scalpel biopsy contains a 4bp deletion which is absent in the 3 o'clock scalpel biopsy and the 3 o'clock scalpel biopsy contains a 5bp deletion which is absent in the 12 o'clock scalpel biopsy. This indicates that there is a population of cells present in the 3 o'clock scalpel biopsy region of the tumour that is not present in the 12 o'clock scalpel biopsy region of the tumour and vice versa.

There may also be evidence of different populations of tumour cells spread throughout the tumour. For the repeat ASTE1 there is a 2bp deletions frequency above 10% in all biopsies except the 12 and 6 o'clock scalpel biopsies (see Figure 6.6 panel C). The 12 o'clock scalpel biopsy has a 1bp deletion frequency of 44.4%. A 2bp deletion of only 4.2% in this biopsy may therefore indicate that the 1 and 2bp deletion are present in different groups of cells. The 2bp deletion in ASTE1 is also likely to be present in a different group of cells than the ones that contain 1bp deletions in the repeats LR32 and GM14 allele2 because these deletions are present at a high frequency in the 12 o'clock scalpel biopsy where the 2bp deletion in ASTE1 is present at a low frequency.

The results above suggests that the tumour PR51896/13 is composed of different sub-clones with one distinct group cells, characterised by mutations in LR46 which have been enriched in the 3 and 9 o'clock scalpel biopsy region of the tumour, while a different population, characterised by 2bp mutations in GM14 allele 2 is enriched in the 12 o'clock needle biopsy region of the tumour.

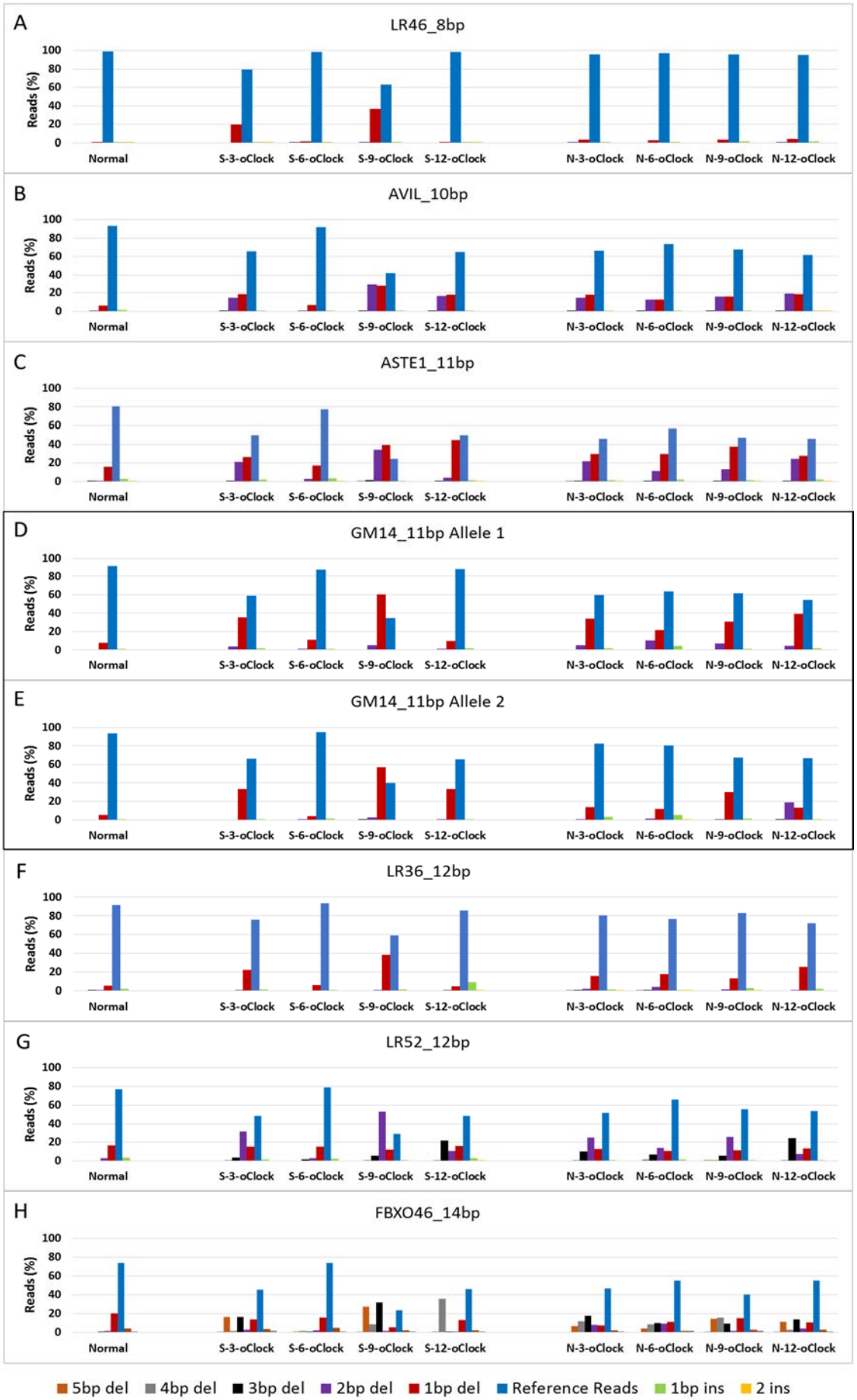

Figure 6.6: Frequencies of variant reads in 7 repeats showing instability for the tumour PR51896/13. Each panel shows the indel frequencies in 8 tumour biopsies and normal mucosa from the same patient. Tumour biopsies were taken from the four quadrants of the tumour with positioning according to the clock face. Normal = Normal Mucosa, S = Scalpel Biopsy (a biopsy from the surface of the tumour), N = Needle biopsy (a biopsy from deeper into the tumour tissue).

6.2.7. The clonal composition of tumour PR10654/14

The tumour PR10654/14 had sixteen out of twenty mononucleotide repeats that showed signs of instability for at least one tumour biopsy. A representative selection of 7 of these repeats can be found in Figure 6.7, the rest of the repeats can be found in the appendix Figure 9.4 and Figure 9.5. As mentioned before the 6 o'clock needle biopsy for tumour PR10654/14 was sequenced to an average read depth of 163 paired end reads per amplicon. As a result, many amplicons had a read depth below 100 paired end reads and were not analysed. The criteria of a minimum of 100 paired end reads was used to prevent a misrepresentation of variant frequencies caused by PCR duplicates. Other samples that were not analysed because of having less than 100 paired end reads were the 12 o'clock scalpel and needle biopsies for the repeat IM49.

For this tumour there is also limited instability seen in the 6 o'clock scalpel biopsy. For the repeats ASTE1, LR17 and FBXO46 there might be instability in the 6 o'clock scalpel biopsy. The repeat ASTE1 has a 2bp deletion frequency that is over 7 times larger in the 6 o'clock scalpel biopsy than in the normal mucosa (see Figure 6.7 panel D). The 6 o'clock scalpel biopsy has a higher frequency of 2bp deletions than the normal mucosa for the repeat LR17 with a 2bp deletion frequency of 3.2% and 0.3% respectively (see Figure 6.7 panel E). The 6 o'clock scalpel biopsy also has 4bp deletion present at a frequency of 5.8% in the repeat FBXO46 (see Figure 6.7 panel I). Because there are no reads in the normal mucosa containing a 4bp deletion for FBXO46 it is highly likely that the 4bp deletion in the 6 o'clock scalpel biopsy is caused by MSI. The results from ASTE1, LR17 and FBXO46 could indicate that there is some instability present in the 6 o'clock scalpel biopsy. It is possible that the low mutation frequencies in this biopsy are caused by contamination from normal tissue.

The 3 o'clock scalpel biopsy differs from many of the other biopsies for repeats LR32, GM07, GM14 and LR52 (see Figure 6.7). For the repeat LR52 the 3 o'clock scalpel biopsy is the only biopsy with a 3bp deletion frequency that notably differs from the normal mucosa, and for the repeat GM14 the 3 o'clock scalpel biopsy is the only biopsy that has a 2bp deletion frequency which notably differs from the normal mucosa on allele 2. This suggests that there is a clonally distinct population of cells, which is highly overrepresented in the 3 o'clock scalpel biopsy region of the tumour.

For the repeat LR32 the 3 o'clock scalpel biopsy and 6 o'clock needle biopsy are the only biopsies with a significantly higher 1bp deletion frequency than the normal mucosa (two-tailed Fisher's exact test p-values: <0.0000001). This is the only repeat where the 3 o'clock scalpel biopsy and 6 o'clock needle biopsy share a mutation which is not present in the other tumour biopsies (see Figure 6.7 panel C). This could suggest that that the 1bp deletion in the 3 o'clock scalpel biopsy and 6 o'clock needle biopsy have arisen separately as opposed to being present in the same population of cells, or that they are present in the same population of cells but these biopsies also contain additional populations of cells that are also present in other biopsies.

For the repeat GM07 allele 2, the 3 o'clock scalpel biopsy and 6 o'clock scalpel biopsy are the only biopsies with a significantly higher 1bp deletion frequency compared to the normal mucosa (two-tailed Fisher's exact test p-values: 0.0012 and <0.0000001 respectively). This could suggest that the cell population that contains the 1bp deletion in the 3 o'clock scalpel biopsy also exist in the neighbouring 6 o'clock scalpel biopsy. It is, however, difficult to analyse the similarities between the 6 o'clock scalpel biopsy and other biopsies because of the low level of instability seen in the 6 o'clock scalpel biopsy.

For the homopolymer GM07 at least 4 different replication mistakes in different groups of cells must have occurred to make the deletion distribution seen in this tumour with both 1bp and 2bp deletions present on both alleles in different biopsies (see Figure 6.7 panels D and E). This repeat has different mutation profiles in different regions of the tumour, which could represent the emergence of different tumour sub-clones. As mentioned before, the 3 o'clock scalpel biopsy and 6 o'clock scalpel biopsy may contain cells from a clonally distinct cell population. Equally, the 9 o'clock biopsies share a distinct mutation pattern with a 2bp deletion frequency above 40% on both alleles. This may indicate a clonally distinct population of cells in the 9 o'clock region. The 12 o'clock biopsies and 3 o'clock needle biopsy share the same mutation profile on both alleles for marker GM07, which may indicate a common clonal decent for cells in these regions.

The repeat FBXO46 appears to be polymorphic with both reference reads and reads containing a 1bp deletion being present at a frequency of 33.5% and 52% respectively in the normal mucosa (see Figure 6.7 panel I).

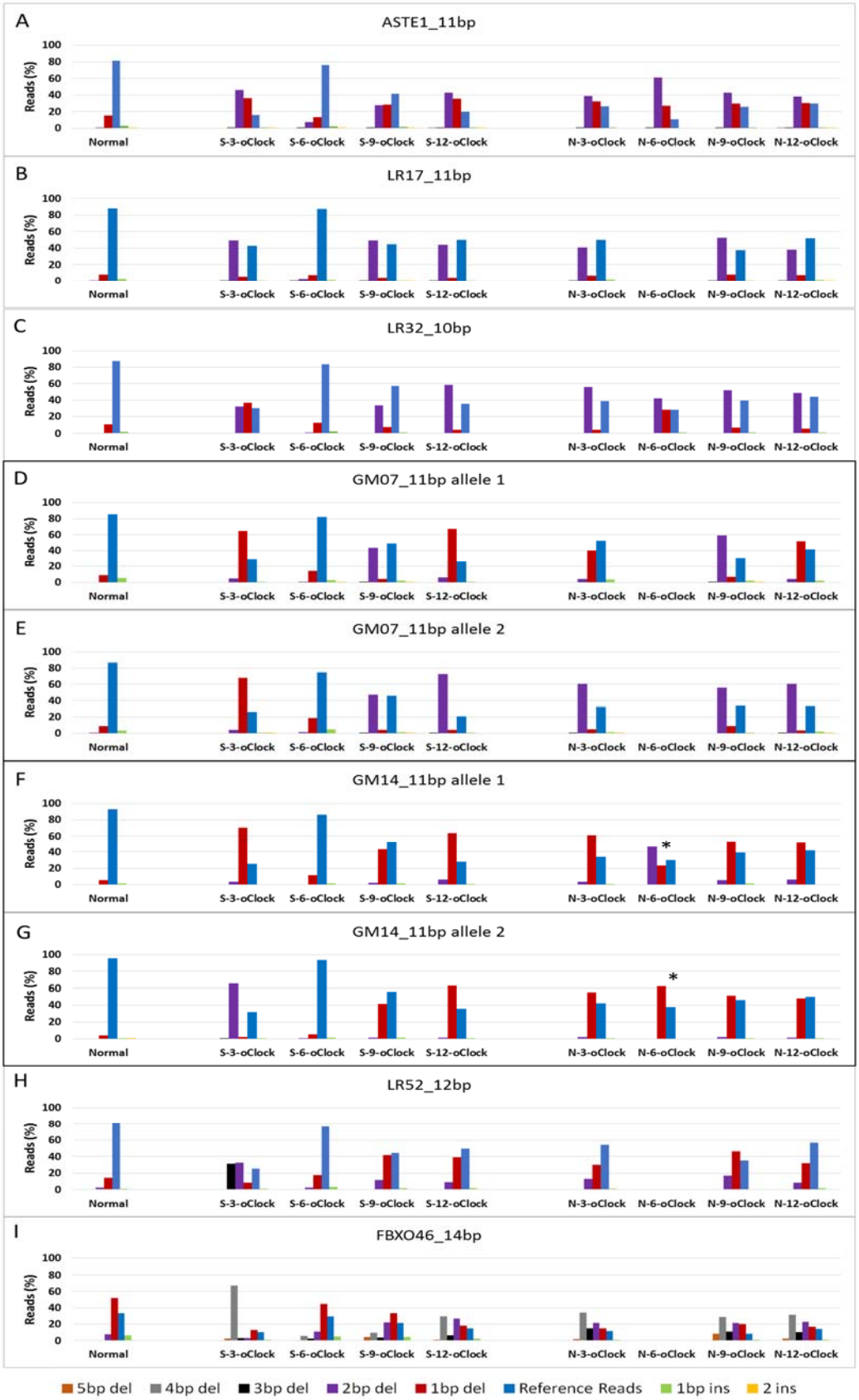

Figure 6.7: Frequencies of variant reads in 7 repeats showing instability for the tumour PR10654/14. Each panel shows the indel frequencies in 8 tumour biopsies and normal mucosa from the same patient. Tumour biopsies were taken from the four quadrants of the tumour according to the clock face. Normal = Normal Mucosa, S = Scalpel Biopsy (a biopsy from the surface of the tumour), N = Needle biopsy (a biopsy from deeper into the tumour tissue). * A total of ≥100 paired end reads for the marker, but less than 100 paired end reads per allele.

Based on the results discussed previously in this section I would conclude that the tumour PR10654/14 could broadly be divided up into 4 distinct regions of clonally different cells (see Figure 6.8). The 6 o'clock scalpel biopsy, which shows limited instability in this tumour, has been designated region 1. The 3 o'clock scalpel biopsy, which has a different deletion pattern to the other tumour biopsies for markers LR32, GM07, GM14 and LR52, has been designated region 2. The 12 o'clock biopsies and the 3 o'clock needle biopsy have the same deletion compositions on each allele for the repeats shown in Figure 6.7. For this reason, the 12 o'clock biopsies and the 3 o'clock needle biopsy have been designated region 3. The two 9 o'clock biopsies have assigned region 4 because these two biopsies were the only biopsies with a 2bp deletion on each allele for the repeat GM07 (see Figure 6.7). Otherwise the two 9 o'clock biopsies are very similar to the biopsies from region 3. Because of the low read depth obtained for the 6 o'clock needle biopsy there were many markers where this biopsy was not analysed. For the markers that were analysed the 6 o'clock needle biopsy, it was similar to the 3 o'clock scalpel biopsy for marker LR32, and was otherwise similar to the biopsies in region 3. I would therefore conclude that the 6 o'clock scalpel biopsy is likely to closely related regions 2 and 3.

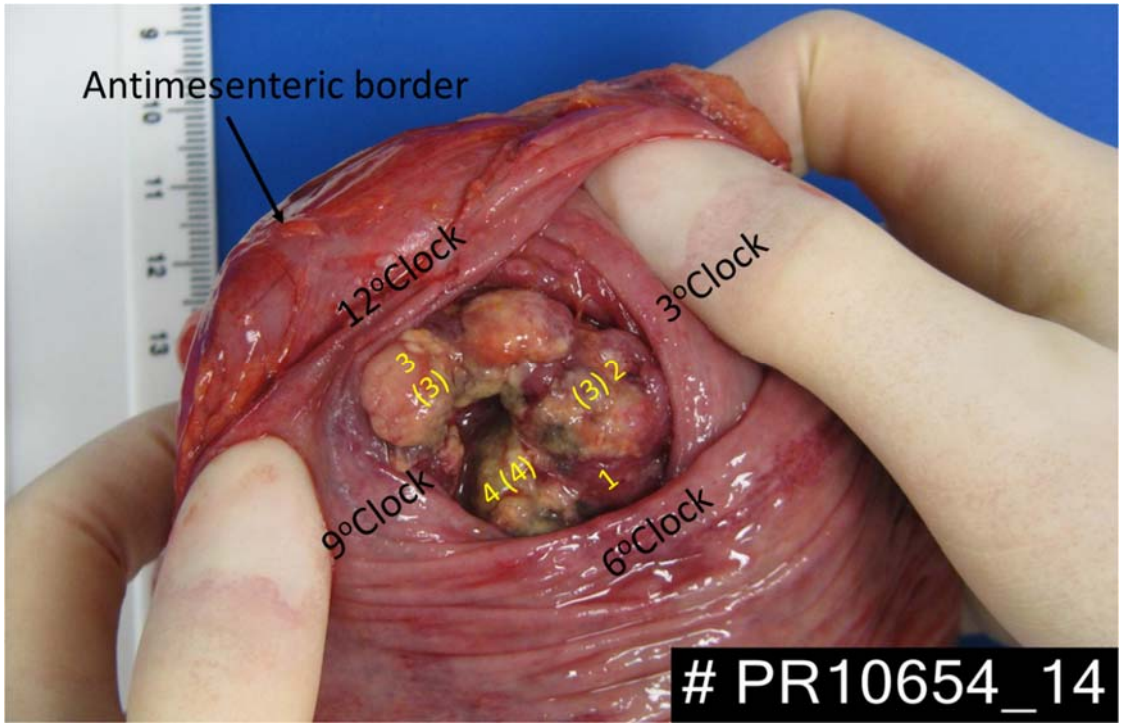

Figure 6.8: Four possible clonal regions for tumour PR10654/14 highlighted in yellow. The number without brackets represent scalpel biopsies and the numbers in brackets represent needle biopsies. The scalpel biopsies were taken from the tumour surface while the needle biopsies were used to sample tissue deeper within the tumour.

6.3. Discussion

The fresh tissue yielded high quantities of DNA even for the fine needle aspiration biopsies. For the fresh tissue there were no problems creating amplicons of 300bp, which has been a challenge in the FFPE tissue used in previous chapters. Creating amplicons from FFPE tissue is a known problem, which arises because preserving tissue in this way causes the fragmentation of DNA. Using DNA obtained from FFPE tissue can therefore be problematic for many diagnostics tests including the current fragment analysis test which is used to diagnose MSI. For this work using DNA from fresh or frozen tissue meant that the number of PCR cycles could be reduced from 35 to 28. A reduction in PCR cycles should give a reduction in PCR error making it easier to distinguish between real mutations and background noise. Using fresh or fresh frozen tissue for the diagnosis of MSI would therefore be an advantage, whether the test is a future sequencing based MSI test or a current fragment analysis test. If standard practice could be changed so that a tissue sample for MSI testing was obtained prior to formalin fixation then MSI diagnosis would be easier with a lower failure rate. There was very low read depth obtained for the PR10654/14 6 o'clock needle biopsy but this is believed to have occurred during the Nextera XT library prep, because all amplicons for this sample were fine when quantified on the QIAxcel prior to starting the Nextera XT library prep.

The needle biopsies consistently showed a high frequency of variant reads for the markers that were unstable in each tumour. This suggests that needle biopsies may be a good method for sampling tumours for MSI analysis. It should however be noted that differences in marker instability across a tumour means that a needle biopsy does not give the complete picture of the MSI across the whole tumour. On the other hand a single needle could be inserted into several regions of the tumour being sampled. Such a technique using needle biopsies could be used as a method for sampling tumour tissue for new point of care devices such as the Q-POC being developed by the company QuantuMDx which is a possible future platform for a sequencing based MSI test.

Short 8bp-14bp microsatellites have made a good tool for identifying different sub-clones using instability in the three MSI-H tumours PR10654/14, PR17848/14 and PR51869/13. Differences between biopsies were observed in all three tumours. Differences between biopsies included; instability in a different number of repeats and different deletion sizes observed in different biopsies from the same tumour. The results

presented here indicated that there is a distinct sub-clone located in the 9 o'clock needle biopsy region of tumour PR17848/14. Results for tumour PR51869/13 suggested that there might be a clonally distinct group of cells in the 3 and 9 o'clock scalpel biopsy region of the tumour and a different clonally distinct group of cells present in the 12 o'clock needle biopsy region of the tumour. The tumour PR10654/14 had a greater number of differences between biopsies than the other two tumours. The results for PR10654/14 suggest that this tumour can be divided up into at least 4 regions with distinct sub-clones. The results for these three tumours suggest that as the tumours have developed they have continued to accumulate mutations. Using short homopolymers to identify distinct clonal regions of MSI-H tumours may be useful as it could, for example, enable metastasis evolution to be tracked.

For the tumour PR51896/13 the 6 o'clock scalpel biopsy showed no sign of instability in any of the 20 repeats tested. This is despite other biopsies from this tumour showing instability in 14 out of the 20 markers tested. There are two possible explanations for why this biopsy exhibits no microsatellite instability. One possible reason is that the 6 o'clock scalpel biopsy contained tumour cells belonging to a clone that arose early in tumour development prior to the knock out of mismatch repair function. Evidence suggests that sporadic MSI-H tumours most likely develop from MSS adenomas (Fearon, 2011, Loukola et al., 1999) so this is a possibility. Information on whether the tumour PR51896/13 is a sporadic MSI-H tumour or not is not available, but the majority of MSI-H tumours occur in patients without a germline mutation so there is a high likelihood that this tumour is a sporadic MSI-H tumour. The other possibility is that the 6 o'clock scalpel biopsy from tumour PR51896/13 has been contaminated by normal tissue. It is peculiar that the 6 o'clock scalpel biopsies from all three tumours contained a lower level of instability compared to the other biopsies. This could suggest this lower level of MSI is something to do with the sampling technique leading to a contamination of normal tissue in the form of blood or normal mucosa.

Using repeats with neighbouring SNPs with a high minor allele frequency proved useful because it provided extra information about the number of different mutations that have occurred in a repeat throughout the tumour. For example in the tumour PR10654/14, being able to see which allele different deletions belonged helped with the analysis of the tumour (see Figure 6.7). For the repeats with neighbouring heterozygous SNPs (GM14 and GM07) there were deletions of the same size on both alleles with frequencies that are unlikely to be caused by PCR and sequencing error. This indicates that there have been

separate replication mistakes on both alleles that have not been rectified by the compromised mismatch repair system in the tumour PR10654/14. It would not have been possible to know this without using the heterozygous SNPs. For example, it was only clear that there had been at least 4 different deletion events in the mononucleotide repeat GM07 for tumour PR10654/14 after analysing the two alleles individually. The extra information provided by being able to distinguish between the two alleles for this tumour did not just show that there were more deletions than expected in these repeats, but also helped identify different sub-clonal regions within the tumour PR10654/14 (see Figure 6.8).

Using repeats with neighbouring heterozygous SNPs also has the potential to help study other aspects of the clonal development of MSI-H tumours. In the study of familial adenomatous polyposis (FAP) a patient with sex chromosome mixoploid mosaicism (XO/XY) helped reveal that some FAP tumour are not of monoclonal origin. For 3 out of the 55 adenomas analysed by Thirlwell et al. (2010) different groups of tumour cells within the same adenoma had the XO and XY genotypes. This showed that the tumour could not have originated from just one cell. Using the same principle, heterozygous SNPs could be used to study the origin of tumours in mosaic patients to determine if the tumours are of monoclonal or polyclonal origin. Identifying SNPs where a minor allele is only present in one of the groups of cells in a mosaic patient could be a powerful tool for analysing the clonal origin of tumours in mosaic patients. Especially if several SNPs are used, some with a minor allele present in one tissue and others with a minor allele present in the other tissue. Although the need for mosaic patients would limit the usefulness of this technique to only a few individuals it would be a more powerful tool than using X chromosome inactivation to determine if tumours are of a monoclonal origin. X chromosome inactivation has been widely used for investigating the clonal origin of tumours (Leedham and Wright, 2008). In early embryonic development one of the X chromosomes in females is inactivated. Which X chromosome is inactivated will differ in different cells. Analysing many tumours showing activation of only one X chromosome have been used as evidence of monoclonal origin in tumours (Fearon et al., 1987). However the discovery of a tumour showing activation of genes on both X chromosomes would not necessary indicate that a tumour was of a polyclonal origin. This is because the reactivation of genes on the inactive X chromosome does occur in tumours (Chaligne et al., 2015). Using heterozygous SNPs in mosaic patients would therefore be better for proving the polyclonal origin of tumours.

6.3.1. Conclusions

In conclusion, there was evidence to suggest that the three MSI-H tumours where multiple biopsies were analysed were all heterogeneous tumours composed of different sub-clones. The use of repeats with neighbouring heterozygous SNPs to identify the allelic origin of deletions also facilitated a more in-depth analysis of the clonal evolution of the three MSI-H tumours.

Chapter 7. MSI test validation and investigation of QuantuMDx's Q-POC platform

7.1. Introduction and aims

7.1.1. Introduction

7.1.1.1. Current MSI testing platform

Recently, it has been reported that current clinical criteria and management guidelines used to identify colorectal cancer (CRC) patients for MSI testing (Amsterdam II criteria and revised Bethesda Guidelines) fail to identify a significant number of Lynch Syndrome patients (Canard et al., 2012, Mills et al., 2014, Perez-Carbonell et al., 2012). This has led to suggestions that all CRC tumours should undergo molecular testing (Vasen et al., 2013, Canard et al., 2012, Mills et al., 2014, Julie et al., 2008). Screening all colorectal cancers for MSI then Lynch Syndrome could be used to detect many of the patients and families with Lynch Syndrome that currently go undetected. Also, to enable future targeted treatment for both sporadic and germline MSI-H CRCs, MSI testing practices should change so that all CRCs are tested for MSI. Because MSI tests are expensive it would, with the current methods, be very expensive to test all CRCs in order to identify the MSI-H cancers.

A Sequence based MSI typing using short mononucleotide repeats could be advantageous in terms of cost and ease of interpretation through automation. This could further lower the cost of an MSI test such that it is more cost effective to test all colorectal cancers for MSI and reduce the time it takes to receive a test result. A sequencing based MSI test could be introduced as a next generation sequencing assay on a platform such the Illumina sequencers or it could be produced even more cheaply on a platform like the one currently being developed by the company QuantuMDx.

7.1.1.2. QuantuMDx's silicon nanowire platform

The company QuantuMDx are developing a cheap and fast DNA testing point of care (Q-POC) device. This device may ultimately allow the rapid diagnosis of many

diseases including MSI testing for colorectal cancer. One of the main aims of this project has been to create an MSI assay that is compatible with the technology being developed by QuantuMDx. As mentioned before, the markers used in current fragment analysis are too long for sequencing because polymerases cannot replicate long homopolymers faithfully. Therefore shorter repeats are needed for a QuantuMDx Q-POC assay. As part of my PhD project I have developed a panel of short homopolymers for MSI detection and also worked on the initial development of some of the components of QuantuMDx's Q-POC device.

QuantuMDx's device consists of four main components. The first component will be a tissue lysis chamber. This may comprise either of a mechanical lysis device or involve chemical lysis using a proteinase k based method. The tissue used can either be FFPE tissue or fresh tissue. QuantuMDx plan to use a needle biopsy style approach to sample fresh tumour tissue. This means that the amount of tissue used will be small and sheared from passing through the needle minimising the time needed to lyse the tissue. This is important for a rapid point of care device. Results in chapter 6 showed that using a needle biopsy to identify microsatellite instability using fresh tumour tissue works well.

The second component is a DNA extraction cassette; this contains a sorbent filter (Q-FILTER™) for the adsorption and removal of cellular components such as proteins, lipids and low molecular weight compounds while DNA is not absorbed. This means that lysed sample and buffer solution can be passed into the filter, and the buffer solution containing purified DNA will pass through the filter ready for PCR amplification while other cellular material is retained.

The third component of QuantuMDx's technology is their micro fluidic PCR cassette (Q-AMP™) which relies on the PCR mixture flowing through different temperature zones to achieve PCR amplification. For the PCR reaction itself there is the possibility of using either a two-step or a three-step continuous flow PCR. A three-step PCR has three heating zones; a denaturation zone, an annealing zone, and an extension zone. A two-step PCR on the other hand only has the denaturation zone and the annealing zone. For the two-step PCR, amplicon extension happens in the brief temperature transition between the annealing zone and denaturation zone. This can be achieved because Taq polymerase can synthesis a new DNA strand at a rate of 60-100 nucleotides per second (Kim et al., 2006). Continuous flow PCR has been shown to produce detectable amounts of PCR product in just 8-30 minutes (Kim et al., 2006). This type of

PCR therefore allows for the rapid production of PCR amplicons, reducing the overall time from sample to detection.

The last component is a silicon nanowire based detection device. The nanowires will be printed with amine terminated DNA or peptide nucleic acid (PNA) probes for each region of interest. The amine group allows the DNA/PNA probes to be attached to the nanowires through a reaction with aldehyde groups on the nanowire surface. DNA features with widths of 100nm can be printed using microarray printing technologies such as Dip-Pen Nanolithography (Demers et al., 2002), allowing different nanowires on the same chip to be printed with different probes. These probes will capture the PCR product, produced by the PCR cassette, and function as primers for a sequencing by synthesis reaction that incorporates negatively charged nucleotides. Silicon nanowires are highly sensitive to the binding of charged molecules and have the advantage of a linear change in conductance with the concentration of charged molecules over a large dynamic range (Cui et al., 2001). This could potentially be used to detect what fraction of DNA molecules contain the base of a mononucleotide repeat and what fraction of molecules contain the base after the repeat enabling the fraction of reads containing an indel as well as indel size to be determined.

QuantuMDx are planning to use proprietary negatively charged nucleotides in their sequencing reaction. These nucleotides comprise of a negatively charged reporter group attached to a dNTP by a cleavable linker (see Figure 7.1). The modified dNTP work as a substrate for polymerases, allowing the negatively charged dNTPs to be added to the nascent strand during a sequencing by synthesis reaction. Upon base incorporation, the negatively charged reported group will create a change in conductance of a nanowire allowing the detection of successful incorporated bases. Using silicon nanowires it is possible to detect DNA hybridising to probes in real time (Gao et al., 2007). It should, therefore, also be possible to detect the incorporation of bases with negatively charged reporter groups in real time. The modified dNTPs also have reversible blocking groups which will ensure that only one base is incorporated at a time in the sequencing reaction. The reversible blocking group is cleaved before the next base can be incorporated. This should allow more accurate sequencing of mononucleotide repeats than is achieved using sequencing technologies such as 454 sequencing or IonTorrent where chain termination is not used and the number of bases in a mononucleotide repeat is inferred by signal intensity (Stranneheim and Lundeberg, 2012, Shendure and Ji, 2008).

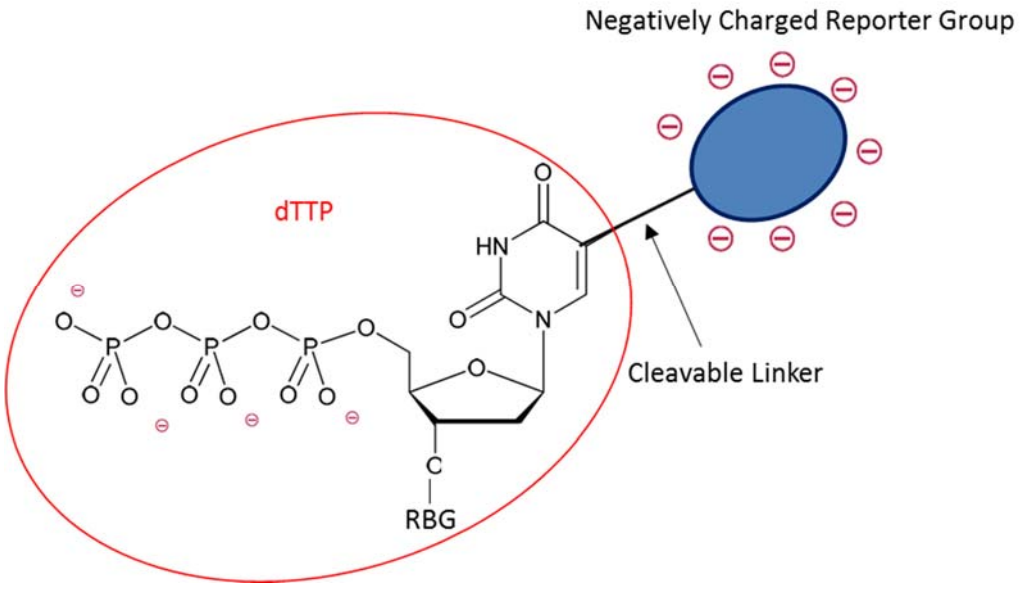

Figure 7.1: A diagram showing an example of one of QuantuMDx's negatively charged bases. Each base will have a reversible blocking group allowing the incorporation of one base at a time, and a negatively charged reporter group which will be detected by the nanowires of QuantuMDx's detector. RBG = reversible blocking group

7.1.2. Aims

The initial screen of 120 homopolymers with neighbouring SNPs, identified from whole genome data, showed a high level of instability in the five MSI tumours sequenced with 40% of the short 7bp-9bp A/T repeats, 80% of the longer 10bp-12bp A/T repeats and 33% of the G/C repeats showing instability in at least one tumour (chapter 5). Markers were arbitrarily defined as unstable if a marker had a deletion frequency >5% and a deletion frequency of at least twice that of any of the control samples for the 7-9bp repeats, or 1.5x that of any of the control samples for the 10-12bp repeats. Using heterozygous SNPs located within 30bp of the repeats I was also able to show that there was an excess of repeats showing allelic bias of reads with deletions in the MSI-H samples. 10 markers from whole genome analysis, which were classed as unstable in at least 60% of the MSI-H samples and also had an AUC of at least 0.9 were chosen for further investigation (chapter 5). The 10 markers taken from the literature, which showed instability in chapter 3 were also selected for further analysis. To further refine the selected panel of 20 repeats it would be advantageous to look at the sensitivity and specificity of the markers at different deletion frequencies to define thresholds for calling instability. To obtain enough data to define reasonable thresholds for each marker, the markers needed to be sequenced in a larger panel of tumours than analysed to date. This

chapter will also outline my contributions towards the hardware development of the QuantuMDx Q-POC device which, when complete, can potentially be used as a platform for a sequencing based MSI test. The work outlined in this chapter will aim to:

- Test a larger panel of CRCs to assess the sensitivity and specificity of the chosen panel of repeats
- Perform an analysis to determine suitable thresholds for calling instability by analysing the sensitivity and specificity of each marker.
- Evaluate the allelic bias in the MSI-H tumours to assess if allelic bias can be used as an additional tool for differentiating between mutations caused by MSI and sequencing and PCR artefacts.
- Develop the QuantuMDx hardware which can be used as a platform for the MSI test.

