Role of secondary mismatch repair (MMR) frameshifts in the evolution of microsatellite instable (MSI) colorectal cancer

Hamzeh Kayhanian

University College London

Supervisor: Marnix Jansen

A thesis submitted for the degree of
Doctor of Philosophy

University College London

June 2021

Declaration

I, Hamzeh Kayhanian confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

[redacted]

Hamzeh Kayhanian
20/06/2021

Abstract

Mismatch repair deficient (MMRd) cancers face a delicate balance. Whilst hypermutation fuels adaptive evolution, it also comes at the cost of immunogenic neoantigens and other deleterious mutations. How MMRd cancers navigate the costs versus benefits of hypermutation is unknown. By visualising the clonal architecture of MMRd colorectal cancer in situ I show that the mismatch repair system unfolds in reversible steps to adapt cellular mutability to immune selection. Mechanistically microsatellite instability unmasks two hypermutable homopolymers in the mismatch repair genes *MSH6* and *MSH3*. Spontaneous frameshift mutation and reversion at these homopolymers allows them to act as a molecular switch, regulating expression of MutS$_{\alpha}$ and MutS$_{\text{ß}}$ respectively. Frameshift switching at these homopolymer sites modulates the rate and spectrum of mutations across the genome. In this manner stochastic mutation bursts combined with stringent immune selection, drive continuous adaptation. This work is supported by a bespoke clonally resolved exome sequencing dataset, validated using two large publicly available genomic datasets and tested in a mathematical model of mutation rate switching. In summary, this work identifies that adaptive mutability associates with increased immune escape and intratumour heterogeneity during mismatch repair deficient cancer evolution. Knowledge of this mechanism of adaptation may inform strategies to target resistance evolution in cancer treatment.

Impact statement

The mismatch repair system is the guardian of the genome and protects against post-replicative and spontaneous mutations. Patients with mismatch repair deficient (MMRd) cancer represent an important group because they are candidates for immune-checkpoint inhibitor therapy. There is a pressing need to better understand the evolutionary trajectory taken by MMRd cancer in evading the host immune system. In this work, I have identified a new and unexpected mechanism controlling mutability in MMRd colorectal cancer which associates with increased genetic diversity and immune escape. Knowledge of this pathway may have both translational and basic science applications.

A key finding of the work is that MMRd colorectal cancers fluctuate mutation rate during growth using reversible frameshifts in the *MSH6* and *MSH3* coding homopolymers. This adaptive mutability accelerates immune escape, but also limits the duration of genotoxic damage from prolonged ultra-hypermutation. Knowledge of this pathway may be important in forecasting immunotherapy resistance which should be investigated in future studies. There may also be applications to therapeutically interrupt this pathway by targeting MSH6 and MSH3 with a view to increasing neoantigen burden and immunotherapy response.

This work also finds that hypermutable microsatellites act as molecular on/off switches in MMRd cancer. Whilst this work focused on the role of the *MSH6* and *MSH3* microsatellites there are likely additional genes which are controlled via a similar mechanism. Microsatellite frameshift switching may represent an important mechanism of resistance evolution in MMRd cancer and should be investigated in future studies.

This work also provides new insights into the operation of the mismatch repair system. We find that loss of the mismatch recognition module MutS (via *MSH6* and *MSH3* frameshifts), independently increases mutability despite existing MutL (*MLH1/PMS2*) deficiency. This finding suggests that MutS has a role beyond its interaction with MutL. A suggested explanation is the potential for cross-talk between MutS and other DNA repair pathways such as base excision repair (Mazurek et al., 2002; Sanders et al., 2021). Future work will need to define the precise mechanisms responsible for these suggested non-canonical roles of MutS in the context MutL deficiency.

In summary, this work advances our understanding of MMRd cancer evolution and provides a mechanism for adaptive mutability in mismatch repair deficient cancer. Knowledge of this pathway may in the future allow strategies to improve treatment options for patients with MMRd cancer.

NEOPRISM clinical trial

This research fellowship has given me the opportunity to gain a detailed understanding of mismatch repair deficient cancer biology. This has allowed me to contribute to the set-up of a phase II neoadjuvant clinical trial investigating the use of anti-PD1 immunotherapy in patients with high-risk early-stage MMRd colorectal cancer. The trial is led by Dr Kai-Keen Shiu and I have been able to contribute to protocol development and planning of translational research. The trial promises to provide a rich translational dataset and will allow tracking of population dynamics of adaptive mutability during immunotherapy.

List of publications arising from this work:

1. Kayhanian et al. Adaptive mutability drives immune escape in mismatch repair deficient cancer. **Under review. June 2021.**
2. Sugrue, Hurst, Gordon, Nassar, Hartridge-Lambert, Ntais, Daly, **Kayhanian**, Ryan, Joharatnam, Rodriguez-Justo, Shiu. A real-world experience of metastatic colorectal cancer patients in the UK: a retrospective chart review. **Under review. International journal of colorectal disease (May 2021).**

Poster presentations:
UCL Cancer Institute annual conference, January 2018.

Oral presentations:
UCL Cancer Institute annual conference, January 2019.

Grant approvals related to this work:
Bowel and Cancer research. Small project grant (May 2019). £45000.

CRUK Clinical research training fellowship. September 2017.

Acknowledgements

This work is dedicated to the patients past and present affected by colorectal cancer. Their difficult journey provides the motivation to conduct research that may improve patient outcomes. I am grateful for the generosity of patients in providing samples for scientific research.

I would like to thank Dr Marnix Jansen for his outstanding supervision throughout the PhD. His thoughtful suggestions, ideas, and ability to see the bigger picture have added significantly to the project. I am also grateful for the time he has dedicated to my development as a clinician scientist. Special thanks also go to Professor Adrienne Flanagan for her support and advice throughout the project.

I am very grateful to our many collaborators with whom it has been a privilege to work with. In particular, I would like to thank, Dr Nischalan Pillay, Dr Christopher Steele, Dr William Cross, Dr Eszter Lakatos, Dominic Patel, Dr Luis Zapata, Dr Giulio Caravagna and Professor Trevor Graham. I thank all members of the Jansen research group and in particular Dr Panagiotis Barmpoutis for his special talents in image analysis. I would also like to thank our clinical collaborators Dr Kai-Keen Shiu and Dr Manuel Rodriguez-Justo.

Finally, I would like to thank my parents for their unwavering and unconditional support. I must also thank my brother Saeed for his helpful perspective and insightful comments throughout this journey.

This work would not have been possible without the generous support of our funders, Cancer Research UK, Rosetrees Trust and Bowel Research UK to whom I am grateful.

Contents

Lists of figures and tables

List of figures

List of tables

Abbreviations

Abbreviation	Description
APM	Antigen presentation machinery
AF	Autofluorescence
BMMRD	Bi-allelic mismatch repair deficiency
Bp	Base pairs
BER	Base excision repair
B2M	Beta-2-microglobulin
CCF	Cancer cell fraction
CPI	Checkpoint inhibitor
CRC	Colorectal cancer
DAB	3,3′-Diaminobenzidine
DBS	Doublet base substitution
DAPI	4′,6-diamidino-2-phenylindole
DNA	Deoxyribonucleic acid
EGFR	Epidermal growth factor receptor
GEL	Genomics England 100,000 genomes project
FDA	Food and Drug Administration
FFPE	Formalin fixed paraffin embedded
H&E	Hematoxylin and eosin
HLA	Human leucocyte antigen
HRP	Horse-raddish peroxidase
ICI	Immune checkpoint inhibitor
IF	Immunofluorescence
IHC	Immunohistochemistry
INDEL	Insertion and deletion
KW	Kruskal wallis
LCM	Laser capture microdissection
MGMT	O^{6}-methylguanine DNA methyltransferase
MMRd	Mismatch repair deficiency
MOBSTER	MOdel Based cluSTering in CancER
MSSS	Microsatellite stable
mTOR	Mammalian target of rapamycin
MSI	Microsatellite instability
MLH1	MutL homolog 1
MSH6	MutS homolog 6
MSH3	MutS homolog 3
MS	Microsatellite
NGS	Next generation sequencing
ORION	FluORescence cell segmentatION workflow
nM	Nano-molar
OS	Overall survival
PFS	Progression free survival
PCR	Polymerase chain reaction
PMS2	Post meiotic segregation increased 2
RNA	Ribonucleic acid
SBS	Single base substitution
SNV	Single nucleotide variant
TCGA	The cancer genome atlas
VAF	Variant allele frequency
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1: Introduction

1.1 Colorectal cancer and mismatch repair deficiency

Colorectal cancer remains a major public health problem and is the second most common cause of cancer death in the UK (CRUK, 2021). Despite advances in screening, precision medicine and genomics, overall survival of patients with metastatic disease remains poor at around 24 months (Biller & Schrag, 2021). In recent years, a major breakthrough in the treatment of certain cancers has been the use of immune checkpoint inhibitors (ICI), with remarkable efficacy observed in non-small cell lung cancer and melanoma (Gandhi et al., 2018; Larkin et al., 2015). Unfortunately, in colorectal cancer the response rate to ICIs has been poor overall (Bendell et al., 2018), but tumours with mismatch repair deficiency (MMRd) have emerged as a subtype deriving significant benefit (Le et al., 2015, 2017), with response rates of around 40% (André et al., 2020). Loss of DNA mismatch repair is observed in approximately 15% (Poynter et al., 2008) of all colorectal cancer cases and leads to a hypermutator phenotype. The relentless accumulation of immunogenic neoantigens renders MMRd tumours responsive to immunotherapy. MMRd tumours thus present a useful model system to study tumour-immune interactions in colorectal cancer.

1.2 Molecular basis for mismatch repair loss in colorectal cancer

Mismatch repair deficient colorectal cancers arise from either a germline or somatic defect in one of the mismatch repair genes; *MLH1*, *PMS2*, *MSH2* or *MSH6* (Kim et al., 2013; Lynch et al., 2009). In most cases (~80%), MMR loss is a sporadic event due to hypermethylation of the *MLH1* promoter (Veigl et al., 1998). *MLH1* hypermethylation is commonly observed in association with somatic *BRAF*$^{\mathrm{V600E}}$ mutation, reflecting the common origin of sporadic cases from serrated adenomas (Kambara et al., 2004; Rustgi, 2013). Alternatively, in germline cases, also known as Lynch syndrome, pathogenic mutations in *MLH1* or *MSH2* are the commonest cause but hereditary defects in any of the MMR genes can be implicated. Rarely germline defects in the gene *EPCAM* which is upstream of MSH2 can also cause Lynch syndrome due to methylation of MSH2 (Huth et al., 2012; Tutlewska et al., 2013). Biallelic germline mismatch repair deficiency (BMMRD) occurs in rare cases and usually presents with colorectal cancer, brain tumours or haematological malignancy at a young age (<20 years) (Durno et al., 2017). A small number of MMRd colorectal cancers also arise due to bi-allelic somatic MMR events (mutation and/or loss of heterozygosity) and are commonly referred to as double-somatic MMR deficient tumours (Mensenkamp et al., 2014).

1.3 Clinical course of MMRd colorectal cancer and immunotherapy response

In early-stage disease, MMRd colorectal cancer is associated with reduced rates of recurrence and improved survival (Gryfe et al., 2000; Lanza et al., 2006). Only around 5% of metastatic colorectal cancer cases display MMR loss and prior to the advent of immune checkpoint inhibitors, metastatic MMRd tumours had a poor overall prognosis(Venderbosch et al., 2014). Immunotherapy has dramatically changed the outlook of metastatic MMRd colorectal cancers, with durable response in a significant proportion of patients (Le et al., 2017). Microsatellite instability (MSI) which is the hallmark of MMR loss, was found to predict immunotherapy response across multiple tumour types resulting in an FDA tissue type agnostic approval of immune checkpoint inhibitors in MSI tumours (FDA, 2017). Initial trials of anti-PD1 checkpoint inhibitors in colorectal cancer, investigated their use in pre-treated metastatic patients who had progressed on chemotherapy. Single agent anti-PD1 therapy with either Pembrolizumab (Keynote 016) or Nivolumab (Checkmate 142) demonstrated response rates of between 30-40%(Le et al., 2015; Overman et al., 2017). Combination immunotherapy using Ipilimumab (anti-CTLA4) and Nivolumab demonstrated a response rate of 55% and progression free survival rate of 71% at 12 months(Overman et al., 2018). More recently efficacy of anti-PD1 therapy in the first-line metastatic setting has also been demonstrated (André et al., 2020). Despite the successes, these studies also show that approximately 40-50% of patients with MMRd colorectal cancer do not achieve long-term benefit and ultimately progress. In a small study of 22 patients, Schrock et al found that tumour mutation burden associated with response to anti-PD1 immunotherapy in metastatic MSI colorectal cancer (Schrock et al., 2019). However, this study was limited by the small sample size. The underlying reason for lack of response in a significant proportion of patients with MMRd colorectal cancer remains unclear.

1.4 Mechanism of action of the mismatch repair system

The mismatch repair system is a highly choreographed process that is conserved between species (Marti et al., 2002). The main mismatch repair proteins have distinctive roles in the MMR pathway. First, mismatch recognition is performed by the MutS heterodimer. MutS can exist either as MutSα due to dimerization between MSH2 and MSH6 or as MutSβ due to dimerization between MSH2 and MSH3. MutSα and MutSβ have partially overlapping roles, but MutSα has stronger affinity for single base substitutions and short (1-2bp) insertion-deletion loops, whilst MutSβ is more involved with recognition of longer insertion-deletion loops (Hsieh & Zhang, 2017). Execution of mismatch repair is performed by the MutL complex comprising of MLH1 and PMS2. The MMR pathway is illustrated in figure 1.1 below.

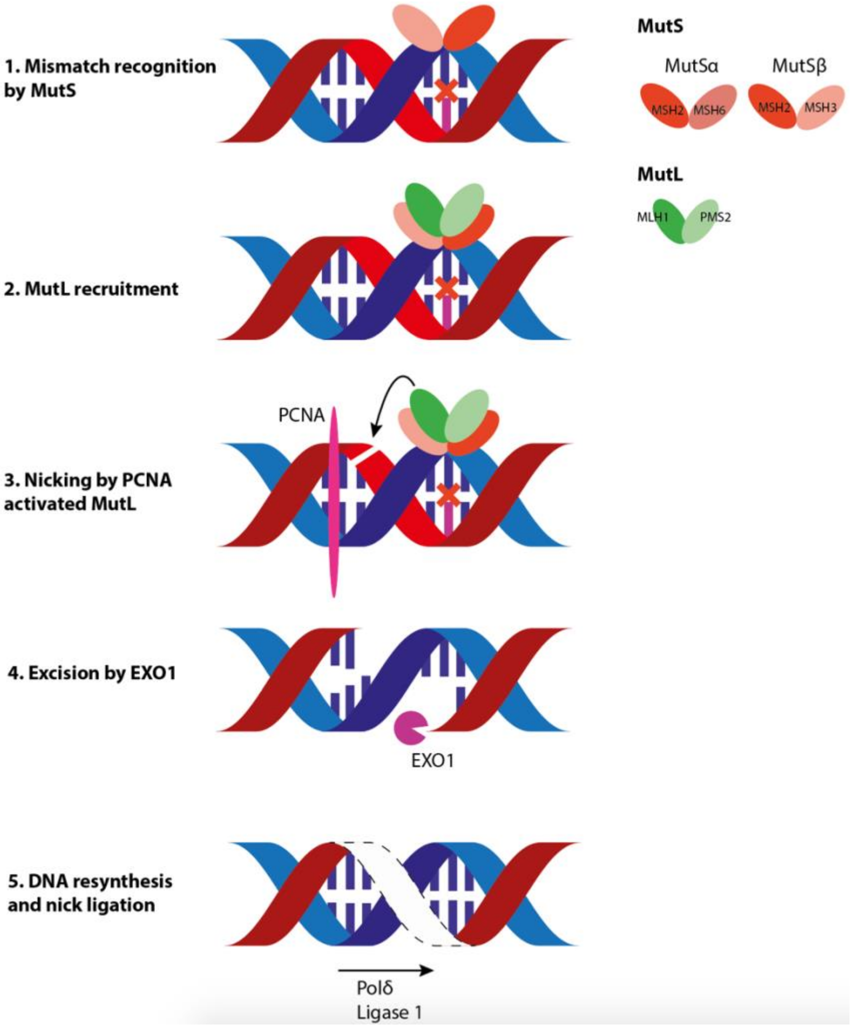

1. Mismatch recognition by MutS
2. MutL recruitment
3. Nicking by PCNA activated MutL
PCNA
4. Excision by EXO1
EXO1
5. DNA resynthesis and nick ligation
Polδ
Ligase 1
MutS
MutSα
MutSβ
MSH2
MSH6
MSH2
MSH3
MutL
MLH1
PMS2

Figure 1.1: The mismatch repair pathway

Mismatch recognition is performed by the MutS heterodimer whilst execution of mismatch repair is performed by MutL.

1.5 Individual MMR gene defects have different phenotypes

Although mismatch repair deficiency is often considered to be a single entity, recent studies indicate that phenotypes vary according to the MMR protein that is lost. For instance, in Lynch syndrome, patients with germline *MSH2* or *MLH1* mutation are considered to have a higher cancer risk and present with colorectal cancer at a younger age than *MSH6* or *PMS2* mutation carriers (Dominguez-Valentin et al., 2020). This suggests that MSH2 and MLH1 which are the obligatory binding partners (Colas et al., 2012) of MutS and MutL respectively, drive a more severe hypermutator phenotype than the minor partners MSH6 or PMS2. Furthermore, advanced adenomas are more frequent in patients with pathogenic germline *MSH2* mutations compared to *MLH1* (Engel et al., 2020), suggesting that MutS loss may have greater malignant potential than MutL. Large sequencing cohorts of MMRd tumours have also found that MutS loss (MSH2) is associated with increased mutation burden compared to MutL (MLH1) loss (Salem et al., 2020). The underlying reasons for these differences have not been studied in detail.

1.6 Evolution of MMRd colorectal cancer

The progression of MMRd colorectal cancer is predictably driven by frameshift mutations at hypermutable coding microsatellites in target genes (see cartoon figure 1.2), reflecting high levels of microsatellite instability (MSI) in these tumours. It should also be noted that MMRd tumours also accumulate large numbers of single nucleotide variants (SNVs) (Kim et al., 2013). Landscape genomic profiling studies have found that mutation burden varies by an order of magnitude across MMRd tumours (Cortes-Ciriano et al., 2017; Hause et al., 2016). Recurrent frameshift events are observed both in tumours of the same tissue type and across all tumours indicating a shared pathogenesis. Many of the main actors involved in MMRd progression are well described and include frameshift events at coding homopolymers of *TGFBR2* (involved in TGFß mediated growth signalling)(Markowitz et al., 1995), *BAX* (involved in apoptotic signalling) (Abdel-Rahman et al., 1999) and *AIM2* (involved in interferon signalling)(Woerner et al., 2007). Table 1.1 below displays the most frequently mutated coding homopolymers in MSI tumours analysed within the TCGA dataset (Cortes-Ciriano et al., 2017).

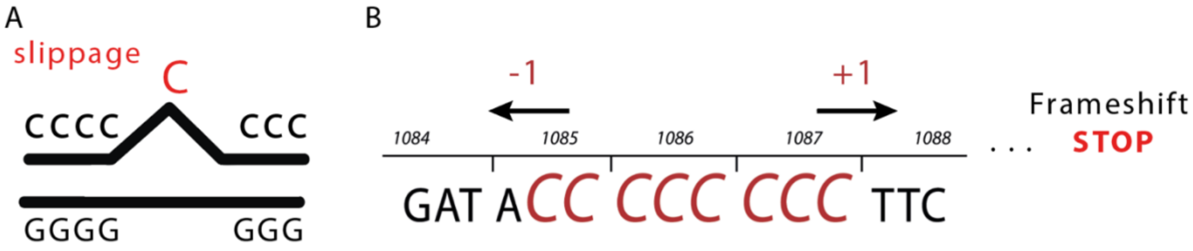


Figure 1.2: Mismatch repair deficiency leads to slippage at microsatellite sequences

A) Cartoon illustrating homopolymer slippage during DNA replication B) Uncorrected slippage at coding microsatellites results in frameshift mutation changing the amino acid reading frame and typically resulting in a premature stop codon.

Gene	Homopolymer	Cases mutated (%)
ACVR2A	A8	52%
KIAA2018	T11	51%
SLC22A9	A11	50%
ASTE1	T11	45%
TGFBR2	A10	44%
NDUFC2	A9	36%
SEC31A	T9	36%
LTN1	T11	36%
AIM2	T10	32%
C18orf34	T10	32%
OR7E24	T11	31%
CCDC150	A11	31%
RPL22	T8	30%
RNF43	C7	30%
MSH3	A8	29%
CASP5	T10	26%
PHACTR4	A10	26%
MIS18BP1	T11	26%
SLC35F5	A10	26%
MLL3	T9	26%

Table 1.1: Top 20 recurrently mutated coding homopolymers in MSI tumours from the TCGA dataset.

Displayed are the gene names, coding homopolymer base length and mutation frequency of top 20 recurrently mutated homopolymers MSI tumours within the TCGA dataset. Source: (Cortes-Ciriano et al., 2017).

Whilst landscape studies have identified recurrent frameshift mutations observed in MMRd cancers, the reason for the wide variation in mutation burden amongst mismatch repair deficient cancers remains poorly understood. Interestingly frameshift disruption of genes involved in DNA repair pathways have been reported in MMRd tumours(Cortes-Ciriano et al., 2017; Hause et al., 2016). This includes secondary frameshifts in genes involved in non-homologous end joining (NHEJ), homologous recombination, base excision repair as well as secondary frameshifts in MMR genes. However, the functional significance of these events during tumour evolution has not been studied in detail. In one study involving cell lines it was found that secondary *MSH6* frameshifts in the context of MLH1 deficient colorectal tumour cells was associated with increased mutation burden(Baranovskaya et al., 2001), but in-vivo validation in the context of a functional immune system has not been demonstrated. Secondary *MSH3* frameshifts in MMRd colorectal cancer have also been observed and in one study were associated with more advanced stage disease but the underlying mechanism for this was not explored (Plaschke et al., 2004).

The elevated mutation rate of MMRd tumours allows rapid adaptation, but also carries costs due to accumulation of neoantigens and other deleterious mutations. Frameshift mutations disrupt the reading frame of target genes resulting in a high rate of neoantigen generation and high levels of immune infiltration(Smyrk et al., 2001). This elevated immune selection pressure drives selection of immune escape events. The common immune evasion pathways, such as mutations in B2M and the HLA complex are well described (Grasso et al., 2018), but the evolutionary pathways underlying these events is not clear. For instance, recent evidence suggests that although B2M mutations, which disrupt HLA mediated antigen presentation are common in MMRd tumours, response to immune checkpoint inhibition is not diminished and may still persist through CD4+ T-cell pathways (Germano et al., 2021). This highlights the complexity of tumour-immune interactions and the arms race that exists between MMRd tumours and the host immune system. Improved understanding of the evolutionary pathway taken by MMRd tumours to manage the costs and benefits associated with hypermutation may help with forecasting patient treatment response.

1.8 Parallels with evolution in hypermutator bacteria

Interestingly hypermutation is also commonly observed during bacterial evolution allowing rapid adaptation to environmental pressure (Maddamsetti & Grant, 2020). Remarkably bacteria have evolved various mechanisms to exploit the benefits of hypermutation whilst limiting the long-term deleterious impact. Broadly these strategies can be divided into three main categories which are to limit hypermutation to specific times, locations within the genome or to specific individuals within the population (Matic, 2019). A particularly relevant example is that the genes involved in bacterial phenotypic variation are often concentrated at repeat sequences called contingency loci (Moxon et al., 2006). These repeat sequences are hypermutable and allow rapid stochastic phenotypic variation. For instance, Haemophilus influenzae uses phase variable repeat sequences to stochastically alter cell wall components to evade immune detection (Bayliss et al., 2001). Phenotypic variation using contingency loci is in many ways akin to microsatellite frameshifts in MMRd cancers allowing generation of phenotypic diversity which natural selection acts on to promote survival. Mutator bacteria due to defective mismatch repair are frequently observed in natural isolates and have been found to outcompete non-mutator strains when the population faces strong selection pressure (Giraud et al., 2001). Following adaptation mutator strains are often counter-selected and in experimental systems anti-mutator mutations have been found to emerge (Wielgoss et al., 2013). Mutation rate fluctuation is therefore an important survival strategy in bacteria, enabling the benefits of population diversity to be balanced against the costs of deleterious mutation accumulation. Whether mismatch repair deficient cancers use similar evolutionary pathways to manage the costs and benefits of hypermutation is unclear.

1.9 Adaptive mutability in cancer evolution

Recent studies have found that targeted therapies, (EGFR and mTOR inhibitors) can fuel cancer resistance evolution by transiently increasing cellular mutation rates (Cipponi et al., 2020; Russo et al., 2019). This adaptive mutability accelerates treatment resistance, whilst limiting the harmful effect of mutagenesis to times of high selection pressure. Work in microsatellite stable (MSS) colorectal cancer has found that anti-EGFR therapy induced transient downregulation of mismatch repair and upregulation of error-prone polymerases to achieve therapy related adaptability (Russo et al., 2019), although the underlying mechanisms for this process are unclear. Adaptive mutability resembles similar survival mechanisms described in bacteria, commonly referred to as stress induced mutagenesis (Bjedov et al., 2003). It is unclear whether adaptive mutability also plays a role in tumours with existing mismatch repair deficiency, where selection pressure from the immune system is already high even in the absence of therapy.

Aims:

The overall aim of this project is to investigate the evolutionary trajectory taken by MMRd colorectal cancer in managing the fitness impact of hypermutation. Of particular interest is to understand how MMRd cancers balance the long-term costs of hypermutation against the adaptive benefits. Specific aims are to:

- Investigate for the presence of secondary regulators of mutation burden and rate during MMRd cancer evolution.
- Characterise the impact of secondary regulators of mutability on the tumour immune microenvironment and immune escape events.
- Model the impact of secondary regulators of mutability on tumour growth dynamics.

Chapter 2: Investigating secondary regulators of hypermutation in MMRd colorectal cancer

2.1 Contribution statement

In this chapter I accessed data from the Genomics England 100,000 whole genome sequencing dataset of colorectal cancers. Identification of microsatellite instable cases was performed by Dr William Cross and all subsequent analysis performed by myself. The figures and data presented here have been used for a manuscript currently under review.

2.2 Introduction

An elevated mutation rate drives the progression of mismatch repair deficient (MMRd) colorectal cancers by generating large numbers of frameshift and SNV mutations for natural selection to act on. A number of canonical driver mutations such as frameshifts in *TGFBR2*, *BAX*, *caspase-5* and others are recognised as involved in MMRd pathogenesis (Abdel-Rahman et al., 1999; Cortes-Ciriano et al., 2017; Kim et al., 2013; Markowitz et al., 1995). SNV mutations in driver genes such as *KRAS and BRAF* are also observed in MMRd colorectal cancer (Chan et al., 2004). However, hypermutation also leads to generation of deleterious mutations. In particular, frequent frameshift mutations in MMRd tumours, result in generation of neoantigens leading to immune recognition (Kloor, 2016), whilst other deleterious mutations can also reduce cellular fitness (McFarland, 2014). Therefore, although hypermutation allows for rapid adaptation, in the longer term, progressive deleterious mutations may reduce cellular fitness; a concept known as Muller's ratchet in evolutionary genetics (Muller, 1964). MMRd cancers given their high mutation rate and lack of recombination may be particularly vulnerable to the effects of progressive irreversible deleterious mutations. Adaptations to overcome Muller's ratchet, such as for example whole genome doubling in lung cancer has been reported (López et al., 2020), but since MMRd cancers are usually chromosomally diploid (Muzny et al., 2012) the mechanisms employed to limit the deleterious impact of hypermutation are unclear.

Interestingly, bacteria commonly develop hypermutator strains to facilitate rapid adaptation, but have evolved mechanisms to limit the long-term mutagenic costs (Matic, 2019). For instance, in E. coli following a period of adaptation, hypermutator strains are often counter-selected and mutation rates restored (Giraud et al., 2001). Bacteria are also known to limit hypermutation to genomic loci where greatest phenotypic diversity can be created. Such 'contingency loci' are repeat sequences where successive insertion and deletion events allow stochastic phenotypic variation. An example of this is phase variation of cell wall components to evade immune recognition (Bayliss and Moxon, 2001). Also of relevance, hypermutator bacteria often arise due to defects in the bacterial mismatch repair system (Maddamsetti & Grant, 2020), indicating MMR is an evolutionarily conserved pathway for rapid adaptation. Hypermutator bacteria therefore appear to share a number of parallels with mismatch repair deficient cancers. Whether MMRd cancers also access similar evolutionary mechanisms to manage the long-term fitness costs associated with hypermutation has not been studied in detail.

One mechanism of bacterial adaptive response, that has been shown to be recapitulated in cancer is stress induced mutagenesis (SIM). Two recent studies found that, in response to stress induced by non-genotoxic targeted therapies (e.g. EGFR inhibitors), tumours underwent a transient increase in mutability facilitating greater adaptability to environmental change (Cipponi et al., 2020; Russo et al., 2019). Mechanistically it was found that the kinase mammalian target of rapamycin (mTOR) operated as a master regulator of cellular stress response leading to transient downregulation of DNA repair pathways including MMR. Therefore, cancers like unicellular bacteria may transiently adjust mutation rates to accelerate genetic diversity until adaptation is achieved. It is unclear if similar pathways operate in MMRd cancers, where mutation rates are already elevated and evolutionary pathways leading to adaptation to hypermutation are unclear. Understanding these pathways will help inform how MMRd tumours survive hypermutation and avoid mutational meltdown.

2.3 *MSH6* and *MSH3* frameshifts associate with increased mutation burden in MSI colorectal tumours

I set out to investigate how MMRd colorectal cancers manage the fitness costs associated with long term hypermutation. In order to screen a large cohort of tumours, I accessed the Genomics England 100,000 genomes project (GEL) dataset of colorectal cancers. From a starting cohort of 992 primary colorectal cancers, 217 (22%) microsatellite instable (MSI) colorectal cancers were identified. A detailed validation process was used to identify MSI cancers. This involved using the bioinformatic tool MSIsensor (version 0.6) to identify MSI cancers followed by confirmation of loss of MMR immunohistochemistry (IHC) where pathological data was available and further validation by confirming the presence of a mismatch repair deficient mutation signature (see methods section 7.6). In addition, cases with tier 1 pathogenic somatic or germline mutations in *POLE* or *POLD1* were excluded from further analysis. Using this validation approach, MSI/ MMRd cancers could be identified, within the context of a dataset where limited IHC data was available. MSIsensor was run using default settings (minimum read depth for MSI analysis=20, minimum homopolymer size=5, maximal homopolymer size=50, maximum repeat motif length=5). The MSIsensor tool involves 2 steps. First the reference human genome is scanned for microsatellite sites. Next a χ^2 test is used to compare the distribution of tumour and germline reads at each identified microsatellite allowing detection of somatically mutated microsatellites. An MSIscore ranging between 0 to 100 is produced indicating the percentage of somatically mutated microsatellite sites. Tumours are classified as MSI if the MSIscore is greater than a cutoff value of 3.5, as per the publication of the MSIsensor tool (Niu et al., 2014). MMR IHC validation was next performed where IHC data was available. MMR IHC status was available in n=101 colorectal cancers (n=20 classified as MMR deficient and n=81 MMR proficient) and this showed 98% agreement with the MSIsensor classification. One discordant case where MMR IHC status was proficient but MSIsensor status was MSI was excluded. Finally COSMIC single base substitution (SBS) signature contribution for each identified MSI tumour was retrieved from the Genomics England main programme cancer analysis table, which are based on the Alexandrov method of mutation signature extraction (Alexandrov et al., 2020). A minimum of 10% contribution from the known MMRd signatures (SBS 6, 15, 21 or 26) was present in identified MSI cases. Figure 2.1 below displays the relative COSMIC SBS signature contribution within each tumour in the cohort of 217 MSI colorectal cancers.

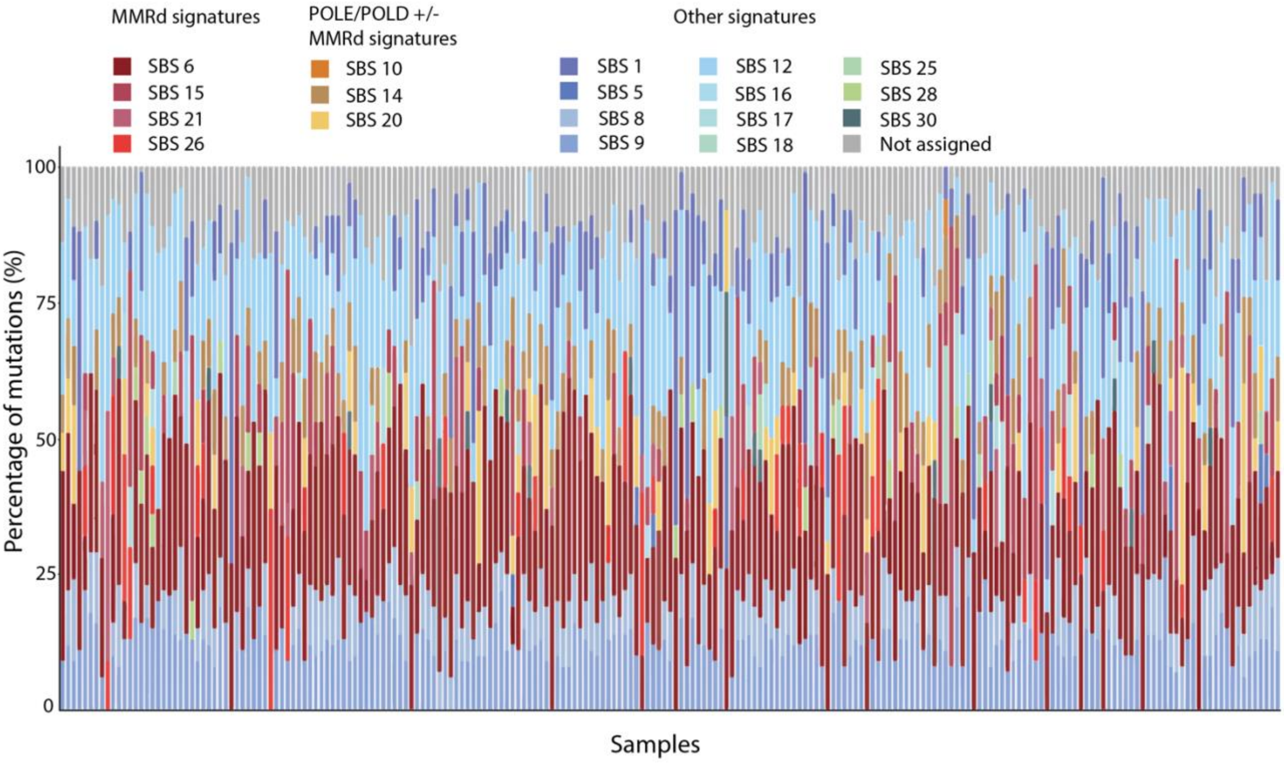


Figure 2.1: COSMIC SBS signature contribution in samples from the Genomics England MSI colorectal cancer cohort.

Stacked barplot showing the relative contribution of COSMIC SBS signatures within each tumour in the GEL MSI colorectal cohort (n=217). Known MMRd signatures are displayed in red shading.

I investigated the primary cause of microsatellite instability in this MSI cohort. *MLH1* promoter methylation is the most common cause of mismatch repair deficiency but methylation data was not available in the Genomics England Cohort. I therefore used the presence of somatic *BRAF*^{V600E} mutation as a surrogate marker for *MLH1* promoter methylation. This is based on the knowledge that in microsatellite instable colorectal cancer, *BRAF*^{V600E} mutation occurs almost exclusively in the context of *MLH1* promoter methylation (Deng et al., 2004; Parsons et al., 2012). This is also the basis of international clinical guidelines (NCCN guidelines in the US and NICE guidelines in the UK), where patients with mismatch repair-deficient colorectal cancer who demonstrate *BRAF*^{V600E} mutation are assumed to have sporadic *MLH1* promoter methylation (Giardiello et al., 2014; National institute for health and care excellence (NICE), 2017). Pathogenic germline and somatic MMR mutations were identified from the Genomics England main programme cancer tiering data tables filtered for tier 1 mutations. Overall, the primary cause of microsatellite instability could be explained in n=151 cases, broken down as follows; n=125 with *MLH1* promoter methylation (evidenced by presence of somatic *BRAF*^{V600E} mutation), n=9 with double somatic pathogenic MMR gene mutations, n=17 with pathogenic germline MMR mutations. In the remaining cases (n=66) where the primary cause of MSI could not be identified, it is possible that these were cases with *MLH1* promoter methylation but absence of somatic *BRAF* mutation as previously reported (Farchoukh et al., 2016). Alternatively previously unknown somatic or germline mutations could also be responsible in some cases. The clinical characteristics of this cohort of MSI tumours is detailed in table 2.1 below.

Genomics England MSI colorectal cancer cohort clinical characteristics	
	Number of cases
Age	
20-40	12 (6%)
41-60	24 (11%)
>=61	181 (83%)
Sex	
Female	127 (59%)
Male	90 (41%)
Primary MMR defect	
MLH1 methylation	125 (58%)
Germline mutation	17 (8%)
Double somatic mutation	9 (4%)
TOTAL	217

Table 2.1: Clinical characteristics of 217 MSI colorectal cancers in the GEL cohort.

Using this cohort of MSI colorectal cancers, I set out to investigate for the presence of secondary regulators of mutability in MMRd cancers. Landscape studies such as the TCGA have found that mutation burden varies by an order of magnitude amongst MMRd tumours (Cortes-Ciriano et al., 2017; Hause et al., 2016), but the reasons for this are unclear. I investigated whether secondary frameshifts in common MSI target genes associate with an increase or decrease in mutation burden, and thus potentially act as secondary regulators of mutability in MMRd colorectal cancer. A set of 20 MSI target genes, all containing coding homopolymers that are frequently mutated in MMRd tumours was selected from the literature (Cortes-Ciriano et al., 2017; Duval & Hamelin, 2002; Hause et al., 2016). This set of genes included known MMRd cancer driver genes such as *TGFBR2*, *BAX* and caspase-5 (Abdel-Rahman et al., 1999; Duval & Hamelin, 2002; Markowitz et al., 1995). *MBD4* was also included given its role in the base excision repair pathway (Bader et al., 1999). *MSH6* and *MSH3* were included given their status as known MMR genes and previous cell line data for *MSH6* as a secondary mutator (Baranovskaya et al., 2001). Other genes were included on the basis that they are commonly mutated in MMRd tumours. Next using multiple linear regression analysis, I assessed the relationship between frameshift mutation in these MSI target genes and total mutation burden in the GEL cohort of 217 MSI colorectal cancers, controlling for tumour purity and patient age. Results of the regression analysis showed that only frameshifts in *MSH6* and *MSH3* associated with a significant increase in total mutation burden. The remaining MSI target genes showed no correlation with mutation burden. These results are illustrated as a volcano plot in figure 2.2 below and also detailed in supplementary table 2A. Furthermore, combined *MSH6* and *MSH3* frameshifts in a tumour had an additive association with mutation burden compared to either mutation alone (supplementary table 2B). The regression model indicated that mutation of *MSH3* or *MSH6* increased mutation burden from a baseline estimate of 161,267 mutations by 88,038 and 63,675 mutations, respectively, whilst mutation of both was associated with an increase of 139,338 mutations. The *MSH6* and *MSH3* frameshifts here both occurred at specific length 8 homopolymers contained within the coding sequence of each gene resulting in the *MSH6*^{F1088FS} and *MSH3*^{K383FS} events respectively. Since both events are homopolymer frameshifts, this is consistent with *MSH6*^{F1088FS} and *MSH3*^{K383FS} being secondary events after the initial loss of mismatch repair in these tumours.

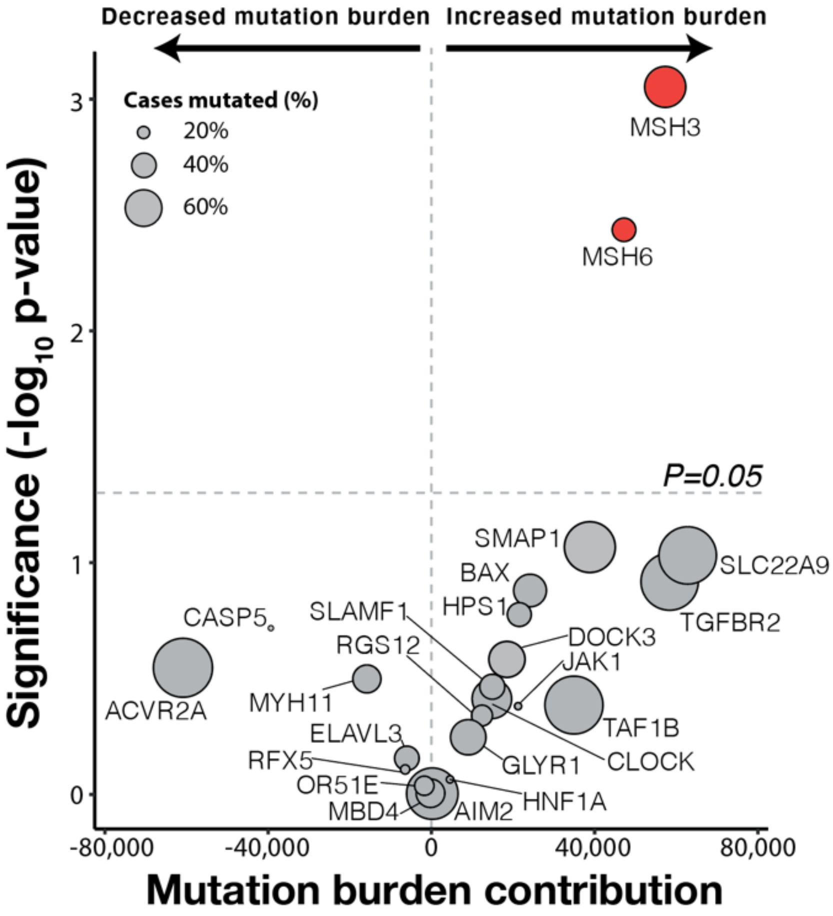


Figure 2.2: Volcano plot showing relationship between frameshift mutations in MSI target genes and total mutation burden.

Volcano plot showing relationship between frameshift mutations in MSI target genes and total mutation burden in multiple linear regression analysis. Only homopolymer frameshifts in MSH6 and MSH3 associate with significantly increased mutation burden.

The regression analysis above shows a strong association between coding frameshift mutations in *MSH6* and *MSH3* and total mutation burden in MMRd colorectal cancer. To further assess this relationship, the regression analysis was repeated using a larger panel of genes containing coding microsatellites. A previous study by Cortes-Cirano et al, reported the most frequently mutated coding microsatellites in MSI tumours (Cortes-Ciriano et al., 2017). From this publication the top 50 recurrently mutated microsatellite targets were selected (see figure 2A Cortes-Ciriano et al., 2017) together with the previously identified *MSH6* and *MSH3* microsatellites and the regression analysis was repeated. This exploratory analysis could also allow identification of additional candidate regulators of mutability. Results of this analysis are presented as a volcano plot in figure 2.3 below and further details are provided in supplementary table 3. *MSH6* and *MSH3* frameshifts remain significantly associated with increased mutation burden in this analysis. Interestingly in this model the presence of an *MSH6* frameshift associated more strongly with increased mutation burden than *MSH3* frameshift. In addition, frameshifts in the genes *UBR5*, *OR52N5* and *MLL3* also show a significant association with increased mutation burden. Of these, frameshift mutation in the histone methyltransferase gene *MLL3* (alias *KMT2C*) showed the strongest association with mutation burden. *MLL3* is responsible for the monomethylation of histone H3 at lysine 4 (H3K4) residues, promoting transcription factor recruitment and open chromatin formation at gene enhancers and promoters (Ford & Dingwall, 2015). *MLL3* has a recognised role in the repair of double strand breaks via the homologous recombination pathway (Rampias et al., 2019). Interestingly an association between *MLL3* mutation and mutation load has previously been described in non-small cell lung cancer (Chang et al., 2021). However, a role for MLL3 or the histone mark H3K4 in the mismatch repair pathway has not been previously described but would be an interesting area of future investigation. The gene *UBR5* has a known role in DNA damage response and double strand break repair, which may explain the association with mutation burden observed. *OR52N5* does not appear to have any previously described role in DNA repair or as a cancer driver and its association here with mutation burden is unclear. *MLL3*, *UBR5* and *OR52N5* frameshifts appear to be interesting targets in MMRd tumours and should be investigated further in future studies. In this thesis I focus on the role of *MSH6* and *MSH3* frameshifts, since both genes are themselves mismatch repair genes and the functional impact of secondary mismatch repair mutations has not been investigated in detail in human cancers.

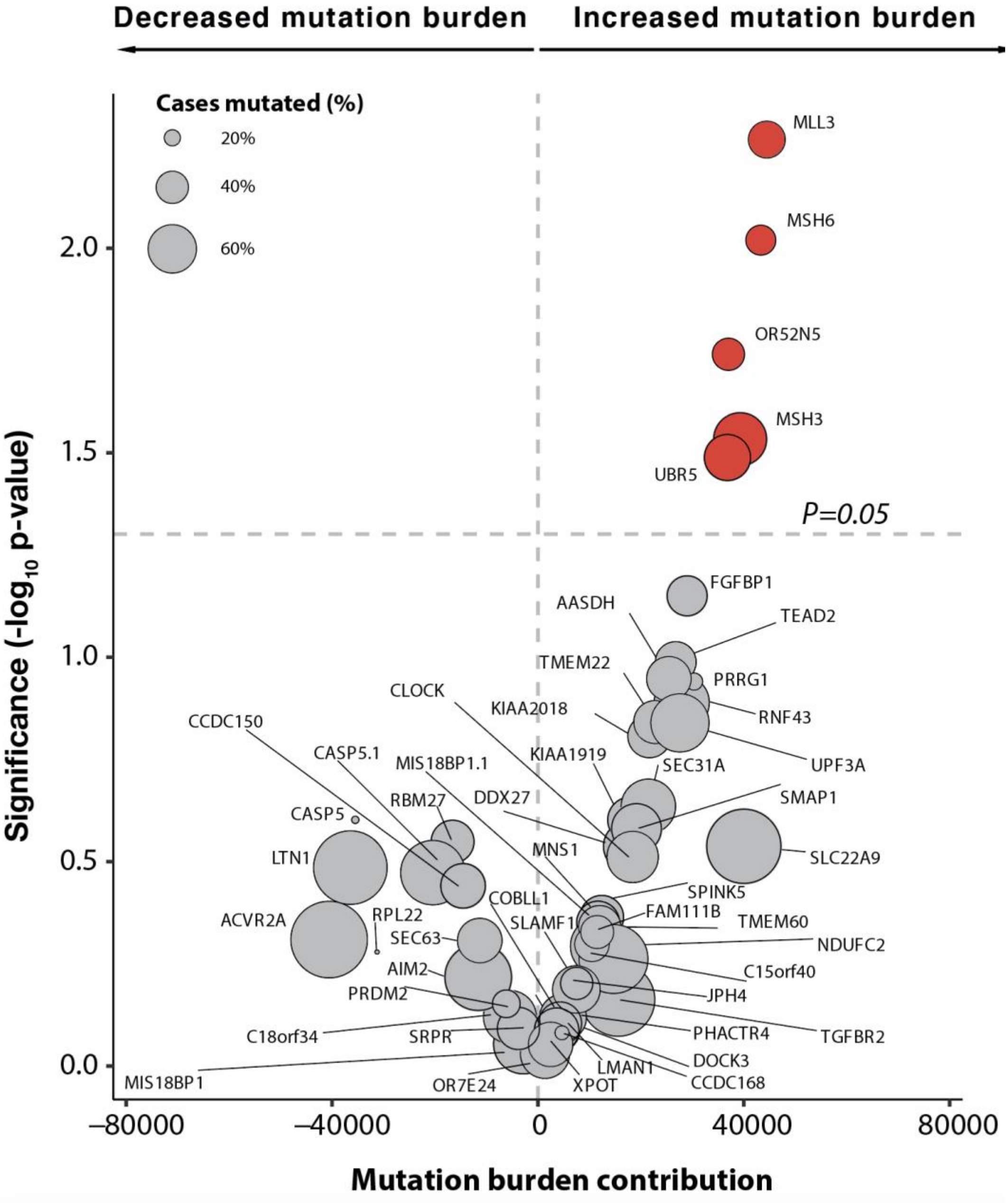


Figure 2.3: Volcano plot showing relationship between frameshift mutations in top 50 recurrently mutated microsatellites and total mutation burden.

Volcano plot showing relationship between frameshift mutations in top 50 recurrently mutated microsatellites in MSI tumours. Size circles indicates proportion of cohort with frameshift in respective gene.

Plotting tumours in order of mutation burden as a waterfall plot with bars coloured according to presence of *MSH6* or *MSH3* frameshifts further illustrated the relationship between secondary MMR frameshift mutation and increased mutation burden (figure

2.4). The primary cause of mismatch repair deficiency where known is also indicated below the waterfall plot.

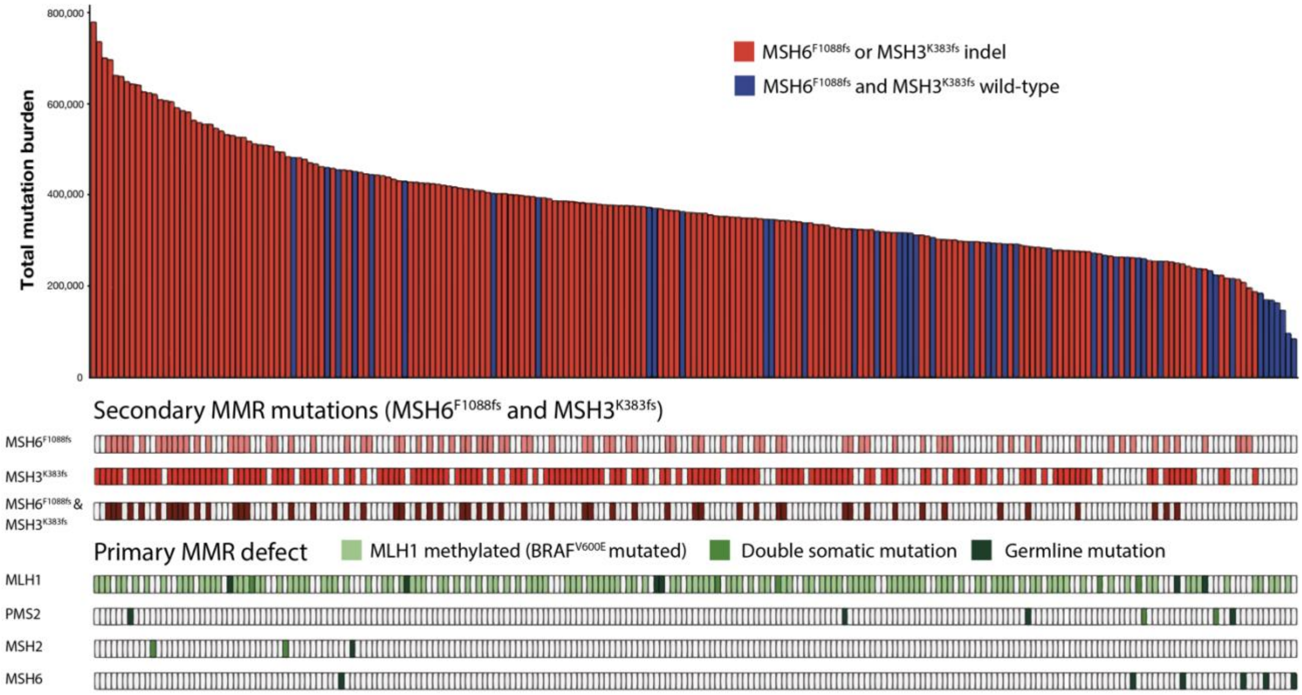


Figure 2.4: Waterfall plot of MSI colorectal tumours from GEL cohort in order of mutation burden.

217 MSI colorectal tumours in order of total mutation burden. Bars are coloured according to the presence of *MSH6* and/or *MSH3* frameshifts. The heatmaps below show breakdown of cases according to *MSH6* and/or *MSH3* frameshifts. The primary MMR defect where known is highlighted in green shading, with cases coloured according to presence of somatic MLH1 promoter methylation (identified through BRAF^{V600E} mutation), germline MMR mutations and double somatic MMR mutations.