7.2. Results

7.2.1. Identification and curation of a panel of colorectal tumours

In previous chapters, the repeat panels have been tested on a small number of MSI-H tumours and controls to identify highly informative markers and assess the impact of length on information content. For the work in this chapter it was important to obtain a large number of tumours to define thresholds for calling instability and determine if the chosen panel of repeats is sufficient for differentiating between MSI-H and MSS tumours.

A total of 92 tumour samples were obtained after ethical review (REC reference 13/LO/1514). These tumours were supplied by Ottie O'Brien (Northern Genetics Service, Newcastle Hospitals NHS Foundation Trust), and Julie Coaker (Institute of Genetic Medicine, Newcastle University) in the form of FFPE wax curls and DNA already extracted FFPE samples.

DNA from the 92 tumours was first assessed to identify how many tumours had a sufficient quantity and quality of DNA to produce amplicons of ~300bp in length for a panel of 20 markers. The size of the panel was chosen because 20 markers should be sufficient to differentiate between MSI-H and MSS tumours and there was insufficient DNA for many of the tumours to amplify a larger panel. For 3 tumours there was too little starting material to be able to amplify 20 repeats. Out of the remaining 89 tumour DNA samples it was possible to amplify 58 of the samples using amplicons of ~300bp.

For 24 tumour samples all the PCR reactions were performed manually. To save time, 8 amplicons (DEPDC2, AL359238, AL954650, AP003532_2, TTK, AL355154, AVIL, ASTE1, EGFR, FBXO46) for 34 tumours were done robotically by NewGene (NewGene Ltd, International Centre for Life, Newcastle Upon Tyne, NE1 4EP, UK). For these 34 tumours the remaining 12 amplicons were produced manually. The PCR protocol and reagents used by NewGene did not differ from the protocol as outlined in methods section 2.3.2.2. The only difference in the protocol for the amplicons produced by NewGene was that after PCR amplification post PCR cleanup was performed by NewGene using Ampure XP beads. NewGene had a high PCR failure rate. 48 out of a total of 272 amplicons produced by NewGene did not produce a sufficient amount of PCR product to give visible products on the gels. The PCR for all of the failed amplicons were repeated manually.

Quantification for all PCR products was done using a Qiagen QIAxcel (Qiagen, Limburg, Netherlands) prior to amplicon pooling. After pooling, all amplicon pools were processed using Agencourt AMPure XP beads (Beckman Coulter, Pasadena, California, United States) to remove residual PCR reagents and Primer dimers. After PCR clean up each amplicon pool was quantified using a Qubit 2.0 fluorometer (Life Technologies, Carlsbad, CA, United States of America) and each amplicon pool was diluted to achieve a DNA concentration of 0.2ng/µl which is the recommended input DNA concentration for the Nextera XT library prep. The Illumina Nextera XT library prep was used to prepare the amplicons for sequencing on the Illumina MiSeq (Illumina, San Diego, CA, United States of America). The sequencing was performed using a MiSeq Reagent Kit v3 (600-cycles) (Illumina, San Diego, CA, United States of America). A flow cell cluster density of 2,068,000mm^2 was obtained for this MiSeq run giving a total read depth of 33,775,992 across all samples. This gave an average read depth of ~10000 paired end reads per amplicon. A Q-Score of over 30 was obtained for 56.8% of the bases sequenced (see Figure 7.2). There was a drop in Q-Score towards the latter cycles (see Figure 7.3). This is believed to be due to reaching the end of some of the amplicons being sequenced.

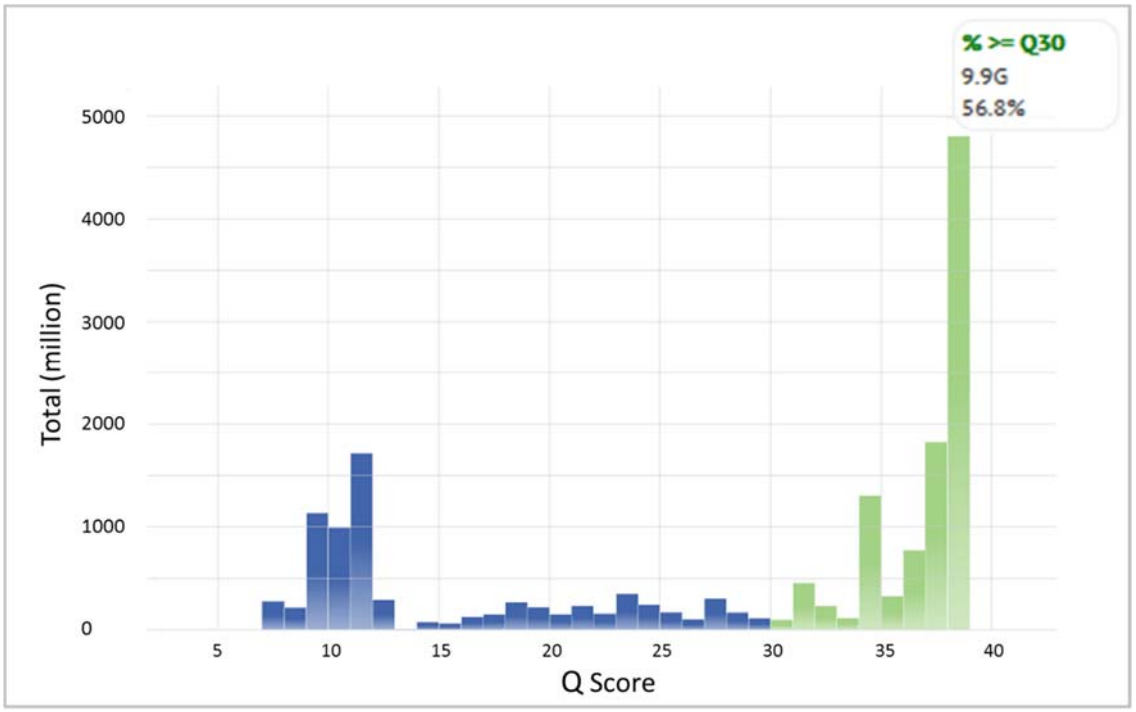

Figure 7.2: The quality score (Q-Score) distribution for the reads generated on the MiSeq. Blue = bases with a Q-Score <30, Green = bases with a Q-Score >30.

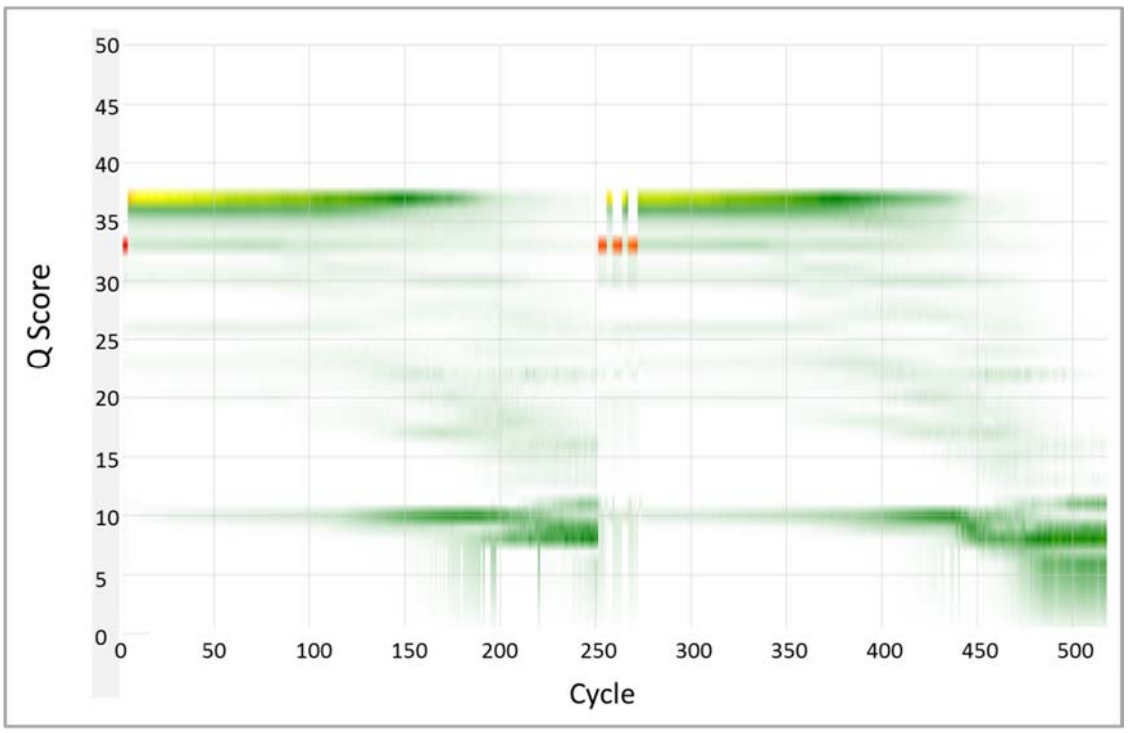

Figure 7.3: The quality score (Q-Score) distribution for each cycle showing a drop in Q-Score towards the later cycles of each read.

Variant calling was performed using COPReC. This is the same variant caller used in chapters 3, 5 and 6 (see methods section 2.8.6.2 for more details).

7.2.1.1. The ability of individual microsatellite markers for detecting MSI-H tumours

166 out of the 224 amplicons produced by NewGene failed to be sequenced, which is believed to be due to these amplicons having undergone PCR cleanup before being quantified using the QIAxcel. This meant that these amplicons were in a solution of dH_2O whereas the amplicons produced manually had not been cleaned up and were therefore in PCR buffer. Due to the amplicons being in different solutions for the quantification it appears that they have not been quantified equally on the QIAxcel, and as a result, the amplicons produced by NewGene were over diluted. This resulted in the amplicons produced by NewGene being under represented in the final library that was sequenced on the MiSeq. As a result these amplicons are underrepresented in the 58 tumours sequenced. Table 7.1 summarises the number of repeats sequenced for each tumour.

Tumour Sample	MSI status	Number of 8bp-12bp Repeats Sequenced	Number of 13bp-14bp Repeats Sequenced
15_S7	MSI-H	18	2
19_S9	MSI-H	17	1
21_S10	MSI-H	16	0
27_S15	MSI-H	18	1
3_S2	MSI-H	15	1
30_S16	MSI-H	18	2
31_S17	MSI-H	18	0
33_S18	MSI-H	15	0
34_S19	MSI-H	12	0
40_S21	MSI-H	9	0
41_S22	MSI-H	15	0
44_S24	MSI-H	15	0
5_S4	MSI-H	18	1
52_S29	MSI-H	16	0
53_S30	MSI-H	15	2
55_S31	MSI-H	15	0
80_S43	MSI-H	12	0
82_S45	MSI-H	15	0
83_S46	MSI-H	14	0
84_S47	MSI-H	16	0
G103_S54	MSI-H	18	0
G135_S55	MSI-H	18	2
G160_S56	MSI-H	18	2
G196_S57	MSI-H	18	2
G21_S51	MSI-H	18	2
G229_S58	MSI-H	18	2
G56_S52	MSI-H	18	2
G73_S53	MSI-H	18	2
13_S6	MSS	18	1
18_S8	MSS	18	0
2_S1	MSS	18	1
22_S11	MSS	15	0
24_S12	MSS	18	0
25_S13	MSS	14	1
26_S14	MSS	14	1
36_S20	MSS	18	1
4_S3	MSS	18	1
43_S23	MSS	14	0
45_S25	MSS	15	0
49_S26	MSS	15	1
50_S27	MSS	15	2
51_S28	MSS	14	0
59_S32	MSS	15	1
60_S33	MSS	12	0
64_S34	MSS	15	0
65_S35	MSS	15	0
69_S36	MSS	16	0
70_S37	MSS	14	0
71_S38	MSS	14	0
72_S39	MSS	13	0
73_S40	MSS	12	0
74_S41	MSS	14	0
79_S42	MSS	14	0
8_S5	MSS	18	1
81_S44	MSS	15	0
88_S48	MSS	13	0
90_S49	MSS	17	0
91_S50	MSS	16	0

Table 7.1: MSI status and number of amplicons sequenced for all 58 tumours.

The ability of each repeat to discriminate between the MSI-H samples and the MSS samples was assessed using the area under the receiver operating characteristic curve (AUC). Dr Mauro Santibanez-Koref (Institute of Genetic Medicine, Newcastle University) performed the AUC calculations. Receiver operating characteristic curves are a method of measuring true positive and false positive rates. In this case the AUC is a measure of how well a given homopolymer can differentiate between the MSI-H and MSS samples. An AUC of 1 is achieved if all the MSI-H samples have a higher deletion frequency than the MSS samples for a given repeat. Any randomly chosen MSI-H sample from the data set would in this case have a 100% chance of having a higher deletion frequency than any randomly chosen MSS sample from the data set. An AUC value of 0.5 would mean that a repeat has no discrimination power because there would be 50-50 chance that any randomly chosen MSI-H sample would have a higher deletion frequency than any randomly chosen MSS sample.

The AUC values for all the homopolymer in the final panel can be found in Table 7.2. On average, the AUC increases with repeat length up to a repeat length of 12bp. This means that the longer repeats, up to a length of 12bp, are better at discriminating between the MSI-H samples and MSS samples. This was expected because longer microsatellites are more prone to microsatellite instability events than shorter repeats. For the shorter repeats there will therefore be more repeats in MSI-H samples that have not been affected by a mutation, decreasing the ability of those repeats to discriminate between MSI-H samples and MSS samples. The 13bp and 14bp repeat have an AUC of 0.9 and 0.722 respectively. These are lower AUC values than seen in all the 12bp and all but one of the 11bp repeats (see Table 7.2). This could indicate that sequencing and PCR error are so high in these repeats that using the frequency of all deletions as a measure of instability is no longer as good for discriminating between MSI-H and MSS samples as it is for the shorter 11bp and 12bp repeats. On the other hand it could be that the chosen 13bp and 14bp repeat are less prone to MSI due to sequence context and there may be many other 13bp and 14bp repeat in the genome that are more unstable than these two. For the 14bp repeat FBXO46 a low AUC could also be due to the presence of a sequence length polymorphism in some of the controls. One of the tumours in chapter 6 had a sequence length polymorphism for this repeat which indicates there is a possibility that FBXO46 could be polymorphic in some samples.

Repeat Name	Size (bp)	Repeat Base	Number of Samples Sequenced	AUC
DEPDC2	8	C	36	0.645
LR46	8	A	58	0.825
AL359238	9	A	53	0.806
AL954650	9	C	29	0.639
AP003532_2	9	A	58	0.896
TTK	9	A	46	0.733
AL355154	10	A	33	0.915
AVIL	10	A	39	0.927
GM29	10	A	57	0.883
LR32	10	A	57	0.910
ASTE1	11	A	41	0.957
GM07	11	A	58	0.968
GM14	11	A	58	0.873
LR11	11	A	55	0.919
LR48	11	A	56	0.988
IM49	12	A	58	0.958
LR36	12	A	58	0.919
LR44	12	A	58	0.994
EGFR	13	A	12	0.900
FBXO46	14	A	23	0.722

Table 7.2: Area under the receiver operating characteristic curve (AUC) for each marker in the final panel of repeats. This table shows the length of each repeat, the repeat unit, and the ability of each repeat to discriminate between MSI-H and MSS samples expressed as the area under the receiver operating characteristic curve.

In the previous chapter I showed that PCR and sequencing error is dependent to some degree on the length of the homopolymer. Therefore different thresholds for calling instability will be needed for different homopolymer lengths. Thresholds for calling a marker unstable can be determined for each repeat length by assessing the sensitivity and specificity of each of the individual markers. Sensitivity and specificity are used to measure test accuracy. Sensitivity is measured as the fraction of patients who have a condition and have a positive test result for it. Specificity is the fraction of patients who don't have a condition and have a negative for that condition. Therefore sensitivity and specificity can be summarised as:

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

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

For this work a tumour was defined as MSI-H if it had previously been classed as MSI-H using a standard Promega MSI test (MSI Analysis System, Version 1.2: Promega, Madison, WI, United States of America). Tumours were classed as MSS if no instability had been detected in any of the five markers from the Promega MSI test. The Promega MSI tests for all tumours were performed by the Northern Genetics service. For each of the short mononucleotide repeats sequenced sensitivity and specificity curves were produced. Each of the sensitivity and specificity curves has the frequency of reads containing deletions on the x-axis. The y-axis of each sensitivity curve is the fraction of

MSI-H samples. The sensitivity curve shows the fraction of MSI-H samples (y-axis) that have a deletion frequency of or below the deletion frequency shown on the x-axis, which is the sensitivity at each given deletion frequency. The y-axis of the specificity curve is the fraction of MSS samples. The specificity curve shows the fraction of MSS samples (y-axis) that have a deletion frequency of or above the frequency shown on the x-axis which is the specificity at each given deletion frequency.

The sensitivity and specificity curves for the 8bp-9bp repeats can be found in Figure 7.4. Of the 8bp repeats, LR46 (extracted from the whole genome analysis) has a higher sensitivity than DEPDC2 (taken from the literature) for deletion frequencies up to 40%. Both repeats have a 100% specificity or no false positives at a deletion frequency of 4.1%. At this deletion frequency LR46 has a sensitivity of 42.9% with 12 out of the 28 MSI-H samples detected, and DEPDC2 has a sensitivity of 26.1% with 6 out of the 23 sequenced MSI-H samples detected.

All of the 9bp repeats have 100% specificity for a 5.5% deletion frequency and above. At a deletion frequency of 5.5% the two repeats AP003532_2 and TTK have the highest sensitivity with 57.1% and 43.5% respectively. The two repeats AL954650 and AL359238 have a sensitivity of 42.1% and 21.7% at this deletion frequency.

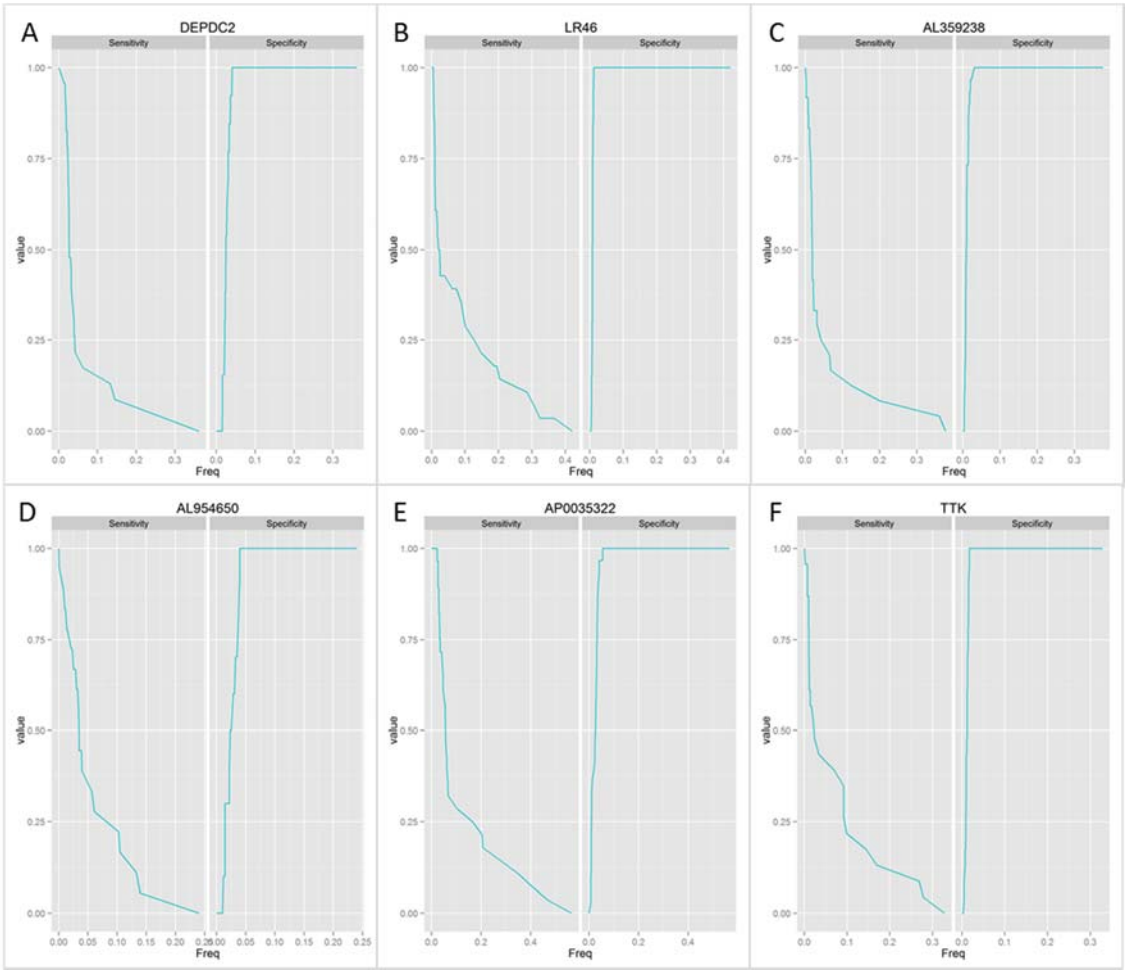

Figure 7.4: Sensitivity and Specificity curves for the 8bp and 9bp homopolymers used in the final panel of repeats. DEPDC2 and LR46 are 8bp repeats while AL359238, AL954650, AP003532_2 and TTK are 9bp repeats. Value = fraction of samples, Freq = deletion frequency

All of the 10bp repeats have a 100% specificity at a deletion frequency of ≥14.2%. For a deletion frequency of 14.2% the repeat LR32 has a sensitivity of 82.1%, which is the highest for any of the 10bp repeats at this deletion frequency. The other 10bp repeats AVIL, AL3551554, GM29 have a sensitivity of 71.4%, 35.3% and 25.9% respectively.

For the 11bp repeats, the repeat ASTE1 had the highest frequency of deletions in the control samples with a deletion frequencies ranging between 11.9% - 19.75%. All of the 11bp repeats have a 100% specificity at a deletion frequency of ≥19.8%.

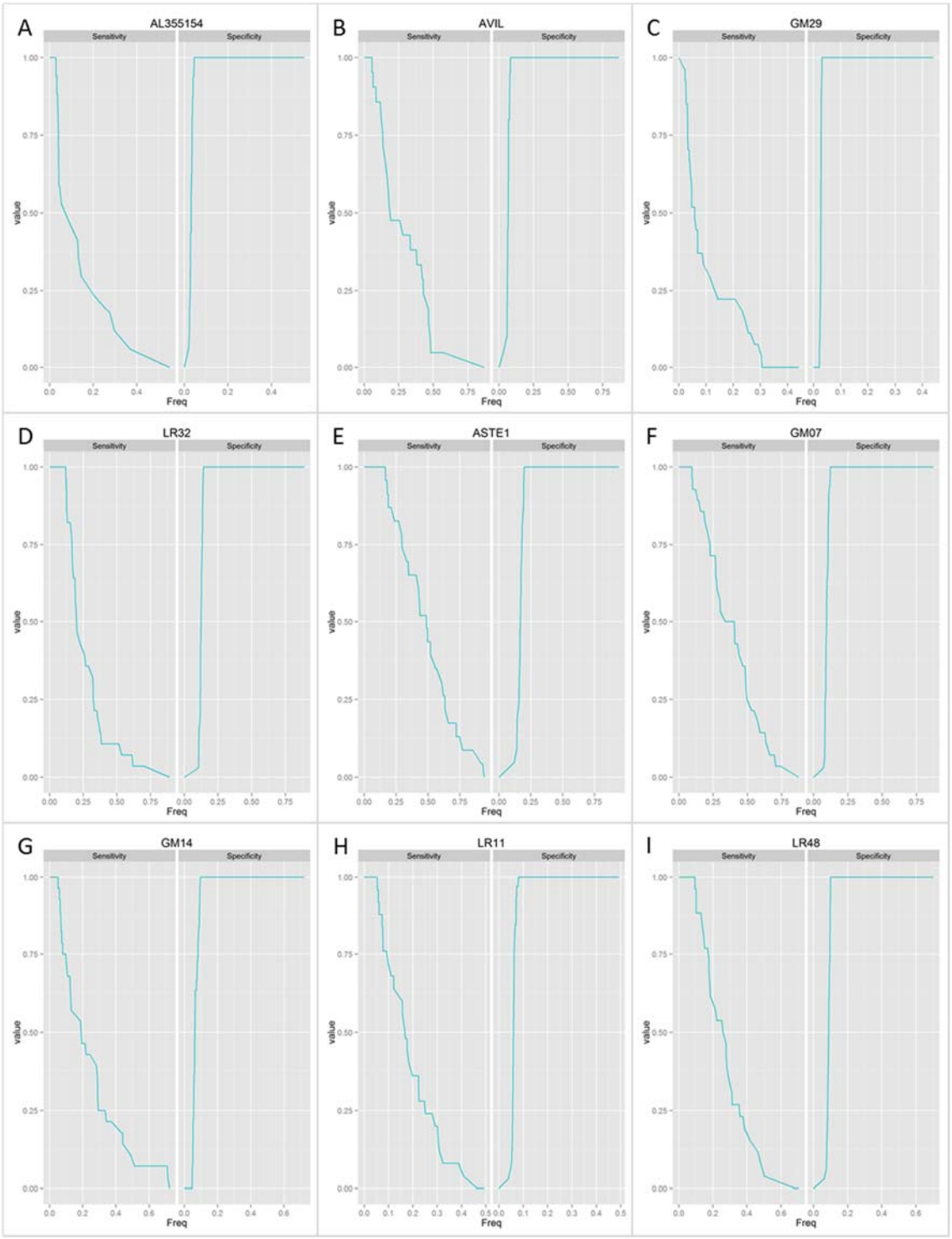

Figure 7.5: Sensitivity and Specificity curves for the 10bp and 11bp homopolymers used in the final panel of repeats. AL355154, AVIL, GM29, and LR32 are 10bp repeats. ASTE1, GM07, GM14, LR11, and LR48 are 11bp repeats. Value = fraction of samples, Freq = deletion frequency

All of the 12bp repeats have a 100% specificity at a deletion frequency of ≥19.4%. At a deletion frequency of 19.4% the repeats LR44, LR36 and IM49 have a specificity of 92.9%, 75% and 64.3% respectively.

The 13bp marker EGFR had a high dropout rate within the sequence data and was only sequenced in 12 of the 58 tumours. Only two out of the 12 tumours that this marker was sequenced in were MSS tumours. EGFR has a 100% specificity at a deletion frequency of ≥24%, but as this is only based on data from 2 MSS samples it is not very dependable.

The 14bp homopolymer FBXO46 was only sequenced in 23 tumours. One of the MSS tumours (26_S14) had a deletion frequency of 88.49% for this tumour. This was the highest deletion frequency seen in any of the tumours, which means that at the point where there is a 100% specificity for this marker there is a 0% sensitivity (see Figure 7.6 panel E). In chapter 6 the repeat FBXO46 was found to have a polymorphism for the patient from which the tumour PR10654/14 was extracted with 52% of the reads in the normal mucosa having a repeat length that was 1bp shorter than the reference sequence. It is possible that the tumour 26_S14 has such a high deletion frequency because it is homozygous for the same polymorphism. Because there is no matching normal tissue for the tumour 26_S14 it is not possible to determine if there is a polymorphism in this patent for the marker FBXO46. The presence of a polymorphism in the tumour PR10654/14 means that this marker is not suitable for the use in an MSI test because the marker being potentially polymorphic means that a high deletion frequency is not necessarily an indication of MSI. Unfortunately the tumour PR10654/14 was sequenced in the same run as the samples discussed in this chapter so I was not aware of the polymorphism prior to the sequencing of the samples discussed here.

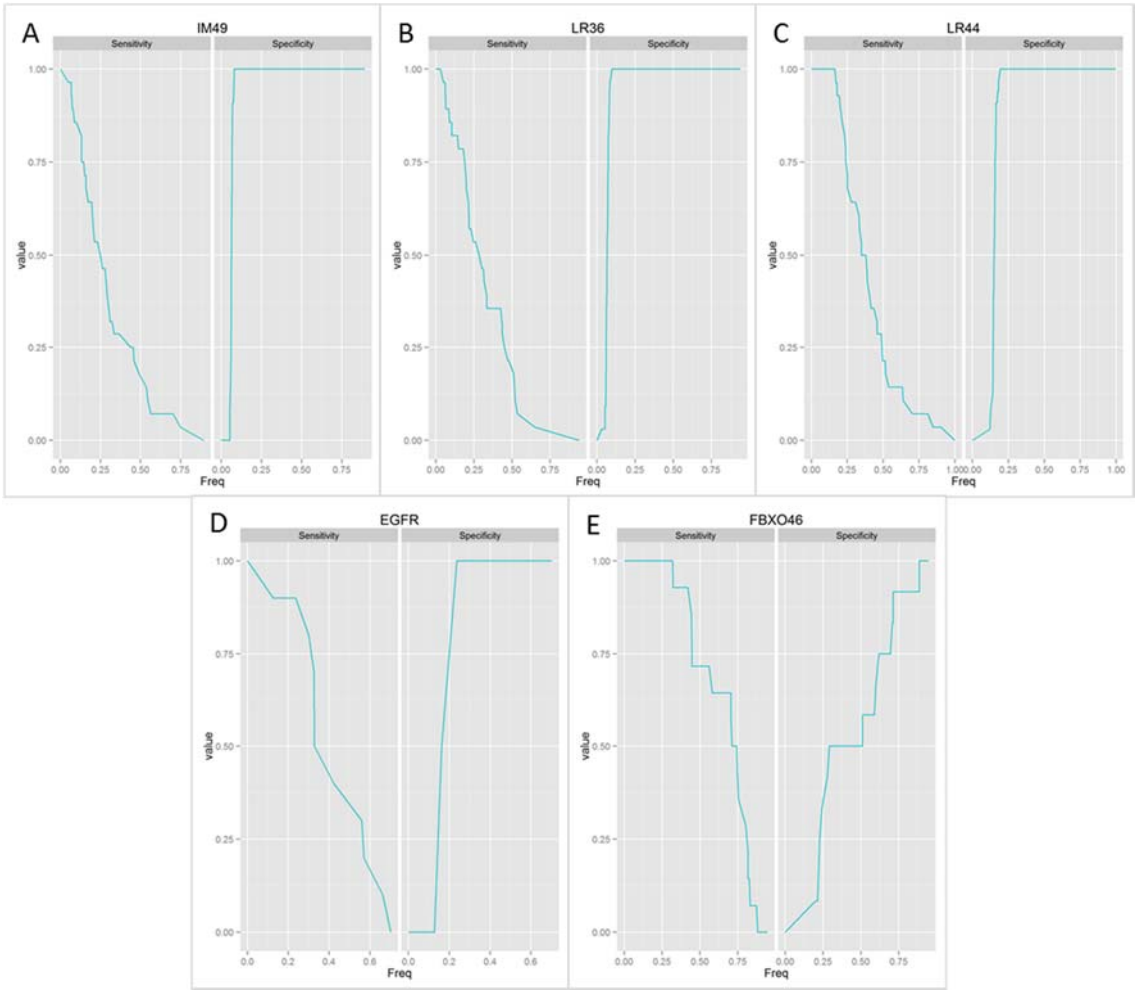

Figure 7.6: Sensitivity and Specificity curves for the 12bp, 13bp, and 14bp homopolymers used in the final panel of repeats. IM49, LR36, and LR44 are 10bp repeats. EGFR and FBXO46 are 13bp and 14bp repeats respectively. Value = fraction of samples, Freq = deletion frequency

7.2.1.2. Optimisation of thresholds for differentiating tumours by MSI status

To assess the performance of the repeats for differentiating between MSI-H tumours and MSS tumours the panel of repeats was evaluated using different deletion frequencies as cut-offs. Two repeats were excluded from this analysis: EGFR because it was only successfully sequenced in two of the MSS samples and there was therefore very little information about the background PCR and sequencing error rates for this repeat. FBXO46 was excluded because this repeat may be polymorphic in some samples. Using a repeat with a repeat length polymorphism is problematic when the deletion frequency of the repeats is being used to classify samples as MSS or MSI-H. This means that the panel of repeats used in the subsequent analysis consist of eighteen 8bp-12bp mononucleotide repeats. Different thresholds were set for each repeat size.