I next looked at single nucleotide variants (SNVs), indels and total mutation burden separately, and again noted a stepwise increase in the presence of *MSH6* and *MSH3* frameshifts (figure 2.5). Tumour purity was also compared between groups, to ensure it was not a factor confounding the results. This confirmed no significant difference in purity between *MSH6* and *MSH3* mutated groups (supplementary figure 1).

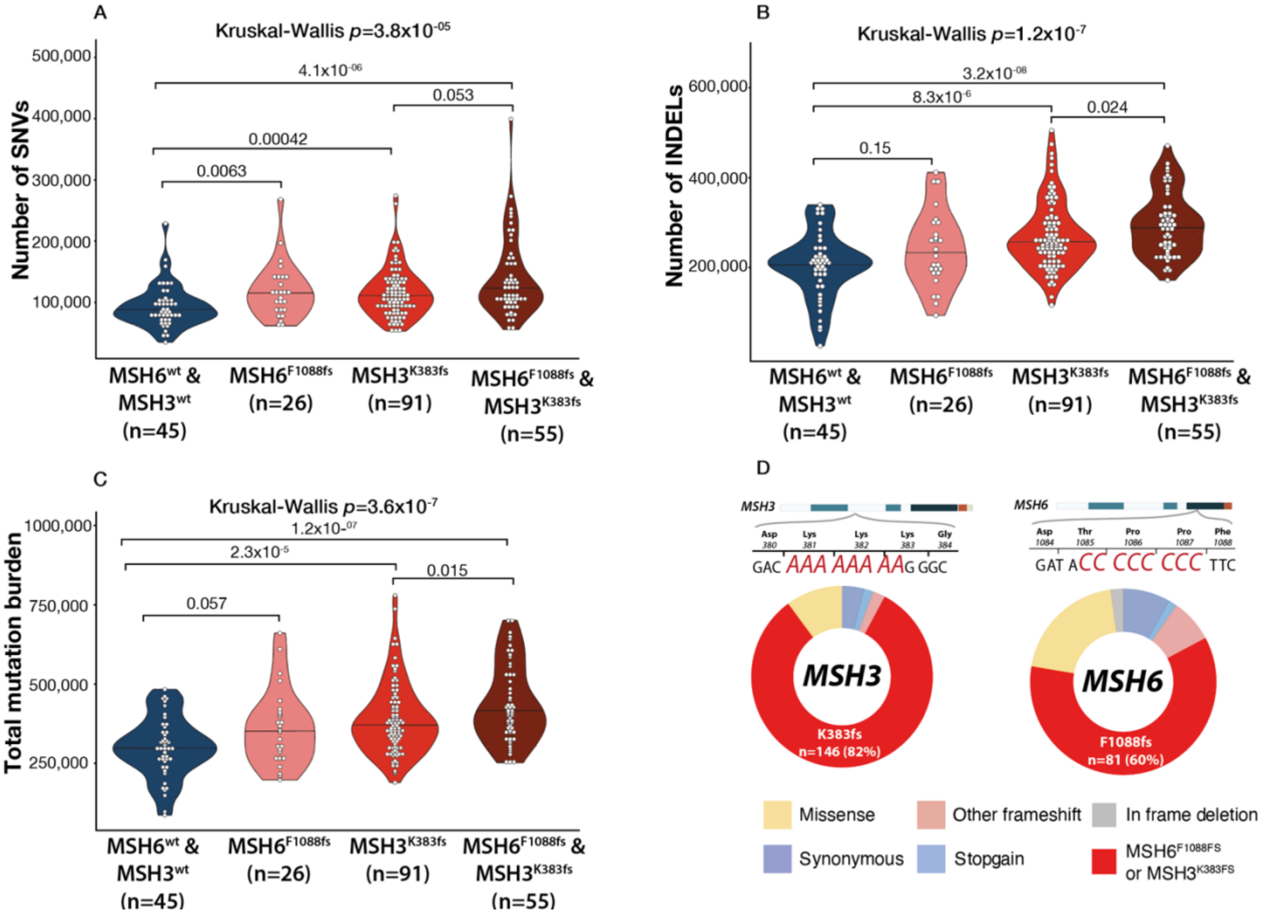


Figure 2.5: Homopolymer frameshifts in MSH6 and MSH3 associate with increased mutation burden.

(A-C) Violin plots displaying number of SNVs, indels and total mutation burden in tumours according to *MSH6* and *MSH3* frameshift mutation status. (D) Cartoon of *MSH3* A8 and *MSH6* C8 coding microsatellites. Pie charts display breakdown of mutation types observed in the *MSH6* and *MSH3* genes in this cohort of tumours. 60% of all *MSH6* mutations and 82% of all *MSH3* mutations occurred at the *MSH6* C8 and *MSH3* A8 homopolymers respectively.

MSH6 and *MSH3* both contain a length 8 coding homopolymer (C8 and A8 respectively) which are hypermutable sites in MMRd tumours. In the GEL cohort the overwhelming majority of *MSH6* and *MSH3* mutations occurred at these sites; 60% of all *MSH6* mutations and 82% of all *MSH3* mutations as detailed in figure 2.5D. I next considered whether these *MSH6* and *MSH3* homopolymer frameshifts occurred more frequently than would be expected by chance. Mutability at microsatellite sequences is known to be influenced by homopolymer length and nucleotide composition (Kim et al., 2013). I therefore compared the frequency of homopolymer frameshifts in *MSH6* and *MSH3* in the cohort against the frequency of all exonic length 8 C:G or A:T homopolymers respectively. Exonic length 8 microsatellites here were extracted using the SciRoKo package (see methods section 7.6). This identified 776 A/T length 8 homopolymers and 164 C/G length 8 homopolymers. There was significant enrichment of frameshifts at both the *MSH6* C8 homopolymer (37.3% v 23.4%; χ^2 p=1.9x10^{-6}) and the *MSH3* A8 homopolymer (67.3% v 9.9%; p<2.2x10^{-16}). These results are detailed in table 2.2 below and suggest there is selection for *MSH6* and *MSH3* frameshifts during MMRd tumour evolution.

(A)	Num. of homopolymers	Wild-type	Mutated	TOTAL	**Percentage mutated (%)**
MSH6 C8 homopolymer	1	136	81	217	**37.3**
Other C/G exonic length 8 homopolymers	163	27109	8262	35371	**23.4**

(B)	Num. of homopolymers	Wild-type	Mutated	TOTAL	**Percentage mutated (%)**
MSH3 A8 homopolymer	1	71	146	217	**67.3**
Other A/T exonic length 8 homopolymers	775	151462	16713	168175	**9.9**

Table 2.2: Homopolymer frameshifts in MSH6 and MSH3 are enriched compared to other exonic homopolymers.

(A) Mutation status of length 8 C/G exonic homopolymers versus *MSH6* C8 homopolymer (B) Mutation status of length 8 A/T exonic homopolymers versus *MSH3* A8 homopolymer.

In the GEL cohort presented here, *MSH6* and *MSH3* frameshifts are presumed secondary events following the initial primary loss of MMR. We are limited in the GEL cohort in that MMR immunohistochemistry status is not available in all cases. To exclude the possibility of confounding due to differences in the primary cause of MMR loss, I next sought to restrict the cohort to cases with confirmed MutLα (MLH1/PMS2) primary loss. Therefore, I restricted the analysis to cases with either somatic *BRAF*^{V600E} mutation (indicating MLH1 promoter methylation) or known germline MLH1 or PMS2 pathogenic mutations (n=135). This provided a subset with confirmed primary MutLα (MLH1/PMS2) deficiency. In this MutLα deficient subset, the relationship between secondary *MSH6* and *MSH3* frameshifts and increased mutation burden was again confirmed across SNVs, indels and total mutation burden (figure 2.6).

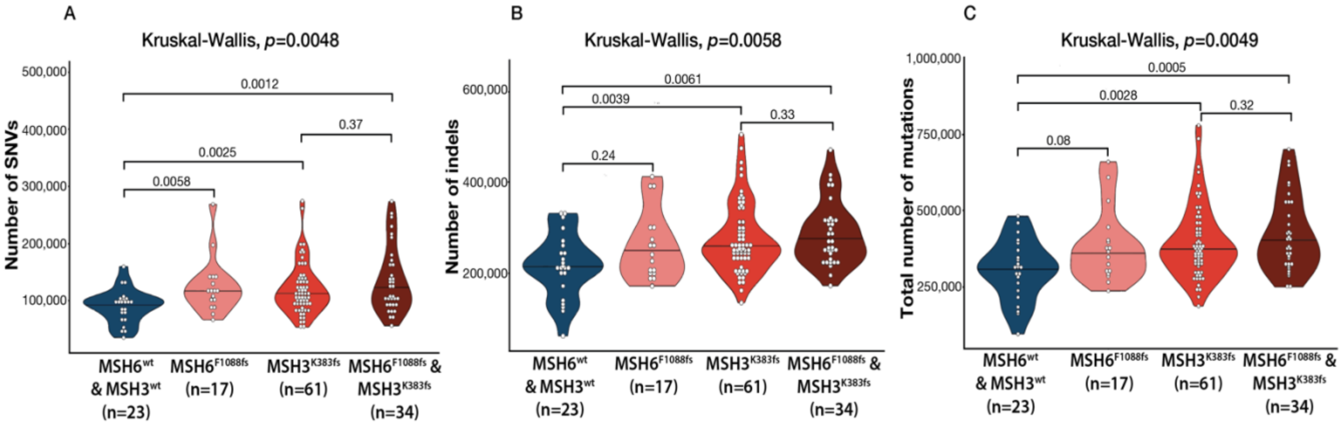


Figure 2.6: GEL cohort subset for cases with confirmed MutLα (MLH1/PMS2) background loss.

(A-C) SNV, INDEL and total mutation burden according to MSH6 and MSH3 frameshift mutation status in MSI tumours with confirmed MutLα (MLH1/PMS2) loss.

2.4 Impact of *MSH6^{F1088fs}* and *MSH3^{K383fs}* clone size on mutation burden

Since *MSH6* and *MSH3* frameshifts are presumed secondary events after the onset of MMR deficiency, the fraction of cells with *MSH6* and *MSH3* frameshifts in a sample may impact the mutation burden. I reasoned that samples containing a higher allelic fraction of *MSH6* or *MSH3* frameshifts, would contain a greater fraction of cells at a higher mutation rate resulting in increased mutation burden. This concept is illustrated in figure 2.8 below. The analysis is complicated by the variable admixture of normal cells from the stromal component within samples. In addition, copy number alterations at the *MSH6* and *MSH3* locus could also impact the mutated fraction. However since MMRd tumours are typically diploid (Muzny et al., 2012), this would be expected to be less of an issue. The Sequenza package was used to derive tumour purity and allele specific copy number estimates for samples. Samples with copy number alterations at the *MSH6* and *MSH3* genomic locus and also cases with known germline mutations in *MSH6* or *MSH3* were excluded (n=6). Purity corrected *MSH6* and *MSH3* variant allele frequency (VAF) for each sample was then calculated as follows:

$$Purity\ adjusted\ VAF = \frac{VAF}{Estimated\ tumour\ purity}$$

To assess the impact of *MSH6* and *MSH3* independently, samples containing both *MSH6* and MHS3 frameshifts were excluded. Then the correlation between *MSH6*^{F1088fs} and *MSH3*^{K383fs} VAF against mutation burden was assessed in samples. The results show that *MSH3*^{K383fs} VAF displays a significant positive correlation with both SNV and indel burden, whilst *MSH6*^{F1088fs} VAF shows a non-significant correlation with SNV burden and no correlation with INDEL burden (figure 2.7). Overall, these results further support the role of *MSH6* and *MSH3* as secondary mutator genes in MMRd tumours.

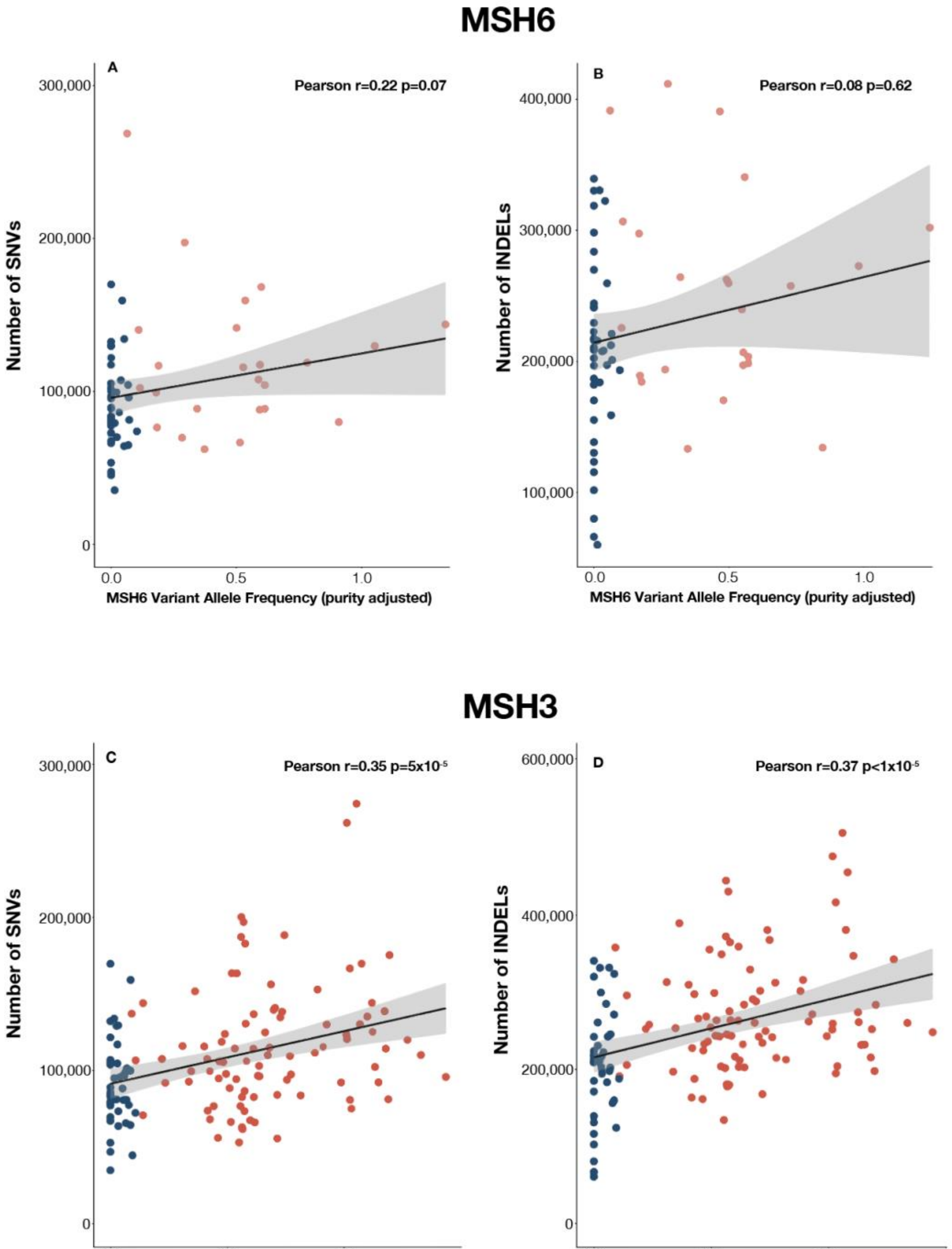


Figure 2.7: Correlation between MSH6 and MSH3 VAF against mutation burden

(A-D) Scatter plots showing correlation between variant allele frequency for *MSH6* and *MSH3* homopolymers and total number of SNVs and indels. Blue points indicate samples wild-type for *MSH6* or *MSH3* (VAF<5%) according to the plot indicated.

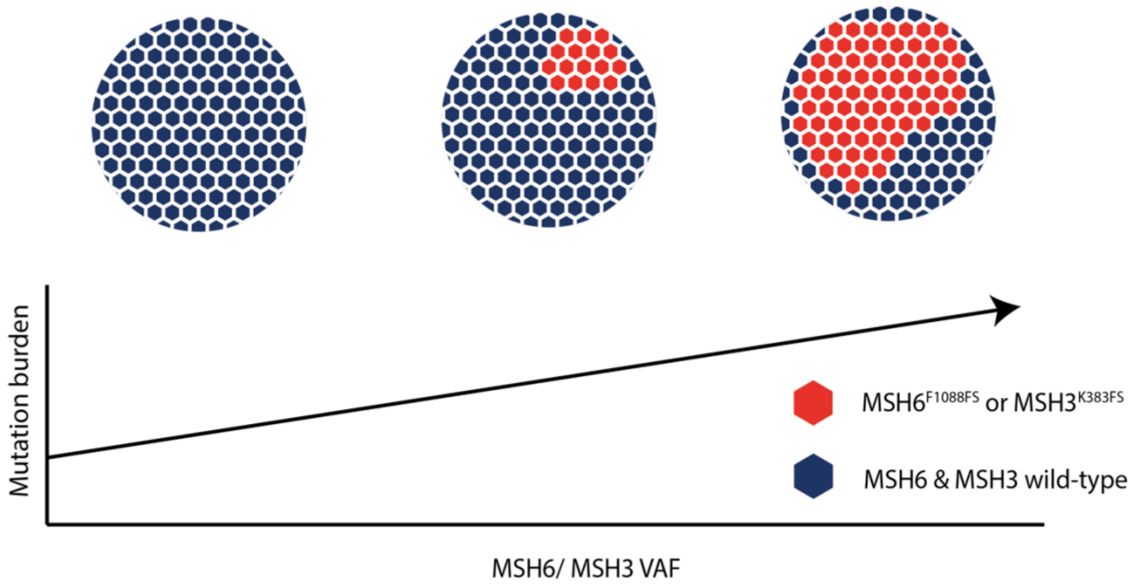


Figure 2.8: Cartoon describing relationship between MSH6^{F1088fs} and MSH3^{K383fs} variant allele frequency on mutation burden

With increasing MSH6^{F1088fs} or MSH3^{K383fs} clone size within a sample, more cells will be at an elevated mutation rate resulting in overall increased mutation burden.

2.5 Independent validation using TCGA dataset

For independent validation, TCGA whole exome sequencing data was analysed. The number of MSI colorectal tumours was smaller in this dataset. To increase the sample size, MSI tumours from other tissue types that commonly display mismatch repair deficiency were included. MSI tumours in the TCGA dataset were identified using the results of a previously published study (Cortes-Ciriano et al., 2017). The final cohort consisted of MSI tumours from colorectal (n=48), uterine (n=67), stomach (n=63) and oesophageal (n=3) tumour types. Tumours were again grouped according to presence or absence of *MSH6* and *MSH3* homopolymer frameshifts. Mutation burden according to SNVs, indels and total mutation count were extracted for each tumour from variant call files (see methods section 7.7). The analysis again confirmed a stepwise increase in mutation burden in tumours with *MSH6* and/or *MSH3* frameshift mutations, including an additive increase in cases with frameshifts in both genes (figure 2.9). This trend was again observed at the level of SNVs, indels and total mutation count. I next compared RNA expression of MSH6 and MSH3 in samples, reasoning that frameshift in these genes would lead to a truncated peptide and reduced RNA transcript stability. I found a clear decrease in MSH6 and MSH3 RNA expression in samples with *MSH6*$^{\mathrm{F1088FS}}$ and *MSH3*$^{\mathrm{K383FS}}$ respectively (figure 2.10).

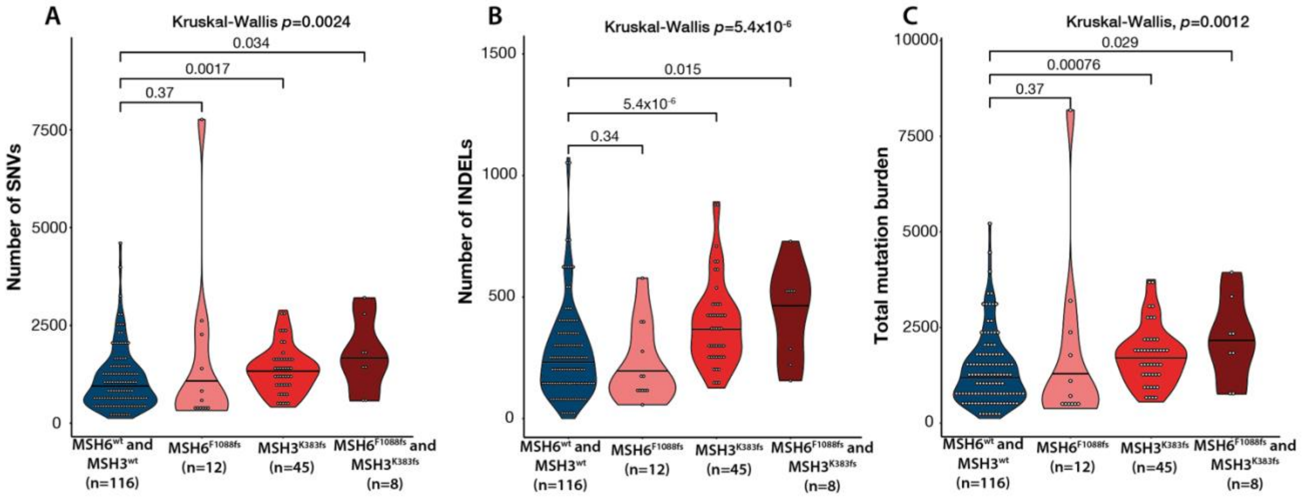


Figure 2.9: TCGA validation cohort

(A-C) Violin plots displaying number of SNVs, indels and total mutation burden in TCGA MSI tumours (n=181) according to MSH6 and MSH3 frameshift mutation status.

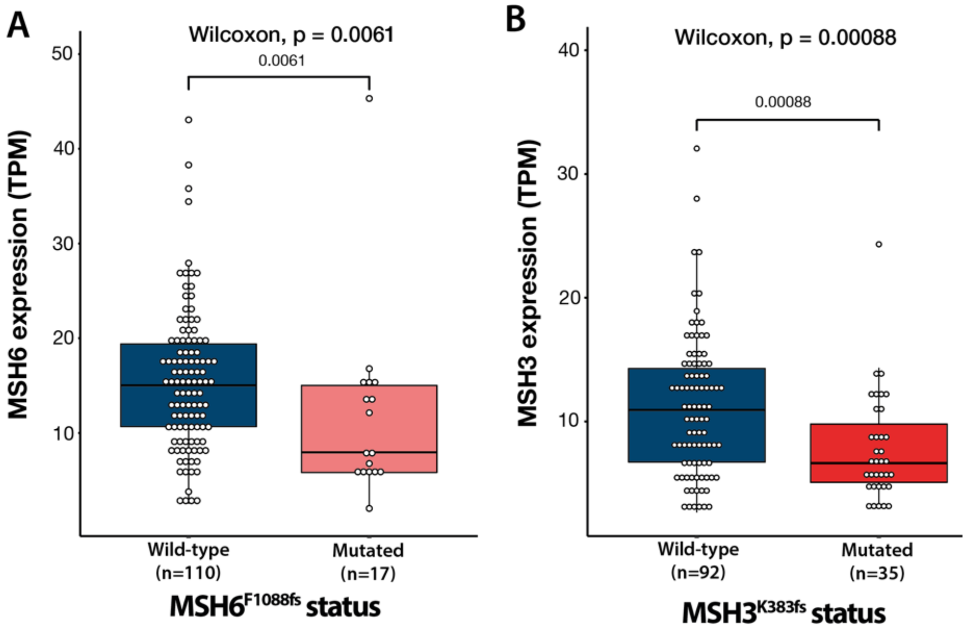


Figure 2.10: RNA expression level of MSH6 and MSH3 in TCGA MSI cohort

(A-B) RNA expression level of MSH6 and MSH3 according to the presence of homopolymer frameshift mutation in each gene.

2.6 Discussion

In this chapter I found that secondary frameshifts in *MSH6* and *MSH3* correlate with increased mutation burden in MMRd cancers. Both *MSH6* and *MSH3* contain a length 8 coding homopolymer which are frequent sites of slippage in the context of microsatellite instability. The analysis found that *MSH6* and *MSH3* frameshifts occur more frequently than would be expected in the absence of selection and frameshift mutation at other common MSI target genes did not associate significantly with mutation burden. Overall, the findings suggest that *MSH6* and *MSH3* frameshifts may increase mutability in MMRd cancers.

This finding raises the question of why secondary mutations in MMR genes increase mutability in the context of existing MMR loss (typically MLH1/PMS2 loss). Consideration of the roles performed by different components of mismatch repair can help explain this phenomenon. *MSH6* and *MSH3* frameshifts disrupt the MutS module of MMR, where both proteins are alternative binding partners of MSH2 forming the MutSɑ and MutSß heterodimers respectively. MutSɑ and MutSß have partially overlapping functions, in the recognition of base-base mismatches and short insertion-deletion loops respectively, but disruption of both proteins causes complete loss of the

MutS mismatch recognition module. Meanwhile, MLH1 and PMS2 form the MutLα heterodimer. MutS and MutLα have differing roles, namely in the recognition and execution of mismatch repair respectively. Whilst both roles are crucial to MMR function, MutS operates earlier in the repair pathway and so may have a more critical role resulting in a more severe phenotype when disrupted. This is demonstrated by the fact that MMRd tumours caused by primary MSH2 loss, display increased mutation burden compared to primary MLH1/PMS2 deficient tumours as shown in figure 2.11B (Salem et al., 2020). Furthermore, studies have reported evidence of cross-talk between MutS and other DNA repair pathways such as base excision repair (BER) (Grin & Ishchenko, 2016; Gu et al., 2002), indicating MutS may have a wider role in DNA repair. For instance cell line experiments have found that MutSα together with the BER glycosylase OGG1 both contribute to repair of 8-oxo-guanine damage, preventing C>A mutations (Mazurek et al., 2002). These additional roles of the MutS heterodimer may explain the increased mutability observed with secondary *MSH6* and *MSH3* frameshifts, reflecting progression towards MutS deficient phenotype. Our data supports this, since combined loss of MSH6 and MSH3, resulted in an additive increase in mutation burden compared to disruption of either gene alone.

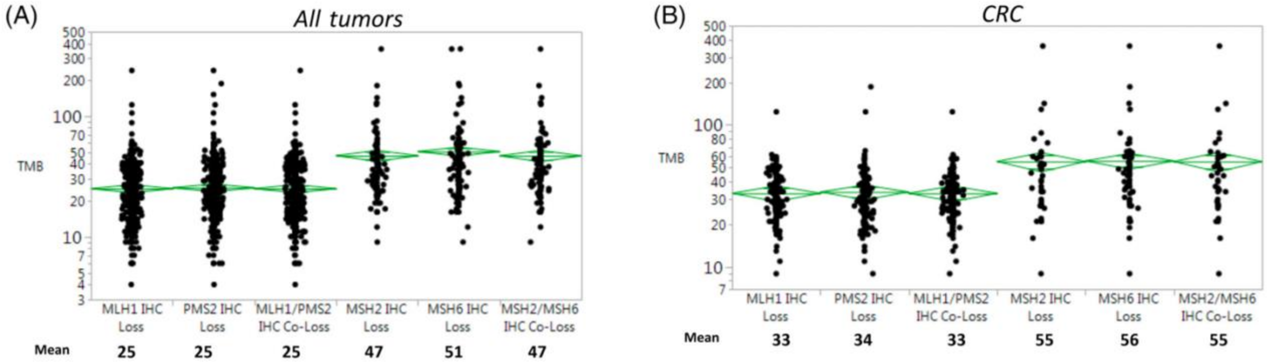


Figure 2.11: Mutation burden according to component of mismatch repair loss by IHC

(A) In all tumour types (B) In colorectal cancer. Figure is from Salem et al.,2020.

Another important consideration is that the data presented in this chapter all came from single biopsy bulk sequencing data. Since secondary *MSH6* and *MSH3* frameshifts are typically subclonal events evidenced by a corrected VAF of less than 0.5, bulk sequencing data may not reveal the full impact of these events. I was able to partially address this by assessing the correlation between *MSH6* and *MSH3* variant allele frequency and mutation burden. A more accurate approach would require taking targeted samples from MSH6/MSH3 deficient subclones. This would require prior knowledge of the clonal structure of a given tumour and will be considered in the next chapter.

Overall, the finding of potential secondary drivers of increased mutation burden in MMRd cancer was a surprising result. MMRd tumours already have a hypermutated phenotype and further increase in mutation rate may exacerbate the impact of deleterious mutations and potentially result in greater immune visibility through neoantigen generation. I was keen to investigate this further to understand the evolutionary pathway underlying these events and how MMRd tumours manage the fitness impact associated with increased mutability.

Conclusions:

- *MSH6* and *MSH3* both contain a length 8 coding homopolymer that is a frequent site of frameshift mutation in MMRd cancers.
- Secondary *MSH6* and *MSH3* frameshifts associate with increased mutation burden in MMRd colorectal cancer.
- *MSH6* and *MSH3* frameshifts occur more frequently than expected by chance compared to other length 8 homopolymers of the same nucleotide base.
- Frameshifts in other MSI target genes do not associate with significant change in mutation burden.

Chapter 3: *MSH6* and *MSH3* homopolymer frameshifts associate with increased mutation burden in MMRd colorectal cancer

3.1 Contribution statement

The analysis presented in this chapter was completed by myself. The work has been submitted as a paper which I wrote together with my supervisor Marnix Jansen. The data and figures presented have been adapted from the paper currently under review.

3.2 Introduction

In the previous chapter secondary *MSH6* and *MSH3* frameshifts were found to associate with increased mutation burden in MMRd colorectal cancers suggesting an increased mutation rate associated with these events. However single biopsy bulk sequencing data provided in the GEL cohort may under-report the impact of secondary mismatch repair frameshifts. Single biopsy data also has limited ability to infer the clonal ordering of mutational events. To obtain more granular data, targeted sampling at clonal resolution according to MSH6/MSH3 status is required. In this chapter I establish the use of MSH6 immunohistochemistry as a lineage tracing tool, allowing monophyletic biopsies to be taken from MSH6 proficient and deficient subclones. This resulted in the multi-region whole exome sequencing dataset that is used throughout the project. Using this independent dataset, I further investigate the functional impact of secondary *MSH6* and *MSH3* frameshifts. I firstly confirm the increase in mutation burden associated with secondary MMR frameshifts. Second, I assess the impact of *MSH6* and *MSH3* frameshifts on mutation bias and signatures. I note that of the known COSMIC mutation signatures, 5 are attributed to mismatch repair deficiency (Alexandrov et al., 2020). The mechanistic basis for these different signatures is unclear but may suggest mutation signature varies according to the component of mismatch repair disrupted. Overall the work in this chapter aims to characterise the functional impact of *MSH6* and *MSH3* frameshifts on the spectrum of mutations across the exome.

3.3 MSH6 immunohistochemistry as a lineage tracing tool

Immunohistochemistry is routinely used to label for MMR proteins in the clinical setting. I reasoned that since frameshift mutation at the *MSH6* homopolymer would disrupt the amino acid reading frame this would lead to loss of immunohistochemical labelling. To assess this, I reviewed the MMR immunohistochemistry of a large cohort of colorectal cancers from the UCLH biobank (detailed below). A subclonal pattern of MSH6 loss was frequently observed and an example case is provided in figure 3.2 below.

3.4 Patients and samples

From the UCL/UCLH biobank of health and disease 546 consecutive colorectal cancers diagnosed between 2014 to 2018 were identified. Of these 88 (17%) were mismatch repair deficient at immunohistochemistry. The mismatch repair deficient tumours assessed here were all from primary surgical resection specimens and patients were naïve to prior systemic anti-cancer therapy. MMR immunohistochemistry was assessed from a single slide per tumour. Of the 88 MMRd tumours, 78 displayed MLH1 and PMS2 loss, 6 displayed PMS2 loss with intact MLH1, 2 displayed MSH2 and MSH6 loss and 2 displayed MSH6 loss alone. Of these tumours, 32 (38%) displayed additional loss of MSH6 in the context of MLH1/PMS2 loss. The pattern of MMR expression is summarised as a heatmap in figure 3.1 below. MSH6 deficient percentage in the heatmap refers to the estimated percentage of tumour cells with loss of expression when visually assessed at a single slide level.

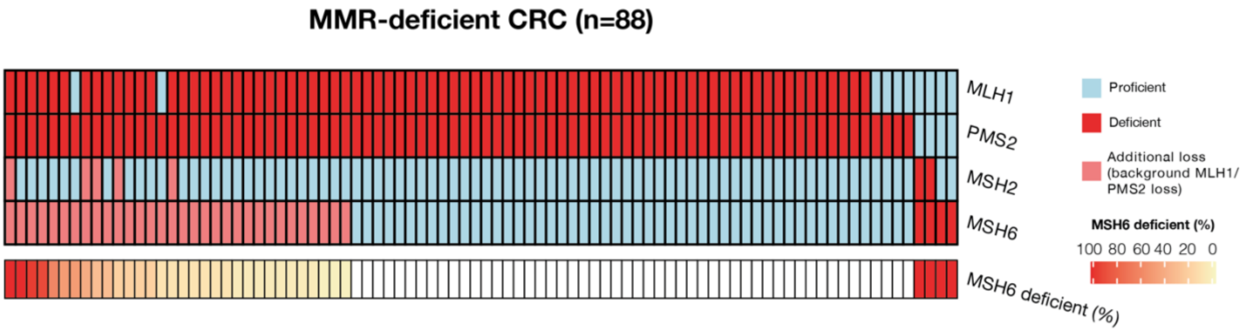


Figure 3.1: Pattern of MMR IHC labelling in MMRd cases within the UCLH colorectal cancer cohort.

Pattern of mismatch repair protein immunohistochemical labelling in 88 colorectal cancers within the UCLH cohort. Cases with additional loss of MSH6 on a background of MLH1/PMS2 loss are indicated in pink. MSH6 deficient (%) refers to the percentage of tumour cells with loss of IHC expression when assessed at whole slide imaging.

3.5 Spatial clonal mapping using MSH6 immunohistochemistry (IHC)

As detailed above, a subclonal pattern of MSH6 loss was commonly observed in colorectal cancers with background MLH1/PMS2 loss. Reviewing the IHC imaging of these cases revealed that the size of MSH6 deficient patches varied between tumours, ranging from isolated tumour glands to whole slide regions. On closer inspection, within MSH6 deficient regions there was often small nested proficient islands. This mosaic appearance suggested dynamic on-off switching of MSH6 expression. An example case of this phenomenon is provided in figure 3.2 below and additional examples are shown in figure S2. I now wanted to confirm that loss of MSH6 expression was due to a frameshift mutation in the hypermutable coding microsatellite contained within MSH6 identified in the previous chapter.

3.6 Sanger sequencing reveals frameshift slippage at the *MSH6* coding microsatellite controls MSH6 expression

A Sanger sequencing experiment was designed to confirm that loss of MSH6 IHC labelling reflected frameshift mutation within the *MSH6* microsatellite. Careful laser capture microdissection (LCM) of background MSH6 proficient tumour, MSH6 deficient tumour and nested MSH6 proficient subclones was performed in 3 separate tumours (all with background MLH1/PMS2 loss). Extracted DNA in each case underwent Sanger sequencing against the known coding microsatellite contained within the *MSH6* gene. The results confirmed that loss of MSH6 expression is associated with frameshift mutation in the *MSH6* homopolymer, whilst the background MSH6 proficient tumour retained the wild-type homopolymer length. In one case, a nested MSH6 proficient subclone within an MSH6 negative region was also obtained revealing return to the wild-type length. These results are illustrated in figure 3.3 below.

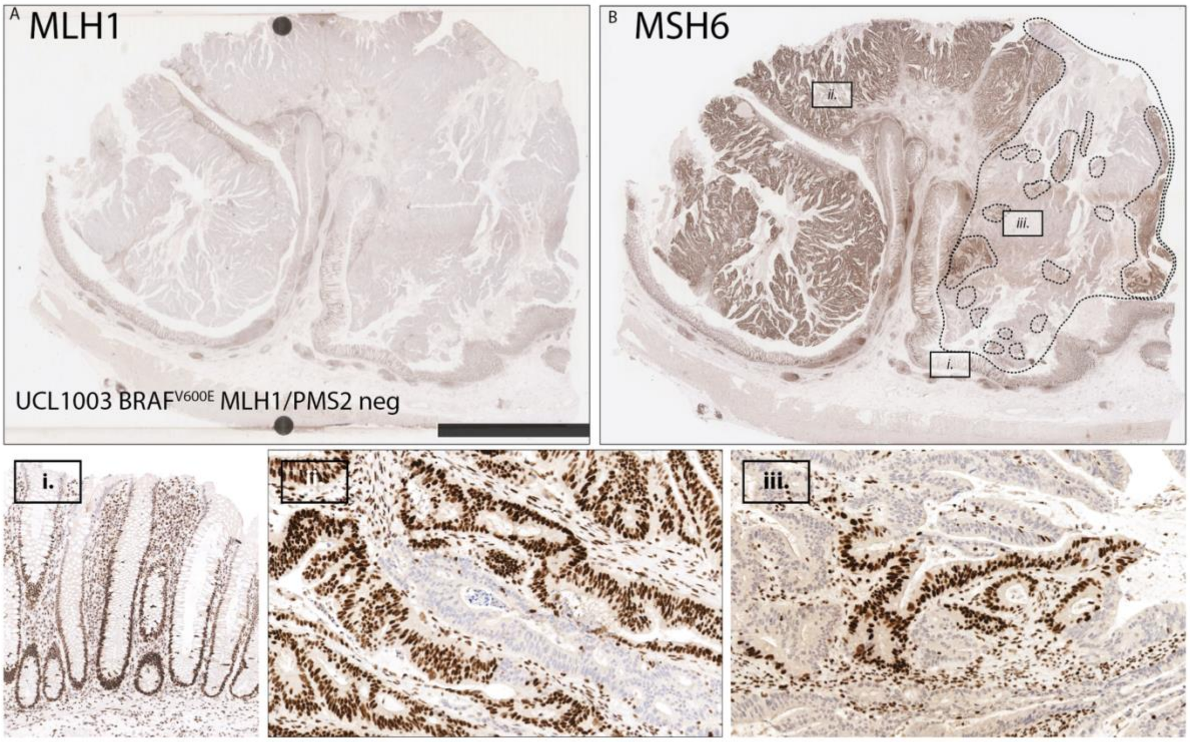


Figure 3.2: Pattern of MSH6 subclonal loss in an example tumour

(A) MLH1 and (B) MSH6 immunohistochemistry in a polypoid colorectal cancer. The right side of tumour is MSH6 deficient. Within the deficient region are numerous nested MSH6 proficient islands. High power photomicrographs display (i) normal colonic mucosa (ii) small MSH6 deficient subclone (iii) nested MSH6 proficient subclone within a deficient region.

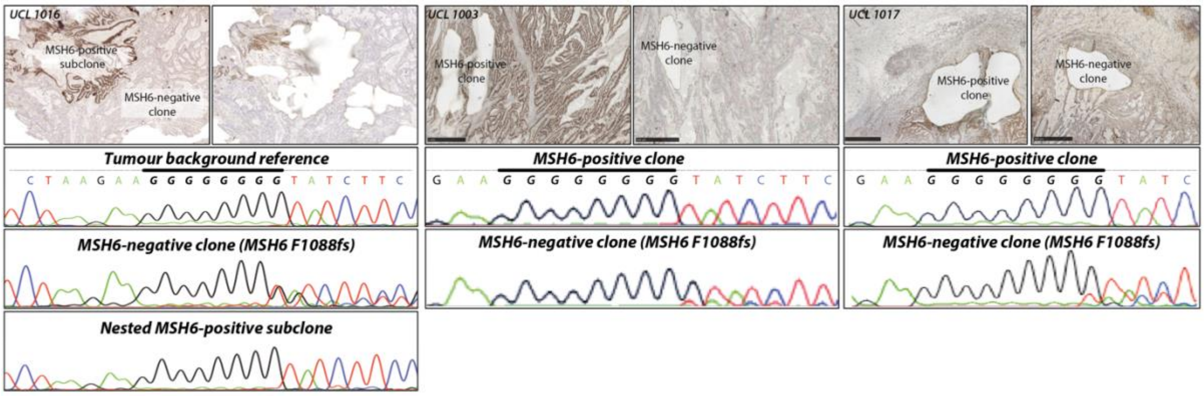


Figure 3.3: Laser capture microdissection of MSH6 proficient and deficient subclones followed by Sanger sequencing of the MSH6 homopolymer

Sanger sequencing results are displayed for 3 tumours; UCL 1016, UCL 1003 and UCL 1017. In each case the MSH6 negative subclone displays an insertion or deletion event at the microsatellite resulting in a frameshift mutation.

These data indicate that the expression of MSH6 is controlled by progressive expansions and contractions within its coding microsatellite resulting in on/off switching of gene expression. This proposed mechanism is illustrated in the cartoon in figure 3.4 below.

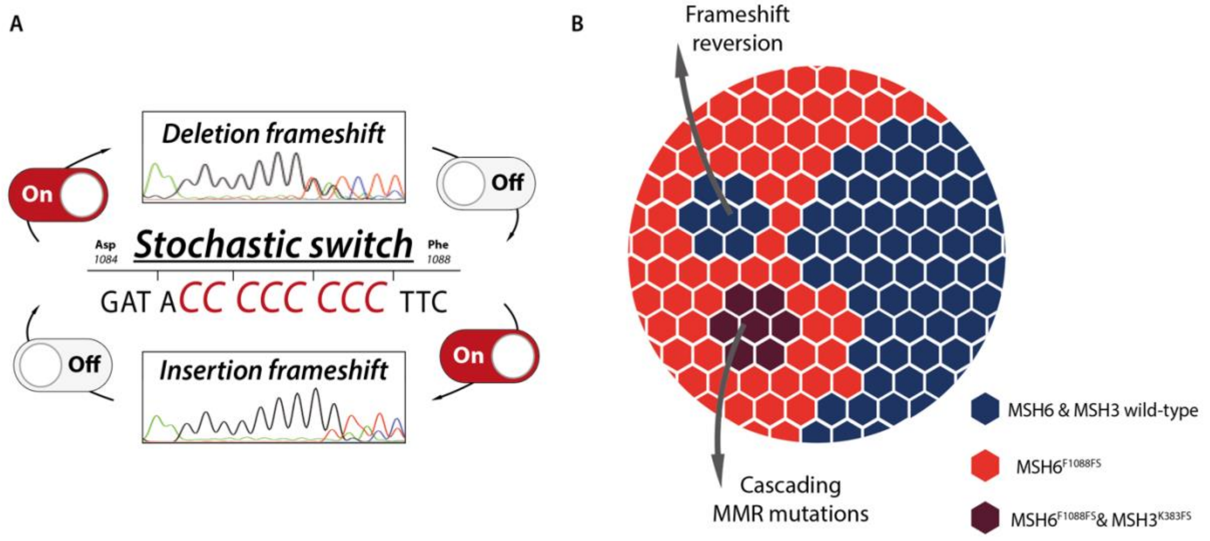


Figure 3.4: Cartoon illustrating proposed mechanism of frameshift switching at the MSH6 and MSH3 coding homopolymers.

(A) Insertion and deletion events at the coding microsatellites of *MSH6* and *MSH3* allow it to act as a molecular switch controlling expression of the gene (B) Cartoon depicting dynamics of *MSH6* and *MSH3* frameshifts in a tumour. Under this model reversion mutations may allow re-expression of the protein within a negative region, creating nested proficient subclones.

3.7 MSH3 immunohistochemistry

Immunohistochemistry was attempted with MSH3 but unfortunately a reliable antibody could not be found. Three separate MSH3 antibody clones were tested and different conditions systematically trialled, but the staining quality remained poor in each case. Occasional cases were of reasonable quality and these were scanned for future use. Review of the literature confirmed that MSH3 immuno-labelling was also typically poor in published studies (Plaschke et al., 2004). Thus, the use of MSH3 as a lineage tracing tool had to be abandoned.

3.8 Whole exome sequencing of MSH6 proficient and deficient subclones

To further investigate the functional impact of secondary mismatch repair frameshifts, I leveraged the use of MSH6 immunohistochemistry as a lineage tracing tool. A multi-region exome sequencing experiment was designed, taking biopsies from tumour regions guided by MSH6 expression status.

From the UCLH MMRd colorectal cancer cohort described above, 11 cancers with subclonal MSH6 loss on a background of MLH1/PMS2 loss were selected. A further 11 control cases with MLH1/PMS2 loss and no evidence of MSH6 loss were also selected. Using Laser Capture Microdissection (LCM), MSH6 proficient and deficient tumour regions were carefully microdissected from formalin fixed paraffin embedded (FFPE) tumour sections. This resulted in 49 samples from 22 tumours, which underwent DNA extraction, library preparation and whole exome sequencing (see methods section 7.2). An FFPE repair step was incorporated into the library preparation protocol to reduce the impact of formalin fixation artifact (discussed further below). In addition, for each tumour a normal tissue block was retrieved from the resection margin (i.e. away from the tumour) and used as the germline sample. Since blood samples were not available from these archival specimens, normal colonic mucosa was the most appropriate source of germline tissue here. H&E sections from the resection margin normal block were reviewed to ensure absence of tumour contamination. The clinical and pathological characteristics of these cases and samples generated is summarised in the table below.

	Immunohistochemistry status						Stage		
Tumour_ID	MLH1	PMS2	MSH2	MSH6	Age	BRAF^{V600E}		MSH6 deficient samples (n)	MSH6 proficient samples (n)
UCL_1018	proficient	deficient	proficient	subclonal	20	NO	T1N2M0	4	2
UCL_1006	deficient	deficient	proficient	subclonal	77	NO	T3N0M0	2	1
UCL_1005	deficient	deficient	proficient	subclonal	34	NO	T4N0M0	1	2
UCL_1004	deficient	deficient	proficient	subclonal	87	YES	T3N0M0	3	3
UCL_1003	deficient	deficient	proficient	subclonal	60	YES	T3N0M0	1	3
UCL_1016	deficient	deficient	proficient	subclonal	85	YES	T3N0M0	2	3
UCL_1007	proficient	deficient	proficient	subclonal	79	NO	T4N2M1	1	1
UCL_1015	deficient	deficient	proficient	subclonal	83	YES	T4N0M0	2	0
UCL_1002	deficient	deficient	proficient	subclonal	63	YES	T4N1M0	2	1
UCL_1014	deficient	deficient	proficient	subclonal	69	YES	T3N0M0	1	1
UCL_1001	deficient	deficient	deficient	subclonal	71	YES	T3N0M0	1	0
UCL_1023	proficient	deficient	proficient	proficient	64	NO	T2N0M0	0	2
UCL_1008	deficient	deficient	proficient	proficient	74	YES	T3N0M0	0	1
UCL_1009	deficient	deficient	proficient	proficient	34	NO	T4N1M0	0	1
UCL_1010	deficient	deficient	proficient	proficient	31	YES	T3N0M0	0	1
UCL_1011	deficient	deficient	proficient	proficient	83	YES	T3N0M0	0	1
UCL_1012	deficient	deficient	proficient	proficient	80	YES	T4N1M0	0	1
UCL_1013	deficient	deficient	proficient	proficient	77	YES	T3N0M0	0	1
UCL_1019	deficient	deficient	proficient	proficient	72	YES	T3N0M0	0	1
UCL_1020	proficient	deficient	proficient	proficient	60	NO	T3N0M0	0	1
UCL_1021	deficient	deficient	proficient	proficient	72	YES	T2N0M0	0	1
UCL_1022	deficient	deficient	proficient	proficient	31	NO	T4N2M1	0	1

Table 3.1: Clinical characteristics of 22 mismatch repair deficient colorectal cancers.

Included are the details of tumours making up the UCLH whole exome sequencing cohort. Subclonal loss of MSH6 indicates cases with tumour regions displaying loss of expression at IHC. All tumours displayed background MLH1/PMS2 loss. The number of samples collected from each tumour using laser capture microdissection is indicated in the final 2 columns.

Processed libraries were sequenced to a median depth of 100x on an Illumina Novaseq instrument (150bp paired end reads). Following Qc checks the sequencing data were processed using a detailed variant calling pipeline (see methods section 7.2). To ensure accurate calling of indels, the consensus of two variant callers Scalpel and Varscan2 were used. As a further validation check, frameshifts in *MSH6* and

MSH3 were checked by examining the length distribution of the microsatellite using the package MSIsensor and any discrepancies manually reviewed using IGV software.

The *MSH6*^{F1088fs} and *MSH3*^{K383fs} mutation status in samples and corresponding MSH6 IHC status is listed in table 3.2 below. All MSH6 deficient samples displayed instability at the MSH6 homopolymer resulting in a *MSH6*^{F1088fs} VAF of greater than 0.05. MSH6 proficient samples also often displayed instability at the MSH6 homopolymer when samples derived from tumours with deficient regions elsewhere in the tumour. Meanwhile MSH6 proficient samples from tumours with no evidence MSH6 deficiency throughout the tumour were all wild-type at the *MSH6* homopolymer. The likely reason for this is that bi-allelic mutations are required for loss of protein expression and in proficient regions with adjacent deficient regions a single allele had undergone frameshift mutation. The mutation status of *MSH6* and *MSH3* presented here was used to group samples as MSH6 and/or MSH3 mutated in downstream analyses.

Tumour ID	**Sample ID**	**MSH6 IHC status**	**MSH6^{F1088FS} status**	**MSH3^{K383FS} status**
UCL_1001	s110	deficient	mutated	mutated
UCL_1002	s30	deficient	mutated	mutated
UCL_1002	s31	proficient	mutated	mutated
UCL_1002	s33	deficient	mutated	mutated
UCL_1003	s6	deficient	mutated	mutated
UCL_1003	s7	proficient	mutated	mutated
UCL_1004	s50	proficient	mutated	mutated
UCL_1004	s51	proficient	mutated	mutated
UCL_1004	s54	proficient	mutated	mutated
UCL_1005	s1	proficient	mutated	mutated
UCL_1005	s5	deficient	mutated	mutated
UCL_1005	s3	proficient	mutated	mutated
UCL_1006	s61	deficient	mutated	mutated
UCL_1007	s66	proficient	wild-type	mutated
UCL_1008	s77	proficient	wild-type	mutated
UCL_1009	s42	proficient	wild-type	mutated
UCL_1003	s8	proficient	wild-type	mutated
UCL_1003	s9	proficient	wild-type	mutated
UCL_1010	s39	proficient	wild-type	mutated
UCL_1011	s40	proficient	wild-type	mutated

UCL_1012	s69	proficient	wild-type	mutated
UCL_1013	s105	proficient	wild-type	mutated
UCL_1014	s111	proficient	mutated	wild-type
UCL_1014	s112	deficient	mutated	wild-type
UCL_1015	s55	deficient	mutated	wild-type
UCL_1015	s56	deficient	mutated	wild-type
UCL_1007	s65	deficient	mutated	wild-type
UCL_1016	s12	deficient	mutated	wild-type
UCL_1016	s13	deficient	mutated	wild-type
UCL_1016	s14	proficient	mutated	wild-type
UCL_1016	s11	proficient	mutated	wild-type
UCL_1016	s10	proficient	mutated	wild-type
UCL_1016	s48	deficient	mutated	wild-type
UCL_1016	s52	deficient	mutated	wild-type
UCL_1016	s53	deficient	mutated	wild-type
UCL_1006	s60	deficient	mutated	wild-type
UCL_1006	s64	proficient	mutated	wild-type
UCL_1018	s22	deficient	mutated	wild-type
UCL_1018	s23	deficient	mutated	wild-type
UCL_1018	s24	deficient	mutated	wild-type
UCL_1018	s21	proficient	mutated	wild-type
UCL_1018	s26	deficient	mutated	wild-type
UCL_1019	s81	proficient	wild-type	wild-type
UCL_1020	s38	proficient	wild-type	wild-type
UCL_1021	s41	proficient	wild-type	wild-type
UCL_1022	s79	proficient	wild-type	wild-type
UCL_1023	s67	proficient	wild-type	wild-type
UCL_1023	s68	proficient	wild-type	wild-type
UCL_1018	s25	proficient	wild-type	wild-type

Table 3.2: MSH6 and MSH3 frameshift mutation status in samples (n=49) from the UCL whole exome sequencing cohort.

Samples and their corresponding tumour ID are provided. MSH6 immunohistochemical status and mutation status at the MSH6 and MSH3 coding homopolymers is stated.