First, thresholds were set so that each group of repeats of the same length had the minimum number of incorrectly classified repeats. If the minimum number of incorrectly classified repeats could be obtained at more than one deletion frequency, then the lowest of these deletion frequencies was used as the threshold. The deletion frequency for each repeat length in which this minimum error rate was achieved can be found in Table 7.3. Assigning thresholds in this way means that there are many instances where a repeat has a deletion frequency above the threshold in the MSS samples, which gives a high false positive rate. For the 8bp repeats (LR46 and DEPDC2) there is a false positive rate of 0.256. This means that for these two repeats the MSS samples have a deletion frequency that meets the threshold for calling instability 25.6% of the time. For the 8bp repeats (LR46 and DEPDC2) there is a false negative rate of 0.235, which means that 23.5% of the time the MSI-H samples have a deletion frequency below the threshold used to call instability. For the false positive and false negative rates for the other repeat sizes see Table 7.3.

Chromosome instability is the most common cause of colon cancer accounting for approximately 85% of CRCs while the other approximately 15% of CRCs have mismatch repair gene defects and are characterized by microsatellite instability (Grady, 2004, Sinicrope and Sargent, 2012). Using this information it is possible to predict how many errors there would be for each repeat size given a panel of tumours which conform to division of MSI-H and MSS tumours that would be expected if all colorectal tumours were tested for MSI. This is done by multiplying the false positive rate by 85 to obtain the percentage of false positive errors and multiplying the false negative rate by 15 to obtain the percentage of false negative errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours. False positive and false negative error rates for each repeat size assuming a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours can be found in Table 7.3.

In the next section, an analysis of repeats based on repeat length by setting thresholds for each repeat length individually and calculating the false positive and false negative error rates for these thresholds is presented. Using these false positive and false negative error rates, the error rates for a panel of tumours consisting of 15% MSI-H tumours and 85% MSS are calculated. The number of unstable repeats for each tumour in the sequenced panel of 58 tumours is also assessed. The equations for calculating false positive and negative error rates can be found in method section 2.9.3. Finally, an evaluation of how allelic bias could be used to augment an MSI test is discussed.

Repeat Length	Deletion Frequency Threshold	Minimum Number of Errors	FPR	FNR	% False Positive Errors (assuming 85% MSS)	% False Negative Errors (assuming 15% MSI-H)
8bp	0.016	23	0.256	0.235	21.7	3.5
9bp	0.041	50	0.011	0.527	0.9	7.9
10bp	0.142	42	0.000	0.452	0.0	6.8
11bp	0.121	40	0.130	0.169	11.1	2.5
12bp	0.164	18	0.033	0.179	2.8	2.7

Table 7.3: Thresholds for each repeat size that minimise the number of misclassified repeats. This table shows the deletion frequency thresholds that give a minimum number of errors for each repeat size. For each threshold the table shows the number of errors, the false positive error rate, the false negative rate, and the percentage of errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours. FPR = false positive error rate, FNR = false negative error rate.

Using the deletion frequency thresholds shown in Table 7.3 the number of repeats passing the threshold for each tumour was plotted using a bar chart (see Figure 7.7). Using these thresholds, every MSI-H tumour had five or more repeats that met the threshold for calling instability. For the MSS samples there were up to three repeats which met the threshold for calling instability. Using these thresholds it is therefore possible to separate the MSI-H tumour and MSS tumours because the panel of 18 repeats is able to correctly classify every MSS and MSI-H cancer using a cut-off of 4 or 5 unstable repeats to classify a sample as MSI-H.

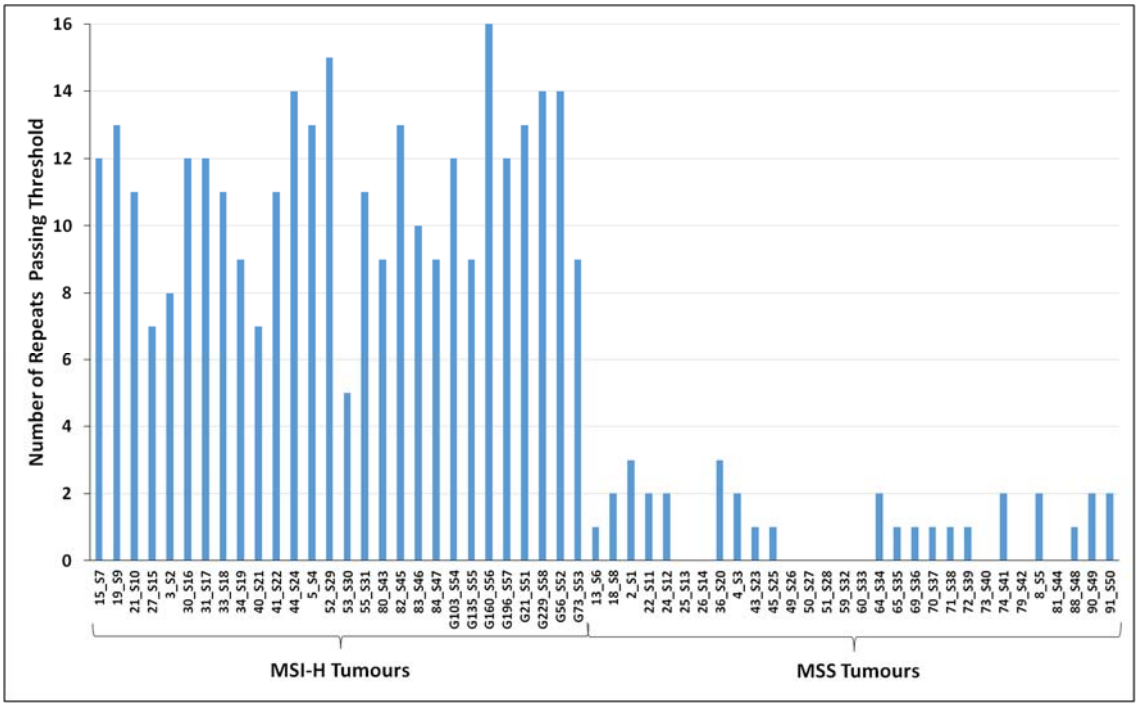

Figure 7.7: Number of 8bp-12bp repeats classed as unstable in each tumour using thresholds for each repeat size that minimise the number of misclassified repeats.

The sensitivity of the marker panel could easily be adjusted by adding more repeats. The specificity is more important because false positives can accumulate. Individual repeats being classed as unstable in MSS samples is therefore more of a problem than individual repeats being classed as stable in MSI-H samples. In fact because replication errors in MSI-H samples occur randomly it is expected that some of the repeats in MSI-H samples will not be affected by replication errors and will therefore remain stable. To better reflect this, different weighting can be placed on false positive and false negative errors. Different weightings of errors were assessed to see how they would affect the false positive and false negative error rates for the sequenced panel of tumours, and the number of unstable repeats in MSI-H and MSS tumour samples.

The weighting of different errors was adjusted so that a false positive error is 1.5x worse than a false negative error and the deletion frequency thresholds for calling a repeat unstable were adjusted to reflect this different cost of the two types of errors. The deletion frequency thresholds were set so that the cost of errors was minimised. This changed the thresholds for the 11bp and 12bp repeats reducing the false positive error rates for these repeats (see Table 7.4).

Repeat Length	Deletion Frequency Threshold	FPR	FNR	% False Positive Errors (assuming 85% MSS)	% False Negative Errors (assuming 15% MSI-H)
8bp	0.016	0.256	0.235	21.7	3.5
9bp	0.041	0.011	0.527	0.9	7.9
10bp	0.142	0.000	0.452	0.0	6.8
11bp	0.174	0.051	0.277	4.3	4.2
12bp	0.194	0.000	0.226	0.0	3.4

Table 7.4: Thresholds for each repeat size that minimise the cost of misclassified repeats given that a false positive error is 1.5x worse than a false negative error. This table shows the deletion frequency thresholds that give a minimum cost of errors for each repeat size. For each threshold the table shows the false positive error rate, the false negative rate, and the percentage of errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours. FPR = false positive error rate, FNR = false negative error rate.

The new deletion frequency thresholds (see Table 7.4) were then used to calculate how many repeats passed the thresholds for each tumour sample. Using the new thresholds all the MSI-H tumours still have 5 or more repeats that are classified as unstable while none of the MSS tumours have more than 2 unstable repeats. The panel of 18 repeats is therefore able to classify every MSS and MSI-H cancer correctly using a cut-off of 3 - 5 unstable repeats to classify a sample as MSI-H (see Figure 7.8). By weighting false positive errors as 1.5 times more costly than false negative errors the panel of 18 repeats is better able to differentiate between the MSI-H and MSS samples.

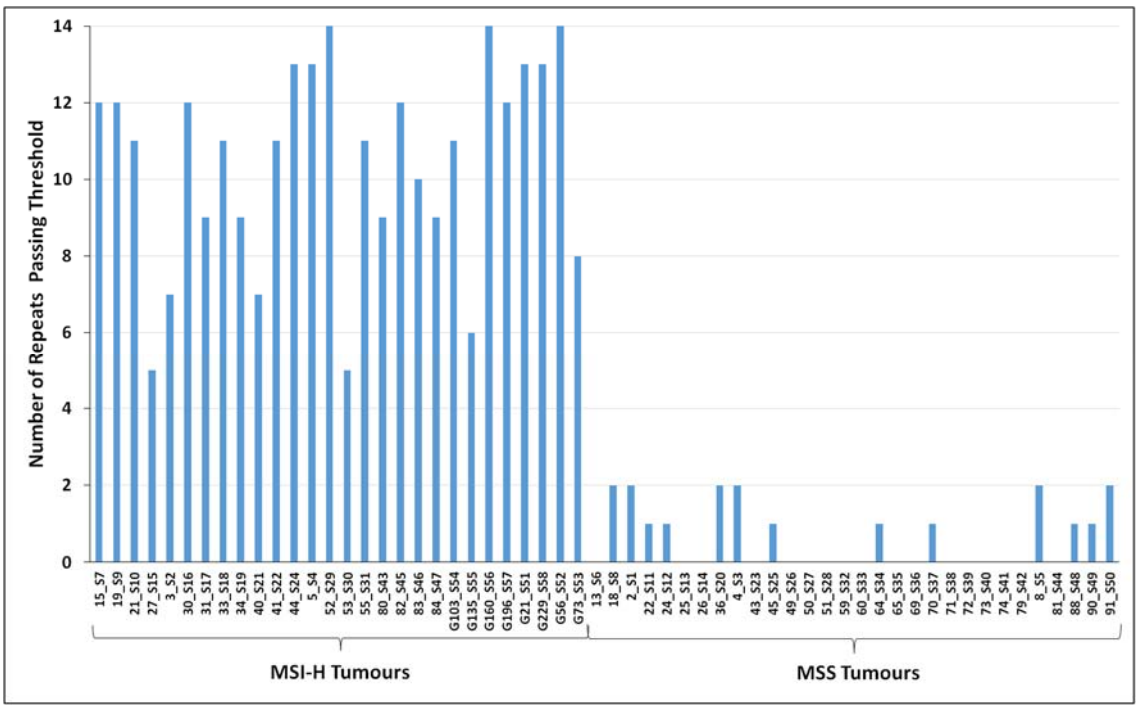

Figure 7.8: Number of 8bp-12bp repeats classed as unstable in each tumour using thresholds for each repeat size where a misclassified repeat in a MSS sample is 1.5x as bad as a misclassified repeat in a MSI-H sample.

The weighting of different errors was adjusted further so that a false positive error is two times worse than a false negative error. The deletion frequency thresholds were adjusted so that cost of errors was minimised. As a result the thresholds for calling a repeat unstable were increased for both the 8bp and 11bp repeats (see Table 7.5). For the 10bp -12bp repeats there are no false positive errors using the current deletion frequency thresholds (see Table 7.5).

Repeat Length	Deletion Frequency Threshold	FPR	FNR	% False Positive Errors (assuming 85% MSS)	% False Negative Errors (assuming 15% MSI-H)
8bp	0.037	0.023	0.608	2.0	9.1
9bp	0.041	0.011	0.527	0.9	7.9
10bp	0.142	0.000	0.452	0.0	6.8
11bp	0.198	0.000	0.369	0.0	5.5
12bp	0.194	0.000	0.226	0.0	3.4

Table 7.5: Thresholds for each repeat size that minimise the cost of misclassified repeats given that a false positive error is 2x worse than a false negative error. This table shows the deletion frequency thresholds that give a minimum cost of errors for each repeat size. For each threshold the table shows the false positive error rate, the false negative rate, and the percentage of errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours. FPR = false positive error rate, FNR = false negative error rate.

The new deletion frequency thresholds found in Table 7.5 were used to analyse the panel of tumours. Using these thresholds has reduced the number of repeats classed as unstable in the MSS tumours to two repeats (see Figure 7.9). One repeat for the tumour 22_S11 and one repeat for the tumour 64_S34. All of the MSI-H tumours have 2 or more repeats which are classed as unstable (see Figure 7.9). The panel of 18 repeats is therefore able to correctly classify all MSS and MSI-H tumours if a cut-off of 2 unstable repeats is used to classify a sample as MSI-H.

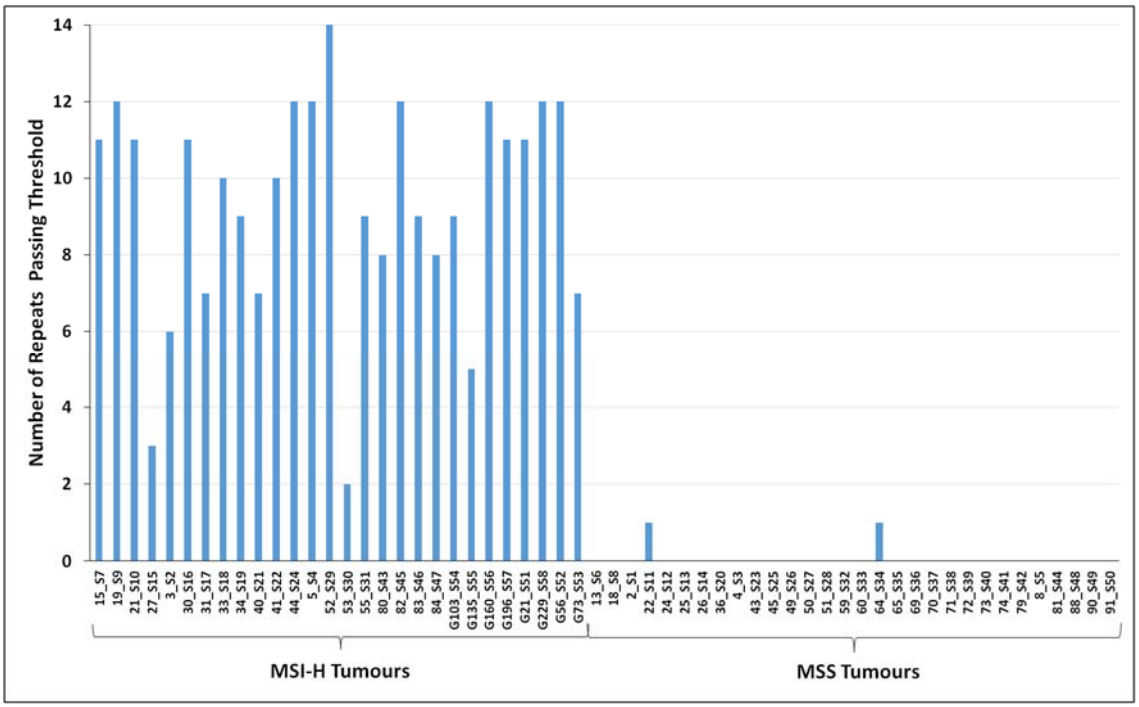

Figure 7.9: Number of 8bp-12bp repeats classed as unstable in each tumour using thresholds for each repeat size where a misclassified repeat in a MSS sample is 2x as bad as a misclassified repeat in a MSI-H sample.

If the weighting of different errors is adjusted so that a false positive error is more than 5 times worse than a false negative error, then the resulting thresholds result in no false positive errors for any repeat size (see Table 7.6). At these thresholds the false negative error rate for the MSI-H samples is between 22.6% for the 12bp repeats and 64.7% for the 8bp repeats. For a panel of tumours which conform to division of 15% MSI-H tumours and 85% MSS tumours the error rate would be between 3.4% and 9.7% for each marker size. All of these errors are false negative errors. Because all 18 markers would be used together for classifying samples as MSI-H the false negative error rate for the full panel of repeats will be much lower than the false negative rate for individual repeat sizes.

Repeat Length	Deletion Frequency Threshold	FPR	FNR	% False Positive Errors (assuming 85% MSS)	% False Negative Errors (assuming 15% MSI-H)
8bp	0.041	0.000	0.647	0.0	9.7
9bp	0.055	0.000	0.581	0.0	8.7
10bp	0.142	0.000	0.452	0.0	6.8
11bp	0.198	0.000	0.369	0.0	5.5
12bp	0.194	0.000	0.226	0.0	3.4

Table 7.6: Thresholds for each repeat size that minimise the cost of misclassified repeats given that a false positive error is >5x worse than a false negative error. This table shows the deletion frequency thresholds that give a minimum cost of errors for each repeat size. For each threshold the table shows the false positive error rate, the false negative rate, and the percentage of errors for a panel of tumours consisting of 85% MSS tumours and 15% MSI-H tumours. FPR = false positive error rate, FNR = false negative error rate.

When the panel of 28 MSI-H tumours and 30 MSS tumours is analysed using the deletion frequency thresholds found in Table 7.6, there are 2 or more repeats classed as unstable in all of the MSI-H tumours. Because the thresholds for each repeat length have been set so that there are no false positive errors the panel of 18 repeats is able to correctly classify all MSS and MSI-H tumours if a cut-off of 1-2 unstable repeats is used to classify a sample as MSI-H.

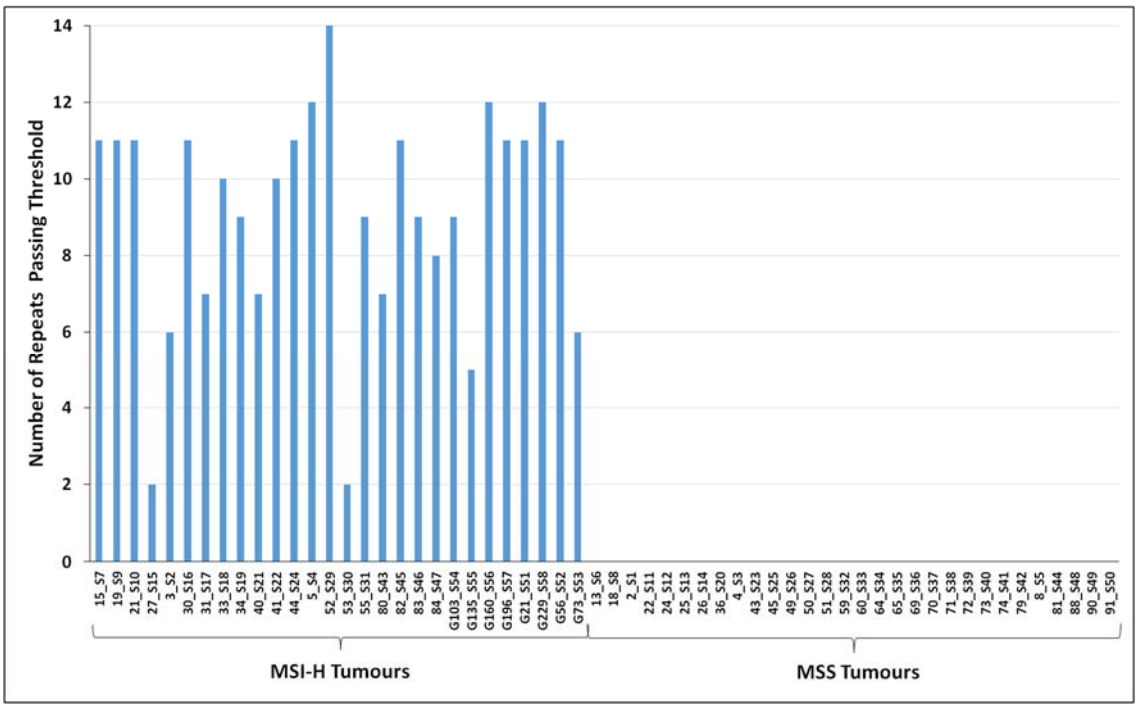

Figure 7.10: Number of 8bp-12bp repeats classed as unstable in each tumour using thresholds for each repeat size where a misclassified repeat in a MSS sample is >5x as bad as a misclassified repeat in a MSI-H sample.

7.2.1.3. Allelic bias in MSI-H tumours

Most of the chosen panel of repeats had neighbouring SNPs with a high minor allele frequency. This was to allow the allelic bias to be analysed in repeats with a heterozygous SNP. Repeats were defined as heterozygous if there were 100 reads spanning both SNP and repeat for each allele and one allele did not have less than 10% of the number of reads compared to the other allele. Heterozygous repeats were identified and a two-tailed Fisher's exact test was performed to determine if there was a bias in deletion frequency between the two alleles using my script FisherTest_AllDeletions.pl (See methods section 2.9.1). A Fisher's exact test was performed for every heterozygous SNP, therefore if there were more than one heterozygous SNP in close proximity to a homopolymer that homopolymer would be analysed using all SNPs, and the data plotted. This method was chosen because different SNPs would have a different number of reads spanning both SNP and repeat. Therefore different repeat and SNP combinations could provide different levels of significance for allelic bias

The results of the two-tailed Fisher's exact test showed an excess of allelic bias in the MSI-H samples compared to the MSS samples (see Figure 7.11). The allelic bias can therefore potentially be used to differentiate between genuine mutations and sequencing artefacts, because low levels of indels caused by sequencing/PCR error tend to affect both alleles equally. There are four of the MSS tumours showing examples statistically significant amount of allelic bias at a Bonferroni corrected p-value of 0.01 (0.01/335= 0.00002). These four MSS tumours consist of 45_S25, 65_S35, 26_S14 and 71_S38. The MSS tumour 45_S25 has allelic imbalance in the repeat LR36. This was measured at two SNPs with a p-value of 1.3×10^{-37} and 7.9×10^{-16} for SNP1 and SNP2 respectively. The MSS tumour 65_S35 shows a significant allelic imbalance for the repeat AL954650 p-value of 2.8×10^{-16}. Tumours 26_S14 and 71_S38 have allelic imbalances for one repeat each at a p-value of 2.1×10^{-6} and 3.0×10^{-5} respectively.

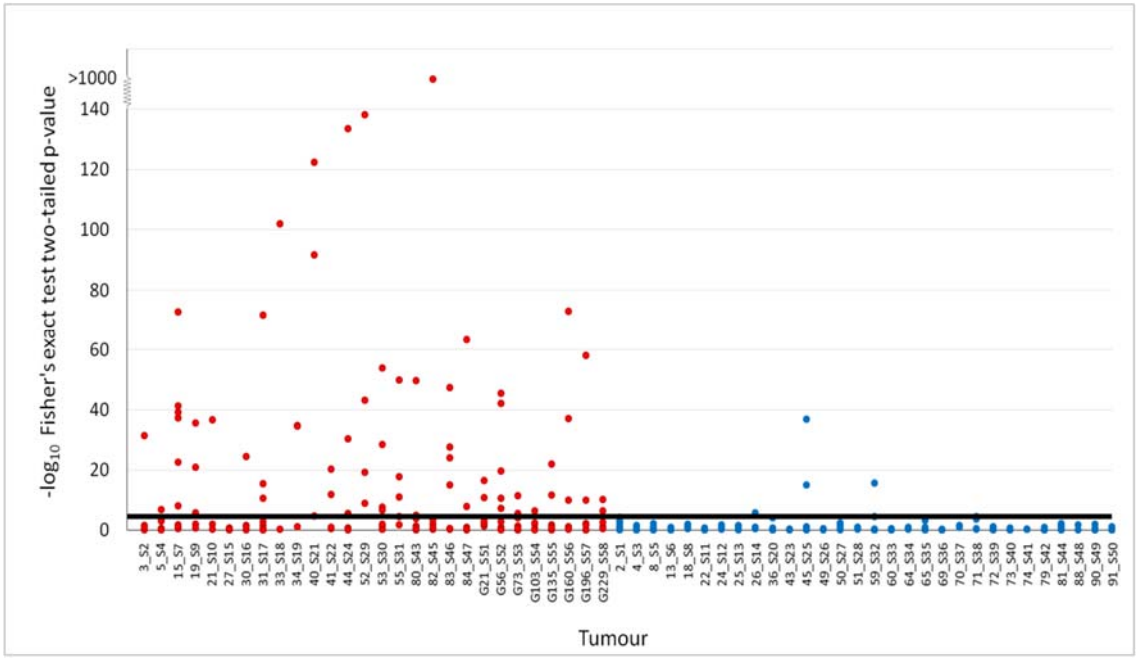

Figure 7.11: Allelic bias in deletion frequency for MSI-H samples and MSS samples measured using the p-value of a two tailed Fisher's exact test. Red = MSI-H samples, Blue = MSS samples. The line corresponds to a Bonferroni corrected p-value of 0.01.

Because an excess of allelic bias is seen in MSI-H tumours compared to MSS tumours it could also be incorporated into an MSI test. This was done by allocating 1 point for each repeat passing the deletion frequency thresholds shown in Table 7.6, and 1.5 points for each repeat that both passes the deletion frequency thresholds and has a statistically significant amount of allelic bias using Bonferroni corrected p-value of 0.01. Using this system of point allocation, each MSI-H tumour has at least two points whereas no points are allocated any of the MSS tumours. A threshold of 1-2 points could therefore be used to classify a tumour as MSI-H. The MSI-H tumour 27_S15, which only has two unstable repeats, is homozygote for SNPs neighbouring both unstable repeats. Adding extra points for allelic bias therefore does not increase the measurable difference between MSI-H tumours and MSS tumours compared to using the same deletion frequency thresholds without extra points for allelic bias.

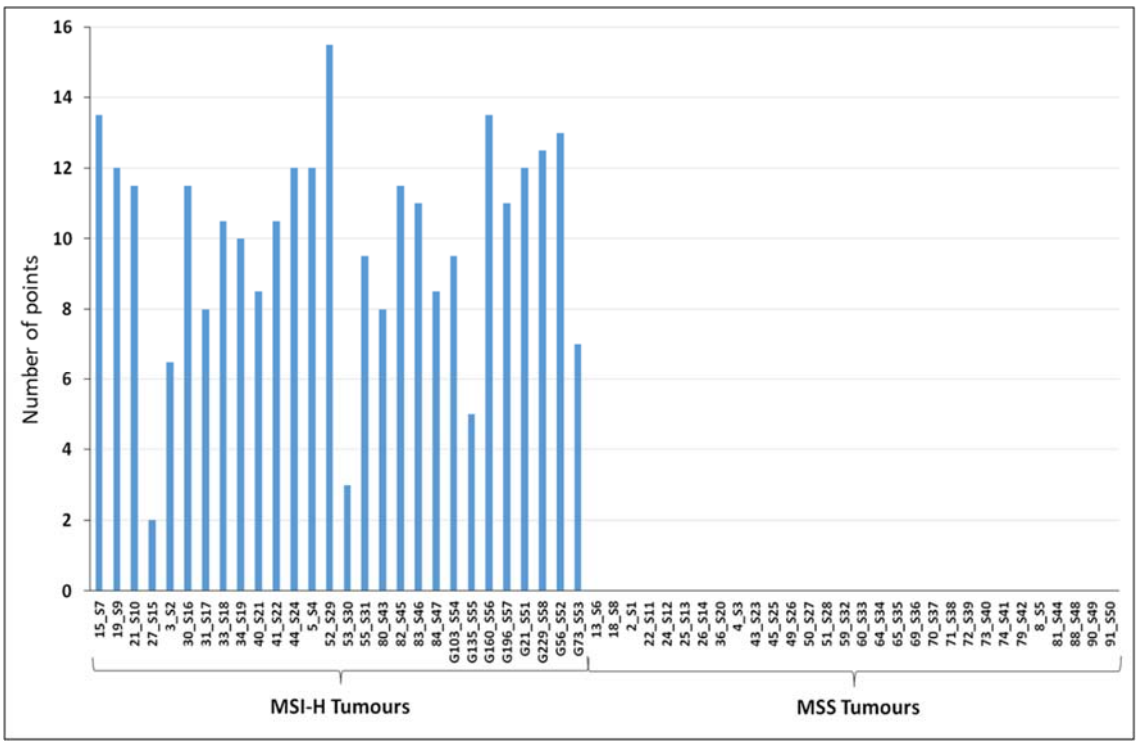

Figure 7.12: Point based MSI assay, with 1 point for each repeat passing a deletion frequency threshold, and 1.5 points for repeats that both pass the deletion frequency threshold and have a statistically significant amount of allelic bias. The thresholds used for each repeat size can be found in Table 7.6 and allelic bias was considered significant at a Bonferroni corrected p-value of 0.01.

7.2.2. Optimization of DNA extraction and amplification using QuantuMDx's microfluidic platform

The company QuantuMDx (QuantuMDx Group Ltd, Times QuantuMDx Square, Newcastle upon Tyne, UK) are producing a point of care device for DNA analysis. My PhD work has been performed in collaboration with this company with the aim that the MSI test I have been working on will be part of a cancer diagnosis assay. As part of my PhD work I have contributed to the development of the QuantuMDx hardware. This work was conducted in 2011-2012 when QMDx's technology was in its infancy. The work I have conducted with QuantuMDx has allowed me to become familiar with the QuantuMDx hardware and how the final MSI assay will work on QuantuMDx's device as well as contribute to the development of the hardware itself. This development work will be described in the next sections of this chapter.

7.2.2.1. DNA extractions from whole blood using the QuantuMDx DNA extraction cassette

DNA extractions from lysed whole blood were performed on QuantuMDx's 2012 prototype DNA extraction cassettes. These extractions from blood were performed as a proof of principle test to establish that the DNA extraction cassette worked, and assess its performance before moving on to attempting DNA extractions of tissue samples. A photo showing the layout of the DNA extraction cassette can be found in Figure 7.13.

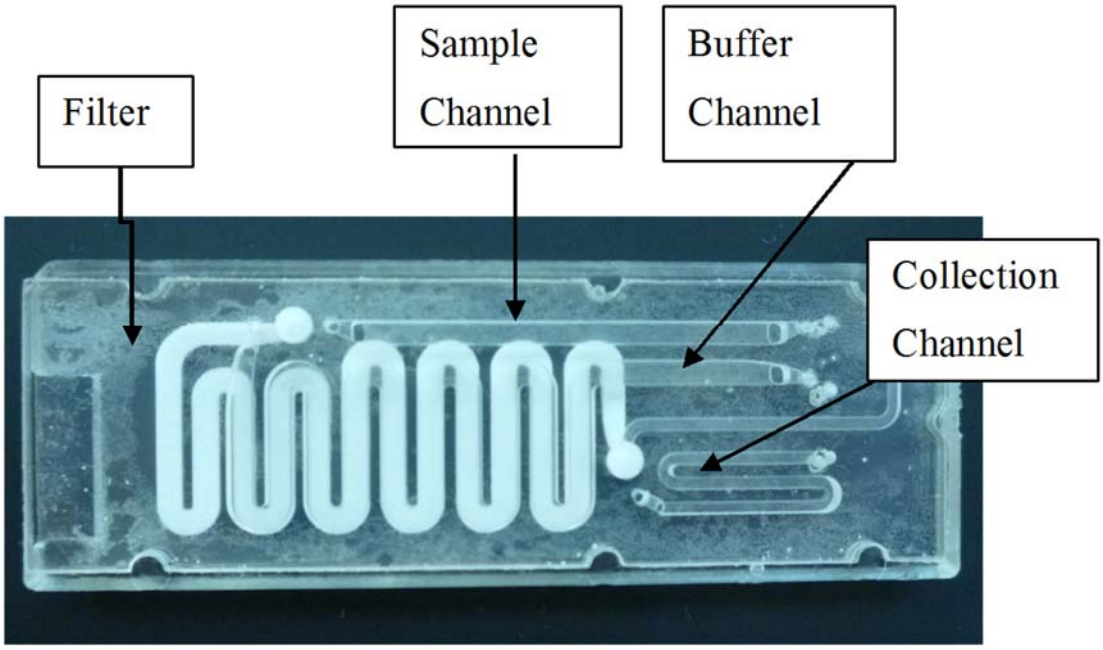

Figure 7.13: QuantuMDx's 2012 prototype DNA extraction cassette.

Prior to DNA extraction, the sample channel was loaded with 200µl of whole blood and the buffer channel was filled with a proprietary buffer. The cassette is then loaded onto QuantuMDx's prototype machine (the MiniChemLab) (see

Figure 7.14). First, the buffer is pushed through the DNA extraction cassette using the syringes of QuantuMDx's prototype machine at a flow rate of 100µl/min for 300 seconds. Once the filter has been soaked through with buffer there is a five minute incubation period while buffer activates the filter. Next the blood is pushed into the filter at a flow rate of 100µl/min for 120 seconds. It is optimal to load ~190µl of the blood into the filter. Leaving some blood behind in the sample channel helps prevent bubbles from entering the filter and creating a channel through the filter by displacing filter particles. Once the blood has been loaded onto the filter buffer from the buffer channel is flowed through the cassette at a flow rate of 50µl/minute, with a pause every 80 seconds to allow

elute to be collected from the collection channel. Due to the limited size of the buffer channel, the buffer channel needs topping up with buffer during this procedure. This is done during one of the pauses when the amount of buffer is low. The buffer pushes the sample through the filter. The filter is meant to retain the cellular components that are passed through it, with the exception of DNA, which is passed out through the collection channel together with the buffer. Each extraction was split into a number of elute fractions of 10-80µl volume each. PCR amplification was then performed to confirm the presence of DNA in each elute and to confirm that the DNA was of a suitable purity to achieve PCR amplification. The PCR products were then visualised on an agarose gel. The gel image in Figure 7.15 shows DNA extracted from blood for two different DNA extraction cassettes. The results showed that QuantuMDx's prototype DNA extraction cassette was able to produce DNA of a sufficient quality to obtain a decent amount of PCR product (see Figure 7.15).