3.9 Reducing sequencing artifacts due to formalin fixation

Sequencing artefacts and extensive DNA fragmentation are recognised problems associated with formalin fixation (Oh et al., 2015). In particular artefactual C>T transitions due to deamination of cytosine are associated with formalin fixation. In this part of the project, I was limited to the use of FFPE tissue due to both sample availability and because the IHC guided clonal sampling technique is better suited to

FFPE material. Current approaches to limit FFPE related artefact can be divided into steps taken during library preparation and bioinformatic approaches to filter out presumed artefact related variants (Guo et al., 2021). Recent reports indicate that most formalin fixation artefacts occur in the low allelic frequency range (<10%) and optimisation steps in the library preparation and data analysis steps can help reduce impact of FFPE artefact (Turnbull et al., 2018). Artefactual lesions can potentially be corrected using a cocktail of glycosylase enzymes such as uracil DNA glycosylase (UDG) and thymine DNA glycosylase (TTG) (Do & Dobrovic, 2015). Commercial FFPE repair kits using such enzymes are now available. In this project I optimised the use of a validated FFPE repair kit produced by New England Biolab (NEB). The manufacturer's publication of this kit reports that the repair mix removes variants produced by cytosine deamination and remaining erroneous variants are in the low frequency (1-5%) range (Chen et al., 2017). Following optimisation of the manufacturer's protocol I implemented FFPE repair during DNA library preparation for this project (see methods section 7.2).

3.10 *MSH6*$^{\mathit{F1088fs}}$ and *MSH3*$^{\mathrm{K383fs}}$ associates with increased mutation burden in UCLH whole exome sequencing cohort

Samples were grouped according to the presence of frameshifts in the known length 8 coding microsatellite of *MSH6* and *MSH3*. The results showed that samples carrying *MSH6*$^{\mathrm{F1088fs}}$ or *MSH3*$^{\mathrm{K383fs}}$ had significantly increased mutation burden and frameshift of both microsatellites had an additive effect resulting in an over two-fold increase in total mutation burden. This increase was also significant at the level of single nucleotide variants and indels (figure 3.5). The magnitude of increase in mutation burden here was larger than observed in the previous GEL cohort, reflecting microdissection enriching for a clonally pure population of cells according to MSH6 status.

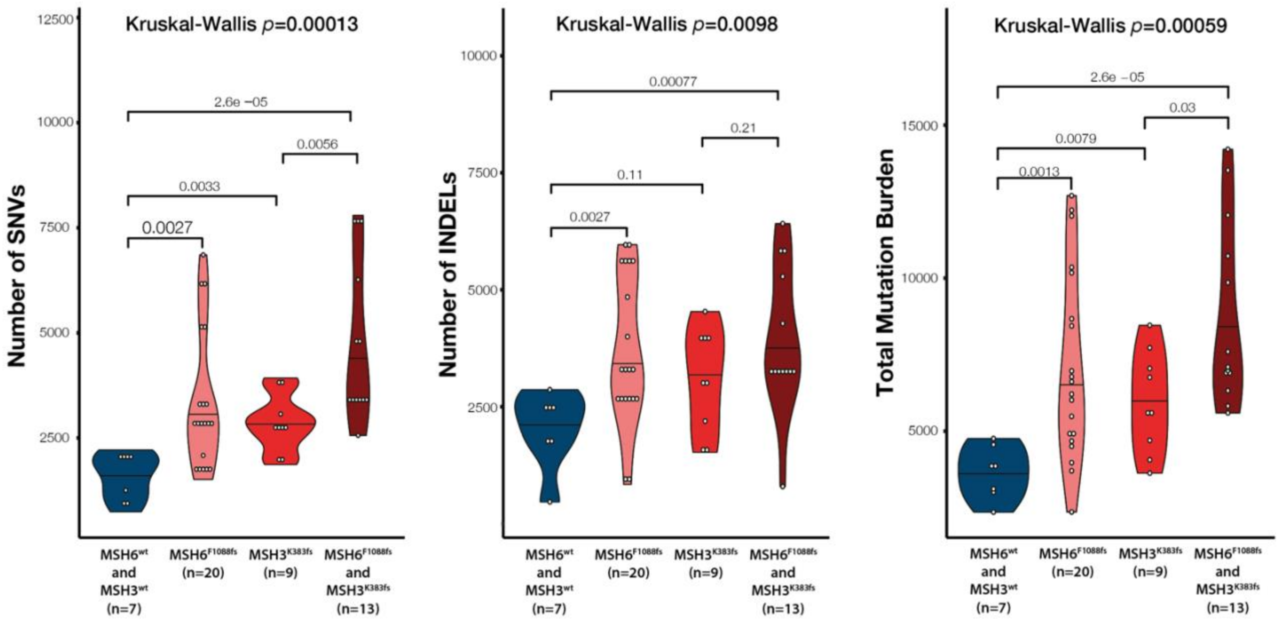


Figure 3.5: Mutation burden in the UCLH cohort according to MSH6 and MSH3 frameshift mutation status.

Violin plots showing SNV, indel and total mutation burden in samples grouped according to *MSH6* and *MSH3* microsatellite frameshift mutation status.

To account for the non-independence of multiple sampling per patient, a linear mixed effects model was created to assess the relationship between *MSH6*/*MSH3* frameshift status on total mutation burden. For this model, individual variation between tumours was entered as a random effect. For fixed effects *MSH6*/*MSH3* frameshift status, age at diagnosis and tumour purity were used. The model confirmed that the increased mutation burden associated with *MSH6* and *MSH3* frameshifts remained significant after accounting for random and fixed effects (p=0.0296, supplementary table S1). As a further check tumour purity was compared between samples according to *MSH6*/*MSH3* grouping confirming no difference between groups (Kruskal-Wallis, p=0.48) (supplementary figure 1). For this analysis tumour purity was estimated using the Sequenza package (see methods section 7.2). Overall, these results confirm that secondary frameshifts in *MSH6* and *MSH3* associate with increased mutation burden in MMRd tumours that have a background of MLH1/PMS2 loss.

3.11 ***MSH6*$^{\mathbf{F1088fs}}$ and *MSH3*$^{\mathbf{K383fs}}$ leads to a qualitative change in mutation bias**

To explore further the functional impact of *MSH6* and *MSH3* frameshifts on the exome I next investigated the mutation bias within samples. Previous studies in model organisms and bacteria have reported a specific mutation bias associated with disruption of different MMR components (Schaaper et al., 1987; Hegan *et al.*, 2006). MutS (MSH2 or MSH6/MSH3) deficiency is associated with a C>T and C>A bias whilst MutL (MLH1/PMS2) deficiency associates with a broader spectrum involving C>T, C>A and T>C.

Remarkably analysis of mutation bias across our tumour cohort was in keeping with what has been described in the literature. Specifically, samples with *MSH6* or *MSH3* frameshifts showed an increase in the proportion of C>T transitions and C>A transversions and a relative decrease in T>C transitions (figure 3.6). The effect was greater in samples with combined *MSH6* and *MSH3* frameshifts. Overall, the change in mutation bias indicates that in the presence of *MSH6*/*MSH3* frameshifts there is a shift towards a MutS mutation bias. This qualitative change in mutation bias provides further evidence for the functional impact of *MSH6* and *MSH3* frameshifts in MLH1/PMS2 deficient tumours.

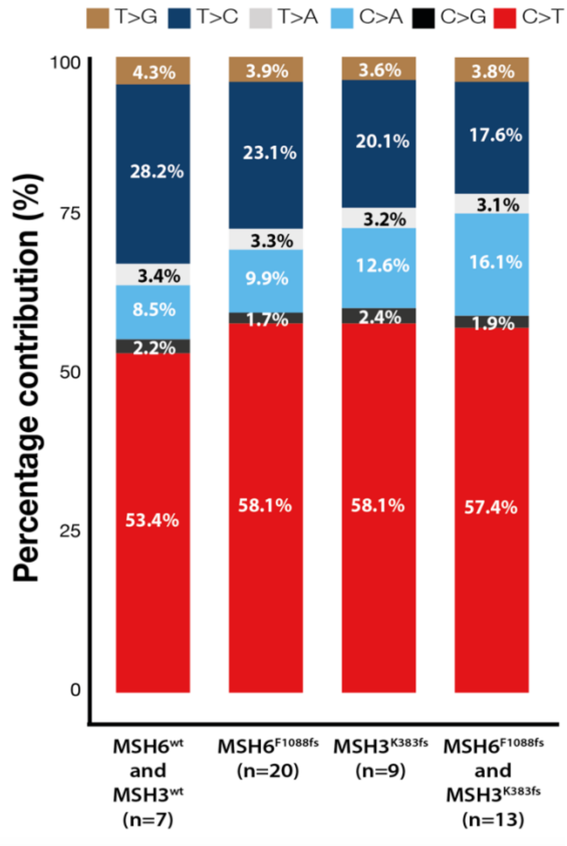


Figure 3.6: Mutation bias in samples grouped according to MSH6 and MSH3 frameshift mutation status.

Stacked barplot showing proportion of each type of single nucleotide variant. Samples are from the UCLH whole exome sequencing dataset ***(n=49 samples)***.

3.12 Mutation signature analysis

The change in mutation bias identified above can be investigated in more detail with mutation signature analysis. Mutation signatures use the genomic context of mutations to identify patterns characteristic of the underlying mutation process (Alexandrov et al., 2020). This will allow further characterisation of the specific bias associated with *MSH6* and *MSH3* frameshifts. Furthermore, signature analysis allows for changes in the indel and doublet base substitution (DBS) patterns to be explored. Interestingly within the COSMIC v3 single bases substitution (SBS) signatures, 5 are attributed to defective mismatch repair (SBS 6,15,21,26 and 44). This raises the possibility that the underlying cause of mismatch repair loss (i.e. MutL vs MutS loss) may impact the observed signature. Given that my work has identified subclonal MSH6 and MSH3 loss in the context of MutL deficiency, the observed signatures may reflect a combination of both processes. I next investigated the mutation signatures operating within these samples by comparing between MSH6/MSH3 grouping of samples. The mutation calls from each sample were pooled into one of 3 groups according to the presence of *MSH6*/*MSH3* frameshifts as follows:

Group 1: *MSH6* & *MSH3* wild-type

Group 2: *MSH6*$^{\mathrm{F1088FS}}$ or *MSH3*$^{\mathrm{K383FS}}$

Group 3: *MSH6*$^{\mathrm{F1088FS}}$ and *MSH3*$^{\mathrm{K383FS}}$

De-novo signatures were extracted from each group of mutations using the package Sigprofiler. Signature extraction was performed with the following settings: number of non-negative matrix factorisation replicates=500, minimum signatures to be extracted=1 and maximum signatures to be extracted=5. Three SBS signatures were extracted (SBS96A, SBS96B and SBS96C). The *MSH6*/*MSH3* wild-type group showed a predominantly SBS96A pattern whilst groups with *MSH6* or *MSH3* frameshifts or both showed SBS96B and SBS96C contribution (figure 3.6). Review of the trinucleotide context showed that SBS96B and SBS96C display a relative increase in C>T contribution especially within a GCG>GTG context and also increased C>A within a CCT>CAT context. In contrast SBS96A displays a relatively higher T>C contribution at multiple peaks. These findings are all in keeping with the previous mutation bias data but with the greater resolution of sequence context.

Cosine similarity analysis was performed to compare the 3 de-novo SBS signatures against the known COSMIC v3 mismatch repair signatures (figure 3.7). Broadly this showed there was moderate similarity to the known MMR signatures. However, since the COSMIC signatures are derived from bulk sequencing data, the effect of subclonal *MSH6*/*MSH3* events are likely to be mixed within existing known signatures.

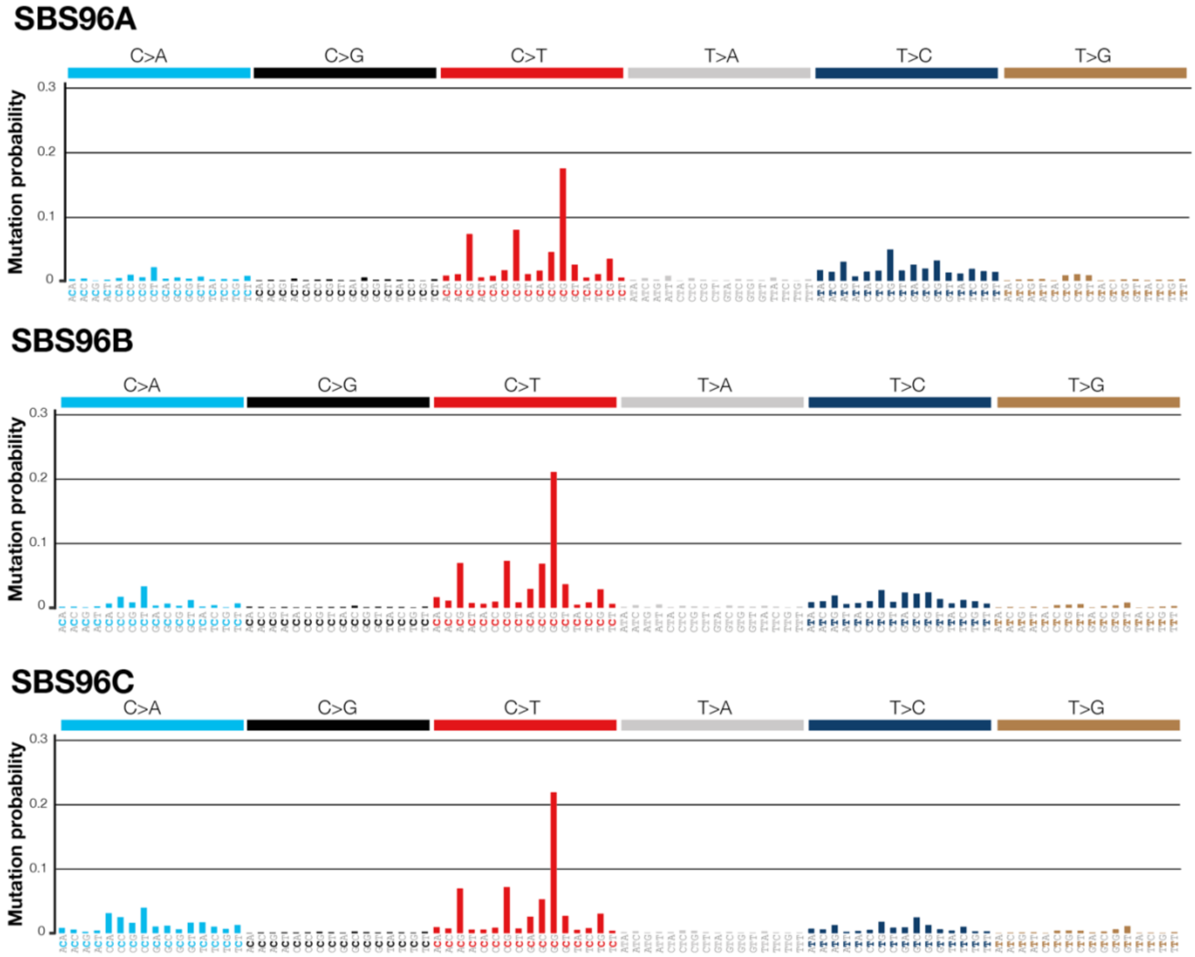


Figure 3.7: Trinucleotide context of de-novo signatures SBS96A, SBS96B and SBS96C.

The proportional contribution of each SNV in its trinucleotide context is displayed for each de-novo signature extracted from samples in ***the*** UCL WES cohort.

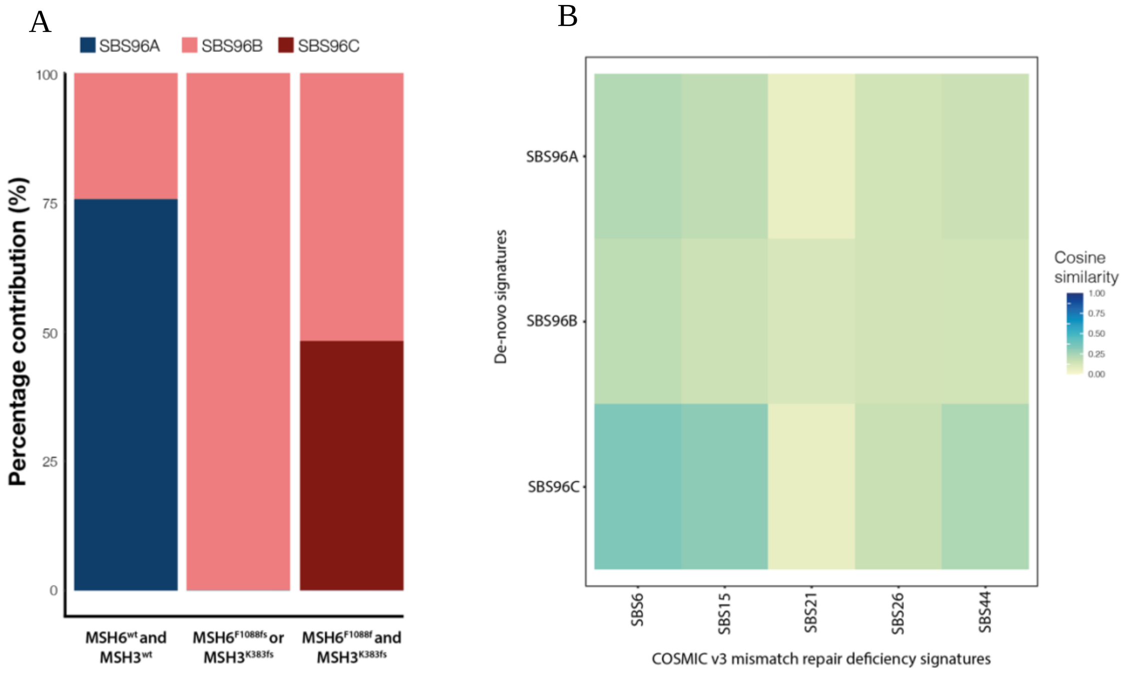


Figure 3.8: Contribution of de-novo SBS signatures in samples according to MSH6/MSH3 frameshift mutation status.

(A) Percentage contribution of de-novo signatures SBS96A, SBS96B and SBS96C in samples grouped according to *MSH6* and *MSH3* microsatellite frameshift status. (B) Cosine similarity between extracted de-novo SBS signatures and known COSMIC v3 mismatch repair deficiency signatures.

Indel and doublet base substitution (DBS) signatures were next investigated and are presented in figures 3.9 and 3.10 respectively. Three de-novo INDEL signatures, ID83A, ID83B and ID83C were identified. The *MSH6*/*MSH3* wild-type group displayed predominantly ID83A, whilst *MSH6* and *MSH3* mutated groups displayed ID83B and ID83C. Homopolymer T deletions were the most frequent type of INDEL in all three signatures. However, ID83B and ID83C displayed an increased proportion of deletions at longer (>1bp) repeat unit lengths. Cosine similarity analysis against the known MMRd signatures found that all 3 signatures showed strong similarity to COSMIC v3 ID2 and moderate similarity to ID7. Overall, these INDEL signature data indicate a broadening of indel repertoire in the presence of *MSH6*/*MSH3* frameshifts with specifically an increased contribution of INDELs at longer repeat unit length.

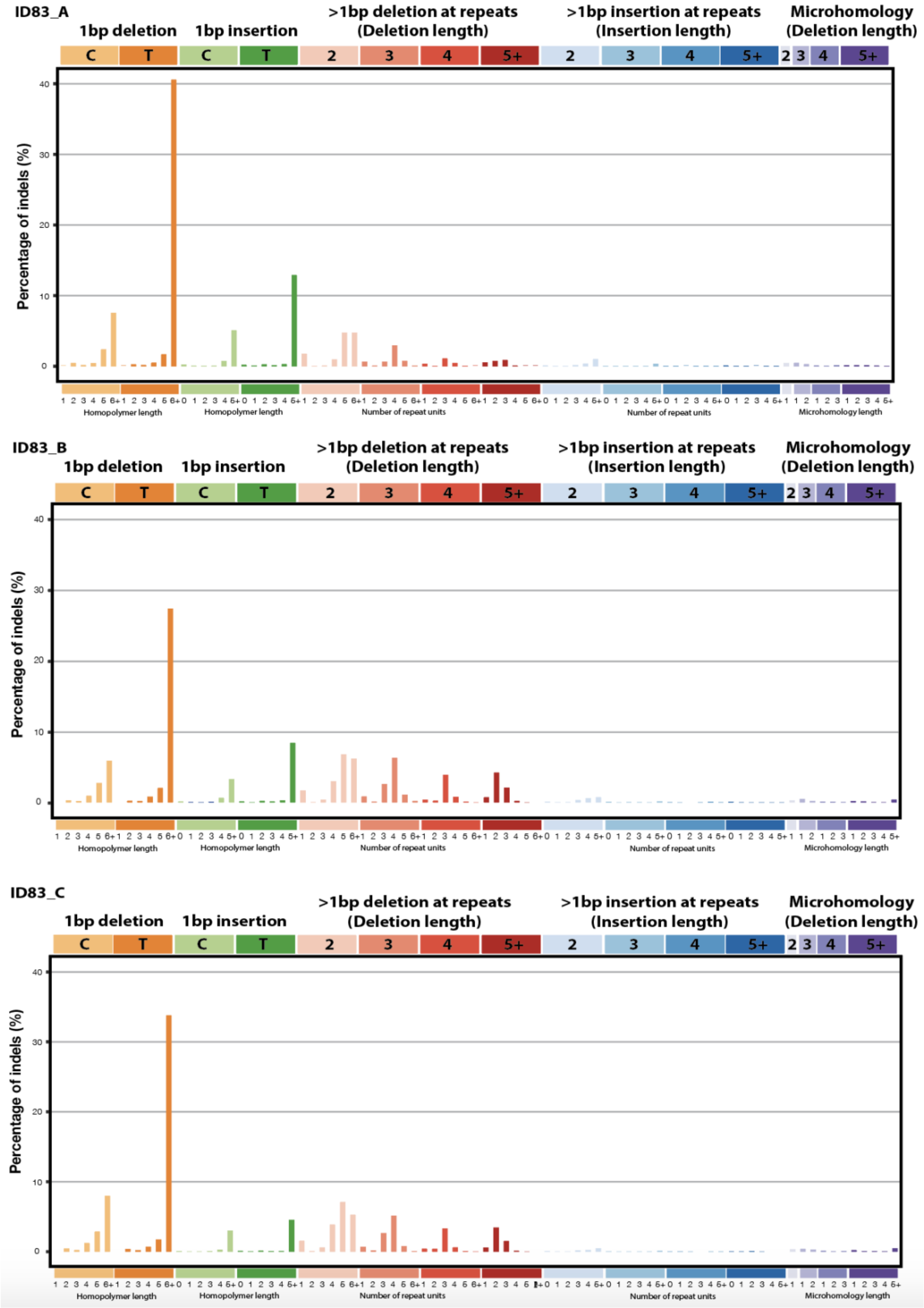


Figure 3.9: Mutation profile of de-novo INDEL signatures ID83A, ID83B and ID83C

Proportional contribution of each INDEL in the 83 class format for de-novo INDEL signatures ID83A-C extracted from samples in the UCL WES cohort.

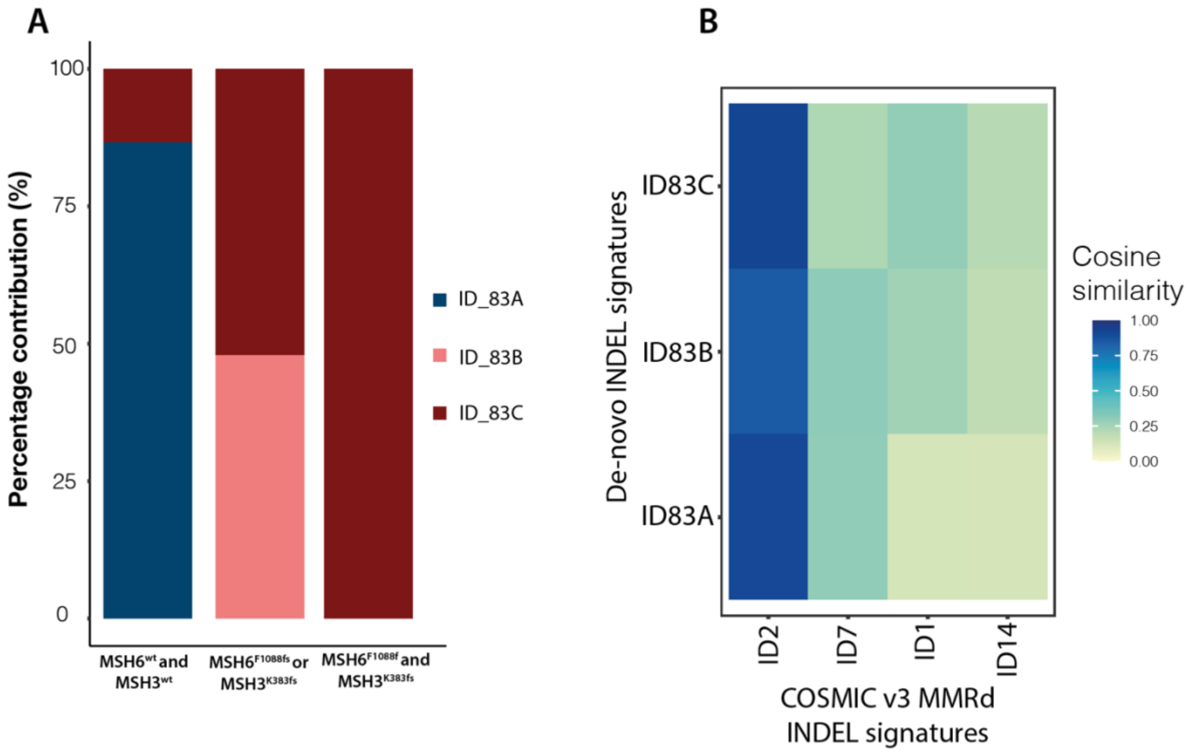


Figure 3.10: Contribution of de-novo indel signatures in samples according to MSH6/MSH3 frameshift mutation status.

(A) Percentage contribution of de-novo INDEL signatures in samples grouped according to *MSH6/MSH3* microsatellite frameshift status. (B) Heatmap displaying cosine similarity between extracted de-novo INDEL signatures and COSMIC v3 mismatch repair deficiency signatures.

One de-novo DBS signature was identified across all 3 groups (figure 3.11). This showed a high proportion of contribution in the CG>TA and GC>AT contexts. The cosine similarity showed this signature to have high similarity to the known DBS MMRd signature DBS10.

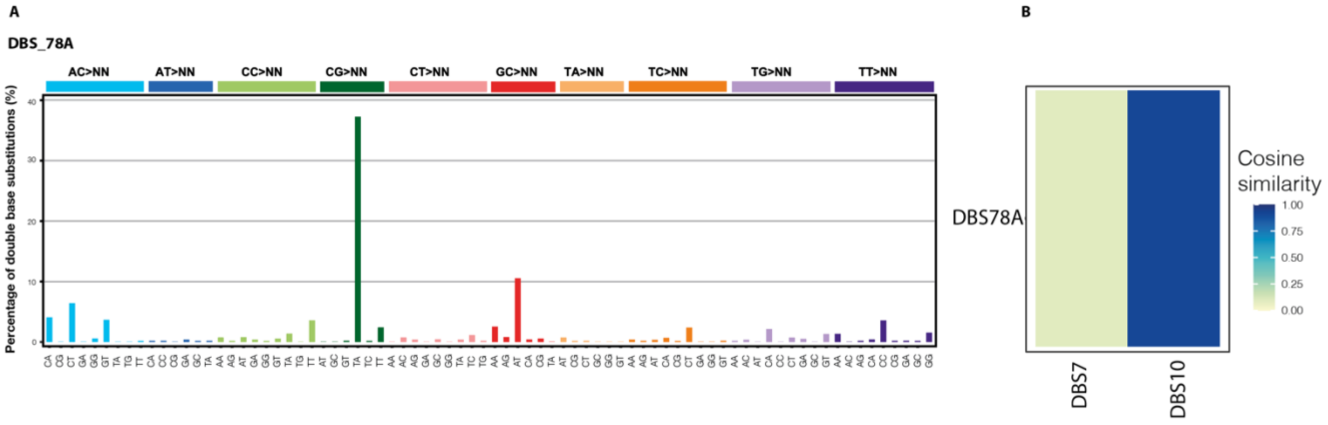


Figure 3.11: De-novo DBS signature extracted from samples in UCL WES cohort.

(A) Mutation types present in de-novo DBS78A, (B) Cosine similarity between DBS78A and COSMIC v3 DBS mismatch repair deficiency signatures.

3.13 COSMIC signature contribution within samples grouped according to *MSH6^{F1088fs}* and *MSH3^{K383FS}*

The previous section focused on identifying de-novo signatures in samples grouped according to *MSH6/MSH3* frameshift mutation status. This approach was taken because the existing COSMIC mutation signatures are derived from bulk sequencing data, whilst *MSH6* and *MSH3* frameshifts are typically subclonal events within MMRd tumours. Despite the advantages of searching for de-novo signatures, it remains important to examine our dataset against the known COSMIC signatures. The COSMIC signatures have been validated in large series of cancers and thus remain an important reference. I set out to determine the contribution of known mutational signatures in our sample groupings. COSMIC signatures were extracted using the Sigprofiler tool. As previously, mutation calls from samples were pooled into three groups according to *MSH6/MSH3* frameshift mutation status. Signature extraction was performed with the following settings: number of non-negative matrix factorisation replicates=500, minimum signatures to be extracted=1 and maximum signatures to be extracted=6. An upper limit of 6 maximum signatures was used to avoid overcalling low prevalence signatures. COSMIC signature contribution for single base substitutions (SBS) and INDELs are displayed in figure 3.12 below.

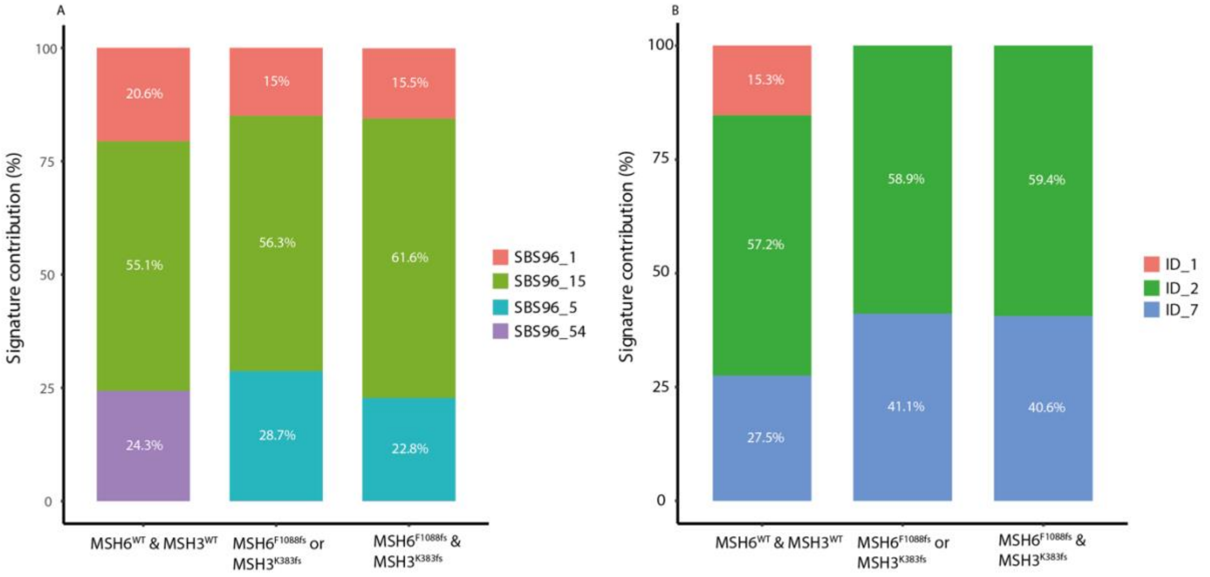


Figure 3.12: Contribution of COSMIC SBS and INDEL signatures in samples from UCL WES cohort grouped according to MSH6/MSH3 frameshift mutation status.

(A) COSMIC SBS signature contribution (B) COSMIC INDEL signature contribution.

For SBS signatures, SBS15 showed the largest contribution in all three sample groupings. SBS15 is a known signature of mismatch repair deficiency. All sample grouping also showed a varying contribution of SBS1. SBS1 is characterised by C>T mutations associated with spontaneous deamination of methyl-cytosine and is also recognised as a signature of aging. The groups with *MSH6* and/or *MSH3* frameshifts also showed contribution from SBS5, which is also an aging/ clock like signature. SBS54 was only observed in the *MSH6/MSH3* wild type group and is considered an artefact related signature. For INDEL signatures, ID2 was the major contributor in all three sample groupings. ID2 is known to be caused by slippage during replication of the template strand and is known to be a significant contributor in samples with mismatch repair deficiency. ID1 was only observed in the *MSH6/MSH3* wild-type group and is known to be associated with slippage during replication of the replicating strand and commonly associated with mismatch repair deficiency. ID7 was observed in all sample groupings but showed increased contribution in the groups with *MSH6* and/or *MSH3* frameshift. ID7 is a signature known to be associated with mismatch repair deficiency. Overall, the COSMIC signature analysis here confirms that known mismatch repair deficiency signatures contribute a significant proportion of mutations observed in this dataset.

Conclusions on mutation signature anaylsis

The de-novo signature analysis found that in samples with *MSH6*/MSH3 frameshifts there is a qualitative change in mutation profile, with relative increase in C>T and C>A SNVs and increased INDELs at longer repeat unit lengths. The potential limitations of this analysis are the small sample size of the dataset, but this is to some extent mitigated by the high mutation frequency observed in MMRd cancers. FFPE related artefact may also impact the signature analysis. Whilst this impact is reduced by having used an FFPE repair protocol during the library preparation, FFPE artefact likely continues to impact the analysis. In the future, a larger cohort of samples, using fresh frozen tissue and whole genome sequencing data would allow for improved signature analysis.

3.14 *MSH6*^{F1088fs} and *MSH3*^{K383fs} are associated with increased proportion of subclonal mutations

As tumours grow, they acquire new mutations which are passed on to daughter cells. Mutations that arise early during tumour formation are shared between all tumour cells and defined clonal, whilst later mutations will only be present in a proportion of tumour cells and are defined subclonal. A key finding of the work so far is that secondary *MSH6* and *MSH3* frameshifts increase mutation burden, suggesting they act by increasing mutation rate. I reasoned that an increase in mutation rate would lead to a greater proportion of subclonal mutations in a given sample, since with each cell division more private mutations would accumulate. To assess this, I compared the proportion of clonal versus subclonal mutations within samples grouped according to *MSH6* and *MSH3* frameshift mutation status. This analysis was performed on the UCL whole exome sequencing dataset of 49 samples, where MSH6 deficient and proficient subclones have been sampled by laser capture microdissection. The LCM technique enriches for a clonally pure population of tumour cells and therefore MSH6 deficient samples are subclones at a tumour level but clones at a sample level. This is in contrast to traditional bulk sequencing data where samples are polyclonal mixtures.

An established formula was used classify all SNVs as either clonal or subclonal at an individual sample level (Letouzé et al., 2017; McGranahan et al., 2015). This calculates the cancer cell fraction (CCF) of each mutation by correcting the variant allele frequency (VAF) for tumour purity and the absolute copy number of the genomic locus of each mutation (see methods section 7.2; Clonality assessment). A 95% confidence interval for each mutation's CCF value is also calculated. Mutations where the CCF 95% C.I. upper boundary is above 0.95 are classed as clonal and subclonal otherwise. Example VAF distribution plots for an MSH6 deficient and proficient sample from the same tumour are provided in figure 3.13B. VAF and CCF distribution plots for all samples are provided in supplementary figure S3. I limited this analysis to SNV mutations only since INDEL mutations can often be biallelic in MMRd cancers, complicating the clonality assessment. The combined results show that samples with *MSH6*^{F1088fs} or *MSH3*^{K383F} or both have significantly increased proportion of subclonal mutations compared to *MSH6*/*MSH3* wild-type samples (figure 3.13A). This result is consistent with *MSH6* and *MSH3* frameshifts increasing mutation rate during tumour evolution, leading to increased subclonal diversification.

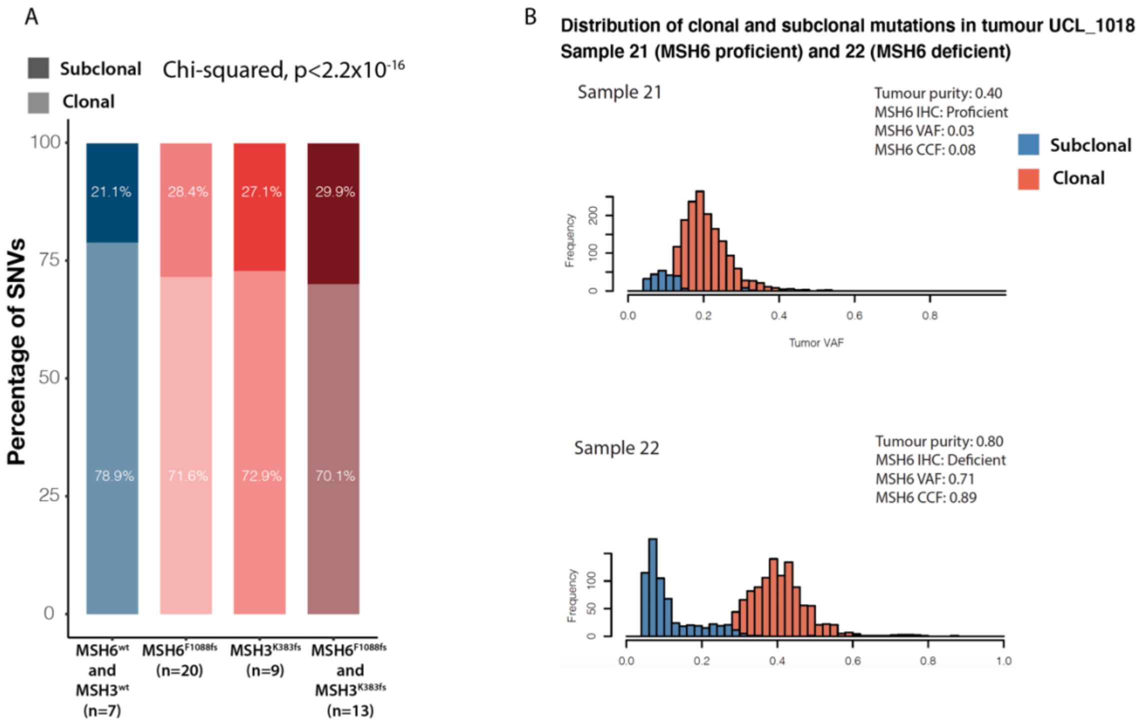


Figure 3.13: MSH6^{F1088fs} and MSH3^{K383fs} increase the proportion of subclonal mutations

A) Proportion of clonal versus subclonal mutations in samples grouped according to *MSH6* and *MSH3* frameshift mutation status. B) Example showing VAF distribution of clonal and subclonal SNVs in a MSH6 proficient (top) and deficient sample (bottom) from the same tumour (UCL_1018).

A limitation of this approach to clonality assessment is that it does not take into account different regions of the same tumour. It is known that clonality assessment using single biopsies may overestimate the number of clonal mutations. This is because there can be the illusion of clonality due to variants present at high VAF in some regions but absent in other regions (de Bruin et al., 2014). Variation in sequencing depth between samples can also impact clonality assessment when using a single biopsy approach. However, in this project we were limited by single biopsies in several tumours and so clonality assessment had to be performed using a single sample approach to ensure consistency. In future work, it will be important to obtain multi-region biopsies from spatially separated regions of tumours such that clonal and subclonal mutations can be more accurately assigned.

3.15 Conclusions

In this chapter two significant milestones have been achieved. Firstly, I was able to establish a technique to spatially map MSH6 deficient subclones within MMRd colorectal tumours. The spatial distribution of subclones together with Sanger sequencing data suggested dynamic switching of MSH6 expression controlled by successive frameshift events at the microsatellite. Using these spatial clonal maps, I was able to take multi-region biopsies of MSH6 proficient and deficient subclones to establish a large whole exome sequencing dataset. This provided a clonally resolved dataset to investigate the impact of secondary mismatch repair frameshifts and provides a useful resource for the rest of the project. Second, the exome sequencing data confirms that secondary frameshifts in *MSH6* and *MSH3* associate with a quantitative increase in mutation burden and a qualitative shift in mutation bias. These findings suggest secondary *MSH6* and *MSH3* frameshifts have a functional role in DNA repair in tumours with background MLH1/PMS2 loss. In the next chapter I will consider the costs and benefits associated with this increased mutation load and the evolutionary pathways underpinning this process.

In summary the key findings in this chapter are:

- *MSH6* and *MSH3* both contain a length 8 coding homopolymer that undergoes frequent frameshift mutations in the context of microsatellite instability leading to loss of protein expression
- Lineage tracing using MSH6 immunohistochemistry identifies frequent nested proficient subclones within deficient regions suggesting dynamic on/off switching of MSH6 expression.
- *MSH6* and *MSH3* frameshifts associate with a quantitative increase in mutation burden in MLH1/PMS2 deficient colorectal cancer.
- *MSH6* and *MSH3* frameshifts associate with a qualitative change in mutation bias and signature with increase in C>T and C>A mutation contribution.
- *MSH6* and *MSH3* frameshifts associate with a broadening of indel signature with greater proportion of deletions at >1bp repeat unit length.

Chapter 4: Costs and benefits associated with secondary *MSH6* and *MSH3* frameshifts

4.1 Contribution statement

The work in this chapter has contributed to a paper currently under review. Bioinformatic analysis of genomic data was completed by myself. The wetlab experiments for multiplex immunofluorescence (MIF) work was completed by myself. Image analysis of the MIF dataset was performed by Dr Panagiotis Barmpoutis. I produced the figures presented which are also used in the paper.

4.2 Introduction

Mutability provides tumours with the opportunity to gain variants that confer a fitness advantage, allowing adaptation to selection constraints. This is particularly important in situations of high selection pressure e.g. during therapy or due to the immune system. Mismatch repair deficient cancers, owing to their high rate of neoantigen generation are under strong immune selection pressure. Multiple mechanisms of immune escape have been described in MMRd cancers, such as mutations in the HLA class I complex, beta-2-microglobulin and components of the antigen processing machinery (Grasso et al., 2018). Whilst these events describe the molecular drivers of immune evasion, the evolutionary dynamics involved in reaching these events in a growing population of tumour cells is unclear. In the previous chapter I found that secondary frameshifts in *MSH6* and *MSH3* lead to both a quantitative increase in mutation burden and a qualitative change in mutation bias. Since increased mutability can lead to both adaptive and deleterious mutations, I now consider the consequences in terms of costs and benefits during tumour growth. I focus on the impact on immune recognition and evasion given the dominant role played by immune system during MMRd evolution.

4.3 Neoantigen burden is increased in the presence of *MSH6*^{F1088fs} and *MSH3*^{K383fs} frameshifts

In the previous chapter it was found that secondary *MSH6* and *MSH3* frameshifts lead to increased tumour mutation burden and a shift in mutation bias. I now explored whether this increased mutability also increased immunogenicity in terms of neoantigen burden. To investigate this neoantigens were predicted using an established pipeline (Schenck et al., 2019), utilising patient specific HLA haplotypes and the NetMHCPan tool to call neoantigens (see methods section 7.2; Neoantigen prediction). The total number of predicted strong binding unique neoantigens were counted for each sample. The results showed a clear increase in neoantigen burden in samples with *MSH6* or *MSH3* frameshifts (figure 4.1A). Again, combined *MSH6* and *MSH3* frameshifts had an additive impact, resulting in a greater than 2-fold increase in neoantigen burden. This finding suggests subclones with *MSH6* and *MSH3* frameshifts have greater capacity to provoke immune recognition.

I was also interested to understand the clonality of the increased neoantigen burden, given the positive association between clonal neoantigen burden on prognosis and immunotherapy response (McGranahan et al., 2016). To assess this, I used the same method used in chapter 3 to assess clonality of overall SNV burden. Briefly, the cancer cell fraction (CCF) of each neoantigen within samples was calculated by correcting the mutation VAF for tumour purity and copy number at the mutation locus (see methods section 7.2; Clonality assessment). Neoantigens present in all tumour cells in a sample are labelled clonal whilst those present in a proportion of tumour cells labelled subclonal. As before only SNV neoantigens were included in this analysis as INDELs can often display biallelic mutations in MMRd tumours complicating the clonality assessment. The results showed a significant increase in the proportion of subclonal neoantigens in samples with *MSH6* and/or *MSH3* frameshifts (figure 4.1B). This finding is in keeping with *MSH6* and *MSH3* frameshifts driving increased genetic diversity and immunogenic variants. As in the previous analysis of clonality, this assessment has been performed using a single biopsy approach to call clonal and subclonal mutations. This approach was taken to ensure consistency in the analysis given that in several tumours (n=11), only single biopsies were available. The limitation of this approach is that it may overcall clonal mutations when a variant is present at high VAF in one region but absent elsewhere.

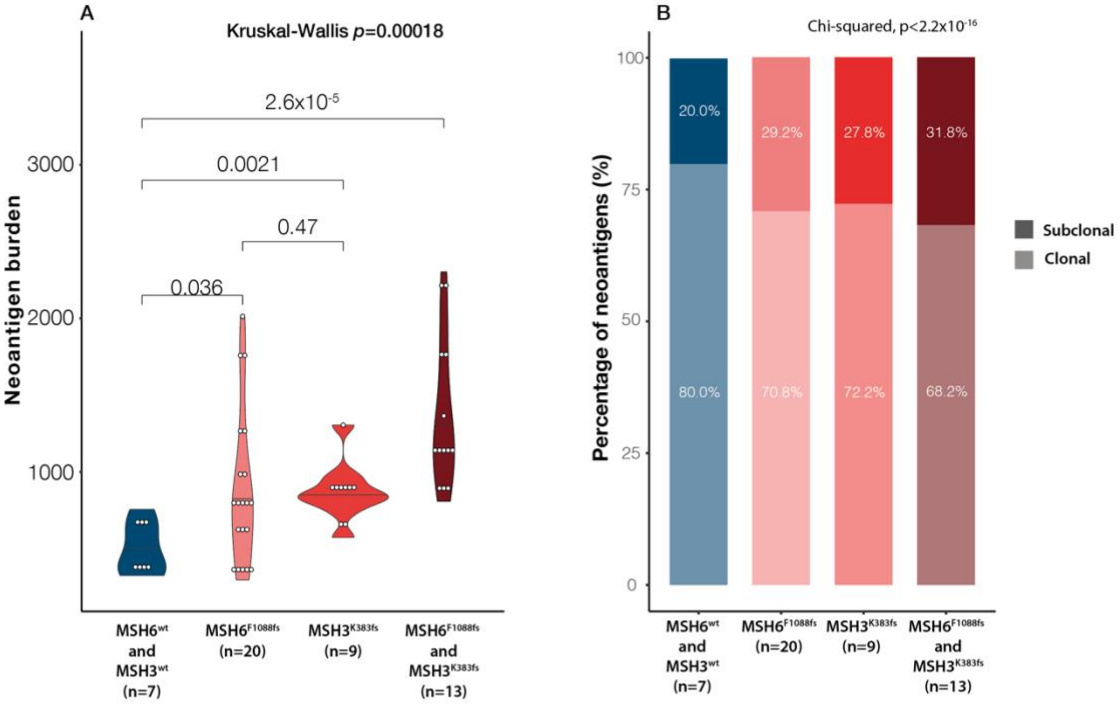


Figure 4.1: Secondary MSH6 and MSH3 frameshifts lead to increased subclonal neoantigen burden

A) Total neoantigen burden and B) Percentage of clonal versus subclonal neoantigens according to MSH6^{F1088fs} and MSH3^{K383fs} grouping of samples.

4.4 Immune infiltration analysis using multiplex immunofluorescence (MIF)

A limitation of the in-silico neoantigen prediction performed in the previous section is that it relies on prediction of mutations likely to have affinity for MHC molecules. However, the immunogenicity of a mutation also depends on various other factors such as, peptide stability, post translational modifications and T-cell receptor diversity. For these reasons identifying genuine neoantigens capable of inducing a T-cell response is challenging. In order to measure the impact of secondary MMR frameshifts on the immune microenvironment in-vivo, I decided to quantify levels of immune infiltration within tumours. This would allow comparison of immune activity both within and between tumours according to MSH6 expression status. A multiplex immunofluorescence (MIF) experiment was designed combining immunolabelling for MSH6 with key immune cell markers, CD8, CD20, CD4, FOXP3 and also the epithelial marker pan-CK (see methods section 7.5). After optimisation, this MIF panel was applied our cohort of MMRd tumours. Tumour sections used here were adjacent sections to the previous genomic analysis, allowing correlation between immune infiltration and genomic data to be performed. After immunofluorescence labelling of tumour sections, high power images of regions were taken. Since immune cells as well as tumour cells can both express MSH6, a bespoke image analysis pipeline was developed to segment tumour cells and immune cells accurately. Neighbourhood analysis was then performed to quantify the number of each immune subset and comparisons between MSH6 proficient and deficient tumour regions performed. This image analysis workflow, named ORION (FluORescence cell segmentatION workflow) is described in figure 4.3A-D below.

4.5 MSH6 deficient subclones have increased immune infiltration

A total of 194 tumour regions across 27 tumours were assessed with multiplex immunofluorescence. This consisted of 21 tumours and matching blocks which had been used in the previous whole exome sequencing dataset plus an additional 5 tumours (all MMRd with background MLH1/PMS2 loss and with or without secondary MSH6 loss). One tumour was discarded due to poor immunolabelling results. 1mm^2 imaging tiles were taken from MSH6 proficient and deficient regions. Where possible the tumour regions that had previously undergone sequencing were captured within imaged tiles. Each imaged tile was classified as MSH6 proficient or deficient and tiles with mixed populations of tumour cells were excluded from the analysis. The frequency of each immune cell subtype was quantified for each imaged tile using neighbourhood analysis (see methods section 7.5). Figure 4.1 below shows the quantified number of each immune cell subtype within each imaged tile for the cohort of tumours. Tumours that were not represented in the previous DNA sequencing work are marked by an asterisk and labelled in grey. There was significantly increased numbers of CD8$^+$ cytotoxic T-cells, CD20$^+$ B lymphocytes, CD4$^+$ T helper cells and FOXP3$^+$ T-regulatory cells within MSH6 deficient regions as compared to proficient regions (figure 4.2A-D). In addition, within individual tumours there was increased CD8 infiltration going from MSH6 proficient to deficient regions (figure 4.2E). This suggested that loss of MSH6 associated with increased immune activity within tumours.

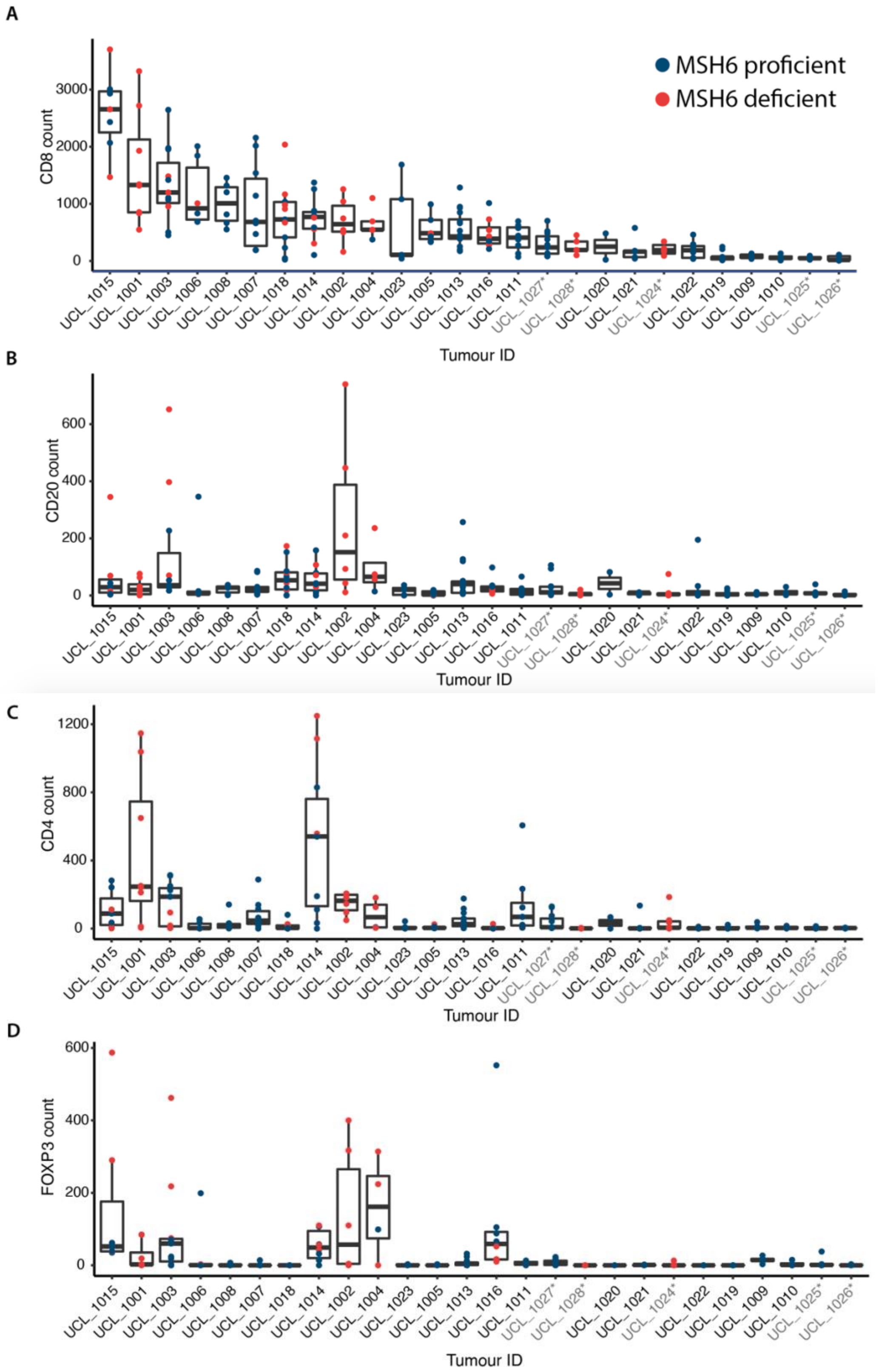


Figure 4.2: Immune infiltration counts per imaged tile across MMRd tumours in UCL cohort.