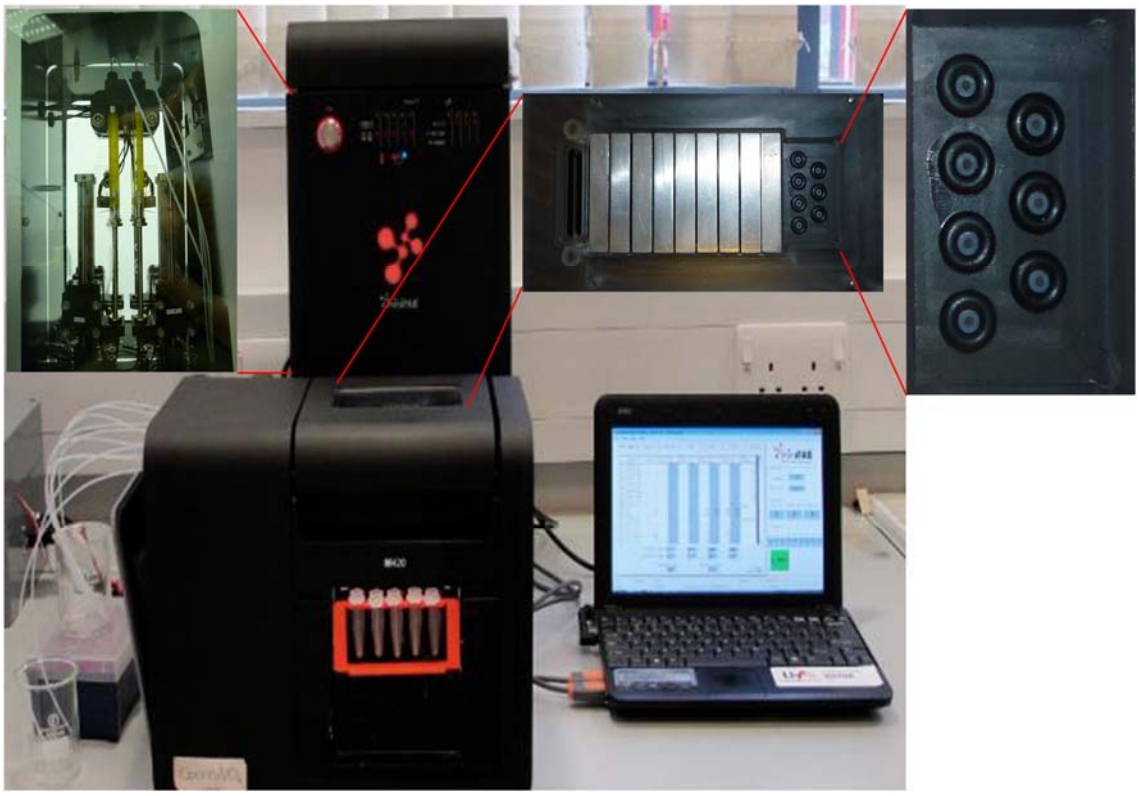

Figure 7.14: The QuantuMDx prototype (The MiniChemLab) with the cassette manifold (right corner) and the syringe pumps (left corner) which are attached to the cassette manifold via plastic tubing.

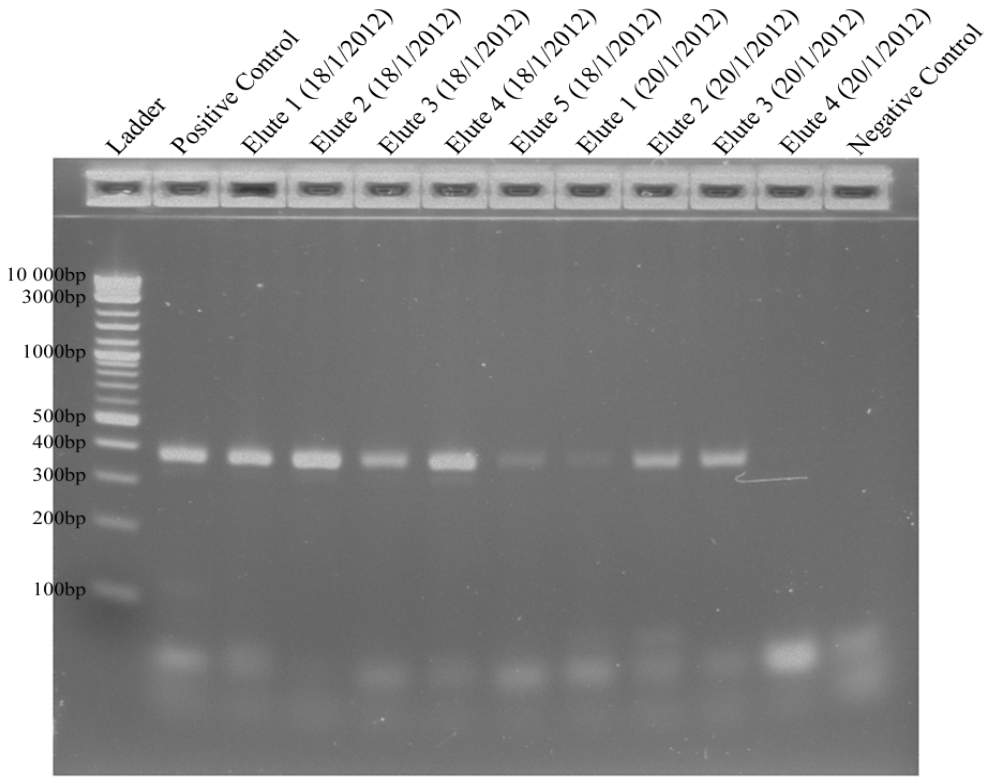

Figure 7.15: Gel image of the PCR results from the blood DNA extraction. DNA was extracted from blood using the DNA extraction cassette. This DNA was collected in elute fractions. To confirm the presence of high quality DNA in the elute fractions BAT26 primers were used to amplify the DNA. The ladder used in this experiment is a Quick-Load 2-log DNA ladder (New England Biolabs).

7.2.2.2. DNA extractions from FFPE tissues using the QuantuMDx DNA extraction cassette

As a control for the DNA extraction of FFPE tissues, use of the Promega ReliaPrep™ FFPE gDNA Miniprep System kit (Promega, Madison, WI, United States of America) was chosen. This kit is a commonly used kit and considered one of the gold standards for the extraction of DNA from FFPE tissues. Prior to performing the experiment using the DNA extraction cassette an experiment was performed to see whether wax curls of a similar size would produce roughly the same amount of DNA. This was done to determine if one could extract DNA from different wax curls of the same size using the DNA extraction cassette and ReliaPrep™ kit and be able to compare the results between the two extraction methods.

Two wax curls of roughly similar size (wax curl 867 and wax curl 902) were extracted using the Promega ReliaPrep™ FFPE gDNA Miniprep System kit. Quantification of the DNA concentration using Quant-iT™ PicoGreen dsDNA Assay (Life Technologies, Carlsbad, CA, United States of America) on an Fluoroskan Ascent

FL (Thermo Scientific, Waltham, Massachusetts, USA) showed that wax curl 867 contained almost 6 times the amount of DNA compared to wax curl 902. Table 7.7 contains the absorbance data and DNA concentrations obtained for this experiment. Figure 7.16 shows the standard curve for the PicoGreen assay.

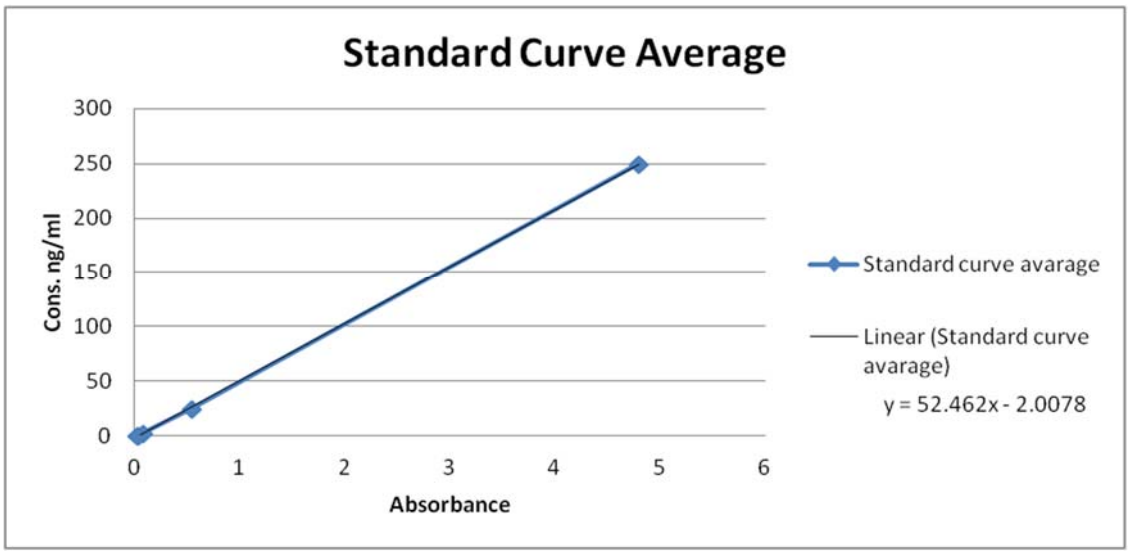

Figure 7.16: Standard curve for the PicoGreen assay used to measure the DNA concentration obtained from the DNA extractions of wax curls 867 and 902.

Sample	Average absorbance	Concentration on plate (ng/ml)	Dilution factor	Concentration of sample (ng/ml)	Initial Volume (µl)	Total yield (ng)
Wax curl 867 DNA	3.53	183	33.3x	6094	30	183
Wax curl 902 DNA	0.673	33	33.3x	1099	30	33

Table 7.7: PicoGreen absorbance readings at 520 nm for wax curls 867 and 902 and the corresponding DNA concentrations.

6.1ng/µl of DNA was obtained from wax curl 867 and 1.1ng/µl of DNA was obtained from wax curl 902. This means that it will not be possible to use wax curls of a similar size to compare yields from different extraction methods. A better method for comparing different DNA extraction methods may be to split the lysate from one wax curl, then extract DNA from the lysate using the DNA extraction cassette and the ReliaPrep™ kit.

Attempts were also made to amplify the DNA obtained from the wax curls 867 and 902. The BAT26 Primers failed to amplify the DNA extracted from the wax curls. One explanation for this may be that the DNA obtained from the wax curls was too fragmented to amplify a 395bp long amplicon. Bioanalyser (Agilent, Santa Clara, CA, United States of America) results confirmed that this may be a possibility. Most of the

DNA obtained from wax curl 867 consisted of 200-1000bp fragments (see Figure 7.17) showing that the genomic DNA obtained from the wax curl is very sheared and the larger the PCR amplicon the less template there will be to start the PCR.

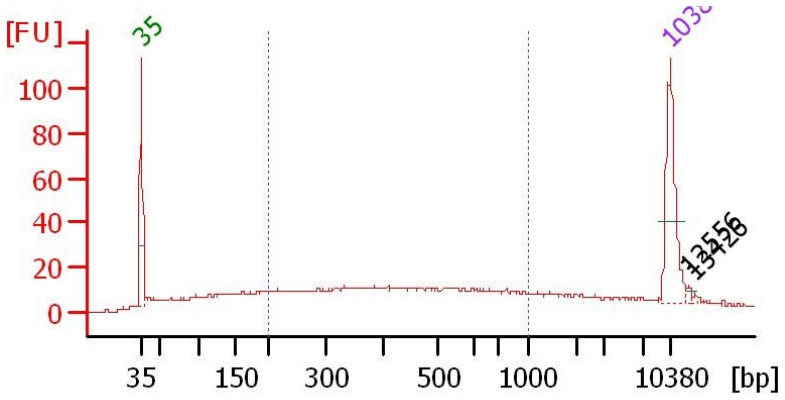

Figure 7.17: Bioanalyser results from the DNA extract obtained from wax curl 867.

Primers for a 150bp amplicon were used to successfully amplify DNA from the wax curl DNA extracts (see Figure 7.18). These results are consistent with the theory that the DNA obtained from the wax curls is very fragmented.

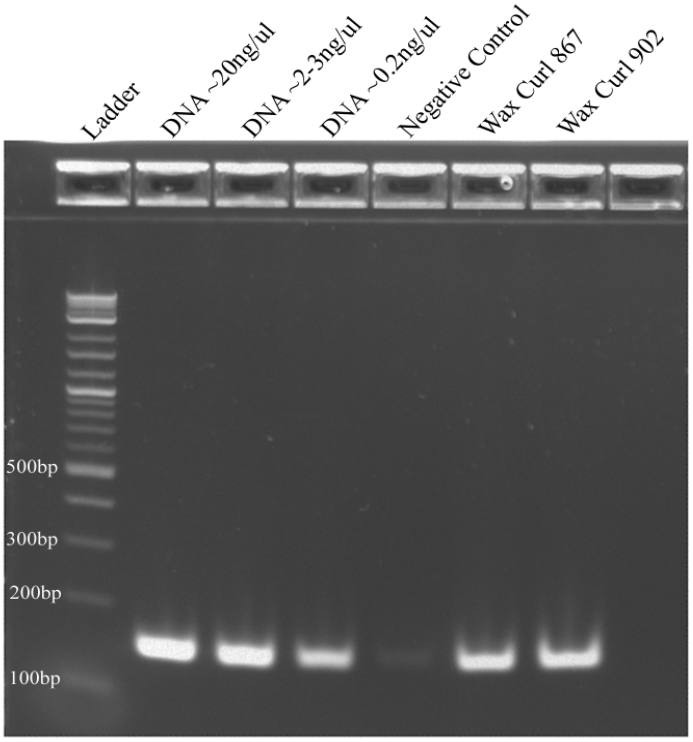

Figure 7.18: PCR amplification of the DNA extract obtained from wax curls 867 and 902. DNA obtained from a blood sample was used as a positive control. The PCR amplification was performed using the CYP2C9 primers which generate a 154bp product.

After the success of extracting DNA from two wax curls using the Promega ReliaPrep™ FFPE gDNA Miniprep System kit, the next steps included determining if DNA could be extracted using the DNA extraction cassette and analysing how these two methods compared to each other. A large wax curl (wax curl number 878) was lysed using the lysis method from the ReliaPrep™ FFPE gDNA Miniprep System kit. 100µl of the lysate was processed using the ReliaPrep™ FFPE gDNA Miniprep System kit and the other 100µl of the lysate was processed using QuantuMDx's DNA extraction cassette (see Figure 7.19). The end product was 40µl of DNA extract from the ReliaPrep™ FFPE gDNA Miniprep System kit and ten 40-50µl fractions from the DNA extraction cassette.

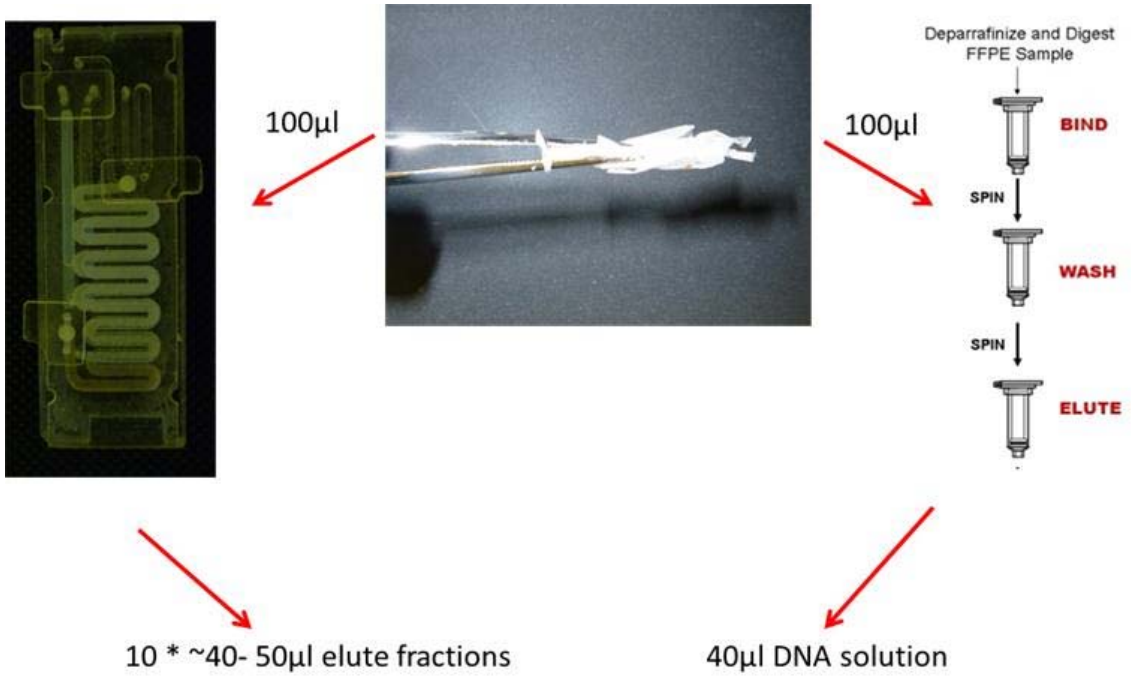

Figure 7.19: Schematic diagram of the DNA extraction of wax curl 878.

The DNA extract obtained both from the Promega spin column and the DNA extraction cassette were analysed using a Quant-iT™ PicoGreen dsDNA Assay on a Fluoroskan Ascent FL. The absorbance readings for these samples can be found in Table 7.8 and the standard curve can be found in Figure 7.20.

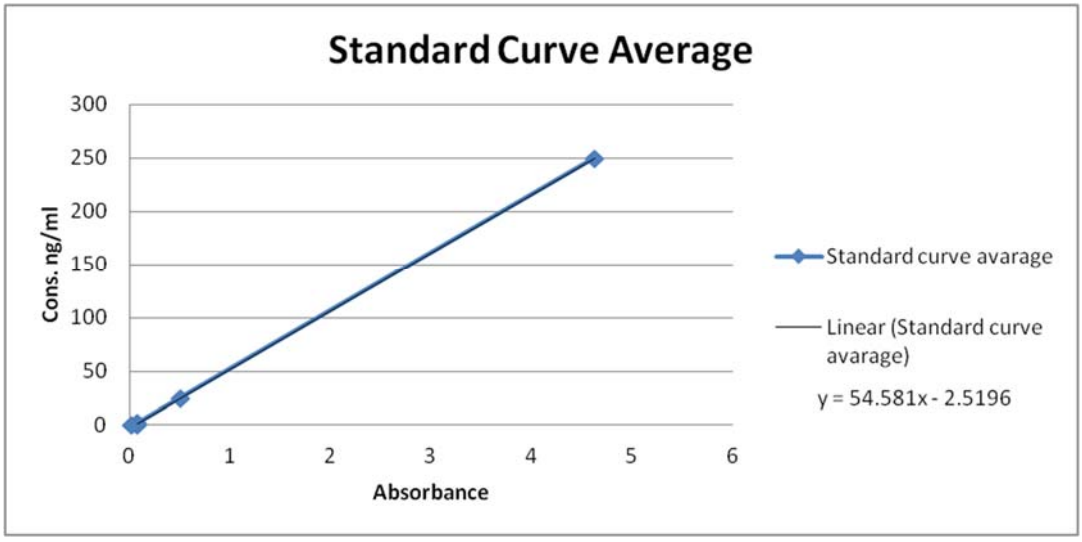

Figure 7.20: Standard curve for the PicoGeen assay showing the correlation between absorbance readings obtained from the Fluoroskan Ascent FL and DNA concentration.

Sample	Average absorbance	Concentration on plate (ng/ml)	Dilution factor	Concentration of sample (ng/ml)	Initial Volume (µl)	Total yield (ng)
Promega sample 10x dilution	1.27	66.6	333x	22.19	40	887.6
Cassette elute1	0.05	0.2	16.6x	0.00	48.4	0.2
Cassette elute 2	0.24	10.6	16.6x	0.18	48.8	8.6
Cassette elute 3	5.98	323.9	16.6x	5.38	49.8	267.7
Cassette elute 4	3.05	163.8	16.6x	2.72	44	119.7
Cassette elute 5	1.23	64.7	16.6x	1.07	43.6	46.8
Cassette elute 6	0.60	30.5	16.6x	0.51	51.6	26.1
Cassette elute 7	0.36	17.1	16.6x	0.28	43.6	12.4
Cassette elute 8	0.23	10.0	16.6x	0.17	47.6	7.9
Cassette elute 9	0.21	9.0	16.6x	0.15	44	6.5
Cassette elute 10	0.16	6.1	16.6x	0.10	43	4.3

Table 7.8: Absorbance values and amount of DNA for the cassette extraction and Promega extraction of wax curl 878. The absorbance values highlighted in red are above the standard curve so any calculations using these are estimates.

After the concentration of DNA had been calculated for each of the elutes from the DNA extraction cassette the amount of DNA in each sample was calculated (see Table 7.8). The DNA recovery rate for the DNA extraction cassette compared to the Promega kit can be found in Table 7.9.

Total amount of DNA from the cassette (ng)	Recovery rate compared to the Promega kit (%)
500.2	56.4

Table 7.9: A comparison of the efficiency of the DNA extraction cassette compared to the Promega kit.

A graphical representation of the DNA output of the DNA extraction cassette can be found in Figure 7.21. This figure shows that most of the DNA exits the cassette in the third elute fraction. The first ~100µl contain very little DNA.

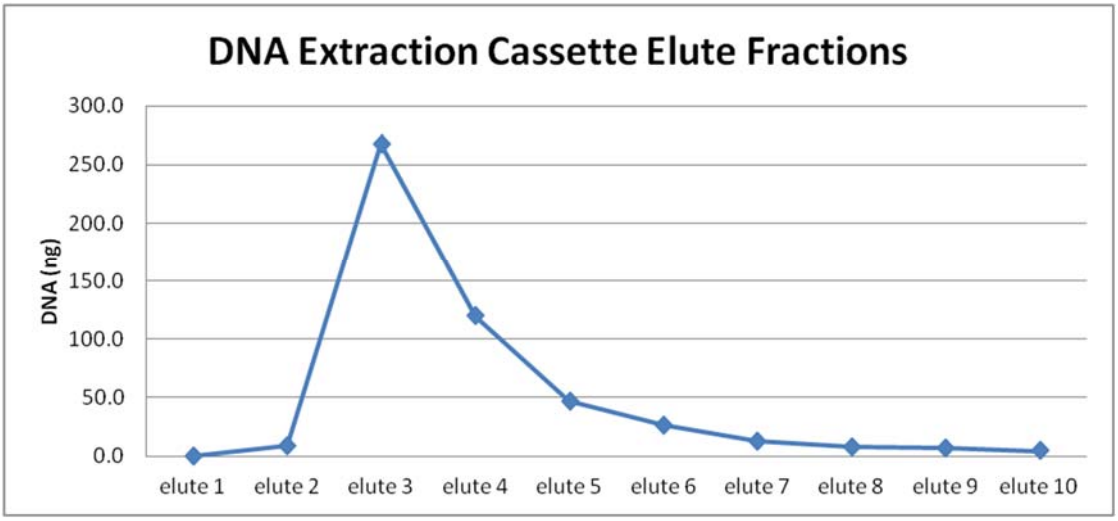

Figure 7.21: The DNA output of the DNA extraction cassette for wax curl number 878.

7.2.2.3. Optimising the QuantuMDx PCR cassette

Before this work was started on optimising the first prototype of QuantuMDx's PCR cassette, PCR amplification had yet to be achieved by QuantuMDx using this cassette. Figure 7.22 shows the PCR cassette and heater layout. PCR experiments were performed on several PCR cassettes in an attempt to optimise the first generation of PCR cassettes. These experiments included changing some of the reagents in the mastermix, the volume of the mastermix to avoid the effects of evaporation, and adding surfactants. The addition of surfactants such as PVP was examined to prevent molecules like Taq polymerase from sticking to the hydrophobic surface of the PCR channels in the QuantuMDx PCR cassette (Kim et al., 2006). BlueJuice tests, which consisted of running 1x BlueJuice (Invitrogen) through the cassettes, were also performed to test the durability of cassettes under different conditions. In the original PCR program the cassettes were pressurised to help maintain a smooth flow of liquid through the PCR cassette and to minimise bubbles. This part of the program needed to be removed to stop the cassettes leaking. The flow rate was shown to still be smooth if the PCR mixture was not too viscous. There was, however, a lot of bubble formation disrupting the liquid column in the PCR tubes. This resulted in lots of small PCR reactions instead of one large PCR reaction. The PCR setup used on the cassette remained a two-step setup throughout all the experiments. Listed below is a short summary of all the cassette experiments performed and the outcome of these experiments (see Table 7.10). Each run on a cassette consisted of 30 PCR cycles.

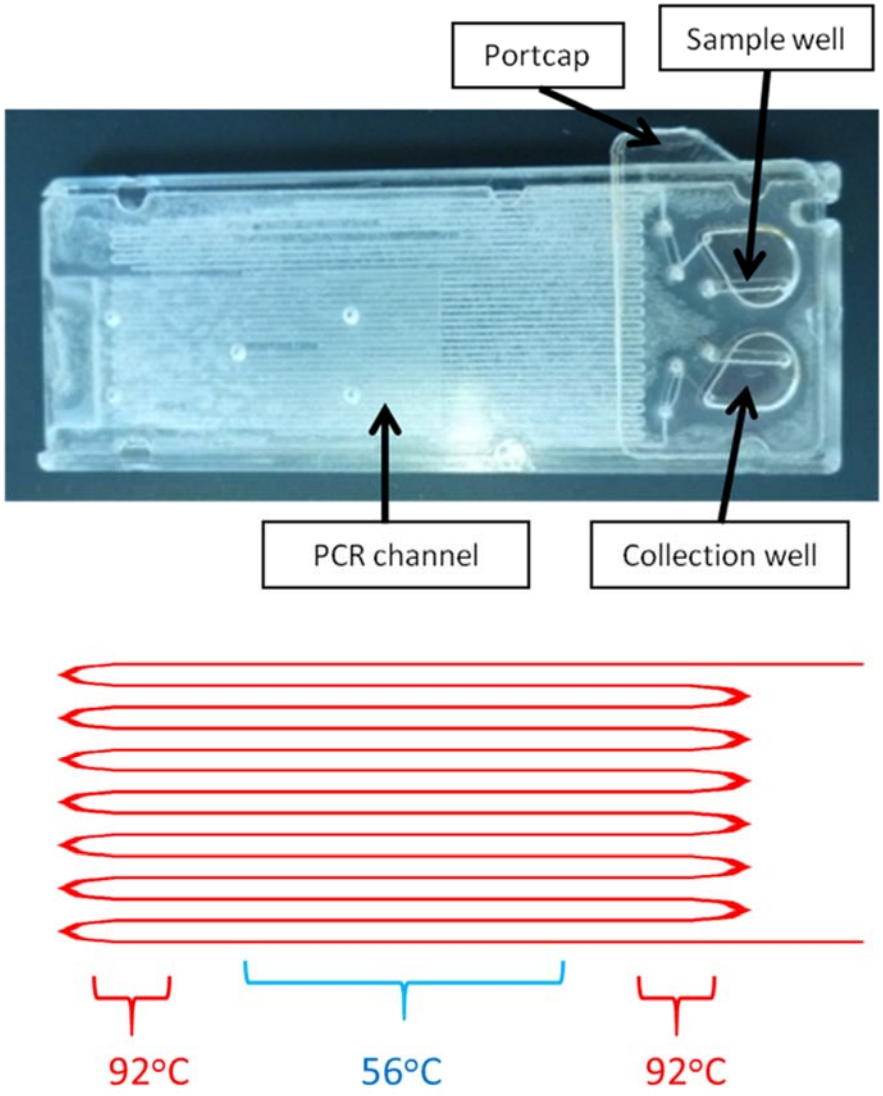

Figure 7.22: PCR Cassette. Panel A: The prototype QuantuMDx PCR cassette. Panel B: Simplified diagram showing how the PCR cassette works when it is placed on the heaters of QuantuMDx's prototype. The PCR channel in one of QuantuMDx's cassettes gives 30 PCR cycles.

Date	Run	Pump rate µl/min	Mastermix Reagents	Input DNA	Surfactant	Pressure	Temp °C	Result
17/2/12	1	10	100µl Mastermix, 10U Taq, BAT26 Primers	gDNA	No Surfactant	Yes	56°C And 90°C	FAIL /leaked
17/2/12	2	10	50µl Mastermix, 5U Taq, BAT26 Primers	gDNA	No Surfactant	Yes	56°C And 90°C	FAIL /leaked
17/2/12	3	10	50µl Mastermix, 5U Taq, BAT26 Primers	gDNA	No Surfactant	Yes	56°C And 90°C	FAIL /leaked
17/2/12	4	10	50µl Mastermix, 5U Taq, BAT26 Primers	gDNA	No Surfactant	Yes	56°C And 90°C	No Product
20/3/12	5	5	BlueJuice Test	No DNA	No Surfactant	Yes	56°C 90°C	FAIL /leaked
20/3/12	6	5	BlueJuice Test	No DNA	No Surfactant	Yes	56°C 90°C	Trial Successful
21/3/12	7	5	50µl Mastermix, 2.5% PVP, 5U Taq, CYP2C9 Primers	H_2O	2.5% PVP 0.1mg/ml BSA, 10 min Soak	Yes	56°C And 90°C	FAIL /leaked
21/3/12	8	5	50µl Mastermix, 2.5% PVP, 5U Taq, CYP2C9 Primers	DNA	2.5% PVP 0.1mg/ml BSA, 20 min Soak	Yes	56°C And 90°C	FAIL /leaked
21/3/12	9	Manual	50µl Mastermix, 2.5% PVP, 5U Taq, CYP2C9 Primers	DNA	2.5% PVP 0.1mg/ml BSA, 10 min Soak	Yes	56°C And 90°C	No Product
23/3/12	10	5	BlueJuice Test	No DNA	No Surfactant	No	56°C 90°C	Trial Successful
23/3/12	11	5	BlueJuice Test	No DNA	No Surfactant	No	56°C 90°C	Trial Successful
23/3/12	12	5	60µl Mastermix, 2.5% PVP, 6U Taq, CYP2C9 Primers	DNA	2.5% PVP 0.1mg/ml BSA, 10 min Soak	No	56°C And 90°C	No Product
23/3/12	13	5	60µl Mastermix, 2.5% PVP, 6U Taq, CYP2C9 Primers	DNA	2.5% PVP 0.1mg/ml BSA, 10 min Soak	No	56°C And 95°C	No Product after 2 runs through cassette
23/3/12	14	5	60µl Mastermix, 2.5% PVP, 6U Taq, CYP2C9 Primers	DNA	No Surfactant	No	56°C And 95°C	No Product
28/3/12	15				2.5% PVP			Leaked
28/3/12	16	5	100µl Mastermix, 2.5% PVP, 10U Taq, CYP2C9 Primers	DNA	2.5% PVP 10 min Soak	No	56°C And 95°C	No Product
28/3/12	17	5	100µl Mastermix, 2.5% PVP, 10U Taq, CYP2C9 Primers	5µl Product from run 16	Same cassette as run 16. No new surfactant was added	No	56°C And 95°C	No Product
28/3/12	18	5	100µl Mastermix, 2.5% PVP, 10U Taq, CYP2C9 Primers	5µl Product from run 17	Same cassette as run 16 and 17. No new surfactant was added	No	56°C And 95°C	No Product. Lots of primer dimers
29/3/12	19	5	100µl Mastermix, 1% PVP, 12.5U Taq, CYP2C9 Primers	1µl 100:1 Amplicon solution	0.1mg/ml BSA, 3 hour Soak	No	56°C And 92°C	Good product
29/3/12	20	5	100µl Mastermix, 1% PVP, 12.5U Taq, CYP2C9 Primers	5µl Product from run 19	Same cassette as run 19. No further surfactant was added	No	56°C And 92°C	Product + Lots of primer dimers
29/3/12	21	5	100µl Mastermix, 1% PVP, 12.5U Taq, CYP2C9 Primers	5µl Product from run 20	Same cassette as run 19 and 20. No new surfactant was added	No	56°C And 92°C	Product + Lots of primer dimers

Table 7.10: List of PCR cassette optimisation experiments that have been performed.

The purpose of these PCR cassette experiments was to show that DNA amplification is achievable on the prototype PCR cassettes. The protocol used to amplify DNA using the PCR cassette is listed in methods section 2.3.1. The result of PCR cassette runs 19, 20 and 21 are shown in Figure 7.23. Lane 2 contains 20µl of the amplicon solution which was used as the input DNA in this experiment and lane 3 shows this amplicon solution diluted to the concentration present in the mastermix. Lane 6 contains the PCR product obtained after one pass through the PCR cassette. Although the band in this lane is dim compared to the positive control in lane 4 it shows that DNA amplification in the microfluidic channels has been achieved. Subsequent passes through the cassette using the PCR product from the previous experiment as template DNA favoured the amplification of primer dimers over the PCR product (see lanes 8 and 10, Figure 7.23). The results of these experiments show that it is possible to achieve PCR amplification using QuantuMDx's prototype cassette. However the fact that detectable PCR amplification was only achieved using diluted PCR product as a template highlighted that a lot more optimisation work was needed before the PCR cassette was preforming well enough to be integrated into a point of care device.

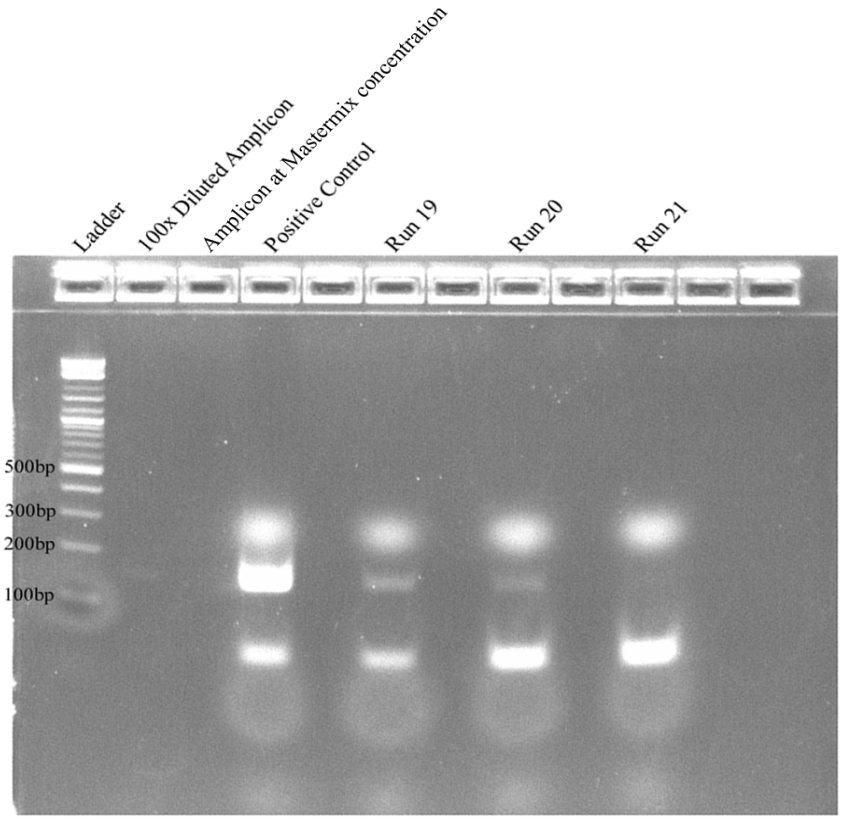

Figure 7.23: Gel image from the PCR cassette experiment. Lane 2 contains 20µl of the amplicon solution of which 1µl was used as the input DNA for run 19. Lane 3 contains the amplicon solution diluted to the same concentration of DNA as was present in the mastermix for run 19 prior to PCR amplification on the PCR cassette. Run 20 used 5µl of product from run 19 as input DNA. Run 21 used 5µl of product from run 20 as input DNA. Both run 20 and 21were performed on the same PCR cassette as run 19.

7.3. Discussion

7.3.1. The feasibility of a sequencing based MSI test using short repeats

Using the panel of homopolymers sequenced in this chapter it was possible to define thresholds that separated all MSI-H and MSS tumours. This panel, therefore, shows good potential for the use as an MSI test for differentiating between MSI-H and MSS tumours. Different thresholds were assessed for their ability to distinguish between MSI-H tumours and MSS tumours. In the first instance, deletion frequency thresholds were set which minimised the number of misclassified homopolymers for each homopolymer length. A homopolymer was defined as misclassified if it was defined as stable in an MSI-H tumour (a false negative error) or unstable in an MSS tumour (a false positive error). The results of this were thresholds where there were a maximum of 3 repeats classed as unstable in any MSS tumour and at least 5 unstable repeats in all the MSI-H samples. For an MSI test, these thresholds could be used with a cut-off of 4 or 5 unstable repeats to call a tumour unstable. This would allow correct classification of all MSI-H and MSS tumours sequenced for this chapter. However, a larger difference between the number of repeats classed as unstable in the MSI-H samples and the MSS samples was achieved when thresholds were adjusted to accommodate a false positive error being 1.5x more costly than a false negative error. With these adjusted thresholds, the maximum number of repeats classed as unstable in a MSS tumour was two, while there was still a minimum of 5 unstable repeats for every MSI-H tumour. For an MSI test, these new thresholds could be used with a cut-off of 3, 4 or 5 unstable repeats to call a tumour unstable. In this case it might be best to go for a cut-off of 4 unstable repeats. This cut-off would allow for more variation in the number of repeats being classified as unstable in both MSI-H tumours and MSS tumours than seen in this chapter before a tumour was misclassified. 28 MSI-H tumours and 30 MSS tumours is a small samples size for testing a panel of repeats. Having a cut-off that allows for more variation in both MSS and MSI-H tumours would be beneficial if the panel is to be used in other tumours, for example, as a routine MSI test.

It may however be better to increase the thresholds so that there are few or no repeats being classed as unstable in the MSS samples. If thresholds are set so that unstable repeats are expected in the MSS samples then there is the risk that in some tumours the numbers of repeats classed as unstable can accumulate without MSI necessarily being the

reason for a large number of repeats being classed as unstable. The sensitivity of the panel of repeats could always be increased by adding more repeats to the panel. It is therefore unnecessary to increase sensitivity by using low thresholds and risk losing specificity. Individual repeats being classed as unstable in MSS tumours is, therefore, more of a problem than individual repeats being classed as stable in MSI-H tumours. The thresholds for calling a repeat unstable can be adjusted so that false positive errors are >5x more costly than a false negative error which means that no repeats are classified as unstable in any of the MSS samples and there are still at least two unstable repeats in any MSI-H tumour. For an MSI test, these thresholds could be used with a cut-off of 1or 2 unstable repeats to call a tumour unstable. It might however be best to use a cut-off of 2 unstable repeats rather than 1 because a low level of microsatellite instability can occur in MSS tumours. For example Yoon et al. (2013) found that mononucleotide repeats with deletions occurred MSS gastric cancer cell lines, but at a lower frequency than in MSI-H sample. Allowing for the odd unstable repeat in a MSS sample would, therefore, be sensible.

It could be beneficial to add more markers to the panel or exchanging some of the shorter markers for longer more unstable ones because for repeat sizes of 8bp-12bp the number of unstable samples detected by each marker generally increased with repeat length. This would allow a cut-off of more than two unstable markers to be used for calling MSI-H. With thresholds set so that no repeats are classed as unstable in the MSS tumours and using two unstable markers as a cut-off for calling instability there is a large risk that the panel of repeats will not be able to cope well with MSI-low samples. Increasing the cut-off to three or more unstable repeats for calling MSI-H tumours cannot be done with the current panel without miss classifying two of the 28 tumours MSI-H tumours giving a false negative rate of 7% for the set of tumours used in this chapter.

There was an excess of homopolymers showing significant allelic bias among MSI-H samples compared to MSS samples. Allelic bias could therefore be used as an indication of whether a variant is a real mutation or the result of sequencing and PCR error. Incorporating allelic bias into the MSI assay was attempted by adding extra points for any unstable repeat with a statistically significant allelic bias. An extra 0.5 points for a significant level of allelic bias was chosen because this gives extra value to an unstable repeat with allelic bias, but one polymorphic repeat would not be enough to misclassify a tumour if a cut-off of 2 points was used to classify a tumour as MSI-H. All 18 markers have no registered polymorphisms as of dbSNP build 173 and therefore should, in theory,

be monomorphic, but there is always the possibility of polymorphisms which have yet to be discovered. For repeats with allelic bias there is also the potential of using the allele without the deletion as an internal control to determine the background PCR and sequencing error for the homopolymer. The usefulness of allelic bias is however limited to repeats with a neighbouring heterozygous SNP. If the SNP is homozygous it is not possible to distinguish the two alleles.

7.3.2. The prospects for QuantuMDx's DNA extraction cassette

QuantuMDx's microfluidic DNA extraction cassette was initially tested on human blood to get an idea of the efficiency of the cassette and to gain experience using the DNA extraction cassette and the MiniChemLab, which was QuantuMDx's prototype. The results from this test were encouraging so the DNA extraction cassette was used to extract DNA from paraffin wax embedded tumour tissue. DNA extraction from paraffin wax embedded tumour tissue was successfully achieved using the lysis method taken from Promega ReliaPrep™ FFPE gDNA Miniprep System kit and DNA extraction using the DNA extraction cassette. The DNA recovery rate for the DNA extraction cassette compared to the Promega ReliaPrep™ FFPE gDNA Miniprep System kit was 56%. The DNA obtained using this extraction method was of a high enough quality for PCR amplification. This is quite good considering that the method of tissue lysis has not been optimised for the DNA extraction cassette and the current filter in the cassette was optimised for blood, not tissue. These experiments indicated that the DNA recovery rate for this cassette could be further improved. However this was only the result of one experiment with no replications. Ideally several replications of this experiment should have been performed, but at the time the DNA extraction and PCR cassette experiments were being performed, QuantuMDx experienced cassette supply problems which limited this aspect of the thesis preparation.

The preliminary results show that the DNA extraction method used by QuantuMDx can be adapted for using FFPE tissues as an input material. Being able to extract DNA from FFPE tissues is of importance for validating a MSI test the QuantuMDx device, because archived FFPE material of known MSI status would be relatively easily available for test validation purposes. QuantuMDx's plan is that the first MSI diagnostic device available on the market will be designed to use FFPE tissues as the input material. This is to allow the MSI test on the Q-POC to easily fit in with the

current practice of fixing tumours just after they have been removed from a patient. After the Q-POC MSI test has become well established, the next step will be to provide a Q-POC MSI test as part of a larger cancer test that can be used in the operating theatre on fresh tissue. A test like this would provide immediate information on MSI status as well as other cancer markers during the operation itself. This could give clinicians information on the importance of polyp resection during endoscopy or help assess a tumour's aggressiveness during surgery. It would also be possible to test resection margins, avoiding either the necessity for further surgery or the inclusion of a large operative margin.

7.3.3. The prospects QuantuMDx's PCR cassette

The initial optimisation of the first generation PCR cassette was successful but further optimisation was indicated. It took quite a while to achieve PCR amplification using the cassettes. Some of these problems were due to a manufacturing fault which resulted in the cassettes being prone to leakage. The other main reason making it difficult to obtain a PCR product from the PCR cassettes was due to problems with coating the cassettes with surfactants. The polycarbonate surface of the cassettes attracts the Taq polymerase enzyme to the hydrophobic surface of the cassette where it is liable to stick and no longer be able to catalyse the PCR reaction (Kim et al., 2006). Using BSA as a surfactant reduces some of this problem, but a better solution was needed. Otherwise the PCR conditions such as temperature and amount of time spent in the two different temperature zones needed to be optimised. In my experiments, a cassette PCR took 33minutes. This was too long for a point of care setting and led to a further and ultimately successful design programme by other colleagues.

Since 2012 when my attachment to the Q-POC team ended, major improvements to their cassette PCR have made multiplex assays feasible, using gDNA as the input. Due to the improvements QuantuMDx have made to their hardware since my contributions, the QuantuMDx Q-POC platform is now starting to look like a promising platform for the MSI test developed in this PhD project.

Future work will aim to reduce the cost of a sequencing based MSI test. Multiplexing will be needed to reduce the cost and effort needed to produce the amplicons for an MSI test. In the future the aim will be to produce primers which will allow for

multiplexing so many or all the repeats can be produced in a single PCR reaction. Multiplexing will be needed whether the final platform consists of using Illumina sequencing or the test is being carried out on QuantuMDx's Q-POC device.

For the Illumina system library preparation is also costly. To reduce this cost our group are currently producing primers with adapter overhangs which will allow Illumina sequencing primers and indexes to be added using a few cycles of PCR on a standard thermocycler. For the QuantuMDx Q-POC platform, library prep is unnecessary as each tumour will be analysed separately. The rapid turn around time of the Q-POC will still allow several tumours to be analysed on the same device over the course of a working day. The aim is that the Q-POC will be able to do testing from sample to result in as little as 15min (Burn, 2013). As well as testing for MSI, other cancer biomarkers such as K-RAS and BRAF will be investigated at the same time. Using QuantuMDx's Q-POC platform the price of an MSI and cancer biomarker test could plummet as low as $20 (Burn, 2013).

7.3.4. Conclusions

A comparison between Quantumdx's DNA extraction cassette and one of the gold standard kits available on the market, the Promega ReliaPrep™ FFPE gDNA Miniprep System kit, showed that the DNA recovery rate of the DNA extraction cassette compared to the Promega kit was 56%.

Experimentation with different surfactants to avoid DNA denaturation on the PCR channel surface highlighted some of the challenges of microfluidics based PCR and data indicating a possible solution was obtained. This project enabled the first QuantuMDx PCR cassettes to function, thus showing that the PCR system was viable.

It was possible to distinguish between the 28 MSI-H tumours and the 30 MSS tumours using the final panel of 18 homopolymers. This suggests that this panel or an extended version of this panel of repeats could be used as sequencing based test for diagnosing MSI. Allelic bias analysis will be a useful adjunct in a next generation sequencing based MSI assay to help differentiate between genuine mutations and sequencing artefacts.

Chapter 8. General discussion and future work

8.1. General discussion

Identifying patients with Lynch Syndrome is important because early intervention can save lives. The use of traditional family history based guidelines for identifying patients with Lynch Syndrome results in many being missed. Molecular testing for all colorectal and endometrial cancers is therefore being recommended (Vasen et al., 2013, Canard et al., 2012, Mills et al., 2014, Julie et al., 2008). MSI testing all colorectal cancers would also be advantageous because new treatments which target MSI-H tumours are being discovered. For example pembrolizumab which blocks the cells' Programmed Death 1 (PD-1) pathway increasing the immune response against cancer cells has been shown to be effective in MSI-H tumours but not MSS tumours (Le et al., 2015). 90% of the patients with a MSI-H colorectal cancer who were given the pembrolizumab monoclonal anti-PD-1 antibody treatment responded to this treatment in this landmark study. To cope with the increase in tumours being MSI tested if a "test all approach" is adopted, it would be advantageous to consider high throughput screening approaches to test for MSI. In the work presented here, next generation sequence typing of short mononucleotide has been developed as a method to identify microsatellite unstable tumours.

In the first results chapter, it has been demonstrated that MSI could be detected using short mononucleotide repeats and an amplicon sequencing approach with the Illumina MiSeq as a sequencing platform. For the mononucleotide repeat lengths investigated, data showed that susceptibility to MSI increases with repeat length, but so does sequencing and PCR error. This is consistent with what has previously been reported in the literature (Fazekas et al., 2010, Flores-Renteria and Whipple, 2011). The longer 10bp-12bp were found to be unstable in more tumours than the shorter repeats. However, in chapters 3 and 5 there were two tumours where the shorter 7bp-10bp repeats were more unstable than the longer 11bp-12bp repeats (U096 tumour in chapter 3, and U312 tumour in chapter 5). This suggested that a panel of repeats consisting of a range of repeat lengths from 7bp-12bp might be the best approach for identifying MSI-H tumours, because it is not known how all MSI-H tumours behave in respect to instability in different repeat lengths. A broad approach with a range of different repeat sizes was considered preferable.

To identify a large number of potential markers for distinguishing between MSI-H and MSS tumours, whole genome sequences were mined for highly unstable 7bp-12bp mononucleotide repeats. This is to our knowledge the first study to analyse whole genome sequences to identify highly variable repeats for panel based MSI identification. The indel frequencies in 7bp-12bp mononucleotide repeats in whole genome sequences of MSI-H CRCs were analysed using matched normal tissue and MSS stable CRC whole genome sequences as controls. One of the limitations of this analysis was the low read depth of the genome sequences available for analysis. For this reason, the reads for each group (MSI-H, matched normal for the MSI-H samples, and MSS samples) were pooled and only repeats with ≥20 reads in each group were analysed. Despite the low read depth of the whole genome sequences used, it was immediately apparent that the MSI-H tumours had a different indel distribution in 7bp-12bp mononucleotide repeats compared to controls. The normal tissue samples and MSS tumours showed the same indel distributions, which suggests that the differences seen in the MSI-H tumours were caused by microsatellite instability. There was a greater excess of deletions in the MSI-H CRCs compared to insertions. This suggested that deletions are more indicative of mismatch repair deficiencies in CRC than insertions for 7bp-12bp mononucleotide repeats.

In chapter 7, the number of mononucleotide repeats was refined down to a panel of eighteen 8bp-12bp repeats consisting of repeats taken from the literature and repeats identified through the whole genome analysis. This panel of repeats was sufficient to distinguish between MSI-H and MMS tumours with a 100% sensitivity and specificity in a sample of 58 tumours (28 MSI-H tumours and 30 MSS tumours) using a range of different deletion frequencies as thresholds and different numbers of unstable repeats to classify tumours as unstable, demonstrating the robustness of the marker panel. The most practical set of thresholds were the ones that allowed no false positive markers in the MSS tumour group. The reason for this is that if thresholds are set so that unstable repeats are expected in the MSS samples then there is the risk that in some tumours the numbers of repeats classed as unstable can accumulate. Using these thresholds there were 2-17 unstable repeats in each of the MSI-H tumours. For an MSI test, a cut-off of 2 unstable repeats to call a tumour MSI-H should be used with this system because the odd unstable repeat can be found in MSS tumours (Yoon et al., 2013).

There were no polymorphisms as of dbSNP build 173 for the 18 markers of the final MSI testing panel, and no repeats showed potential polymorphism in the MSS tumours used to test these repeats. All repeats should therefore be monomorphic, which

means that the panel of repeats can be used without the need for a comparison between tumour and normal tissue. However, it is possible that polymorphisms in some of these repeats may be discovered in the future. This is another reason why a cut-off of 2 unstable repeats for calling a tumour MSI-H would be wise. It is however conceivable that it may not be possible to define a clear cut-off for identifying all MSI-H tumours because at the lower end of the spectrum there may be a continuum of instability levels between MSI-H, MSI-L and MSS tumours.

Another advantage of the MSI test described in this thesis is that the test can be automated, reducing the need to use valuable staff time to determine the MSI status of tumours. The monomolecular nature of next generation sequencing provides a quantitative approach to measuring deletion frequencies allowing automation. The approach of using deletion frequencies as thresholds for calling unstable markers lends itself well to automation, in contrast to the current Promega MSI test where fragment analysis traces are subjectively analysed.

Recently, there have been a couple of next generation sequencing panel based MSI test approaches published. These tests are the first next generation sequencing MSI tests to use a small panel of repeats and amplicon sequencing, which is also the strategy used in this thesis. This highlights the current relevance and importance of the work.

Gan et al. (2015) used a series of 5 long mononucleotide and dinucleotide markers (BAT25, BAT26, BAT34c4, D18S55, D5S346) in their next generation sequencing based MSI test. Tumours were defined as MSI-H if the repeat length with the most reads had a deviation of ≥2bp and ≥4bp for mononucleotide and dinucleotide repeats respectively compared to the repeat length with the most reads in matching normal tissue. This method does have the disadvantage that matching normal tissue is needed. This assay may also have a reduced sensitivity compared to the currently popular Promega fragment analysis assay for tumours such as the U312 tumour analysed in chapter 5. The Promega pentaplex assay identified this tumour as MSI-H because of extra stutter peaks, but the highest peak was the same in the normal and tumour tissue for all 5 markers. In this thesis, short 7bp-12bp repeats were sensitive enough to identify MSI in this tumour, which could indicate that the methodology used in this thesis would have a higher sensitivity than the method proposed by Gan et al. (2015).

A paper by Hempelmann et al. (2015) has suggested a similar approach to MSI testing as the one devised in this work. This showed that it can be cost effective to use

amplicon sequencing on a platform such as the Illumina MiSeq to identify MSI-H tumours. However, Hempelmann et al. (2015) used 11 mononucleotide repeats of 12bp-28bp in length as their MSI marker panel, rather than the short (7-12bp repeats) analysed here. The MSI test they have developed (MSIplus) is performed using the software mSINGS developed by Salipante et al. (2014). This software required the input of a set of MSS reference samples to establish baseline values for the run. Samples are subsequently classed as MSI-H if the number of variant read lengths exceed the mean number of read lengths + 3x the standard deviation as calculated using the set of reference samples. Hempelmann et al. (2015) analysed a total of 81 tumours using MSIplus. Interpretable results were obtained for 96% of the tumours leaving 3 tumours which could not be typed. For the tumours with interpretable results a 97.1% sensitivity and 100% specificity was achieved. This assay has the advantages that the PCR reactions are performed in multiplex and Illumina adapters are added using a second PCR reaction priming off the amplicon specific primers. This helps reduce the cost of the assay. Other advantages include; that primers for mutations in *KRAS*, *NRAS*, and *BRAF* are included in the assay, so these can be typed using the same MiSeq run, and mSINGS operates without the need for matched normal tissue. Disadvantages include the use of long microsatellites (12bp-28bp), which means high levels of PCR and sequencing error, and as a result the assay requires the establishment of assay specific baseline values for each sequencing run.

Advantages of the MSI assay presented in this thesis compared to MSIplus is the use of shorter repeats which means less sequencing and PCR error, and sequencing error is further reduced through paired end read sequencing and the use of an indel caller that only analyses concordant paired end reads. For the panel of repeats devised in this thesis the aim will be to use the same threshold values for calling instability for different runs, assuming the same methodology is used. In the assay presented in this thesis it was also possible to differentiate between all MSI-H and MSS tumours; in contrast some of the tumours could not be typed by MSIplus and one tumour was misclassified. On the other hand, a larger panel of tumours was analysed using MSIplus compared to the numbers analysed in this thesis. This work has been limited by the availability of MSI-H tumour samples. Challenges have included obtaining tumour samples and obtaining DNA of usable quality from tumours preserved by formalin fixing and paraffin embedding.

Short mononucleotide repeats could also be used to investigate the clonal evolution of MSI-H tumours. In chapter 5 short repeats were used to show that the latter

of two tumours derived from patient U179 was unlikely to be a reoccurrence of the first tumour. This is because several repeats found to be unstable in the tumour extracted in 2003 were stable in the tumour extracted in 2012. In chapter 6 I used twenty 8bp-14bp repeats to evaluate the clonal composition of three MSI-H tumours. There was evidence of different sub-clones in all three MSI-H tumours analysed. For each tumour 8 different biopsies were analysed using a biopsy of normal mucosa from the same patient as a reference point. For the tumour PR17848/14 there was evidence from three separate repeats to support the existence of a clonally distinct group of cells in the 9 o'clock needle biopsy region of the tumour. Results for tumour PR51869/13 suggested that there might be a clonally distinct group of cells in the 3 and 9 o'clock scalpel biopsy region of the tumour and a different clonally distinct group of cells present in the 12 o'clock needle biopsy region of the tumour. In contrast with the results from tumour PR17848/14 the evidence for each of the different clonally distinct group of cells identified was only substantiated by one repeat. Results from several different repeats for the third tumour, PR10654/14, suggested there were distinct sub clones in at least 4 regions of the tumour. It is possible that more variation could have been detected in this tumour if a greater number of biopsies had been used. Plans have been made to study these three tumours further by using immunohistochemistry to identify different morphological regions which can be biopsied and sequenced.

We believe this is the first time that short mononucleotide repeats have been used to assess the clonal evolution of MSI-H tumours. The addition of neighbouring SNPs allowed individual alleles to be analysed separately. The added information provided by being able to distinguish between the two alleles of repeats meant it was often easier to identify multiple deletion events in repeats with a heterozygous SNP, making neighbouring SNP a valuable asset.

8.2. Future work

The panel of repeats outlined here has subsequently been expanded upon. This work has including calculating threshold values for the new repeats. The hope is that this will increase the minimum number of unstable markers detected in MSI-H samples and result in larger difference between MSI-H and MSS samples. Future work consists of validating the improved marker panel and on a larger number of tumours. The extended marker panel is currently being tested on a cohort of 220 colorectal tumours of which

roughly 140 are MSI-H. This work will enable the validation of markers and thresholds devised in this thesis to be assessed and refined if necessary.

Instead of the costly Nextera XT library prep, a two stage PCR approach is being used where amplicon specific primers are used to amplify the targets of interest. Overhangs consisting of partial Illumina sequencing adaptors are attached to the amplicon specific primers allowing the Illumina sequencing adaptors from the Nextera XT index kit to be added in a second PCR reaction. All amplicons for each individual tumour will be pooled prior to performing the second PCR amplification step, saving further time and cost. Using this sample prep method, a cost of ~£26 pounds per sample can be achieved for a MiSeq run containing 96 samples (see Appendix Table 9.4 for a breakdown of costs). Because the need for the Nextera XT kit has been eliminated 300bp amplicons are no longer necessary and amplicons have been redesigned to a length of ~100bp. This has allowed better amplification from FFPE tissue. Further refinement to this system will include optimising a multiplex PCR so that each mononucleotide repeat need not be amplified separately. This will make the assay more cost effective and reduce the turnaround time from sample receipt to sequencing. A further reduction in the cost per sample can be achieved by increasing the number of samples per MiSeq run once the required read depth per amplicon has been established for this new library prep method.

Once the assay has been optimised using the 220 tumours described above, the sequencing based MSI assay will be run in parallel with MSI testing performed by the Northern Genetics Service. This will enable the new method to be trialled before being put into routine practice.

The MSI test will be developed and trialled using Illumina as a sequencing platform. This is one possible platform for the MSI test, but another possibility is the QuantuMDx Q-POC platform. The assay has been designed so it is compatible with QuantuMDx's technology, and a future aim is to transfer the assay to the QuantuMDx platform after the development of this platform is complete. If the QuantuMDx device lives up to expectations, the MSI assay can potentially be performed on a chip with other cancer biomarkers at a cost under £20 (Burn, 2013). It has also been estimated that the QuantuMDx device may cost as little as £500 and the turnaround time for assays may be as little as 15min (Burn, 2013). This could enable the test to be performed in the operating theatre, giving information about prognosis during operations. This could allow surgeons to make decisions regarding the operation based on the tumour's genetic profile. Lymph

nodes and resection margins could also be tested during surgery enabling the removal of more tissue if necessary and eliminating the need for another surgery, which is often needed if resection margins test positive for tumour content.

Running the MSI marker panel in parallel with other tumour markers in the same MiSeq run, or on the same QuantuMDx chip, would eliminate the need for many separate diagnostic tests and reduce testing costs. In the future the MSI test will also be automated eliminating the subjective analysis currently needed to analyse MSI test fragment analysis traces. This will save man-hours and cost. The reduced cost could potentially allow the screening of all colorectal cancers in countries that currently rely on the Amsterdam II Criteria and revised Bethesda Guidelines to identify patient with Lynch Syndrome. The revised Bethesda guidelines and Amsterdam II Criteria fail to identify a significant number of Lynch Syndrome patents (Canard et al., 2012, Mills et al., 2014, Perez-Carbonell et al., 2012). Screening of all colorectal cancers for Lynch Syndrome will save lives, ensuring appropriate surveillance and identifying relatives who have also inherited a mismatch repair mutation so they can be monitored. Gene carriers can be offered prophylactic medication (like aspirin) to reduce Lynch Syndrome tumour rates. Molecular testing for mismatch repair defect in all colorectal and endometrial tumours for the identification of Lynch Syndrome patients is supported by the literature (Vasen et al., 2013, Canard et al., 2012, Mills et al., 2014, Julie et al., 2008).

Another reason for MSI testing all colorectal cancer would be to enable the future use of specific treatments targeted at MSI-H tumours. The landmark study of Le et al (2015) is likely to result in a major shift towards identification of MSI high tumours in order to deploy PD1 blockade as a primary intervention. Hopefully the work described in this thesis will help bring the promise of these new agents to early fruition.

Chapter 9. Appendix

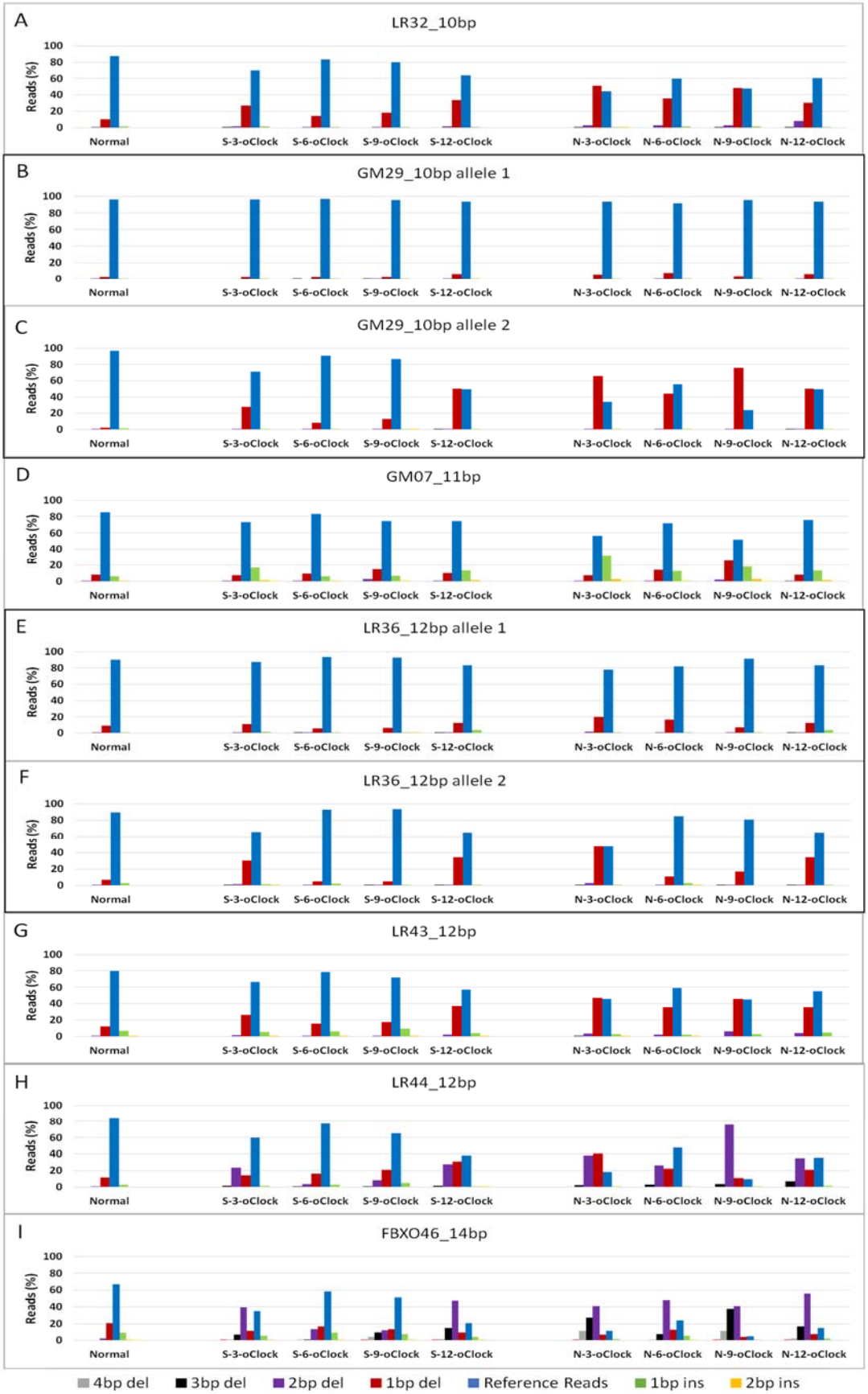

Figure 9.1: Repeats for tumour PR17848/14 which were not included in chapter 6. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat.

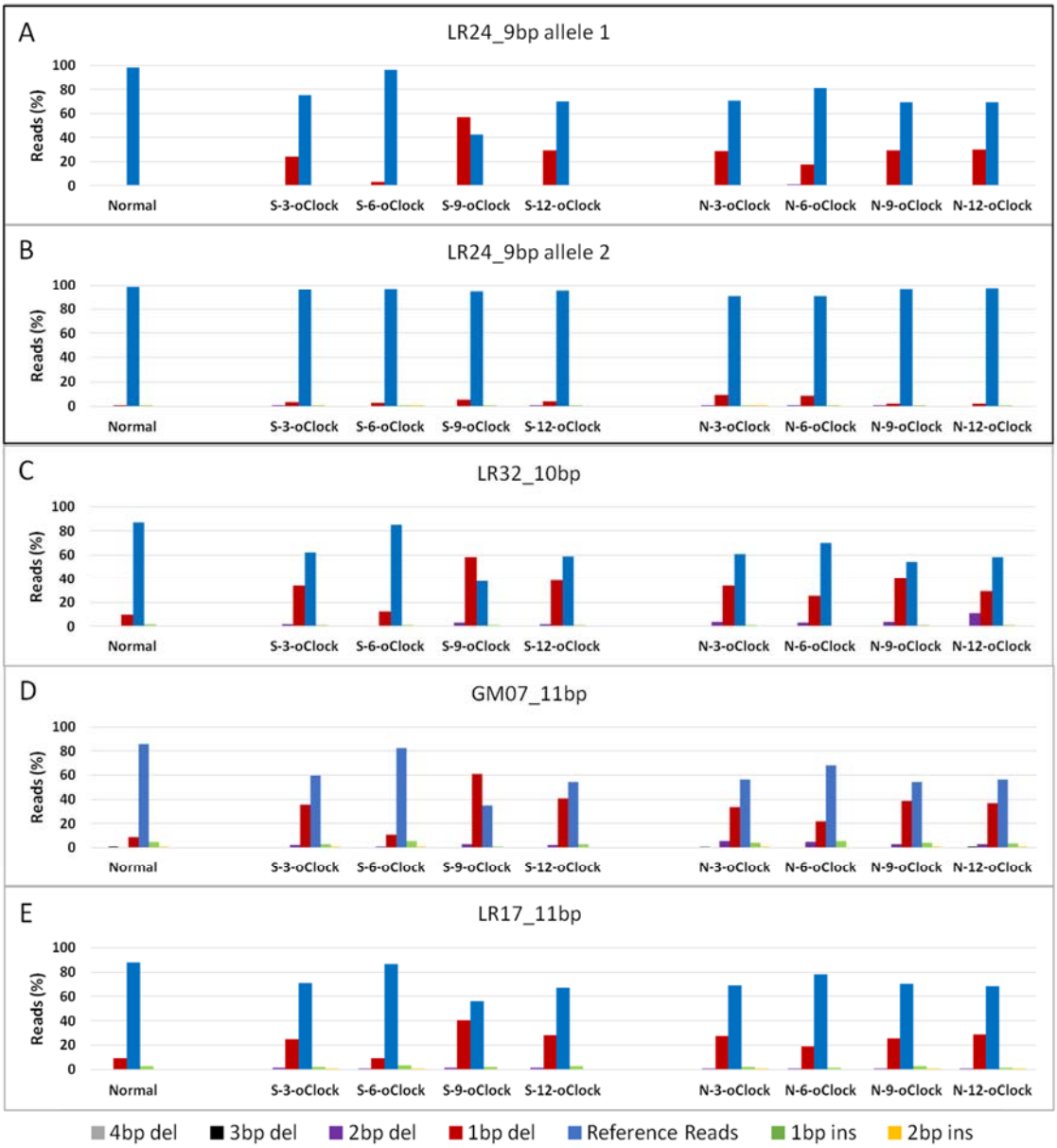

Figure 9.2: 9bp-11bp repeats for tumour PR51896/13 which were not included in chapter 6. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat.