(A-D) Shows results of neighbourhood analysis quantifying levels of CD8, CD20, CD4 and FOXP3 immune cells per imaged tile respectively. MSH6 proficient and deficient tiles are labelled blue and red respectively. Tumours that were not represented in the previous genomic sequencing dataset are marked grey and with an asterisk.

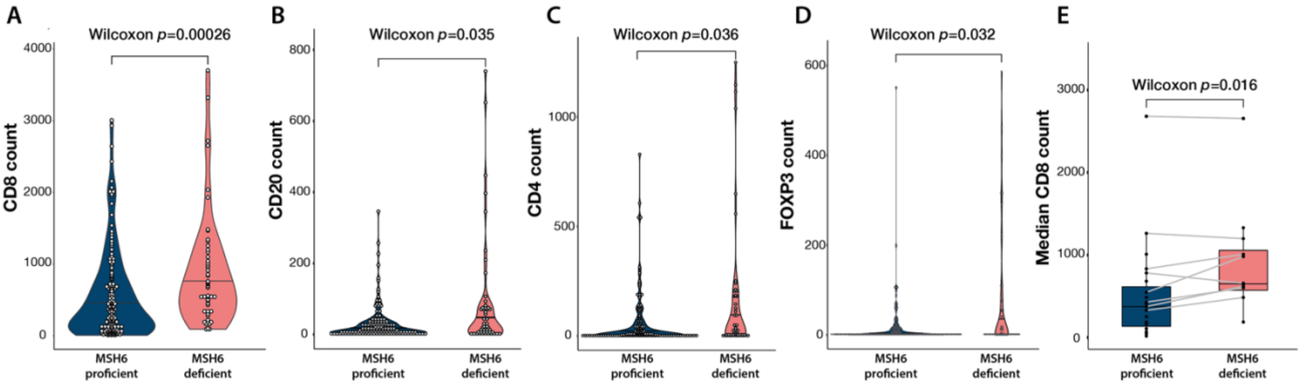


Figure 4.3: Immune infiltration is increased within MSH6 deficient subclones

A-D) Infiltration levels of CD8+, CD20+, CD4+ and FOXP3+ cells within the neighbourhood of MSH6 proficient and deficient subclones. Counts refer to the total number of each immune cell identified within a 100um radius of tumour cells within the imaged tile. E) Ladder plot showing change in median CD8 count within individual tumours going from MSH6 proficient to deficient regions.

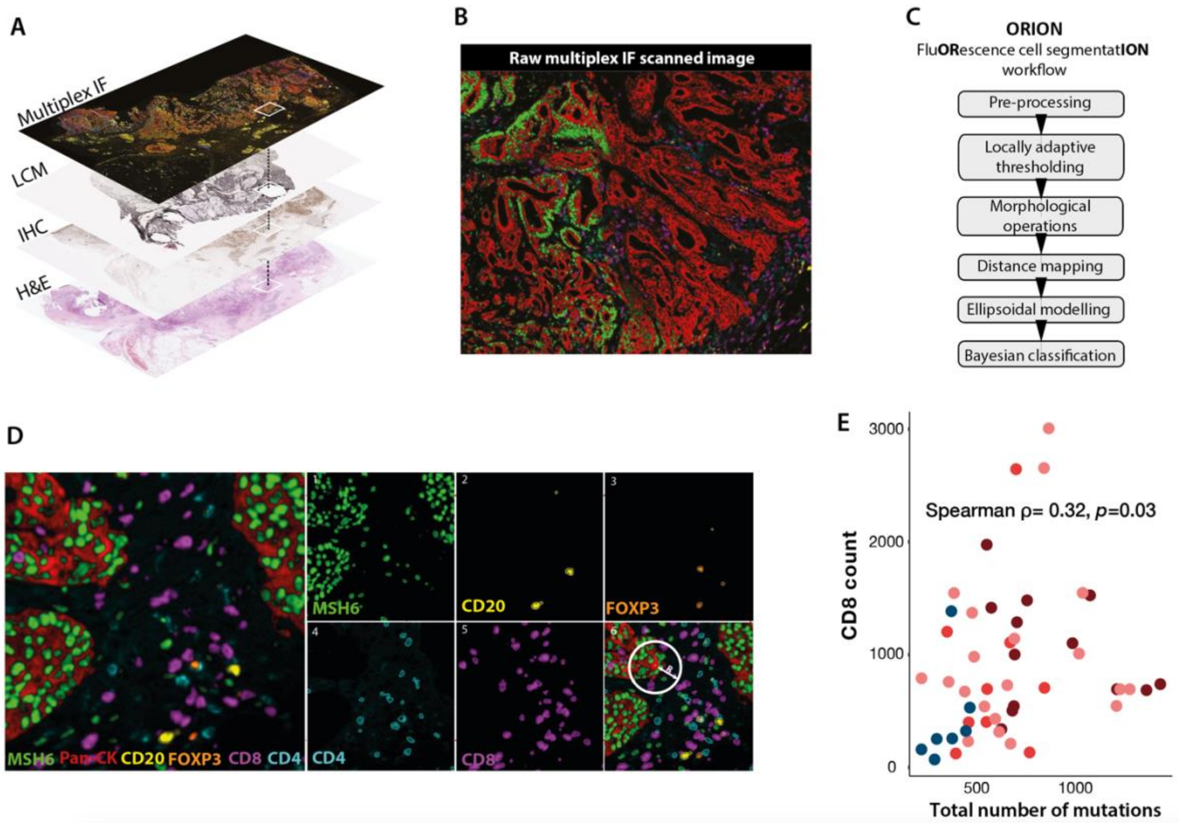


Figure 4.4: Workflow for immune infiltration analysis using multiplex immunofluorescence

A-D) Consecutive serial sections allow integration of multiplex immunofluorescence and exome sequencing data B) Raw multiplex immunofluorescence image C) ORION cell segmentation workflow D) Example high power imaging tile with constituent individual fluorescence channels displayed E) Scatter plot showing correlation between CD8^{+} T cell count and total mutation burden. Colour scheme of points as previously.

Next, correlation analysis between immune infiltration and our previous genomic data was performed. The cases with asterisks in figure 4.1, lacking genomic data were excluded from this analysis. Here the region of the tumour sequenced was matched to the immune infiltration level identified from the corresponding adjacent section. A positive correlation between mutation burden and CD8 infiltration level was observed (figure 4.4E) suggesting that immune activity is increased as a result of increased mutation burden. Clonal neoantigens are known to be especially important in anti-tumour immune response (McGranahan et al., 2016). Therefore, I explored the association between clonal and subclonal neoantigen burden versus CD8 infiltration level (figure 4.5 below). There was no significant correlation between either clonal neoantigen burden (Spearman rho=0.16, p=0.26) or subclonal neoantigen burden (Spearman rho=0.04, p=0.78) and CD8 infiltration level. This may reflect a more complex relationship between immune infiltration and neoantigen burden. The limitations of this analysis should also be noted here. Accurate prediction of neoantigens is known to be challenging. There could also be heterogeneity between the tumour section used in genomic sequencing and the adjacent section used in immunofluorescence analysis. Finally, use of single biopsies to call clonal and subclonal mutations remains a limitation of this analysis. In future work, multi-region biopsies allowing more accurate calling of clonal versus subclonal mutations could explore the relationship between clonal neoantigens and immune infiltration further.

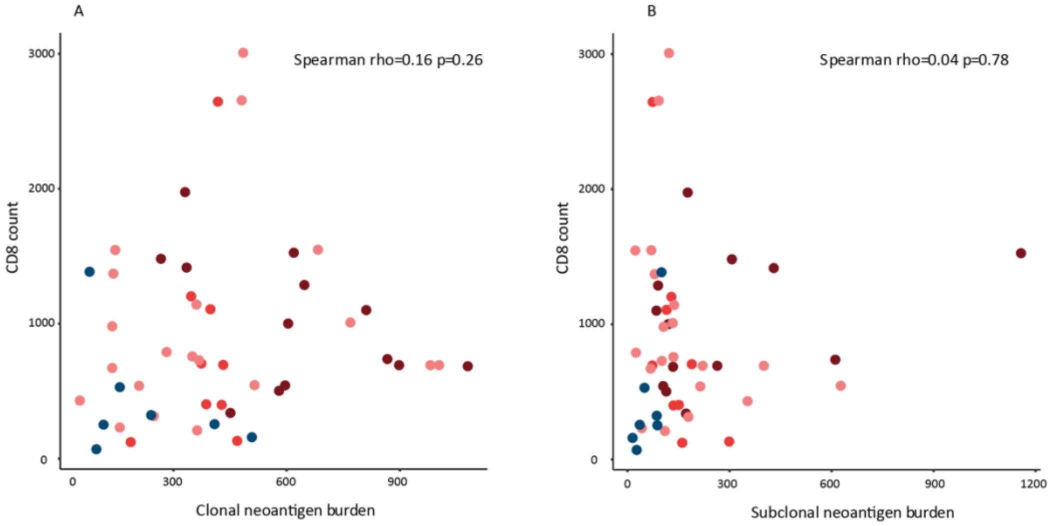


Figure 4.5: Scatter plot showing correlation between clonal and subclonal neoantigen burden versus CD8 infiltration count.

Scatter plots showing correlation between both A) Clonal and B) Subclonal neoantigen burden and CD8 infiltration counts from neighbourhood analysis.

4.6 Secondary *MSH6* and *MSH3* frameshifts associate with increased genetic diversity

The increased mutation and neoantigen burden observed with *MSH6* and *MSH3* frameshifts suggested an increase in underlying mutation rate. Increased immune visibility is therefore a fitness cost associated with this increased mutability. I reasoned that increased mutability would also lead to a general increase in genetic diversity. I decided to measure genetic diversity within samples focussing on microsatellites given they are hypermutable sites in MMRd tumours. A widely used measure of diversity in ecology is the Shannon entropy index (Roswell et al., 2021). Shannon entropy provides a robust measure of diversity because it accounts for both the number and abundance of species. I applied Shannon diversity to microsatellite homopolymers by considering each read length at a given microsatellite as a separate species as detailed in equation 1 below.

$$Shannon\ diversity = -\sum_{i=1}^{R} \ [p_i ln\ (p_i)\ ] \qquad \text{Equation 1}$$

Where p_i = the proportion of total reads represented by the ith microsatellite length
R= total number of read lengths present at a microsatellite

Using the above formula, Shannon diversity was calculated for each exonic length 8 microsatellite and averaged to obtain a Shannon diversity score for each sample. Figure 4.6 summarises this method of measuring Shannon microsatellite diversity in samples. A Shannon diversity score was calculated for each sample in the UCL whole exome sequencing dataset (n=49 samples from 22 tumours).

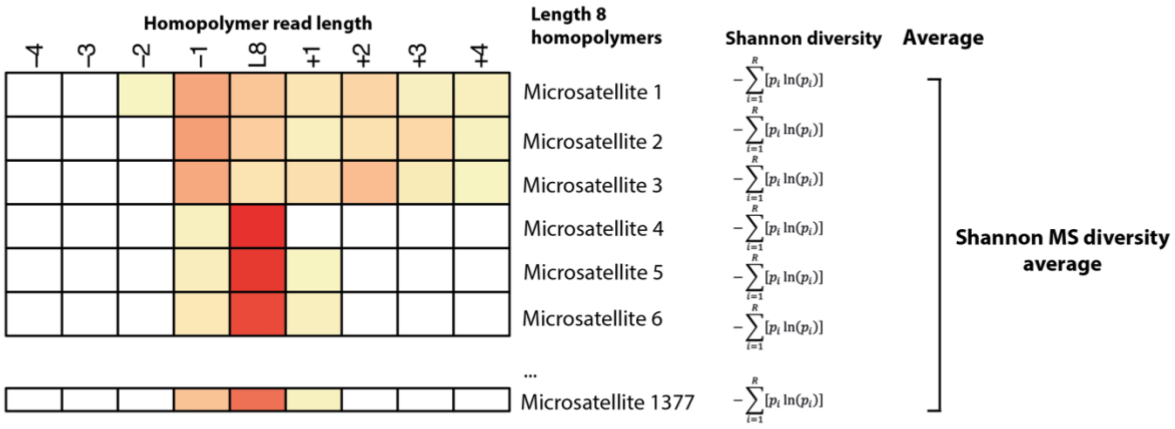


Figure 4.6: Methodology for calculating Shannon microsatellite diversity

Cartoon illustrating the calculation of the Shannon microsatellite diversity for each sample. Shannon diversity is calculated for each exonic homopolymer of a given length (length 8 in this example), based on the proportion of reads observed at each length. An average is then taken of the calculated Shannon diversity result of each microsatellite to give a sample level Shannon diversity average.

There was a clear increase in Shannon diversity in samples with *MSH6* and *MSH3* frameshifts (figure 4.7A). There was also strong positive correlation between Shannon microsatellite diversity and total mutation burden (figure 4.7B), supporting the hypothesis that secondary MMR frameshifts drive increased subclonal diversity. I then measured Shannon diversity across other homopolymer lengths in the range of 6-11 and again found a similar pattern with increased microsatellite diversity in the presence of *MSH6* and *MSH3* frameshifts (figure 4.8A-C). Shannon diversity was generally larger at longer homopolymer lengths, but the difference between *MSH6*/*MSH3* groups tended to diminish at longer homopolymer lengths. Shannon diversity of samples showed no correlation with tumour purity or median microsatellite read depth confirming these results were not a consequence of sequencing artifact (supplementary figure 3). Overall, these findings are in keeping with increased mutation rate and genetic diversification following secondary *MSH6* and *MSH3* frameshifts in MMRd tumours.

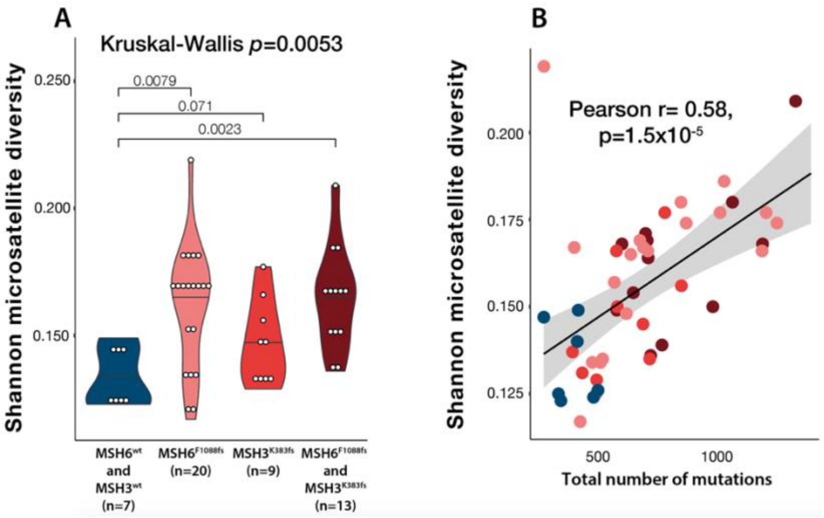
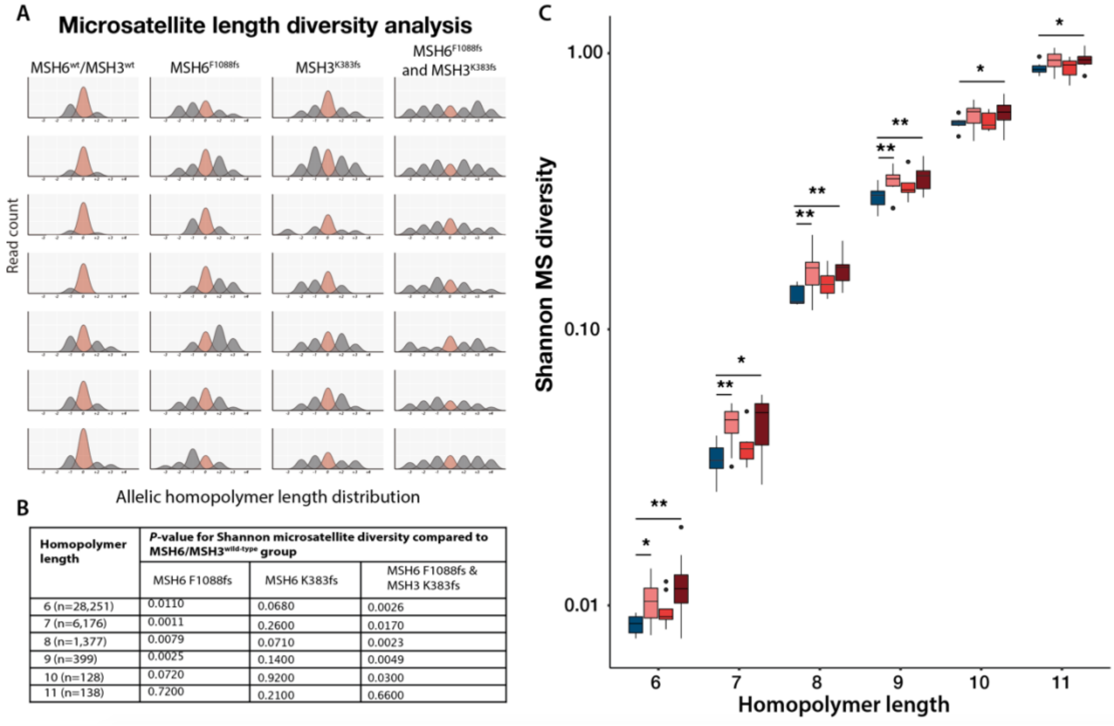


Figure 4.7: Shannon microsatellite diversity in samples from the UCL WES cohort according to MSH6 and MSH3 frameshift mutation status.

A) Shannon microsatellite diversity in samples (n=49) grouped according to *MSH6* and *MSH3* frameshift mutation status B) Total mutation burden versus Shannon microsatellite diversity. Point colours as previously. Average Shannon microsatellite diversity at length 8 homopolymers is used in both plots.



Homopolymer length	*P*-value for Shannon microsatellite diversity compared to MSH6/MSH3$^{wild\text{-}type}$ group		
	MSH6 F1088fs	MSH6 K383fs	MSH6 F1088fs & MSH3 K383fs
6 (n=28,251)	0.0110	0.0680	0.0026
7 (n=6,176)	0.0011	0.2600	0.0170
8 (n=1,377)	0.0079	0.0710	0.0023
9 (n=399)	0.0025	0.1400	0.0049
10 (n=128)	0.0720	0.9200	0.0300
11 (n=138)	0.7200	0.2100	0.6600

Figure 4.8: Shannon microsatellite diversity according to homopolymer length

A) Cartoon showing microsatellite diversity is increased in the presence of *MSH6*^{F1088fs} and *MSH3*^{K383fs}. B) Table showing the p-values for comparisons made in plot to the right. C) Plot showing that *MSH6* and *MSH3* homopolymer frameshifts result in increased MS diversity but the effect is lost at longer homopolymer lengths.

4.7 Increased genetic diversity suggests a trade-off between neoantigen burden and immune escape

Genetic diversity provides the substrate for resistance evolution to develop. I hypothesised that the increased genetic diversity observed with secondary *MSH6* and *MSH3* frameshifts, would provide increased opportunity for immune escape events, thereby mitigating the impact of increased neoantigen burden.

I assessed this by investigating for mutations in the HLA class I antigen presentation complex which has been previously reported in MMRd cancers (Grasso et al., 2018). Using the Polysolver package each patient's individual HLA haplotype was determined and used as a reference to identify non-synonymous HLA mutations. Both SNVs and indel events were identified in HLA class I genes. Synonymous SNVs were excluded in this analysis. The analysis revealed that samples with *MSH6* and/or *MSH3* frameshifts more frequently had multiple mutated HLA alleles (figure 4.9A), although this difference was not statistically significant. However, a significant positive correlation was found between Shannon microsatellite diversity of samples and the number of HLA mutations (figure 4.9B). This data supports a model where secondary MMR frameshifts increase genetic diversity providing opportunity for greater immune escape events to occur.

A potential limitation of the HLA mutation analysis is that whilst frameshift and stop-gain mutations disrupt gene function, the pathogenicity of non-synonymous SNVs cannot be assumed. In this cohort, frameshift and stop-gain mutations accounted for 38% (n=28/73) of HLA class I gene mutations, with the remainder of mutations being non-synonymous SNVs. Interpreting the pathogenicity of non-synonymous SNVs in HLA alleles is complicated due to the already polymorphic nature of the HLA alleles. Nevertheless, this data finds frequent coding mutations in genes involved in antigen presentation, reflecting a likely route to immune escape in these tumours.

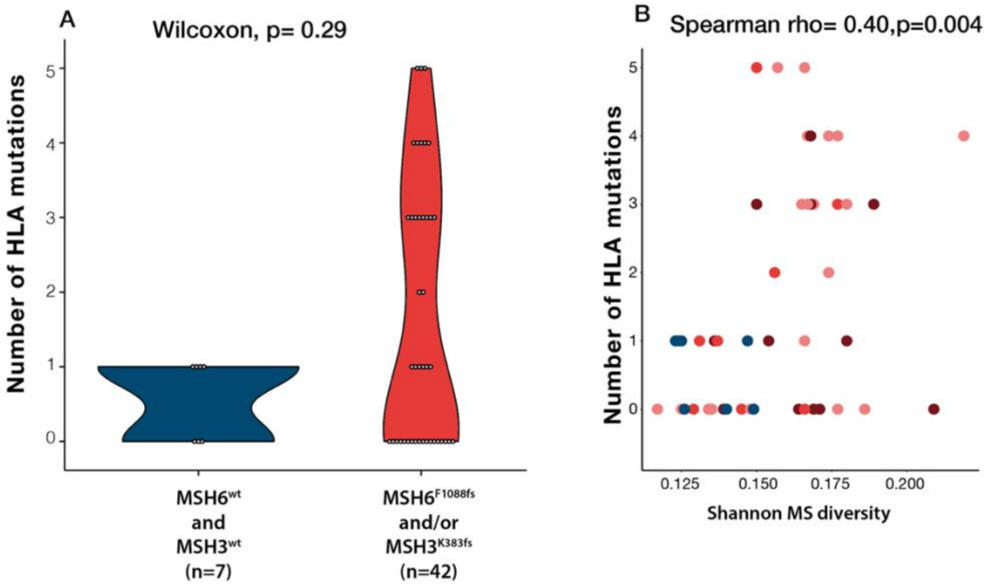
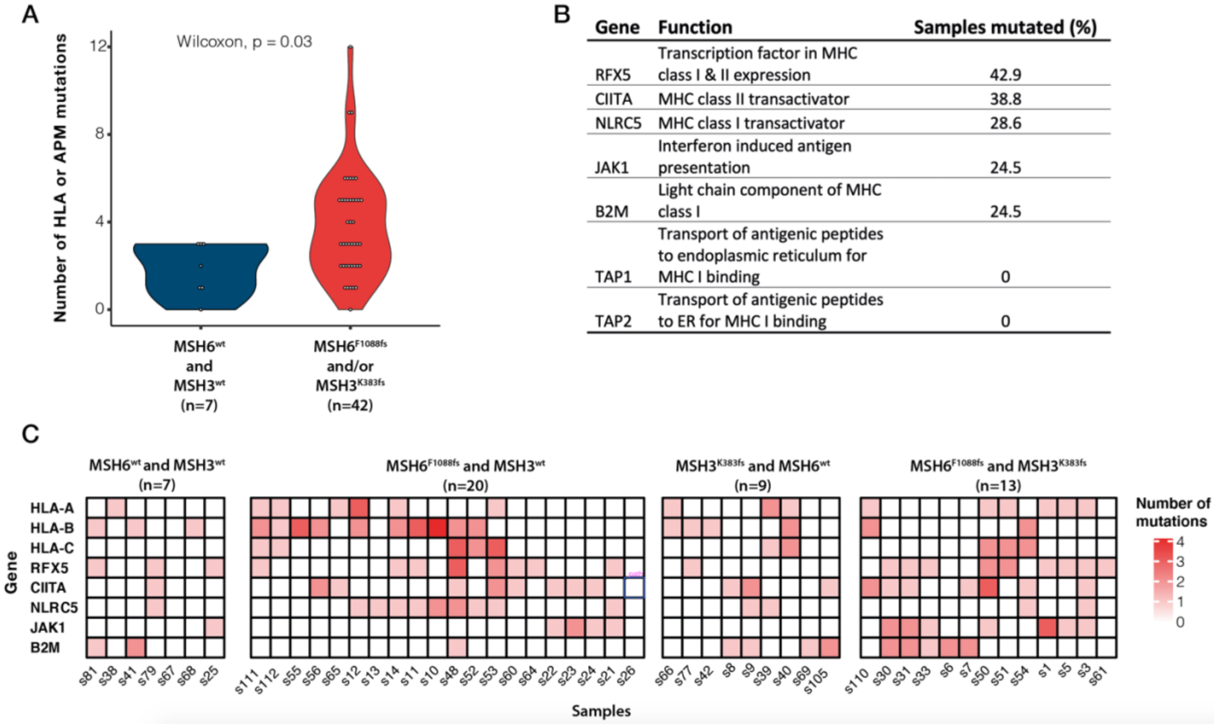


Figure 4.9: Analysis of non-synonymous mutations in HLA class I genes

A) Number of HLA class I mutations in samples from UCL WES cohort (n=49) grouped according to *MSH6* and *MSH3* frameshift mutation status B) Shannon microsatellite diversity versus number of coding HLA class I mutations. Points coloured as previously.