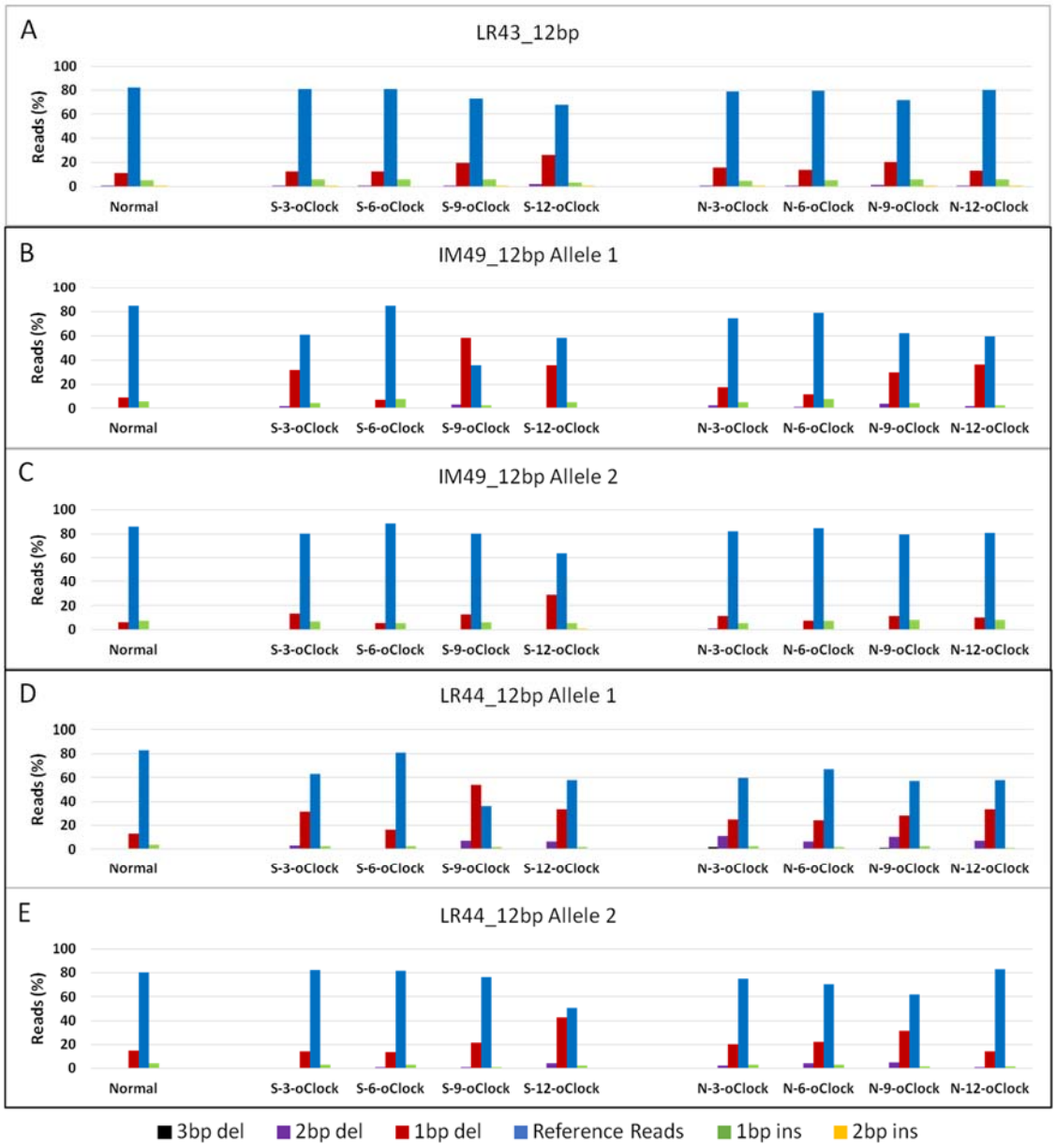

Figure 9.3: 12bp repeats for tumour PR51896/13 which were not included in chapter 6. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat.

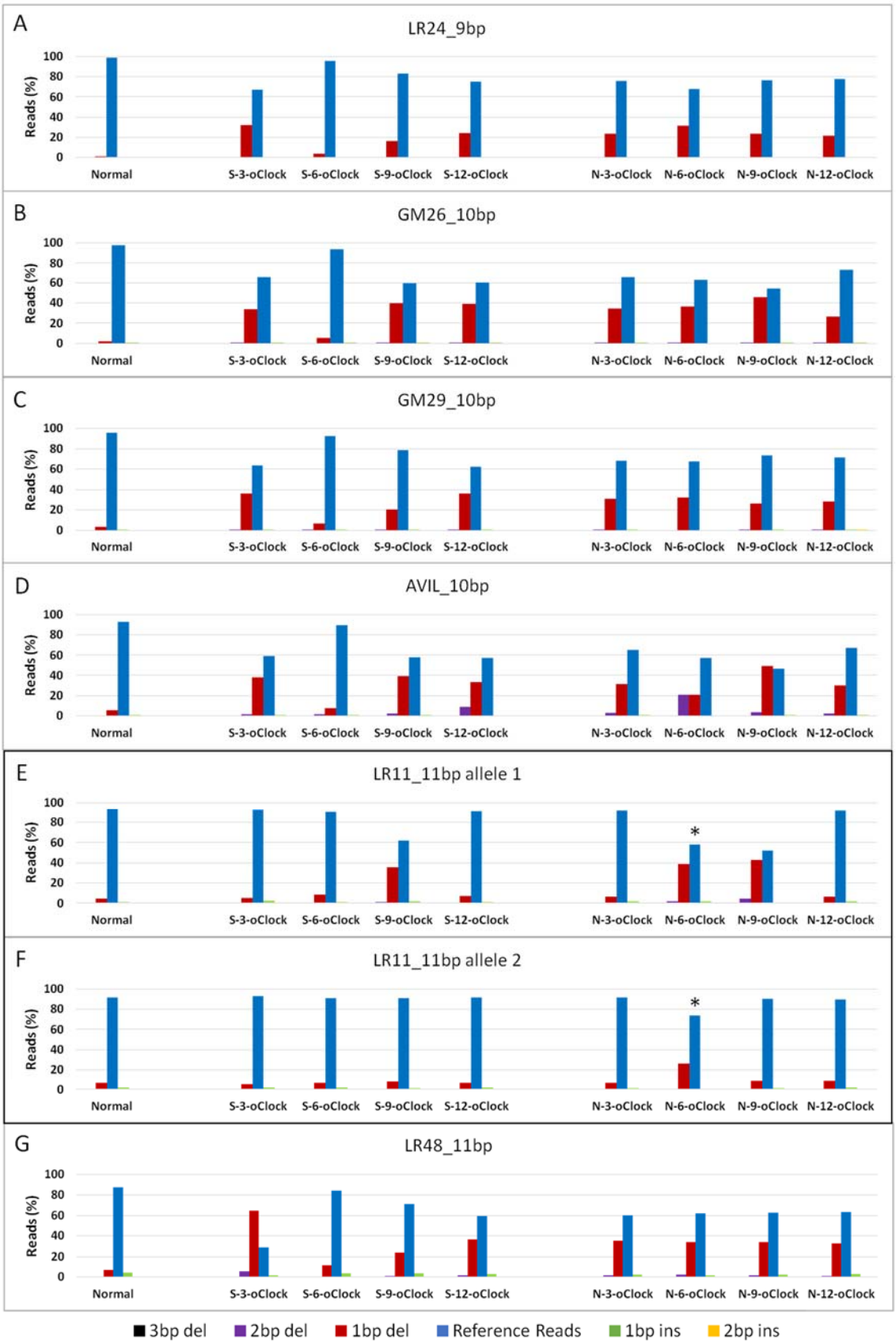

Figure 9.4: 9bp-11bp repeats for tumour PR10654 which were not included in chapter 6. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat. * A total of ≥100 paired end reads for the marker, but less than 100 paired end reads per allele.

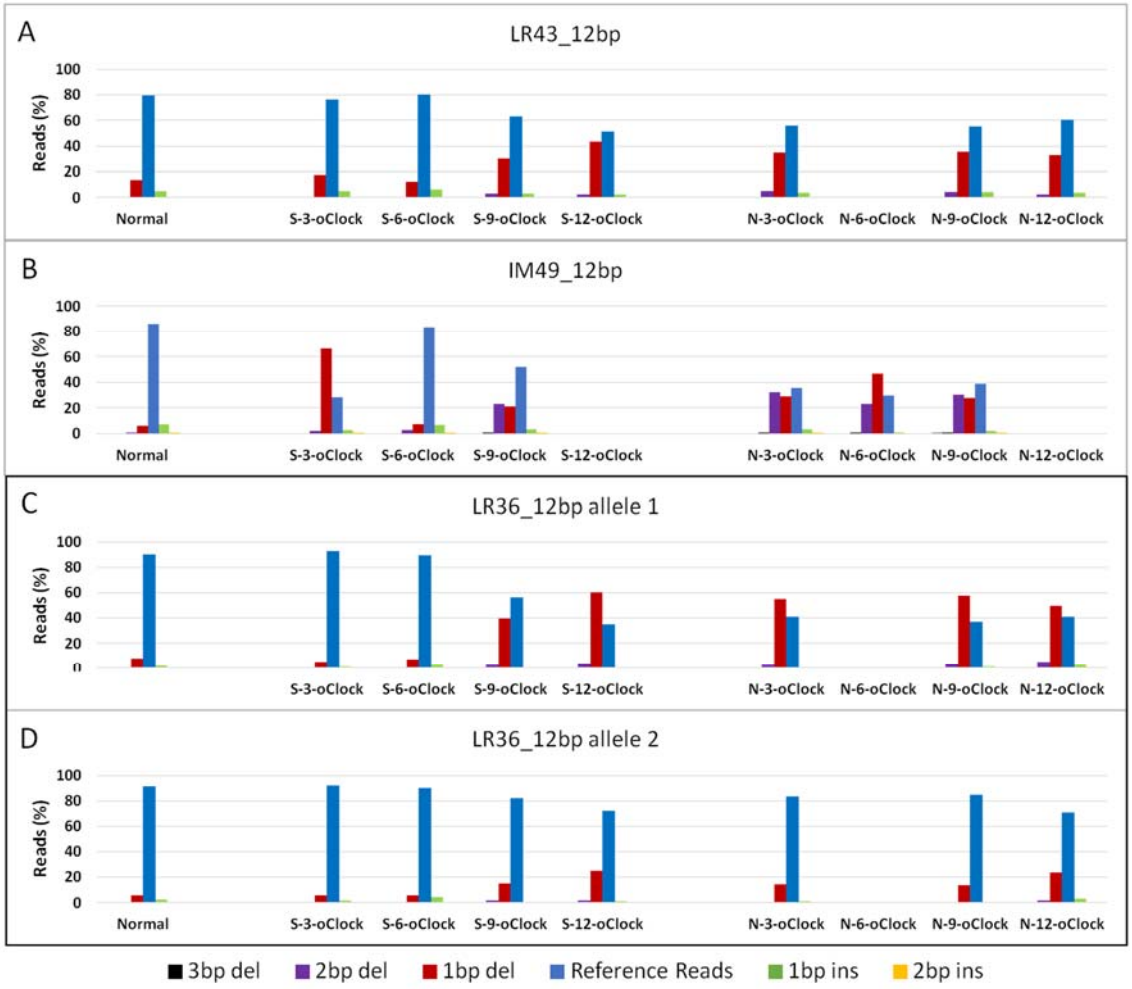

Figure 9.5: 12bp repeats for tumour PR10654 which were not included in chapter 6. Repeats were only analysed if there were ≥100 paired end reads spanning the repeat.

Amplicon Name	Repeat length (bp)	Repeat Unit	Amplicon position	Forward Primer	Reverse Primer	SNP1	SNP2	SNP3
CYP2C9	N/A	N/A	chr10:96740990-96741143	TGCATGCAAGACAGGAGCC	GGAGAAACAAACTTACCTTGGGAA			
BAT26	26	A	chr2:47641351-47641743	CTTTAGAACTGGATCCAGTGG	AAAAAGTGGAGTGGAGGAGG			
Axin2	7	C	chr17:63532406-63532719	AACCCAGTTTCTTTCCTTCTG	GCCTCAACCTAGGACCCTTC	rs35415678		
AL590078	8	A	chr9:26468834-26469145	TCACCACTGGGGACTTTTTC	TGAGCACACCAAGTCATTCTG	rs10967352		
MX1	8	C	chr21:42825925-42826244	TAGAGGCAGCAGGCTCTCAG	ACCCCACAAACCATGAAATC	rs35138081		
HPS1	8	C	chr10:100186775-100187078	CACAGCCCATTCCTGGAC	GCCATTGCTTACATCTCATGG	rs12571249	rs12571245	
IL1R2	8	C	chr2:102626258-102626576	AGGACTCTGGCACCTACGTC	TCGCAAGGAAACTACAGCAG	rs2282747		
DEPDC2	8	C	chr8:68926559-68926888	TCTGGGAAAAAGCCCATAAC	ACAACACCCTCTCACCCAAC	rs4610727		
APBB2	8	C	chr4:41034386-41034688	TGACTATGACAGGAGCTTAAAACTG	CCCACACCACATTGTATGTAGAC	rs4861359		
SLC4A3	8	C	chr2:220493959-220494271	GGCACACCAGGAGAAAGAGG	GCCCCGACCTACCATACAG	rs597306		
AC079893	9	A	chr7:109669372-109669697	CGTTTTTGTGGAAGCATACG	CCAAATGGCAAATAAAAGAAGG	rs4591959		
AL390295	9	A	chr13:35354677-35355008	CATGATATGCCCATGTAGGG	ATTGGTGAAGGAACCAGCAG	rs9572382		
AL359238	9	A	chr14:83421969-83422285	CAGCTGAAACCGAAGTGAAG	TTGATGATCCTTTTGACACCAC	rs72703572		
AP003532_2	9	A	chr11:127624900-127625216	CCCTTTACACCACATCAATGC	GCAGGGCCCATCATACAG	rs10893736		
TTK	9	A	chr6:80751710-80752026	TTCCCAACTGTAAGAACAAGAGAG	CACTTCAGAGTGATGTTGTCTTCA	rs17254634		
C4orf6	9	A	chr4:5526980-5527306	TCTTCCTTATGACAACCCACAC	GAGCACCTTCCGACTCACTC	rs886532	rs113971480	
AL954650	9	C	chr1:191926696-191927019	TGCCAATATTTCAATTTTTCTCC	AGACTATGCCTTGCCCAGAG	rs77489859		
AL355154	10	A	chr13:82018382-82018682	TGCCAATATTTCAATTTTTCTCC	AGACTATGCCTTGCCCAGAG	rs9545694		
AVIL	10	A	chr12:58202332-58202663	CTGCAGAGCCACCCATTC	AGATGAACCAAGCCAGAAGC	rs2277326		
ASTE1	11	A	chr3:130732912-130733215	TGGAGGCCTCACTATGTTCC	CTGGTGCACGGACTATGC			
MRPL2	12	C	chr6:43021823-43022132	GTGGGGACAGACCCAGTG	GGGCAAGAGGCCTAACAGTG	rs58470539		
EGFR	13	A	chr7:55273419-55273760	CACAGACTGGTTTTGCAACG	CTTGTGCTCCTTGCTCACAG			
FBXO46	14	A	chr19:46214532-46214834	CTCCAGCGAGAAAGAATTGG	ATTGATCCCTCACCGGAAC	rs34505186		
FTO	15	A	chr16:54147638-54147956	TTTGTTATATCCCATTAGGTGCC	ATCACGAGGTTGAGATCGAG	rs77984007	rs11348169	

GM01	10	A	chr11:28894282-28894553	TCAAGGCCAGGCAATTAATCAG	ACTTGCTGAATGTCCAAGGTG	rs7951012		
GM02	11	A	chr1:116245990-116246244	GTGCTACATGAGATAGCTGGGA	CTCTTCTGGCCAGTTCTATGTGT	rs10802173	rs148789685	
GM03	8	A	chr4:120206298-120206557	TGGAGTAAGACCCTTTAGGCAG	AGACTCTGGAAGCAAATGGCA	rs17050454	rs10032299	
GM04	7	A	chr13:92677409-92677684	CCTTTTGGCCAGAATATGCC	GGCATGAGGAAGTGAAGGGA	rs9560900		
GM05	9	A	chr2:216770642-216770900	AGGTGTCAAGCAAGGACTCAG	AGGCGTTTTCACGTTGGAGG	rs6704859		
GM06	9	A	chr16:77496387-77496667	AGAGGCAGAATGTGGAAAAGTC	GCATTCTCCCACAGCACAAT	rs6564444	rs143453795	rs145573459
GM07	11	A	chr7:93085548-93085828	GGAGGGACATGTGTTTCCAAAT	CACAATGAGCCAAGTCTCACA	rs2283006		
GM08	8	A	chr21:36574923-36575189	AGCAACCTCTTAAATCCAGTACT	TGGGCTTTCTTGACTTTGGA	rs2834837	rs115025058	
GM09	8	A	chr20:6836843-6837099	TTTCTCAGGACAAAGAGCAAGGT	CTGGGTTCCATCTTGTGGGG	rs6038623		
GM10	9	A	chr1:59891529-59891795	ATCAGCTGACTCCTTACCCT	TGGGGTGAGAGATGGACATG	rs946576	rs182557762	
GM11	9	A	chr5:166099809-166100081	CTCATGGTTAATACAATTAGGCACA	ACATGGTGTGCTACCTTTCA	rs347435		
GM13	11	A	chr12:107492450-107492711	TTCTTCAGGGCCCATTATTGT	TGAGGAATGTGCAGTTGACAC	rs34040859	rs77265275	rs201488736
GM14	11	A	chr3:177328721-177329014	AGCTTGGCCATATTTGTGCA	ACTTGATAGGGTTAAATGTCCGT	rs6804861		
GM15	9	A	chr7:97963570-97963830	TGCCTTCGAGTTTAAATGCCT	GCCTCGTTATTTTGTGTGCC	rs6465672		
GM16	8	A	chr6:100743524-100743782	GCCACACTGACTTTGAACCTT	ACAGCTTCTTCCTCACTCTACT	rs7765823		
GM17	9	A	chr11:95550977-95551231	TCCCTAGAAAGAGAACGACAACA	AAATGCCCACCAAGATTGTAAAA	rs666398		
GM18	12	A	chr10:8269462-8269727	GGGGAGAAGACGGTTGAACT	ACTGGTTCACTGGCCTTTTG	rs113251670	rs189036006	rs533236
GM19	7	A	chr11:114704247-114704523	AGGTAAAGTCAGACACAATCCCA	ACCCTCATGTTTCCCACCTCA	rs142833335	rs190597109	rs10502196
GM20	8	A	chr7:142597420-142597679	GCAATCACATTTGCATTGGTTTT	TGACTATGAGCTCCACAAACGTA	rs6961869	rs6961877	
GM21	9	A	chr3:142695286-142695560	TTCTCCATTGGAAGTATTTGGGA	TGTGTATTCAGGGTCCAGGG	rs185182		
GM22	10	A	chr14:43400950-43401207	TCATAACCAAGAGCACCACCT	TGTGATAGGGAAACACACGGA	rs58274313		
GM23	9	A	chr5:11345800-11346075	CAGCATAAATCCAATGGCTATG	TCAGATTGCAAAGGGGTACA	rs184237728	rs32123	
GM24	7	A	chr10:117432031-117432299	AAACATTTCGACTGGTGCAA	TTCTTCTTTCCCCCAAATGA	rs2532728		
GM25	7	A	chr3:110871894-110872161	TGGGATTAGGGAAGGGAGAG	GGCCCTCCCCAACTAAAAT	rs74593281	rs6437953	rs188039266
GM26	10	A	chr14:49584656-49584913	CCTTCCTTTGATCCGCAAGC	CTGCCACCTAGGAACTGGAG	rs187027795	rs11628435	
GM27	7	A	chr11:85762061-85762349	TTTTTGTTGCCCATTTCCTC	AGGGTACTGACCCTAGCTCCA	rs669813	rs181565251	rs146406522
GM28	9	A	chr5:29209275-29209526	CTCAGACAAAGACATACGAAGCC	TTGGTTCTACAGTAATTGTGCTTCT	rs4130799		

GM29	10	A	chr3:70905468-70905731	CCCTCCCAAATGTCAAGTGT	CCCACCCACACTCTTTTGTT	rs2687195		
GM30	7	A	chr14:53111446-53111710	TCAATGCTATTGGCCTATAAAGAGT	ATGCATTTCCTTCTGGCCTA	rs12880534		
IM07	10	A	chr6:100701756-100702050	TCACCATCATCACCATGCTT	TCTGGCAAACTCTTCACTGG	rs189035042	rs6915780	
IM12	10	A	chr8:23602751-23603036	AGTGGAGAAAACGGTTGTGG	GAAGGCAGACAAGGGATTCA	rs389212		
IM13	7	A	chr2:235496873-235497180	GTGACCGCACAAAGTCACAC	TCCAACAATCACAGTCCATGA	rs6721256	rs183025093	rs187312036
IM14	7	A	chr7:80104285-80104624	TCAAGACTCAGCCATTTCCA	GGAAGCTGAGAGCAGGTTTTT	rs11760281		
IM15	8	A	chr6:91455016-91455307	TCGTCAGGCTCTGCAACTAC	CGATGGGATTGAATTTGGAT	rs1231482		
IM16	9	A	chr18:1108609-1108894	AGGACCTCGAGCTTCTCTTT	TTCTTTTGCTTCCGTGTGTG	rs114923415	rs73367791	rs59912715
IM17	9	A	chr13:31831349-31831705	TGCAACCAGAGGTTTTAATCG	CTCAATTCAGCAACAGGTCA	rs932749		
IM19	7	A	chr9:82474924-82475277	CAACCACAGTTTGCCAGCTA	TCCTTGCTATCATTTGGAGAGA	rs72736428	rs186539440	rs4877153
IM20	7	A	chr13:57644542-57644833	CCAGTTTCACATTTCGCTTGT	TGGCAACAAAACAGTAACAGGA	rs6561918		
IM21	8	A	chr1:215136329-215136605	AGTGAATGGGCTTTGGACTG	AACTGGAGTGGGTGAACCTG	rs181787229	rs1901621	rs1901620
IM22	7	A	chr7:90135380-90135698	CACCAGCTTTTCTCCCTTCA	TGGCACTCAATACCAAACTGG	rs10487118	rs10487117	rs139214151
IM23	7	A	chr6:72729441-72729714	GGTTTCTGTGCTGAATCTTGG	AACCCCAGTTTTCTGCCTCT	rs557365		
IM25	8	A	chr12:24568297-24568575	CCATGGTACCACTGTGGAGT	TAGAGGGGGCTTGAATGTTG	rs10771087		
IM26	7	A	chr3:166053374-166053712	GGGCTCGACTTGATTTACGA	GGGAAGCAATCTCATGGCTA	rs2863375		
IM27	7	A	chr7:35079029-35079302	ACGCATGGAAAAAGAGGTTC	CAAGGCTGGTATGGGTCAAT	rs4723393	rs112516918	
IM28	11	A	chr9:5122829-5123102	TGTGGAATCCCTCCTGAAAT	CCGCTGGTGGACTTTTACTC	rs10815163		
IM32	11	A	chr18:42045361-42045640	GCCAAAATGCCTAACTCCAA	GGACTCGGATGGAAGACAAA	rs8087346		
IM33	10	A	chr8:25731833-25732120	AGGGTATGATTTGGGGGTGT	GTGGACCAAAGGAGCAGAAG	rs202225742	rs35644463	rs113180202
IM34	10	A	chr7:83714549-83714816	TGAGGGTGGATGCTTCATTT	CAGGATATTCCTCAGTTCAGTTCC	rs1524881		
IM35	10	A	chr11:84425027-84425322	TCAAATGCAGACTCAACATGA	AGCAGAGGAGCCATCAATTC	rs67283158	rs10792775	rs116387070
IM37	10	A	chr17:50813421-50813720	CAGGCACACACACTTTCGTT	TTCTCATGCAGTCAACCATTG	rs2331498		
IM39	8	A	chr2:103233602-103233932	AGACGTCCAAAGGTCGCTAA	CCCTCACTGCCTGTAAACCT	rs76771828	rs190979688	rs187315716
IM40	8	A	chr4:84074695-84074985	ATCACAAAAACAGGGGCCTA	CCTTGTCTGGCTCAATCACC	rs10516683		
IM41	8	A	chr6:147948700-147949027	CTGCTCCACATTCCCATTCT	TGGCAGGAAACATCTGTTCA	rs1944640	rs112075239	
IM42	9	A	chrX:96502491-96502781	TGGCTGAGTAAAATGGTGACA	GCTTGGGGGAATTTCTTGAT	rs1409192		

IM43	7	A	chr21:32873526-32873866	CAGAAGGTCAGGACCACACA	ATTTGGTGGGTTCCAGTGAG	rs9981507		
IM44	9	A	chr12:9796844-9797182	CCTCCTAGCATTCCATAGCAC	TGCAACCTCGTAAGCTCATTT	rs201750704	rs4763716	
IM45	11	A	chr4:99545274-99545564	GCCACATTTGCTGGTATTCA	TTTTTCCTCTGGGAAACCAT	rs189419054	rs2178216	
IM47	12	A	chr21:22734257-22734517	TGGTTCAGACATACACGTACAGG	ATAACAGGCACAAGGGTGGA	rs2588655	rs149325240	rs232496
IM49	12	A	chr3:56681883-56682149	CCTGGCAAATGATGCTTTAGA	CCTCCCTCCTAGGCTCAAGT	rs7642389		
IM50	12	A	chr20:37047920-37048224	CGAGGCGGGTATTTACTTGA	GGAGTTGGGGCAAAAATCAC	rs1739651	rs145870165	
IM51	12	A	chr5:128096936-128097255	CAAACCCCCGAGACACAC	AACGTGGCTCTTTATCCCATT	rs4836397		
IM52	11	A	chr21:22846659-22846944	GATGGAGGGCCCTTTAATTT	CGATGAAGTGGTTGATGTGAG	rs74462385	rs9982933	rs2155801
IM53	11	A	chr9:20662482-20662766	GACAACTCCGAAGGGCAATA	AGTTTGGGTTGCAAGACGTT	rs182630429	rs140426089	rs12352933
IM54	11	A	chr21:33709922-33710213	GCAACATTGAAATGCTGGAA	TAACATTTGGGAGGGGGAAT	rs13046776		
IM55	7	A	chr3:143253627-143253930	GCTGAATAGCGGGATCAAAA	GGAATTAGGTACCAGATCTCCTTT	rs13099818		
IM57	8	A	chr3:81209863-81210156	GATTATCAGCCCAGGGAGGT	ATGGCAGCACTGGGAAATTA	rs35085583		
IM59	8	A	chr8:108358809-108359137	TATGGCTGCAGCATTACCAG	GCCAGAGTCCACAGACTCAA	rs10156232		
IM61	7	A	chr12:73576301-73576606	GAGCAAGGCATTTGAATCTG	ATATGAGGCGCTCTCTCTCG	rs34696106		
IM63	8	A	chr3:115815913-115816216	TGCCTTTGGTTGTACCTTTG	TCAAGTGAGCCTTGTGGAAA	rs34764455		
IM64	12	A	chr16:14215981-14216240	CCTTCCCCGTTCTTTCTCTT	AAGGTAGGTGACCGGCTGAT	rs201451896	rs112858435	rs75477279
IM65	11	A	chr13:25000797-25001149	GCATCTCAAACTGTGCCTGT	CACGGGTCTAACTGTCCTCA	rs7324645	rs9511253	
IM66	7	C	chr17:48433883-48434148	CCACTCCAGCAAGTCTCCAG	CAAGGGCCTGCTGTATGTCA	rs147847688	rs141474571	rs4794136
IM67	7	C	chr7:22290637-22290990	AGCCCATGTTTTCCACAGAA	TACCAGGTGCCCTAAACAGG	rs67082587	rs57484333	
IM68	8	C	chr12:129289515-129289789	TTCTAGACACAGACGCACACG	GGGACTGCCACTAGTAGCTCA	rs10847692		
IM69	7	C	chr9:92765658-92765989	TGGGGGCAGTTTCTATTCTG	ATCAGTTTTCGATGGGGAGA	rs1036699		
LR01	11	A	chr13:97387292-97387567	TTGGATGCTGGATTTTGACA	CTCATATCCCCCTCCCAGAA	rs1924584	rs4771258	
LR02	8	C	chr4:134947615-134947875	TATTGGCCAGGAATTTTTGC	GGAGCTCACGCTAATGACCT	rs189671825	rs192703656	rs1494978
LR04	7	C	chr1:4676948-4677234	CCCCAAGCTGTTTCCTCCAT	GCTGGGGCAAGAAATTCAGC	rs113646106	rs2411887	
LR05	9	C	chr2:10526489-10526814	GAGCTGCCTACTCGCTGACT	GCCACTGATGACAACCTCCT	rs111286197	rs13431202	
LR06	7	C	chr18:20089314-20089588	CATCTAGCATTCTCTCATTTCAGC	TGCCAAAACCAAAGACAAGG	rs501714		
LR08	7	C	chr11:56546008-56546315	GGCTGCTTAAGGGAAAGTGC	CGTGTTTTGGTCAAAGTTGTG	rs181578273	rs7117269	

LR10	9	A	chr1:81591297-81591555	ATGTTTGGTGCATGAAATCTG	TGAGTTCCACATGGCTCTTG	rs111814302	rs1768398	rs1768397
LR11	11	A	chr2:217217726-217218005	TATTCCCCTTGTGTGGGAGA	CAAAGAGAATGGGTGGGAGT	rs13011054	rs147392736	rs139675841
LR12	11	A	chr14:47404086-47404346	GGTGAGGAAAGCACAAGGTC	CCGTGGAATTTCTTCTGCAC	rs187434561	rs144159314	
LR13	7	A	chr8:21786845-21787107	TCCTCGTCCTCTCAGATGTGT	TCAGGACTTAGCACCAGGAAA	rs2127206		
LR14	9	A	chr17:69328365-69328640	CCCGTTTTCAGACCAAGTGT	TTGGAACAGGATGGGTGAAT	rs9895642		
LR15	7	A	chr8:92077118-92077383	TGATTCGGGCTTGGACTTAG	GTCAATCACTTTGCCTGCTC	rs56084507		
LR16	11	A	chr3:8522305-8522590	GTTTGATCTCTGGCCCTGTC	GCCTCCTTAATCTCCTCCATC	rs148171413	rs6770049	
LR17	11	A	chr14:55602913-55603194	AGACCACCCCTTAGGCAAAC	AGTGCAGCAAGGCAGATGAG	rs79618905	rs77482253	rs1009977
LR18	8	A	chr1:220493800-220494106	TGGGGAGGGAACCTCATTAC	CAGTGCCTGTTGAGTAGAACC	rs191265856	rs199830128	rs74940412
LR19	8	A	chr12:29508532-29508843	TGAGTGCTGCTCATATTTTTCC	GGGGCTTCAGTCTCAGGATAG	rs10843391	rs186762840	
LR20	8	A	chr1:64029521-64029836	TCAGCCTATGAAGATCCTCTG	AAGGAAGACGGGGAAGACTG	rs146973215	rs191572633	rs217474
LR21	9	A	chr15:50189339-50189607	TGGGTACAAAGCTCAAGTCAAC	TCTCCAAAGGCTTCTCCTTG	rs182900605	rs80237898	rs2413976
LR23	11	A	chr2:142013847-142014151	TGTAGCCTAGGTAAAGAGGACAA	CATTTAGCATTTTGCCATTCC	rs434276	rs146141768	
LR24	9	A	chr1:153779290-153779565	TATGCCTTCTGGAGGAGTGG	TGGAATAGCGGTAAGGCTTG	rs192329538	rs1127091	
LR25	7	A	chr16:63209414-63209676	TTAACCTGCCAGCTCAGTTC	GCTTCCACTCATTTGCATTG	rs76192782	rs79880398	rs4949112
LR26	10	A	chr16:80050164-80050433	TGCATAGGCAGACCTCAAAAC	GAAAGCCTGATGTTTGACACC	rs4889066	rs187883346	
LR27	8	A	chr4:72877320-72877604	TTTGGTCATTGCTGTCATGG	CAACAAGGAATTGAATGATGC	rs55894427	rs74733006	
LR28	9	A	chr12:81229619-81229925	TGAGTCCCTTTTGAAATGTTG	GCCAACCAATGGAGTTTTAAG	rs185642078	rs28576612	rs10862196
LR29	10	A	chr6:78198189-78198498	CAATGTTTGATTAACCATGACG	GCACTTTTCTCACACAATTTGG	rs1778257		
LR30	10	A	chr11:105444906-105445201	GCAGGAATTCATTCTGAAGC	AACGCAGTGAGGAACAAAGG	rs7933640		
LR31	8	A	chr3:62995387-62995657	TGGATTTGCATCTGTGAATTG	TTTTGATGGCTTTTACTTTTCC	rs183248146	rs2367592	
LR32	10	A	chr19:37967035-37967313	CTGCCTATGCCAAACAAATG	AGCACAAGCCTTTTGTCAGC	rs7253091		
LR33	11	A	chr4:138498516-138498782	GAATAGCGGGAAGAACTGGA	TGCATTCGAATCAGGAATGA	rs200714826	rs4637454	rs111688169
LR34	9	A	chr3:115376990-115377261	CCCATCCTTAGACCCCAGAC	GAAAATGAGACGCGAAAAGG	rs187521190	rs192106258	rs9883515
LR35	10	A	chr8:130384312-130384584	AAAGCTTGTGGGTGATGGAG	TGCTTGGAATAGGATGCTTTG	rs4733547		
LR36	12	A	chr4:98999555-98999845	TCCCCAGGACCCTAGTCTTC	GGTGGCAAGCACTTTTGTAAG	rs182020262	rs17550217	
LR39	10	A	chr17:66449171-66449485	AGCATGGGAATAACGACAGG	TCGTTGTGTTGGAGGTAGAGC	rs2302784		

LR40	9	A	chr2:13447304-13447570	AAATGAACACTATGCATGTCAGG	TTGCCTCTTGCAACTGATTG	rs6432372		
LR41	12	A	chr4:34073929-34074197	CATGGACCGCTGATCTCTG	GGAGGGATCTAGCCACCAC	rs190518698	rs6852667	
LR43	12	A	chr5:86198899-86199207	GGCAACAGCCTCATAACTGC	GCTGTCTCCTGGCTCTAACC	rs201282399	rs10051666	rs6881561
LR44	12	A	chr10:99898182-99898454	TTTGGCTGGGCCTGGTAG	CAGAGTGCACCTCAGTGACC	rs78876983	rs7905388	rs7905384
LR45	7	A	chr2:226937965-226938246	TGCAGAGAAGAGATACAGAAAGC	TGCAAAAATCCCAGATTGAAG	rs180896305	rs1522818	rs144175764
LR46	8	A	chr20:10659968-10660261	GAGTGTGGGAGAAGTCCTACG	TTCAGGAGATGAAAAGGCTTG	rs143884078	rs182346625	rs6040079
LR47	7	A	chr10:20506574-20506830	TCCCTGAAGGAAGGAAAAATC	GTGATTGTGAAGTTGGATTTGC	rs11597326	rs12256106	
LR48	11	A	chr12:77988002-77988288	ATTACCCATGGGGGATGTTG	AGTTGGGGAACATTCCTTCC	rs11105832		
LR49	7	A	chr15:93618885-93619163	ATCTGTAAGGATCGGGCTGA	CAACACAACGCCATACTGCT	rs80323298	rs201097746	rs12903384
LR50	7	A	chr2:76556173-76556470	TTCCCCATTTGATGATCCTG	AGAGTTTTCCCCACTCAGCA	rs925991	rs144630203	
LR51	7	A	chr10:51026570-51026831	TGAATATGCCTCAAGCACCA	AATGCAAACCTCCTAGGTTAAAA	rs8474		
LR52	12	A	chr16:63861273-63861586	GTGCTCTGCATCTCATACGC	CCTCCTTGGCTAACTTGCTC	rs2434849		

Table 9.1: List containing amplicon/repeat name, amplicon position (genome build hg19), primers, and SNP rs numbers for SNPs in close proximity to mononucleotide repeats.

CAPP patient Number	Gene containing germline mutation	Tissue Type	Tumour Block Number	Chapters where the samples were used
U029	MSH2	Tumour	H10/7014 A18	5
U096	MLH1	Tumour	R06038/03-1E	3
U096	MLH1	Normal Mucosa*	R0603F/03-1C	5
U179	MLH1	Tumour	H03-19031-1 B42	3 and 5
U179	MLH1	Tumour	H12/4786 A6	5
U184	MLH1	Tumour	8.9.05 6cFT	3
U303	MLH1	Tumour	07/1615-1B	3 and 5
U312	MSH2	Tumour	07/3480-1C	3
U312	MSH2	Tumour	07/3480-1B	5

Table 9.2: CAPP Lynch Syndrome patient tumour samples used in the work described in this thesis. * Block containing the distal resection margin which was erroneously supplied instead of tumour tissue.

patient	Tissue Type	SRA Sample Accession Number	Sample Id	Analysis Id
TCGA-AA-3516	MSI-H Tumour	SRS130750	TCGA-AA-3516-01A-02D-1167-02	69e9e641-fa2e-4bd9-848e-be5f507660a2
TCGA-AA-3672	MSI-H Tumour	SRS097008	TCGA-AA-3672-01A-01D-0957-02	9b01e1d4-2cca-49ef-9672-e8692ae621be
TCGA-AA-3715	MSI-H Tumour	SRS097080	TCGA-AA-3715-01A-01D-0957-02	23acfb6f-8071-47cf-9c62-67c22c63e0ec
TCGA-AA-3966	MSI-H Tumour	SRS130791	TCGA-AA-3966-01A-01D-1109-02	4b50f9fa-0fb6-4293-afc5-adc0350f4ed2
TCGA-AA-A00R	MSI-H Tumour	SRS153954	TCGA-AA-A00R-01A-01D-A077-02	300eea0f-bc14-4253-a544-fbc53243ecce
TCGA-AA-A01P	MSI-H Tumour	SRS130846	TCGA-AA-A01P-01A-21D-A079-02	60b884be-5009-495f-aec8-bd3be4bf7597
TCGA-AA-A01Q	MSI-H Tumour		TCGA-AA-A01Q-01A-01D-A077-02	d8e8f805-00c6-467e-a8fc-a2674f1ac38e
TCGA-AA-A02R	MSI-H Tumour	SRS154223	TCGA-AA-A02R-01A-01D-A077-02	1d75d53d-94a2-497e-9c4e-f45cf27912d9
TCGA-AZ-4313	MSI-H Tumour	SRS157354	TCGA-AZ-4313-01A-01D-1405-02	a0a2a333-708c-46dd-ab19-c6e7e033d724
TCGA-AZ-4615	MSI-H Tumour	SRS157387	TCGA-AZ-4615-01A-01D-1405-02	ae769553-f8cb-407b-b41a-524adb07282d
TCGA-CK-4951	MSI-H Tumour	SRS159294	TCGA-CK-4951-01A-01D-1405-02	ef054dd4-e5ed-4143-80ef-08beffa04d1b
TCGA-CM-4746	MSI-H Tumour	SRS159316	TCGA-CM-4746-01A-01D-1405-02	1878a6ba-0f5c-40b4-a018-7508c8fa3dc2
TCGA-AA-3516	Matched Normal	SRS130751	TCGA-AA-3516-10A-01D-1167-02	e1dbd1cc-89ad-4f93-97f1-982b4ac7f7f3
TCGA-AA-3672	Matched Normal	SRS097012	TCGA-AA-3672-10A-01D-0957-02	99a5462d-3cb8-464b-98c6-cec13491994c
TCGA-AA-3715	Matched Normal	SRS097084	TCGA-AA-3715-10A-01D-0957-02	6c0543fd-e91d-4cf1-a64b-89ad9e63cf71
TCGA-AA-3966	Matched Normal	SRS130801	TCGA-AA-3966-10A-01D-1109-02	078906fa-6e88-4626-b01f-9b529b969460
TCGA-AA-A01P	Matched Normal	SRS130854	TCGA-AA-A01P-11A-11D-A079-02	74847765-b70c-47c8-9eb4-d5eae2e4c704
TCGA-AA-A01Q	Matched Normal		TCGA-AA-A01Q-10A-01D-A078-02	0e3b9a0b-8fd8-4726-bcd8-d5b8563bb630
TCGA-AA-A02R	Matched Normal		TCGA-AA-A02R-10A-01D-A078-02	0f21aa03-df30-4b29-b908-daf9584088d6
TCGA-AZ-4313	Matched Normal	SRS157361	TCGA-AZ-4313-10A-01D-1405-02	7c0a3b4d-0fc0-4b9a-b4c4-d075b8117f42
TCGA-AZ-4615	Matched Normal	SRS157394	TCGA-AZ-4615-10A-01D-1405-02	af33c9f9-02a6-4e4a-9f1d-52f04bfa6116
TCGA-CK-4951	Matched Normal	SRS159301	TCGA-CK-4951-10A-01D-1405-02	5ac4252e-6cb6-4226-a8cd-489a9986c61e
TCGA-CM-4746	Matched Normal	SRS159323	TCGA-CM-4746-10A-01D-1405-02	a9540af8-5f10-4d06-a31d-1e25a69e31bd
TCGA-AA-3509	MSS Tumour	SRS156892	TCGA-AA-3509-01A-01D-1405-02	4e552949-246f-4788-a760-9b6a23d89bf3
TCGA-AA-3555	MSS Tumour	SRS196934	TCGA-AA-3555-01A-01D-1637-02	875b31fd-9e8a-49d2-89b4-d8256f89ef5a
TCGA-AA-3558	MSS Tumour	SRS130763	TCGA-AA-3558-01A-01D-1167-02	1b21ee51-605f-478b-8a82-e25a3fa9b678
TCGA-AA-3685	MSS Tumour	SRS130776	TCGA-AA-3685-01A-02D-1167-02	25f4344f-ea46-48ac-b2a9-74f9222fb8aa
TCGA-AA-3693	MSS Tumour	SRS097064	TCGA-AA-3693-01A-01D-0957-02	3a9f1142-0d5b-4583-a570-4da8e1455e0c
TCGA-AA-3968	MSS Tumour	SRS130808	TCGA-AA-3968-01A-01D-1167-02	3f6440b7-1298-4892-a133-3d48eb885eda
TCGA-AA-3970	MSS Tumour	SRS130814	TCGA-AA-3970-01A-01D-1109-02	d67dcadf-7893-42fd-afd7-e2ed9a1aa33d
TCGA-AA-A00U	MSS Tumour	SRS153966	TCGA-AA-A00U-01A-01D-A077-02	285ce8fc-dc1b-4188-bb8e-723573a9545a
TCGA-AY-4070	MSS Tumour	SRS133582	TCGA-AY-4070-01A-01D-1109-02	acec5f0d-3fa8-45ca-bdda-3058c14bbcc0
TCGA-AY-4071	MSS Tumour	SRS133599	TCGA-AY-4071-01A-01D-1109-02	1b3451c6-b020-4a04-b454-018ceb86da2f
TCGA-CA-5256	MSS Tumour	SRS159111	TCGA-CA-5256-01A-01D-1405-02	1cad444d-0ed8-437d-8f26-66e103379160
TCGA-CM-4748	MSS Tumour	SRS159338	TCGA-CM-4748-01A-01D-1405-02	bc5d8cfa-f666-4b3a-9117-5b60f92d480e

Table 9.3: Sample identifiers for whole genome sequences obtained from The Cancer Genome Atlas (TCGA) group.

Reagent	Total cost (£)	Number of reactions per sample	Estimated number of samples the product can be used for	Cost per sample (£)
Amplicon specific primer with Illumina overhang adapters 18 primer pairs (Synthesis scale 0.04µmol)	291.60	18	400	0.73
Nextera® XT Index Kit (96 Indices, 384 Samples)	662.00	1	384	1.72
Herculase II Fusion DNA Polymerase 400Rxn (800 reactions with a PCR volume of 25µl)	273.00	19	42	6.48
AMPure XP - 60ml (45µl per sample for each cleanup)	721.02	2	660	1.09
QIAxcel DNA Screening Kit (2400)	517.00	18	130	3.98
Qubit® dsDNA HS Assay Kit (500 assays kit)	161.20	2	250	0.64
MiSeq Reagent Kit v3 (600 cycles)	1100.00	1	96	11.46

Table 9.4: The estimated cost per sample for a sequencing based MSI assay composed of 18 markers. The costs are estimated for a sample prep where amplicon specific primers with Illumina overhang adaptor sequences are used, enabling Illumina sequencing primers to be added in a second PCR reaction. The second PCR reaction will be performed after all 18 amplicons have been pooled at an equal concentration. Sequencing 96 samples per MiSeq run gives an average read depth of ~10000 paired end reads per amplicon if a read depth comparable to what was obtained for the MiSeq run in chapter 7 is achieved.

References

AALTONEN, L. A., PELTOMAKI, P., LEACH, F. S., SISTONEN, P., PYLKKANEN, L., MECKLIN, J. P., JARVINEN, H., POWELL, S. M., JEN, J., HAMILTON, S. R. & ET AL. 1993. Clues to the pathogenesis of familial colorectal cancer. *Science,* 260**,** 812-6.

AARNIO, M., SANKILA, R., PUKKALA, E., SALOVAARA, R., AALTONEN, L. A., DE LA CHAPELLE, A., PELTOMAKI, P., MECKLIN, J. P. & JARVINEN, H. J. 1999. Cancer risk in mutation carriers of DNA-mismatch-repair genes. *Int J Cancer,* 81**,** 214-8.

ALBERS, C. A., LUNTER, G., MACARTHUR, D. G., MCVEAN, G., OUWEHAND, W. H. & DURBIN, R. 2011. Dindel: accurate indel calls from short-read data. *Genome Res,* 21**,** 961-73.

ALHOPURO, P., PHICHITH, D., TUUPANEN, S., SAMMALKORPI, H., NYBONDAS, M., SAHARINEN, J., ROBINSON, J. P., YANG, Z., CHEN, L. Q., ORNTOFT, T., MECKLIN, J. P., JARVINEN, H., ENG, C., MOESLEIN, G., SHIBATA, D., HOULSTON, R. S., LUCASSEN, A., TOMLINSON, I. P., LAUNONEN, V., RISTIMAKI, A., ARANGO, D., KARHU, A., SWEENEY, H. L. & AALTONEN, L. A. 2008. Unregulated smooth-muscle myosin in human intestinal neoplasia. *Proc Natl Acad Sci U S A,* 105**,** 5513-8.

ALHOPURO, P., SAMMALKORPI, H., NIITTYMAKI, I., BISTROM, M., RAITILA, A., SAHARINEN, J., NOUSIAINEN, K., LEHTONEN, H. J., HELIOVAARA, E., PUHAKKA, J., TUUPANEN, S., SOUSA, S., SERUCA, R., FERREIRA, A. M., HOFSTRA, R. M., MECKLIN, J. P., JARVINEN, H., RISTIMAKI, A., ORNTOFT, T. F., HAUTANIEMI, S., ARANGO, D., KARHU, A. & AALTONEN, L. A. 2012. Candidate driver genes in microsatellite-unstable colorectal cancer. *Int J Cancer,* 130**,** 1558-66.

ALLEGRETTA, C. 1999. GNU nano. Nano Core Development Team.

BALMANA, J., CASTELLS, A. & CERVANTES, A. 2010. Familial colorectal cancer risk: ESMO Clinical Practice Guidelines. *Ann Oncol,* 21 Suppl 5**,** v78-81.

BAM2FASTQ SOFTWARE [HTTP://GSL.HUDSONALPHA.ORG/INFORMATION/SOFTWARE/BAM2FASTQ].

BARKER, N., VAN DE WETERING, M. & CLEVERS, H. 2008. The intestinal stem cell. *Genes Dev,* 22**,** 1856-64.

BARRATT, P. L., SEYMOUR, M. T., STENNING, S. P., GEORGIADES, I., WALKER, C., BIRBECK, K. & QUIRKE, P. 2002. DNA markers predicting benefit from adjuvant fluorouracil in patients with colon cancer: a molecular study. *Lancet,* 360**,** 1381-91.

BARROW, P., KHAN, M., LALLOO, F., EVANS, D. G. & HILL, J. 2013. Systematic review of the impact of registration and screening on colorectal cancer incidence and mortality in familial adenomatous polyposis and Lynch syndrome. *Br J Surg,* 100**,** 1719-31.

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., WIRTZ, H. C., FISHEL, R. & RUSCHOFF, J. 1997. Microsatellite instability analysis: a multicenter study for reliability and quality control. *Cancer Res,* 57**,** 4739-43.

BOLAND, C. R. 2007. Clinical uses of microsatellite instability testing in colorectal cancer: an ongoing challenge. *J Clin Oncol,* 25**,** 754-756.

BOLAND, C. R. 2012. Lynch syndrome: new tales from the crypt. *Lancet Oncol,* 13**,** 562-4.

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. & SRIVASTAVA, S. 1998. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Res,* 58**,** 5248-57.

BORSTNIK, B. & PUMPERNIK, D. 2002. Tandem repeats in protein coding regions of primate genes. *Genome Res,* 12**,** 909-15.

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. *Arch Pathol Lab Med,* 138**,** 363-70.

BRAY, F., JEMAL, A., GREY, N., FERLAY, J. & FORMAN, D. 2012. Global cancer transitions according to the Human Development Index (2008-2030): a population-based study. *Lancet Oncol,* 13**,** 790-801.

BRUECKL, W. M., MOESCH, C., BRABLETZ, T., KOEBNICK, C., RIEDEL, C., JUNG, A., MERKEL, S., SCHABER, S., BOXBERGER, F., KIRCHNER, T., HOHENBERGER, W., HAHN, E. G. & WEIN, A. 2003. Relationship between microsatellite instability, response and survival in palliative patients with colorectal cancer undergoing first-line chemotherapy. *Anticancer Res,* 23**,** 1773-7.

BUBB, V. J., CURTIS, L. J., CUNNINGHAM, C., DUNLOP, M. G., CAROTHERS, A. D., MORRIS, R. G., WHITE, S., BIRD, C. C. & WYLLIE, A. H. 1996. Microsatellite instability and the role of hMSH2 in sporadic colorectalcancer. *Oncogene,* 12**,** 2641-9.

BUHARD, O., CATTANEO, F., WONG, Y. F., YIM, S. F., FRIEDMAN, E., FLEJOU, J. F., DUVAL, A. & HAMELIN, R. 2006. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. *J Clin Oncol,* 24**,** 241-51.

BURN, J. 2013. Company profile: QuantuMDx group limited. *Pharmacogenomics,* 14**,** 1011-5.

BURN, J., GERDES, A. M., MACRAE, F., MECKLIN, J. P., MOESLEIN, G., OLSCHWANG, S., ECCLES, D., EVANS, D. G., MAHER, E. R., BERTARIO, L., BISGAARD, M. L., DUNLOP, M. G., HO, J. W., HODGSON, S. V., LINDBLOM, A., LUBINSKI, J., MORRISON, P. J., MURDAY, V., RAMESAR, R., SIDE, L., SCOTT, R. J., THOMAS, H. J., VASEN, H. F., BARKER, G., CRAWFORD, G., ELLIOTT, F., MOVAHEDI, M., PYLVANAINEN, K., WIJNEN, J. T., FODDE, R., LYNCH, H. T., MATHERS, J. C. & BISHOP, D. T. 2011. Long-term effect of aspirin on cancer risk in carriers of hereditary colorectal cancer: an analysis from the CAPP2 randomised controlled trial. *Lancet,* 378**,** 2081-7.

CANARD, G., LEFEVRE, J. H., COLAS, C., COULET, F., SVRCEK, M., LASCOLS, O., HAMELIN, R., SHIELDS, C., DUVAL, A., FLEJOU, J. F., SOUBRIER, F., TIRET, E. & PARC, Y. 2012. Screening for Lynch syndrome in colorectal cancer: are we doing enough? *Ann Surg Oncol,* 19**,** 809-16.

CANCER GENOME ATLAS NETWORK 2012. Comprehensive molecular characterization of human colon and rectal cancer. *Nature,* 487**,** 330-7.

CHALIGNE, R., POPOVA, T., MENDOZA-PARRA, M. A., SALEEM, M. A., GENTIEN, D., BAN, K., PIOLOT, T., LEROY, O., MARIANI, O., GRONEMEYER, H., VINCENT-SALOMON, A., STERN, M. H. & HEARD, E.

2015. The inactive X chromosome is epigenetically unstable and transcriptionally labile in breast cancer. *Genome Res,* 25**,** 488-503.

CHEN, W. S., CHEN, J. Y., LIU, J. M., LIN, W. C., KING, K. L., WHANG-PENG, J. & YANG, W. K. 1997. Microsatellite instability in sporadic-colon-cancer patients with and without liver metastases. *Int J Cancer,* 74**,** 470-4.

CHOI, S. W., LEE, K. J., BAE, Y. A., MIN, K. O., KWON, M. S., KIM, K. M. & RHYU, M. G. 2002. Genetic classification of colorectal cancer based on chromosomal loss and microsatellite instability predicts survival. *Clin Cancer Res,* 8**,** 2311-22.

CHUNG, H., LOPEZ, C. G., HOLMSTROM, J., YOUNG, D. J., LAI, J. F., REAM-ROBINSON, D. & CARETHERS, J. M. 2010. Both microsatellite length and sequence context determine frameshift mutation rates in defective DNA mismatch repair. *Hum Mol Genet,* 19**,** 2638-47.

CLARKE, L. A., REBELO, C. S., GONCALVES, J., BOAVIDA, M. G. & JORDAN, P. 2001. PCR amplification introduces errors into mononucleotide and dinucleotide repeat sequences. *Mol Pathol,* 54**,** 351-3.

COLOMBINO, M., COSSU, A., MANCA, A., DEDOLA, M. F., GIORDANO, M., SCINTU, F., CURCI, A., AVALLONE, A., COMELLA, G., AMORUSO, M., MARGARI, A., BONOMO, G. M., CASTRIOTA, M., TANDA, F. & PALMIERI, G. 2002. Prevalence and prognostic role of microsatellite instability in patients with rectal carcinoma. *Ann Oncol,* 13**,** 1447-53.

CONSTANTIN, N., DZANTIEV, L., KADYROV, F. A. & MODRICH, P. 2005. Human mismatch repair: reconstitution of a nick-directed bidirectional reaction. *J Biol Chem,* 280**,** 39752-61.

CROCKETT, S. D., SNOVER, D. C., AHNEN, D. J. & BARON, J. A. 2015. Sessile serrated adenomas: an evidence-based guide to management. *Clin Gastroenterol Hepatol,* 13**,** 11-26.e1.

CUI, Y., WEI, Q., PARK, H. & LIEBER, C. M. 2001. Nanowire nanosensors for highly sensitive and selective detection of biological and chemical species. *Science,* 293**,** 1289-92.

CURRAN, B., LENEHAN, K., MULCAHY, H., TIGHE, O., BENNETT, M. A., KAY, E. W., O'DONOGHUE, D. P., LEADER, M. & CROKE, D. T. 2000. Replication error phenotype, clinicopathological variables, and patient outcome in Dukes' B stage II (T3,N0,M0) colorectal cancer. *Gut,* 46**,** 200-4.

DALGAARD, J. Z. 2012. Causes and consequences of ribonucleotide incorporation into nuclear DNA. *Trends Genet,* 28**,** 592-7.

DECHERING, K. J., CUELENAERE, K., KONINGS, R. N. & LEUNISSEN, J. A. 1998. Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. *Nucleic Acids Res,* 26**,** 4056-62.

DEMERS, L. M., GINGER, D. S., PARK, S. J., LI, Z., CHUNG, S. W. & MIRKIN, C. A. 2002. Direct patterning of modified oligonucleotides on metals and insulators by dip-pen nanolithography. *Science,* 296**,** 1836-8.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M., MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet,* 43**,** 491-8.

DEVARAJ, B., LEE, A., CABRERA, B. L., MIYAI, K., LUO, L., RAMAMOORTHY, S., KEKU, T., SANDLER, R. S., MCGUIRE, K. L. & CARETHERS, J. M. 2010. Relationship of EMAST and microsatellite instability among patients with rectal cancer. *J Gastrointest Surg,* 14**,** 1521-8.

DIEP, C. B., THORSTENSEN, L., MELING, G. I., SKOVLUND, E., ROGNUM, T. O. & LOTHE, R. A. 2003. Genetic tumor markers with prognostic impact in Dukes' stages B and C colorectal cancer patients. *J Clin Oncol,* 21**,** 820-9.

DOMINGO, E., LAIHO, P., OLLIKAINEN, M., PINTO, M., WANG, L., FRENCH, A. J., WESTRA, J., FREBOURG, T., ESPIN, E., ARMENGOL, M., HAMELIN, R., YAMAMOTO, H., HOFSTRA, R. M., SERUCA, R., LINDBLOM, A., PELTOMAKI, P., THIBODEAU, S. N., AALTONEN, L. A. & SCHWARTZ, S., JR. 2004. BRAF screening as a low-cost effective strategy for simplifying HNPCC genetic testing. *J Med Genet,* 41**,** 664-8.

DORARD, C., DE THONEL, A., COLLURA, A., MARISA, L., SVRCEK, M., LAGRANGE, A., JEGO, G., WANHERDRICK, K., JOLY, A. L., BUHARD, O., GOBBO, J., PENARD-LACRONIQUE, V., ZOUALI, H., TUBACHER, E., KIRZIN, S., SELVES, J., MILANO, G., ETIENNE-GRIMALDI, M. C., BENGRINE-LEFEVRE, L., LOUVET, C., TOURNIGAND, C., LEFEVRE, J. H., PARC, Y., TIRET, E., FLEJOU, J. F., GAUB, M. P., GARRIDO, C. & DUVAL, A. 2011. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. *Nat Med,* 17**,** 1283-9.

ELLEGREN, H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet,* 5**,** 435-45.

ELSALEH, H. & IACOPETTA, B. 2001. Microsatellite instability is a predictive marker for survival benefit from adjuvant chemotherapy in a population-based series of stage III colorectal carcinoma. *Clin Colorectal Cancer,* 1**,** 104-9.

EVERTSON, S., WALLIN, A., ARBMAN, G., RUTTEN, S., EMTERLING, A., ZHANG, H. & SUN, X. F. 2003. Microsatellite instability and MBD4 mutation in unselected colorectal cancer. *Anticancer Res,* 23**,** 3569-74.

FAZEKAS, A., STEEVES, R. & NEWMASTER, S. 2010. Improving sequencing quality from PCR products containing long mononucleotide repeats. *Biotechniques,* 48**,** 277-85.

FEARON, E. R. 2011. Molecular genetics of colorectal cancer. *Annu Rev Pathol,* 6**,** 479-507.

FEARON, E. R., HAMILTON, S. R. & VOGELSTEIN, B. 1987. Clonal analysis of human colorectal tumors. *Science,* 238**,** 193-7.

FEARON, E. R. & VOGELSTEIN, B. 1990. A genetic model for colorectal tumorigenesis. *Cell,* 61**,** 759-67.

FEELEY, K. M., FULLARD, J. F., HENEGHAN, M. A., SMITH, T., MAHER, M., MURPHY, R. P. & O'GORMAN, T. A. 1999. Microsatellite instability in sporadic colorectal carcinoma is not an indicator of prognosis. *J Pathol,* 188**,** 14-7.

FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D. M., FORMAN, D. & BRAY, F. 2015. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer,* 136**,** E359-86.

FISHEL, R., LESCOE, M. K., RAO, M. R., COPELAND, N. G., JENKINS, N. A., GARBER, J., KANE, M. & KOLODNER, R. 1993. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell,* 75**,** 1027-38.

FLORES-RENTERIA, L. & WHIPPLE, A. V. 2011. A new approach to improve the scoring of mononucleotide microsatellite loci. *Am J Bot,* 98**,** e51-3.

GAFA, R., MAESTRI, I., MATTEUZZI, M., SANTINI, A., FERRETTI, S., CAVAZZINI, L. & LANZA, G. 2000. Sporadic colorectal adenocarcinomas with high-frequency microsatellite instability. *Cancer,* 89**,** 2025-37.

GALLUZZI, L., SENOVILLA, L., ZITVOGEL, L. & KROEMER, G. 2012. The secret ally: immunostimulation by anticancer drugs. *Nat Rev Drug Discov,* 11**,** 215-33.

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 (Basel),* 6**,** 46-59.

GAO, Z., AGARWAL, A., TRIGG, A. D., SINGH, N., FANG, C., TUNG, C. H., FAN, Y., BUDDHARAJU, K. D. & KONG, J. 2007. Silicon nanowire arrays for label-free detection of DNA. *Anal Chem,* 79**,** 3291-7.

GILBERT, M. T., HASELKORN, T., BUNCE, M., SANCHEZ, J. J., LUCAS, S. B., JEWELL, L. D., VAN MARCK, E. & WOROBEY, M. 2007. The isolation of nucleic acids from fixed, paraffin-embedded tissues-which methods are useful when? *PLoS One,* 2**,** e537.

GIUFFRE, G., MULLER, A., BRODEGGER, T., BOCKER-EDMONSTON, T., GEBERT, J., KLOOR, M., DIETMAIER, W., KULLMANN, F., BUTTNER, R., TUCCARI, G. & RUSCHOFF, J. 2005. Microsatellite analysis of hereditary nonpolyposis colorectal cancer-associated colorectal adenomas by laser-assisted microdissection: correlation with mismatch repair protein expression provides new insights in early steps of tumorigenesis. *J Mol Diagn,* 7**,** 160-70.

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.

GONZALEZ-GARCIA, I., MORENO, V., NAVARRO, M., MARTI-RAGUE, J., MARCUELLO, E., BENASCO, C., CAMPOS, O., CAPELLA, G. & PEINADO, M. A. 2000. Standardized approach for microsatellite instability detection in colorectal carcinomas. *J Natl Cancer Inst,* 92**,** 544-9.

GRADY, W. M. 2004. Genomic instability and colon cancer. *Cancer Metastasis Rev,* 23**,** 11-27.

GREAVES, M. & MALEY, C. C. 2012. Clonal evolution in cancer. *Nature,* 481**,** 306-13.

GRYFE, R., KIM, H., HSIEH, E. T., ARONSON, M. D., HOLOWATY, E. J., BULL, S. B., REDSTON, M. & GALLINGER, S. 2000. Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *N Engl J Med,* 342**,** 69-77.

GUIDOBONI, M., GAFA, R., VIEL, A., DOGLIONI, C., RUSSO, A., SANTINI, A., DEL TIN, L., MACRI, E., LANZA, G., BOIOCCHI, M. & DOLCETTI, R. 2001. Microsatellite instability and high content of activated cytotoxic lymphocytes identify colon cancer patients with a favorable prognosis. *Am J Pathol,* 159**,** 297-304.

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

GULLAND, A. 2014. Global cancer prevalence is growing at "alarming pace," says WHO. *Bmj,* 348**,** g1338.

HAMPEL, H., FRANKEL, W. L., MARTIN, E., ARNOLD, M., KHANDUJA, K., KUEBLER, P., CLENDENNING, M., SOTAMAA, K., PRIOR, T., WESTMAN, J. A., PANESCU, J., FIX, D., LOCKMAN, J., LAJEUNESSE, J., COMERAS, I. & DE LA CHAPELLE, A. 2008. Feasibility of screening for Lynch syndrome among patients with colorectal cancer. *J Clin Oncol,* 26**,** 5783-8.

HARFE, B. D. & JINKS-ROBERTSON, S. 2000. Sequence composition and context effects on the generation and repair of frameshift intermediates in mononucleotide runs in Saccharomyces cerevisiae. *Genetics,* 156**,** 571-8.

HAUGEN, A. C., GOEL, A., YAMADA, K., MARRA, G., NGUYEN, T. P., NAGASAKA, T., KANAZAWA, S., KOIKE, J., KIKUCHI, Y., ZHONG, X., ARITA, M., SHIBUYA, K., OSHIMURA, M., HEMMI, H., BOLAND, C. R. & KOI, M. 2008. Genetic instability caused by loss of MutS homologue 3 in human colorectal cancer. *Cancer Res,* 68**,** 8465-72.

HEINEMANN, V., STINTZING, S., KIRCHNER, T., BOECK, S. & JUNG, A. 2009. Clinical relevance of EGFR- and KRAS-status in colorectal cancer patients treated with monoclonal antibodies directed against the EGFR. *Cancer Treat Rev,* 35**,** 262-71.

HEMMINKI, A., MECKLIN, J. P., JARVINEN, H., AALTONEN, L. A. & JOENSUU, H. 2000. Microsatellite instability is a favorable prognostic indicator in patients with colorectal cancer receiving chemotherapy. *Gastroenterology,* 119**,** 921-8.

HEMPELMANN, J. A., SCROGGINS, S. M., PRITCHARD, C. C. & SALIPANTE, S. J. 2015. MSIplus: Integrated Colorectal Cancer Molecular Testing by Next-Generation Sequencing. *J Mol Diagn*.

HEMPEN, P. M., ZHANG, L., BANSAL, R. K., IACOBUZIO-DONAHUE, C. A., MURPHY, K. M., MAITRA, A., VOGELSTEIN, B., WHITEHEAD, R. H., MARKOWITZ, S. D., WILLSON, J. K., YEO, C. J., HRUBAN, R. H. & KERN, S. E. 2003. Evidence of selection for clones having genetic inactivation of the activin A type II receptor (ACVR2) gene in gastrointestinal cancers. *Cancer Res,* 63**,** 994-9.

HENDRIKS, Y. M., WAGNER, A., MORREAU, H., MENKO, F., STORMORKEN, A., QUEHENBERGER, F., SANDKUIJL, L., MOLLER, P., GENUARDI, M., VAN HOUWELINGEN, H., TOPS, C., VAN PUIJENBROEK, M., VERKUIJLEN, P., KENTER, G., VAN MIL, A., MEIJERS-HEIJBOER, H., TAN, G. B., BREUNING, M. H., FODDE, R., WIJNEN, J. T., BROCKER-VRIENDS, A. H. & VASEN, H. 2004. Cancer risk in hereditary nonpolyposis colorectal cancer due to MSH6 mutations: impact on counseling and surveillance. *Gastroenterology,* 127**,** 17-25.

HOANG, J. M., COTTU, P. H., THUILLE, B., SALMON, R. J., THOMAS, G. & HAMELIN, R. 1997. BAT-26, an indicator of the replication error phenotype in colorectal cancers and cell lines. *Cancer Res,* 57**,** 300-3.

HOUNIET, D. T., RAHMAN, T. J., AL TURKI, S., HURLES, M. E., XU, Y., GOODSHIP, J., KEAVNEY, B. & SANTIBANEZ KOREF, M. 2015. Using population data for assessing next-generation sequencing performance. *Bioinformatics,* 31**,** 56-61.

HSIEH, P. & YAMANE, K. 2008. DNA mismatch repair: molecular mechanism, cancer, and ageing. *Mech Ageing Dev,* 129**,** 391-407.

HUTCHINS, G., SOUTHWARD, K., HANDLEY, K., MAGILL, L., BEAUMONT, C., STAHLSCHMIDT, J., RICHMAN, S., CHAMBERS, P., SEYMOUR, M., KERR, D., GRAY, R. & QUIRKE, P. 2011. Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. *J Clin Oncol,* 29**,** 1261-70.

HUXLEY, R. R., ANSARY-MOGHADDAM, A., CLIFTON, P., CZERNICHOW, S., PARR, C. L. & WOODWARD, M. 2009. The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence. *Int J Cancer,* 125**,** 171-80.

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**,** 558-61.

IYER, R. R., PLUCIENNIK, A., BURDETT, V. & MODRICH, P. L. 2006. DNA mismatch repair: functions and mechanisms. *Chem Rev,* 106**,** 302-23.

JASS, J. R. 2004. HNPCC and sporadic MSI-H colorectal cancer: a review of the morphological similarities and differences. *Fam Cancer,* 3**,** 93-100.

JERNVALL, P., MAKINEN, M. J., KARTTUNEN, T. J., MAKELA, J. & VIHKO, P. 1999. Microsatellite instability: impact on cancer progression in proximal and distal colorectal cancers. *Eur J Cancer,* 35**,** 197-201.

JIN, M., HAMPEL, H., ZHOU, X., SCHUNEMANN, L., YEARSLEY, M. & FRANKEL, W. L. 2013. BRAF V600E mutation analysis simplifies the testing algorithm for Lynch syndrome. *Am J Clin Pathol,* 140**,** 177-83.

JOHANNSDOTTIR, J. T., BERGTHORSSON, J. T., GRETARSDOTTIR, S., KRISTJANSSON, A. K., RAGNARSSON, G., JONASSON, J. G., EGILSSON, V. & INGVARSSON, S. 1999. Replication error in colorectal carcinoma: association with loss of heterozygosity at mismatch repair loci and clinicopathological variables. *Anticancer Res,* 19**,** 1821-6.

JULIE, C., TRESALLET, C., BROUQUET, A., VALLOT, C., ZIMMERMANN, U., MITRY, E., RADVANYI, F., ROULEAU, E., LIDEREAU, R., COULET, F., OLSCHWANG, S., FREBOURG, T., ROUGIER, P., NORDLINGER, B., LAURENT-PUIG, P., PENNA, C., BOILEAU, C., FRANC, B., MUTI, C. & HOFMANN-RADVANYI, H. 2008. Identification in daily practice of patients with Lynch syndrome (hereditary nonpolyposis colorectal cancer): revised Bethesda guidelines-based approach versus molecular screening. *Am J Gastroenterol,* 103**,** 2825-35; quiz 2836.

KEMPERS, M. J., KUIPER, R. P., OCKELOEN, C. W., CHAPPUIS, P. O., HUTTER, P., RAHNER, N., SCHACKERT, H. K., STEINKE, V., HOLINSKI-FEDER, E., MORAK, M., KLOOR, M., BUTTNER, R., VERWIEL, E. T., VAN KRIEKEN, J. H., NAGTEGAAL, I. D., GOOSSENS, M., VAN DER POST, R. S., NIESSEN, R. C., SIJMONS, R. H., KLUIJT, I., HOGERVORST, F. B., LETER, E. M., GILLE, J. J., AALFS, C. M., REDEKER, E. J., HES, F. J., TOPS, C. M., VAN NESSELROOIJ, B. P., VAN GIJN, M. E., GOMEZ GARCIA, E. B., ECCLES, D. M., BUNYAN, D. J., SYNGAL, S., STOFFEL, E. M., CULVER, J. O., PALOMARES, M. R., GRAHAM, T., VELSHER, L., PAPP, J., OLAH, E., CHAN, T. L., LEUNG, S. Y., VAN KESSEL, A. G., KIEMENEY, L. A., HOOGERBRUGGE, N. & LIGTENBERG, M. J. 2011. Risk of colorectal and endometrial cancers in EPCAM deletion-positive Lynch syndrome: a cohort study. *Lancet Oncol,* 12**,** 49-55.

KENT, W. J. 2002. BLAT--the BLAST-like alignment tool. *Genome Res,* 12**,** 656-64.

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 Res,* 12**,** 996-1006.

KIKUCHI, H., PINO, M. S., ZENG, M., SHIRASAWA, S. & CHUNG, D. C. 2009. Oncogenic KRAS and BRAF differentially regulate hypoxia-inducible factor-1alpha and -2alpha in colon cancer. *Cancer Res,* 69**,** 8499-506.

KIM, J. A., LEE, J. Y., SEONG, S., CHA, S. H., LEE, S. H., KIM, J. J. & PARK, T. H. 2006. Fabrication and characterization of a PDMS–glass hybrid continuous-flow PCR chip. *Biochemical Engineering Journal,* 29**,** 91-97.

KLOOR, M. pers. Comm.

KLOOR, M., HUTH, C., VOIGT, A. Y., BENNER, A., SCHIRMACHER, P., VON KNEBEL DOEBERITZ, M. & BLAKER, H. 2012. Prevalence of mismatch repair-deficient crypt foci in Lynch syndrome: a pathological study. *Lancet Oncol,* 13**,** 598-606.

KOBOLDT, D. C., CHEN, K., WYLIE, T., LARSON, D. E., MCLELLAN, M. D., MARDIS, E. R., WEINSTOCK, G. M., WILSON, R. K. & DING, L. 2009. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics,* 25**,** 2283-5.

KOCHHAR, R., HALLING, K. C., MCDONNELL, S., SCHAID, D. J., FRENCH, A. J., O'CONNELL, M. J., NAGORNEY, D. M. & THIBODEAU, S. N. 1997. Allelic imbalance and microsatellite instability in resected Duke's D colorectal cancer. *Diagn Mol Pathol,* 6**,** 78-84.

KORONA, D. A., LECOMPTE, K. G. & PURSELL, Z. F. 2011. The high fidelity and unique error signature of human DNA polymerase epsilon. *Nucleic Acids Res,* 39**,** 1763-73.

KUNKEL, T. A. 2004. DNA replication fidelity. *J Biol Chem,* 279**,** 16895-8.

KUNKEL, T. A. & ERIE, D. A. 2005. DNA mismatch repair. *Annu Rev Biochem,* 74**,** 681-710.

LANG, G. I., PARSONS, L. & GAMMIE, A. E. 2013. Mutation rates, spectra, and genome-wide distribution of spontaneous mutations in mismatch repair deficient yeast. *G3 (Bethesda),* 3**,** 1453-65.

LE, D. T., URAM, J. N., WANG, H., BARTLETT, B. R., KEMBERLING, H., EYRING, A. D., SKORA, A. D., LUBER, B. S., AZAD, N. S., LAHERU, D., BIEDRZYCKI, B., DONEHOWER, R. C., ZAHEER, A., FISHER, G. A., CROCENZI, T. S., LEE, J. J., DUFFY, S. M., GOLDBERG, R. M., DE LA CHAPELLE, A., KOSHIJI, M., BHAIJEE, F., HUEBNER, T., HRUBAN, R. H., WOOD, L. D., CUKA, N., PARDOLL, D. M., PAPADOPOULOS, N., KINZLER, K. W., ZHOU, S., CORNISH, T. C., TAUBE, J. M., ANDERS, R. A., ESHLEMAN, J. R., VOGELSTEIN, B. & DIAZ, L. A. 2015. PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine,* 372**,** 2509-2520.

LEACH, F. S., NICOLAIDES, N. C., PAPADOPOULOS, N., LIU, B., JEN, J., PARSONS, R., PELTOMAKI, P., SISTONEN, P., AALTONEN, L. A., NYSTROM-LAHTI, M. & ET AL. 1993. Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell,* 75**,** 1215-25.

LEEDHAM, S. J. & WRIGHT, N. A. 2008. Human tumour clonality assessment--flawed but necessary. *J Pathol,* 215**,** 351-4.

LI, H. 2014. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics,* 30**,** 2843-51.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIANG, J. T., CHANG, K. J., CHEN, J. C., LEE, C. C., CHENG, Y. M., HSU, H. C., CHIEN, C. T. & WANG, S. M. 1999. Clinicopathologic and carcinogenetic appraisal of DNA replication error in sporadic T3N0M0 stage colorectal cancer after curative resection. *Hepatogastroenterology,* 46**,** 883-90.

LIANG, J. T., HUANG, K. C., LAI, H. S., LEE, P. H., CHENG, Y. M., HSU, H. C., CHENG, A. L., HSU, C. H., YEH, K. H., WANG, S. M., TANG, C. & CHANG, K. J. 2002. High-frequency microsatellite instability predicts better chemosensitivity to high-dose 5-fluorouracil plus leucovorin chemotherapy for stage IV sporadic colorectal cancer after palliative bowel resection. *Int J Cancer,* 101**,** 519-25.

LIPTON, L. R., JOHNSON, V., CUMMINGS, C., FISHER, S., RISBY, P., EFTEKHAR SADAT, A. T., CRANSTON, T., IZATT, L., SASIENI, P., HODGSON, S. V., THOMAS, H. J. & TOMLINSON, I. P. 2004. Refining the Amsterdam Criteria and Bethesda Guidelines: testing algorithms for the prediction of mismatch repair mutation status in the familial cancer clinic. *J Clin Oncol,* 22**,** 4934-43.

LOUKOLA, A., SALOVAARA, R., KRISTO, P., MOISIO, A. L., KAARIAINEN, H., AHTOLA, H., ESKELINEN, M., HARKONEN, N., JULKUNEN, R., KANGAS, E., OJALA, S., TULIKOURA, J., VALKAMO, E., JARVINEN, H., MECKLIN, J. P., DE LA CHAPELLE, A. & AALTONEN, L. A. 1999. Microsatellite instability in adenomas as a marker for hereditary nonpolyposis colorectal cancer. *Am J Pathol,* 155**,** 1849-53.

LU, Y., SOONG, T. D. & ELEMENTO, O. 2013. A novel approach for characterizing microsatellite instability in cancer cells. *PLoS One,* 8**,** e63056.

LUCERI, C., DE FILIPPO, C., GUGLIELMI, F., CADERNI, G., MESSERINI, L., BIGGERI, A., MINI, E., TONELLI, F., CIANCHI, F. & DOLARA, P. 2002. Microsatellite instability in a population of sporadic colorectal cancers: correlation between genetic and pathological profiles. *Dig Liver Dis,* 34**,** 553-9.

LUJAN, S. A., WILLIAMS, J. S., CLAUSEN, A. R., CLARK, A. B. & KUNKEL, T. A. 2013. Ribonucleotides are signals for mismatch repair of leading-strand replication errors. *Mol Cell,* 50**,** 437-43.

LUKISH, J. R., MURO, K., DENOBILE, J., KATZ, R., WILLIAMS, J., CRUESS, D. F., DRUCKER, W., KIRSCH, I. & HAMILTON, S. R. 1998. Prognostic significance of DNA replication errors in young patients with colorectal cancer. *Ann Surg,* 227**,** 51-6.

LYNCH, H. T. 1985. Aldred scott warthin, m.d., ph.d. (1866-1931). *CA: A Cancer Journal for Clinicians,* 35**,** 345-347.

LYNCH, H. T. & SMYRK, T. 1996. Hereditary nonpolyposis colorectal cancer (Lynch syndrome): An updated review. *Cancer,* 78**,** 1149-1167.

LYONS, S. M. & SCHENDEL, P. F. 1984. Kinetics of methylation in Escherichia coli K-12. *J Bacteriol,* 159**,** 421-3.

MA, X., ROGACHEVA, M. V., NISHANT, K. T., ZANDERS, S., BUSTAMANTE, C. D. & ALANI, E. 2012. Mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Rep,* 1**,** 36-42.

MALEY, C. C., GALIPEAU, P. C., FINLEY, J. C., WONGSURAWAT, V. J., LI, X., SANCHEZ, C. A., PAULSON, T. G., BLOUNT, P. L., RISQUES, R. A., RABINOVITCH, P. S. & REID, B. J. 2006. Genetic clonal diversity predicts progression to esophageal adenocarcinoma. *Nat Genet,* 38**,** 468-73.

MARTIN-LOPEZ, J. V. & FISHEL, R. 2013. The mechanism of mismatch repair and the functional analysis of mismatch repair defects in Lynch syndrome. *Fam Cancer,* 12**,** 159-68.

MASSA, M. J., INIESTA, P., GONZALEZ-QUEVEDO, R., DE JUAN, C., CALDES, T., SANCHEZ-PERNAUTE, A., CERDAN, J., TORRES, A. J., BALIBREA, J. L. & BENITO, M. 1999. Differential prognosis of replication error phenotype and loss of heterozygosity in sporadic colorectal cancer. *Eur J Cancer,* 35**,** 1676-82.

MCGIVERN, A., WYNTER, C. V., WHITEHALL, V. L., KAMBARA, T., SPRING, K. J., WALSH, M. D., BARKER, M. A., ARNOLD, S., SIMMS, L. A., LEGGETT,

B. A., YOUNG, J. & JASS, J. R. 2004. Promoter hypermethylation frequency and BRAF mutations distinguish hereditary non-polyposis colon cancer from sporadic MSI-H colon cancer. *Fam Cancer,* 3**,** 101-7.

MERLO, L. M., PEPPER, J. W., REID, B. J. & MALEY, C. C. 2006. Cancer as an evolutionary and ecological process. *Nat Rev Cancer,* 6**,** 924-35.

MESSERINI, L., CIANTELLI, M., BAGLIONI, S., PALOMBA, A., ZAMPI, G. & PAPI, L. 1999. Prognostic significance of microsatellite instability in sporadic mucinous colorectal cancers. *Hum Pathol,* 30**,** 629-34.

MILLS, A. M., LIOU, S., FORD, J. M., BEREK, J. S., PAI, R. K. & LONGACRE, T. A. 2014. Lynch syndrome screening should be considered for all patients with newly diagnosed endometrial cancer. *Am J Surg Pathol,* 38**,** 1501-9.

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 Biol,* 12**,** R112.

NELSON, H. C., FINCH, J. T., LUISI, B. F. & KLUG, A. 1987. The structure of an oligo(dA).oligo(dT) tract and its biological implications. *Nature,* 330**,** 221-6.

NEUMAN, J. A., ISAKOV, O. & SHOMRON, N. 2013. Analysis of insertion-deletion from deep-sequencing data: software evaluation for optimal detection. *Brief Bioinform,* 14**,** 46-55.

NICK MCELHINNY, S. A., KISSLING, G. E. & KUNKEL, T. A. 2010. Differential correction of lagging-strand replication errors made by DNA polymerases α and δ. *Proc Natl Acad Sci U S A,* 107**,** 21070-5.

NIU, B., YE, K., ZHANG, Q., LU, C., XIE, M., MCLELLAN, M. D., WENDL, M. C. & DING, L. 2014. MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. *Bioinformatics,* 30**,** 1015-6.

O'RAWE, J., JIANG, T., SUN, G., WU, Y., WANG, W., HU, J., BODILY, P., TIAN, L., HAKONARSON, H., JOHNSON, W. E., WEI, Z., WANG, K. & LYON, G. J. 2013. Low concordance of multiple variant-calling pipelines: practical implications for exome and genome sequencing. *Genome Med,* 5**,** 28.

OVERBEEK, L. I., LIGTENBERG, M. J., WILLEMS, R. W., HERMENS, R. P., BLOKX, W. A., DUBOIS, S. V., VAN DER LINDEN, H., MEIJER, J. W., MLYNEK-KERSJES, M. L., HOOGERBRUGGE, N., HEBEDA, K. M. & VAN KRIEKEN, J. H. 2008. Interpretation of immunohistochemistry for mismatch repair proteins is only reliable in a specialized setting. *Am J Surg Pathol,* 32**,** 1246-51.

PABINGER, S., DANDER, A., FISCHER, M., SNAJDER, R., SPERK, M., EFREMOVA, M., KRABICHLER, B., SPEICHER, M. R., ZSCHOCKE, J. & TRAJANOSKI, Z. 2014. A survey of tools for variant analysis of next-generation genome sequencing data. *Brief Bioinform,* 15**,** 256-78.

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

PEDERSEN T. https://metacpan.org/pod/Text::NSP::Measures::2D::Fisher::twotailed. Text::NSP::Measures::2D::Fisher::twotailed - Perl module implementation of the two-sided Fisher's exact test.

PELTOMAKI, P., AALTONEN, L. A., SISTONEN, P., PYLKKANEN, L., MECKLIN, J. P., JARVINEN, H., GREEN, J. S., JASS, J. R., WEBER, J. L., LEACH, F. S. & ET AL. 1993. Genetic mapping of a locus predisposing to human colorectal cancer. *Science,* 260**,** 810-2.

PEREZ-CARBONELL, L., RUIZ-PONTE, C., GUARINOS, C., ALENDA, C., PAYA, A., BREA, A., EGOAVIL, C. M., CASTILLEJO, A., BARBERA, V. M., BESSA, X., XICOLA, R. M., RODRIGUEZ-SOLER, M., SANCHEZ-FORTUN, C., ACAME, N., CASTELLVI-BEL, S., PINOL, V., BALAGUER, F., BUJANDA,

L., DE-CASTRO, M. L., LLOR, X., ANDREU, M., CARRACEDO, A., SOTO, J. L., CASTELLS, A. & JOVER, R. 2012. Comparison between universal molecular screening for Lynch syndrome and revised Bethesda guidelines in a large population-based cohort of patients with colorectal cancer. *Gut,* 61**,** 865-72.

PICARD [HTTP://PICARD.SOURCEFORGE.NET].

PINO, M. S. & CHUNG, D. C. 2010. The chromosomal instability pathway in colon cancer. *Gastroenterology,* 138**,** 2059-72.

POGUE-GEILE, K., YOTHERS, G., TANIYAMA, Y., TANAKA, N., GAVIN, P., COLANGELO, L., BLACKMON, N., LIPCHIK, C., KIM, S. R., SHARIF, S., ALLEGRA, C., PETRELLI, N., O'CONNELL, M. J., WOLMARK, N. & PAIK, S. 2013. Defective mismatch repair and benefit from bevacizumab for colon cancer: findings from NSABP C-08. *J Natl Cancer Inst,* 105**,** 989-92.

POPAT, S., HUBNER, R. & HOULSTON, R. S. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. *J Clin Oncol,* 23**,** 609-18.

POURHOSEINGHOLI, M. A., VAHEDI, M. & BAGHESTANI, A. R. 2015. Burden of gastrointestinal cancer in Asia; an overview. *Gastroenterol Hepatol Bed Bench,* 8**,** 19-27.

PUKKILA, P. J., PETERSON, J., HERMAN, G., MODRICH, P. & MESELSON, M. 1983. Effects of high levels of DNA adenine methylation on methyl-directed mismatch repair in Escherichia coli. *Genetics,* 104**,** 571-82.

QIU, R., DEROCCO, V. C., HARRIS, C., SHARMA, A., HINGORANI, M. M., ERIE, D. A. & WENINGER, K. R. 2012. Large conformational changes in MutS during DNA scanning, mismatch recognition and repair signalling. *Embo j,* 31**,** 2528-40.

R CORE TEAM R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing.

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., TU, D., REDSTON, M. & GALLINGER, S. 2003. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *N Engl J Med,* 349**,** 247-57.

ROBINSON, J. T., THORVALDSDOTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nat Biotech,* 29**,** 24-26.

ROSE, O. & FALUSH, D. 1998. A threshold size for microsatellite expansion. *Mol Biol Evol,* 15**,** 613-5.

ROZEN, S. & SKALETSKY, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol,* 132**,** 365-86.

SALAHSHOR, S., KRESSNER, U., FISCHER, H., LINDMARK, G., GLIMELIUS, B., PAHLMAN, L. & LINDBLOM, A. 1999. Microsatellite instability in sporadic colorectal cancer is not an independent prognostic factor. *Br J Cancer,* 81**,** 190-3.

SALIPANTE, S. J., SCROGGINS, S. M., HAMPEL, H. L., TURNER, E. H. & PRITCHARD, C. C. 2014. Microsatellite instability detection by next generation sequencing. *Clin Chem,* 60**,** 1192-9.

SAMMALKORPI, H., ALHOPURO, P., LEHTONEN, R., TUIMALA, J., MECKLIN, J. P., JARVINEN, H. J., JIRICNY, J., KARHU, A. & AALTONEN, L. A. 2007. Background mutation frequency in microsatellite-unstable colorectal cancer. *Cancer Res,* 67**,** 5691-8.

SAMOWITZ, W. S., CURTIN, K., MA, K. N., SCHAFFER, D., COLEMAN, L. W., LEPPERT, M. & SLATTERY, M. L. 2001. Microsatellite instability in sporadic colon cancer is associated with an improved prognosis at the population level. *Cancer Epidemiol Biomarkers Prev,* 10**,** 917-23.

SAMOWITZ, W. S., SWEENEY, C., HERRICK, J., ALBERTSEN, H., LEVIN, T. R., MURTAUGH, M. A., WOLFF, R. K. & SLATTERY, M. L. 2005. Poor survival associated with the BRAF V600E mutation in microsatellite-stable colon cancers. *Cancer Res,* 65**,** 6063-9.

SARGENT, D. J., MARSONI, S., MONGES, G., THIBODEAU, S. N., LABIANCA, R., HAMILTON, S. R., FRENCH, A. J., KABAT, B., FOSTER, N. R., TORRI, V., RIBIC, C., GROTHEY, A., MOORE, M., ZANIBONI, A., SEITZ, J. F., SINICROPE, F. & GALLINGER, S. 2010. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *J Clin Oncol,* 28**,** 3219-26.

SCHOFIELD, L., WATSON, N., GRIEU, F., LI, W. Q., ZEPS, N., HARVEY, J., STEWART, C., ABDO, M., GOLDBLATT, J. & IACOPETTA, B. 2009. Population-based detection of Lynch syndrome in young colorectal cancer patients using microsatellite instability as the initial test. *Int J Cancer,* 124**,** 1097-102.

SCHWITALLE, Y., KLOOR, M., EIERMANN, S., LINNEBACHER, M., KIENLE, P., KNAEBEL, H. P., TARIVERDIAN, M., BENNER, A. & VON KNEBEL DOEBERITZ, M. 2008. Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers. *Gastroenterology,* 134**,** 988-97.

SHENDURE, J. & JI, H. 2008. Next-generation DNA sequencing. *Nat Biotech,* 26**,** 1135-1145.

SHERRY, S. T., WARD, M. H., KHOLODOV, M., BAKER, J., PHAN, L., SMIGIELSKI, E. M. & SIROTKIN, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res,* 29**,** 308-11.

SIDDLE, H. V., KREISS, A., ELDRIDGE, M. D., NOONAN, E., CLARKE, C. J., PYECROFT, S., WOODS, G. M. & BELOV, K. 2007. Transmission of a fatal clonal tumor by biting occurs due to depleted MHC diversity in a threatened carnivorous marsupial. *Proc Natl Acad Sci U S A,* 104**,** 16221-6.

SIEGEL, R. L., MILLER, K. D. & JEMAL, A. 2015. Cancer statistics, 2015. *CA: A Cancer Journal for Clinicians,* 65**,** 5-29.

SILVA, F. C., VALENTIN, M. D., FERREIRA FDE, O., CARRARO, D. M. & ROSSI, B. M. 2009. Mismatch repair genes in Lynch syndrome: a review. *Sao Paulo Med J,* 127**,** 46-51.

SIMONS, B. D. & CLEVERS, H. 2011. Stem cell self-renewal in intestinal crypt. *Exp Cell Res,* 317**,** 2719-24.

SINICROPE, F. A. & SARGENT, D. J. 2012. Molecular pathways: microsatellite instability in colorectal cancer: prognostic, predictive, and therapeutic implications. *Clin Cancer Res,* 18**,** 1506-12.

SNOWSILL, T., HUXLEY, N., HOYLE, M., JONES-HUGHES, T., COELHO, H., COOPER, C., FRAYLING, I. & HYDE, C. 2014. A systematic review and economic evaluation of diagnostic strategies for Lynch syndrome. *Health Technol Assess,* 18**,** 1-406.

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**,** 313.

STOFFEL, E., MUKHERJEE, B., RAYMOND, V. M., TAYOB, N., KASTRINOS, F., SPARR, J., WANG, F., BANDIPALLIAM, P., SYNGAL, S. & GRUBER, S. B. 2009. Calculation of Risk of Colorectal and Endometrial Cancer Among Patients with Lynch Syndrome. *Gastroenterology,* 137**,** 1621-7.

STRANNEHEIM, H. & LUNDEBERG, J. 2012. Stepping stones in DNA sequencing. *Biotechnology Journal,* 7**,** 1063-1073.

STRAUSS, B. S. 1999. Frameshift mutation, microsatellites and mismatch repair. *Mutat Res,* 437**,** 195-203.

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**,** 1804-11.

TERDIMAN, J. P., GUM, J. R., JR., CONRAD, P. G., MILLER, G. A., WEINBERG, V., CRAWLEY, S. C., LEVIN, T. R., REEVES, C., SCHMITT, A., HEPBURN, M., SLEISENGER, M. H. & KIM, Y. S. 2001. Efficient detection of hereditary nonpolyposis colorectal cancer gene carriers by screening for tumor microsatellite instability before germline genetic testing. *Gastroenterology,* 120**,** 21-30.

THIBODEAU, S. N., BREN, G. & SCHAID, D. 1993. Microsatellite instability in cancer of the proximal colon. *Science,* 260**,** 816-9.

THIRLWELL, C., WILL, O. C., DOMINGO, E., GRAHAM, T. A., MCDONALD, S. A., OUKRIF, D., JEFFREY, R., GORMAN, M., RODRIGUEZ-JUSTO, M., CHIN-ALEONG, J., CLARK, S. K., NOVELLI, M. R., JANKOWSKI, J. A., WRIGHT, N. A., TOMLINSON, I. P. & LEEDHAM, S. J. 2010. Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology,* 138**,** 1441-54, 1454.e1-7.

THORSTENSEN, L., LIND, G. E., LOVIG, T., DIEP, C. B., MELING, G. I., ROGNUM, T. O. & LOTHE, R. A. 2005. Genetic and epigenetic changes of components affecting the WNT pathway in colorectal carcinomas stratified by microsatellite instability. *Neoplasia,* 7**,** 99-108.

TOMLINSON, I., HALFORD, S., AALTONEN, L., HAWKINS, N. & WARD, R. 2002. Does MSI-low exist? *J Pathol,* 197**,** 6-13.

UMAR, A. 2004. Lynch syndrome (HNPCC) and microsatellite instability. *Dis Markers,* 20**,** 179-80.

UMAR, A. 2006. Lynch syndrome (HNPCC) and microsatellite instability analysis guidelines. *Cancer Biomark,* 2**,** 1-4.

UMAR, A., BOLAND, C. R., TERDIMAN, J. P., SYNGAL, S., DE LA CHAPELLE, A., RUSCHOFF, J., FISHEL, R., LINDOR, N. M., BURGART, L. J., HAMELIN, R., HAMILTON, S. R., HIATT, R. A., JASS, J., LINDBLOM, A., LYNCH, H. T., PELTOMAKI, P., RAMSEY, S. D., RODRIGUEZ-BIGAS, M. A., VASEN, H. F., HAWK, E. T., BARRETT, J. C., FREEDMAN, A. N. & SRIVASTAVA, S. 2004. Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute,* 96**,** 261-268.

VAN DER POST, R. S., KIEMENEY, L. A., LIGTENBERG, M. J., WITJES, J. A., HULSBERGEN-VAN DE KAA, C. A., BODMER, D., SCHAAP, L., KETS, C. M., VAN KRIEKEN, J. H. & HOOGERBRUGGE, N. 2010. Risk of urothelial bladder cancer in Lynch syndrome is increased, in particular among MSH2 mutation carriers. *J Med Genet,* 47**,** 464-70.

VASEN, H. F., BLANCO, I., AKTAN-COLLAN, K., GOPIE, J. P., ALONSO, A., ARETZ, S., BERNSTEIN, I., BERTARIO, L., BURN, J., CAPELLA, G., COLAS, C., ENGEL, C., FRAYLING, I. M., GENUARDI, M., HEINIMANN, K., HES, F. J., HODGSON, S. V., KARAGIANNIS, J. A., LALLOO, F., LINDBLOM, A., MECKLIN, J. P., MOLLER, P., MYRHOJ, T., NAGENGAST, F. M., PARC, Y., PONZ DE LEON, M., RENKONEN-SINISALO, L., SAMPSON, J. R., STORMORKEN, A., SIJMONS, R. H., TEJPAR, S.,

THOMAS, H. J., RAHNER, N., WIJNEN, J. T., JARVINEN, H. J. & MOSLEIN, G. 2013. Revised guidelines for the clinical management of Lynch syndrome (HNPCC): recommendations by a group of European experts. *Gut,* 62**,** 812-23.

VASEN, H. F., MOSLEIN, G., ALONSO, A., ARETZ, S., BERNSTEIN, I., BERTARIO, L., BLANCO, I., BULOW, S., BURN, J., CAPELLA, G., COLAS, C., ENGEL, C., FRAYLING, I., RAHNER, N., HES, F. J., HODGSON, S., MECKLIN, J. P., MOLLER, P., MYRHOJ, T., NAGENGAST, F. M., PARC, Y., PONZ DE LEON, M., RENKONEN-SINISALO, L., SAMPSON, J. R., STORMORKEN, A., TEJPAR, S., THOMAS, H. J., WIJNEN, J., LUBINSKI, J., JARVINEN, H., CLAES, E., HEINIMANN, K., KARAGIANNIS, J. A., LINDBLOM, A., DOVE-EDWIN, I. & MULLER, H. 2010. Recommendations to improve identification of hereditary and familial colorectal cancer in Europe. *Fam Cancer,* 9**,** 109-15.

VASEN, H. F., WATSON, P., MECKLIN, J. P. & LYNCH, H. T. 1999. New clinical criteria for hereditary nonpolyposis colorectal cancer (HNPCC, Lynch syndrome) proposed by the International Collaborative group on HNPCC. *Gastroenterology,* 116**,** 1453-1456.

VILKKI, S., LAUNONEN, V., KARHU, A., SISTONEN, P., VASTRIK, I. & AALTONEN, L. A. 2002. Screening for microsatellite instability target genes in colorectal cancers. *J Med Genet,* 39**,** 785-9.

WANG, C., SCHROEDER, K. B. & ROSENBERG, N. A. 2012. A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics,* 192**,** 651-69.

WANG, C., VAN RIJNSOEVER, M., GRIEU, F., BYDDER, S., ELSALEH, H., JOSEPH, D., HARVEY, J. & IACOPETTA, B. 2003. Prognostic significance of microsatellite instability and Ki-ras mutation type in stage II colorectal cancer. *Oncology,* 64**,** 259-65.

WARD, R. L., CHEONG, K., KU, S. L., MEAGHER, A., O'CONNOR, T. & HAWKINS, N. J. 2003. Adverse prognostic effect of methylation in colorectal cancer is reversed by microsatellite instability. *J Clin Oncol,* 21**,** 3729-36.

WARTHIN, A. S. 1985. Classics in oncology. Heredity with reference to carcinoma as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. By Aldred Scott Warthin. 1913. *CA Cancer J Clin,* 35**,** 348-59.

WATANABE, T., WU, T. T., CATALANO, P. J., UEKI, T., SATRIANO, R., HALLER, D. G., BENSON, A. B., 3RD & HAMILTON, S. R. 2001. Molecular predictors of survival after adjuvant chemotherapy for colon cancer. *N Engl J Med,* 344**,** 1196-206.

WHYTE, S., CHILCOTT, J. & HALLORAN, S. 2012. Reappraisal of the options for colorectal cancer screening in England. *Colorectal Dis,* 14**,** 1463-1318.

WILLIAMS, D. S., BIRD, M. J., JORISSEN, R. N., YU, Y. L., WALKER, F., ZHANG, H. H., NICE, E. C. & BURGESS, A. W. 2010. Nonsense mediated decay resistant mutations are a source of expressed mutant proteins in colon cancer cell lines with microsatellite instability. *PLoS One,* 5**,** e16012.

WISTUBA, II 2007. Genetics of preneoplasia: lessons from lung cancer. *Curr Mol Med,* 7**,** 3-14.

WOERNER, S. M., BENNER, A., SUTTER, C., SCHILLER, M., YUAN, Y. P., KELLER, G., BORK, P., DOEBERITZ, M. & GEBERT, J. F. 2003. Pathogenesis of DNA repair-deficient cancers: a statistical meta-analysis of putative Real Common Target genes. *Oncogene,* 22**,** 2226-35.

WOERNER, S. M., GEBERT, J., YUAN, Y. P., SUTTER, C., RIDDER, R., BORK, P. & VON KNEBEL DOEBERITZ, M. 2001. Systematic identification of genes

with coding microsatellites mutated in DNA mismatch repair-deficient cancer cells. *Int J Cancer,* 93**,** 12-9.

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 Res,* 38**,** D682-9.

WRIGHT, C. M., DENT, O. F., BARKER, M., NEWLAND, R. C., CHAPUIS, P. H., BOKEY, E. L., YOUNG, J. P., LEGGETT, B. A., JASS, J. R. & MACDONALD, G. A. 2000. Prognostic significance of extensive microsatellite instability in sporadic clinicopathological stage C colorectal cancer. *Br J Surg,* 87**,** 1197-202.

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., PARK, J. H., KIM, H. M., LEE, S. Y., PARK, J., KIM, W. H., BHAK, J., YANG, H. K. & KIM, S. J. 2013. Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. *Genome Res,* 23**,** 1109-17.

YUAN, Z., SHIN, J., WILSON, A., GOEL, S., LING, Y. H., AHMED, N., DOPESO, H., JHAWER, M., NASSER, S., MONTAGNA, C., FORDYCE, K., AUGENLICHT, L. H., AALTONEN, L. A., ARANGO, D., WEBER, T. K. & MARIADASON, J. M. 2009. An A13 repeat within the 3'-untranslated region of epidermal growth factor receptor (EGFR) is frequently mutated in microsatellite instability colon cancers and is associated with increased EGFR expression. *Cancer Res,* 69**,** 7811-8.

ZHANG, L., YU, J., WILLSON, J. K., MARKOWITZ, S. D., KINZLER, K. W. & VOGELSTEIN, B. 2001. Short mononucleotide repeat sequence variability in mismatch repair-deficient cancers. *Cancer Res,* 61**,** 3801-5.

ZHOU, X. P., HOANG, J. M., LI, Y. J., SERUCA, R., CARNEIRO, F., SOBRINHO-SIMOES, M., LOTHE, R. A., GLEESON, C. M., RUSSELL, S. E., MUZEAU, F., FLEJOU, J. F., HOANG-XUAN, K., LIDEREAU, R., THOMAS, G. & HAMELIN, R. 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 Cancer,* 21**,** 101-7.