The process of immune escape is not a discrete event and multiple mechanisms are involved. I next investigated for mutations in the antigen processing machinery (APM) pathway. Mutations of APM genes have been previously reported in MMRd tumours (Grasso et al., 2018). As expected, in our cohort frequent mutations in these genes were also observed and are summarised in the heatmap in figure 4.10C below. Samples with *MSH6* and *MSH3* frameshifts more frequently carried multiple mutations in APM or HLA genes (figure 4.10A). Mutations in genes involved in HLA class II expression, CIITA and RFX5 were particularly prevalent in samples with both *MSH6* and *MSH3* frameshifts. Analysis of mutation bias of immune evasion variants further showed that in samples with *MSH6*/*MSH3* frameshifts, C>A mutations in a C[C>A]T and G[C>A]T context made a significant contribution whilst they were not observed in *MSH6*/*MSH3* wild-type samples (figure 4.11). This was consistent with the previous signature analysis and suggests that the broadening of mutation repertoire associated with secondary MMR frameshifts also induces immune escape events. These findings overall support a model whereby the increased genetic diversity associated with secondary MMR frameshifts leads to an evolutionary trade-off between neoantigen load and immune escape.



Gene	Function	Samples mutated (%)
RFX5	Transcription factor in MHC class I & II expression	42.9
CIITA	MHC class II transactivator	38.8
NLRC5	MHC class I transactivator	28.6
JAK1	Interferon induced antigen presentation	24.5
B2M	Light chain component of MHC class I	24.5
TAP1	Transport of antigenic peptides to endoplasmic reticulum for MHC I binding	0
TAP2	Transport of antigenic peptides to ER for MHC I binding	0

Figure 4.10: Combined analysis of immune escape mutations

Total number of coding mutations in HLA and antigen presentation machinery (APM) genes in samples grouped according to MSH6 and MSH3 frameshift mutation status. B) APM genes and percentage of samples mutated. C) Heatmap displaying per sample number of HLA and APM coding mutations.

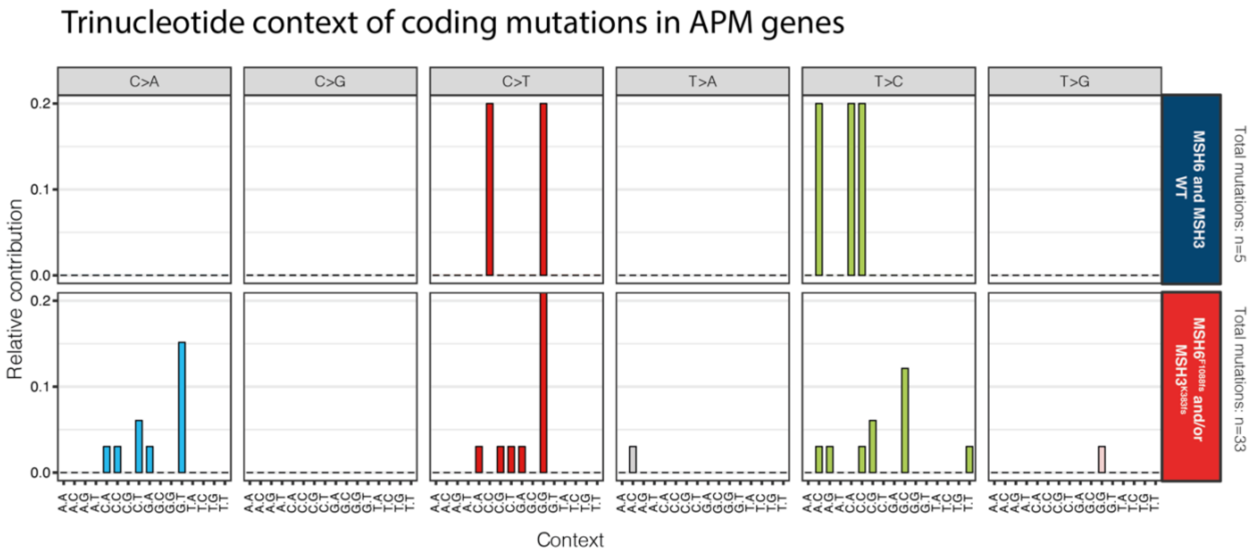


Figure 4.11: Trinucleotide context of mutations in antigen presentation machinery genes

Trinucleotide context of coding mutations in antigen presentation machinery genes in samples grouped according to presence or absence of *MSH6* or *MSH3* frameshifts.

4.8 Discussion

MMRd cancers, due to their elevated mutation burden are under strong selection pressure from the immune system which shapes their evolution. In this chapter I considered the impact of secondary *MSH6* and *MSH3* frameshifts on tumour immune interactions. I found that the increased mutability resulting from *MSH6* and *MSH3* frameshifts, associates with increased genetic diversity, creating both adaptive and deleterious mutations. Genetic diversity was observed both in terms of subclonal neoantigens and diversity within microsatellites, indicating increased population diversity for natural selection to act on. Exploration of the costs and benefits of genetic diversity identified that whilst immune evasion events were more frequent in the presence of *MSH6* and *MSH3* frameshifts, this came at the cost of increased neoantigen burden and immune infiltration. This suggests that increased mutability creates a trade-off between immunogenic neoantigens and immune evasion events. It is interesting to note that secondary *MSH6/MSH3* frameshifts associate with both an increase in immune infiltration and frequency of immune escape events. This initially seems counterintuitive but may reflect the ongoing co-evolution of tumour and immune system following increased mutability. Increased mutability leads to increased neoantigen generation and immune infiltration, but also facilitates greater opportunity for immune escape events that may undergo selection to overcome immune recognition. This may lead to an ongoing process, or 'arms race', whereby tumour and immune system are constantly adapting to overcome one another. It should also be noted that the tumours in this cohort all came from single time point surgical resection specimens. Immune infiltration and immune escape events are likely to change over time and our data may reflect the past history of tumours. In future work it will be important to study biopsies from multiple time points to better understand changes in immune selection over time.

Key findings:

- Secondary *MSH6* and *MSH3* frameshifts associate with increased genetic diversity in MMRd tumours.
- This supports a model whereby increased mutability may create a trade-off between immune escape events and immunogenic neoantigens.

- Immune escape events were observed as coding mutations in HLA class I and antigen presentation machinery genes.
- Immunogenicity was observed through neoantigens and infiltrating lymphocytes.

Chapter 5: *MSH6* and *MSH3* homopolymers modulate mutation rate through spontaneous frameshift and reversion events

5.1 Contribution statement

The sequencing data used in this chapter was obtained from wetlab experiments I performed. I constructed tumour phylogenies from multi-region sequencing data together with advice obtained from Dr William Cross. Computational modelling work was performed by Dr Eszter Lakatos with my input on model parameters. Immune peptidome dN/dS analysis was performed by Dr Luis Zapata, followed by generation of plots by myself. Mutation rate analysis using the MOBSTER tool was performed by Dr Giulio Caravagna. The work in this chapter has contributed to a paper currently under review which I wrote together with my supervisor Dr Marnix Jansen.

5.2 Introduction

Cancer growth often displays the loss of constraints of multicellularity in favour of unicellularity (Trigos et al., 2018). Parallels with bacterial evolution observed in cancer therefore reflect survival strategies favouring unicellular behaviour. Hypermutable microsatellites are frequently utilised in bacteria as a source of variation. Such microsatellites, known as 'contingency loci', are concentrated in genes involved in environmental adaptation(Moxon et al., 2006). An example of this is phase variation of surface molecule expression through expansion and contraction of microsatellite sequences allowing evasion of host immune response(Bayliss et al., 2001). A key feature of bacterial phase variation is the reversibility of microsatellite frameshift mutations allowing frequent switching of gene expression(Moxon et al., 2006).

In striking similarity, this work has identified hypermutable microsatellites in *MSH6* and *MSH3* that are unmasked by microsatellite instability and frameshifts at these homopolymers associates with increased total mutation burden. Immunohistochemical labelling combined with Sanger sequencing has revealed the spatial distribution of *MSH6* frameshifts and showed frequent nested proficient subclones within deficient regions. This nested pattern suggested that successive expansions and contractions may occur at the *MSH6* and *MSH3* homopolymers allowing these sites to act as molecular on/off switches, regulating mutation rate during tumour growth. Whilst homopolymer frameshift switching is well described in bacteria, it has not been previously reported in cancer. I hypothesised that frameshift switching at the *MSH6* and *MSH3* microsatellites, provides MMRd tumours with the ability to increase mutation rate in response to selection pressure followed by restoration DNA repair protein expression after adaptation. In this chapter I set out to obtain evidence to support our observations for *MSH6*/*MSH3* homopolymer frameshift switching over time in MMRd tumours. Furthermore, I evaluate the dynamics of hypermutability at the *MSH6/MSH3* homopolymers to understand the impact of frameshift switching on MMRd cancer evolution.

5.3 Constructing phylogenetic trees from multi-region whole exome sequencing data

To investigate microsatellite frameshift switching over time, phylogenetic trees were constructed from the multi-region WES SNV data and combined with MSH6 IHC labelling and microsatellite length distribution data. I reasoned that subclones that had undergone *MSH6* frameshift reversion would be located on the same clade as subclones with *MSH6* frameshifts, whilst unrelated subclones without a history of *MSH6* frameshift would be located on a different branch in any given tumour. Phylogenetic trees were created for tumours with multi-region sequencing data available (see methods section 7.2; phylogenetic trees). Briefly, SNV calls were converted into a binary presence/absence matrix and the Paup software package used to construct the most parsimonious tree using a previously established workflow (Cross et al., 2018). Each tree was subsequently labelled with the MSH6 IHC expression status of subclones and any identified immune evasion events. In 2 out of 10 tumours with multi-region data, there was evidence to support frameshift reversion leading to re-expression of MSH6. These are discussed in detail below.

5.4 Tumour phylogenies identify *MSH6* frameshift reversion events

<u>Tumour UCL_1002:</u>

This tumour consisted of two main branches, with one branch represented by both an MSH6 deficient and proficient patch and a second branch derived from an MSH6 deficient patch. Phylogenetic ordering revealed that the MSH6 proficient patch in this case was closely related to the MSH6 deficient patch on the same branch. Analysis of *MSH6* homopolymer read length distribution showed that the MSH6 proficient lineage carried a +3 insertion responsible for expression of the protein. In other words the

MSH6 homopolymer had undergone progressive nucleotide insertion until the reading frame had been restored. Of note this ‘reverter’ lineage also carried a B2M mutation, suggesting an immune escape event was responsible for its clonal expansion. Also, of interest in this case, all three lineages carried an *MSH3* -1 deletion resulting in loss of MSH3 labelling. MSH2 labelling was lost in the region that had lost both MSH6 and MSH3, but not in the MSH6 proficient patch. This is in keeping with the fact that MSH3 and MSH6 are alternative binding partners for MSH2 in the MutS heterodimer, but loss of both proteins leads to loss of MSH2. Overall, this case provides a clear example MSH6 reversion through a +3 insertion. This is likely an unusual event, as in most cases successive -1/+1 frameshifts would leave no record of the reversion event. It is accepted that protein function in the case of a +3 insertion may not be entirely normal due to the gain of an additional amino acid but it does nevertheless restore reading frame.

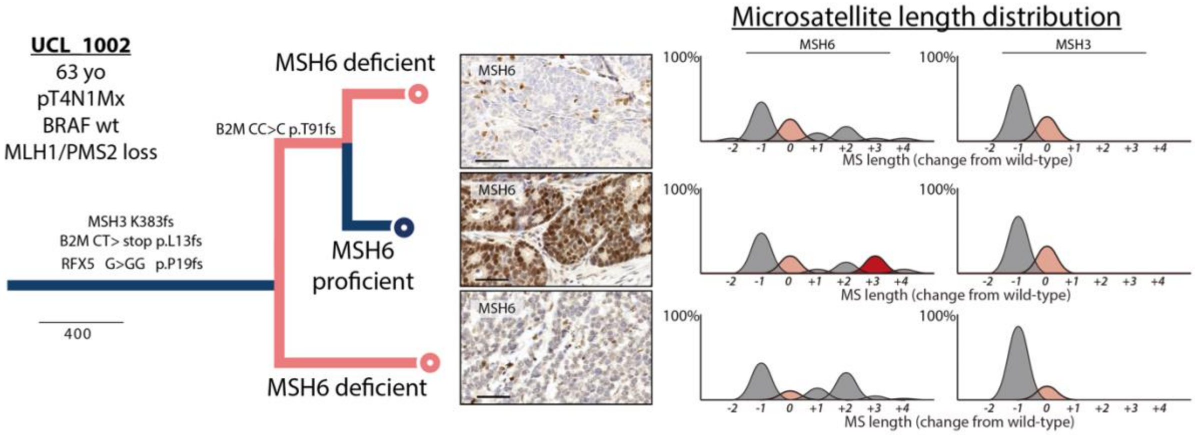


Figure 5.1: Phylogenetic tree for tumour UCL_1002

Branches are coloured according to MSH6 IHC labelling. Blue indicating MSH6 proficient and pink indicating MSH6 deficient lineages. High power images show MSH6 immunohistochemistry. Plots show allelic frequency of microsatellite length distribution for *MSH6* and *MSH3*. Peaks are coloured beige for wild-type C8 or A8 homopolymer length, grey for expanded or contracted alleles and red for +3 frameshift. The tree is labelled with immune escape events.

Tumour UCL_1018

This tumour included an MSH6 deficient clade derived from three MSH6 deficient patches and a terminal branch that was a MSH6 proficient patch. This is an example where an *MSH6* reversion event has taken place within an MSH6 deficient background. The *MSH6* homopolymer read length distribution in the MSH6 deficient lineages all carry a -1 deletion, whilst the MSH6 proficient region does not.

Interestingly the MSH6 proficient ‘reverter’ lineage in this case carried a non-synonymous C>T mutation in the HLA class II regulatory protein RFX5, possibly explaining its clonal expansion. This case also carried a separate MSH6 proficient branch and separately another MSH6 deficient branch derived from a lymph node metastasis. The lymph node metastasis had a shorter branch length, possibly reflecting increased immune predation in its microenvironment.

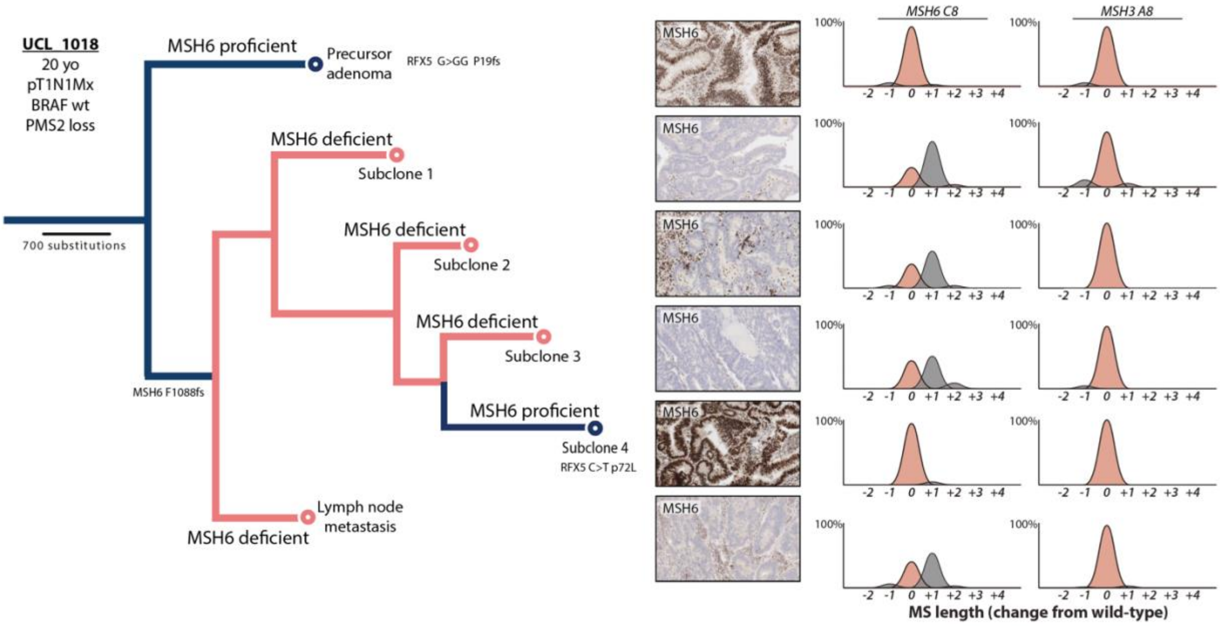


Figure 5.2: Phylogenetic tree for tumour UCL_1018.

High power images show MSH6 immunohistochemistry. Tree and plots labelled as before.

Phylogenetic trees for the remaining cases are shown in figure 5.3 and 5.4 below. These cases did not show clear evidence of *MSH6* or *MSH3* frameshift reversion, however there was extensive subclonal diversification as previously reported in MMRd tumours(von Loga et al., 2020). In one case (UCL_1004, figure 5.3 below), one MSH6 deficient branch showed a small population of reads containing +3 insertion event, suggesting a small nested MSH6 positive subclone could be present within this sample. This branch showed 5 separate HLA mutations suggesting extensive clonal diversification.

The tumour phylogenies presented here provide evidence for frameshift reversion events at the *MSH6* homopolymer, supporting the hypothesis that successive expansions and contractions at the homopolymer allow it to act as a molecular on/off switch. A limitation of the tumour phylogenies presented here is that they are based on binary presence/absence matrices of SNVs. An alternative approach used in the field is to cluster mutations according to cancer cell fraction (CCF) values and infer phylogenies using the pigeon-hole principle (Dentro et al., 2017). Both CCF approaches and the binary matrix approach to phylogeny construction have been compared in the literature(Miura et al., 2020). In general, CCF based approaches are considered to produce more accurate clonal ordering and tree topology. However, CCF clustering relies on accurate assessment of purity, ploidy and absolute copy number which can be problematic in FFPE derived sequencing data due to the inherent noise present in the data. Given that here the binary matrix approach was used to infer phylogenies this could potentially impact the accuracy of clonal ordering and identification of *MSH6* reversion events. In future, multi-region sequencing data from fresh frozen material could allow more accurate CCF based phylogeny construction. In addition, the availability of longitudinal samples from different time points in the same patient would be important in understanding changes in *MSH6/MSH3* mutated clone size over time.

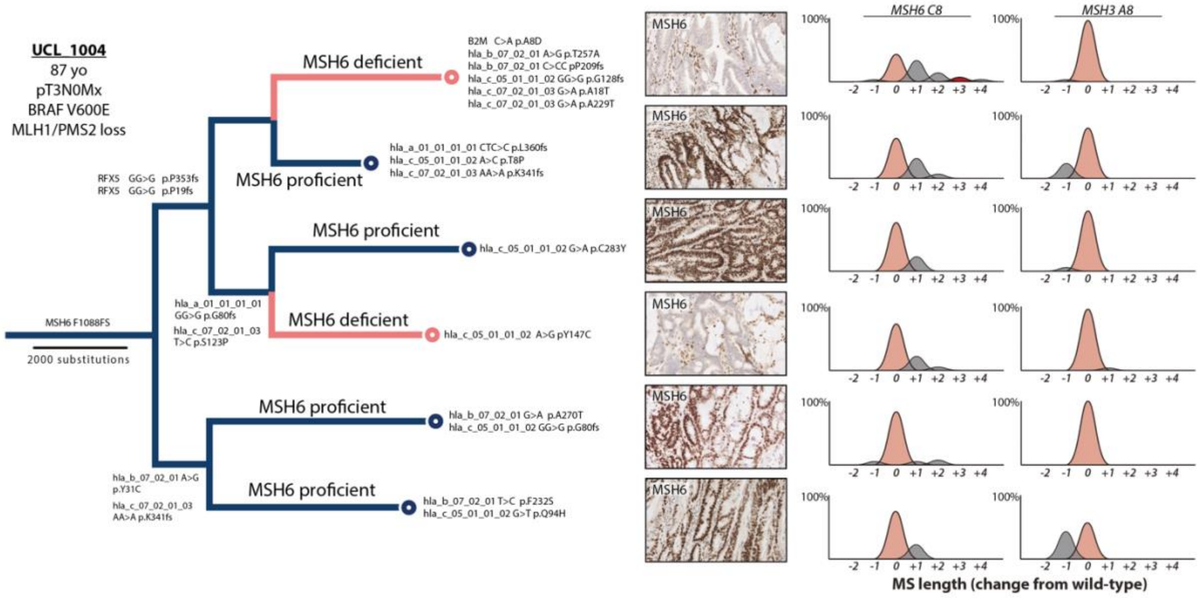

UCL_1004
87 yo
pT3N0Mx
BRAF V600E
MLH1/PMS2 loss
MSH6 F1088FS
2000 substitutions
RFX5 GG>G p.P353fs
RFX5 GG>G p.P19fs
hla_a_01_01_01_01 GG>G p.G80fs
hla_c_07_02_01_03 T>C p.S123P
hla_b_07_02_01 A>G p.Y31C
hla_c_07_02_01_03 AA>A p.K341fs
MSH6 deficient
B2M C>A p.A8D
hla_b_07_02_01 A>G p.T257A
hla_b_07_02_01 C>CC pP209fs
hla_c_05_01_01_02 GG>G p.G128fs
hla_c_07_02_01_03 G>A p.A18T
hla_c_07_02_01_03 G>A p.A229T
MSH6 proficient
hla_a_01_01_01_01 CTC>C p.L360fs
hla_c_05_01_01_02 A>C p.T8P
hla_c_07_02_01_03 AA>A p.K341fs
MSH6 proficient
hla_c_05_01_01_02 G>A p.C283Y
MSH6 deficient
hla_c_05_01_01_02 A>G pY147C
MSH6 proficient
hla_b_07_02_01 G>A p.A270T
hla_c_05_01_01_02 GG>G p.G80fs
MSH6 proficient
hla_b_07_02_01 T>C p.F232S
hla_c_05_01_01_02 G>T p.Q94H
MSH6
MSH6 C8
MSH3 A8
100%
-2 -1 0 +1 +2 +3 +4
MS length (change from wild-type)

Figure 5.3: Phylogenetic tree for tumour UCL_1004.

High power images show MSH6 immunohistochemistry. Tree and plots labelled as before.

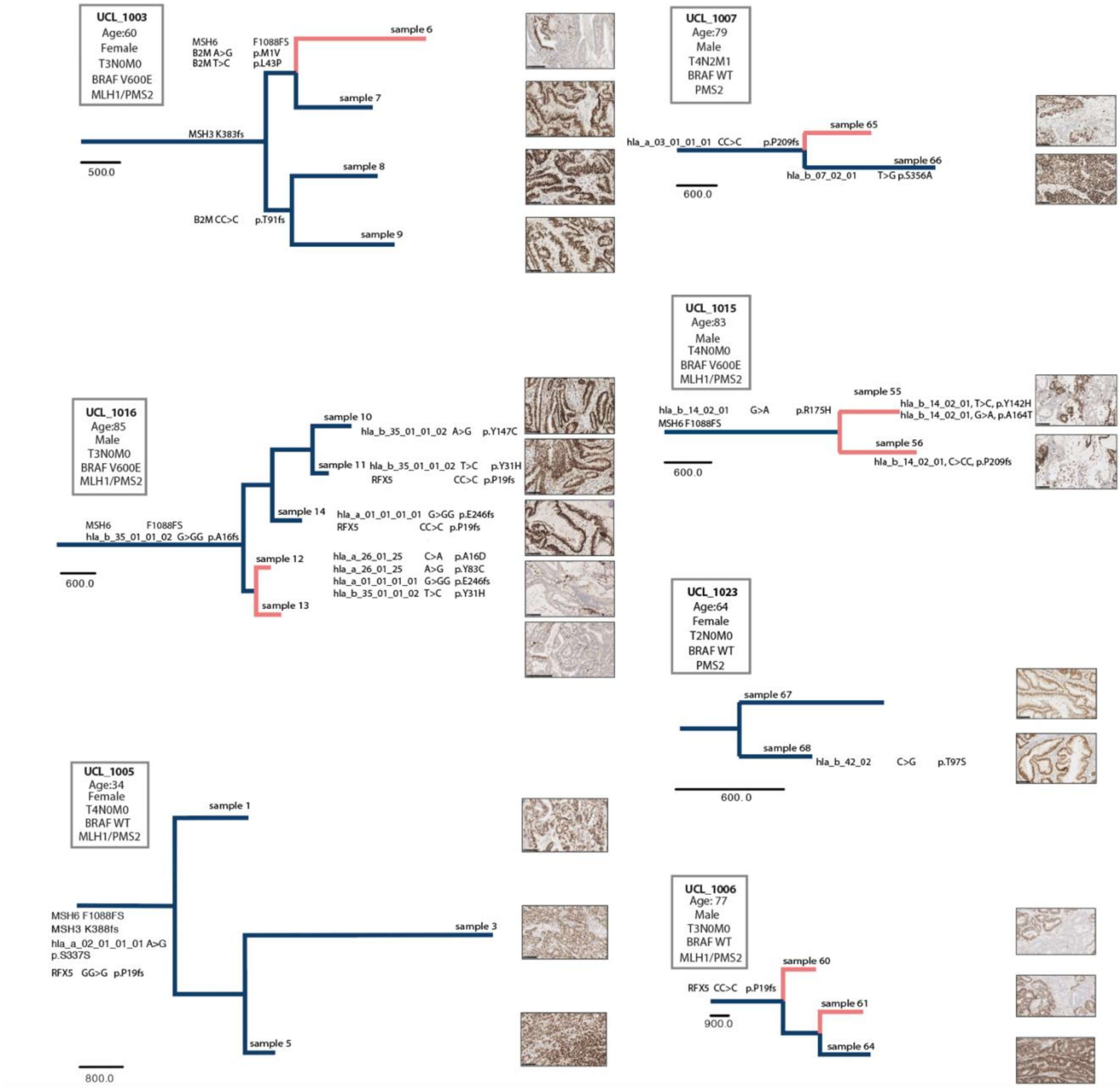


Figure 5.4: Phylogenetic tree for tumours without evidence of MSH6 homopolymer frameshift reversion.

Phylogenies displayed for 7 tumours. High power images show MSH6 immunohistochemistry. Tree and plots labelled as before.

5.5 Computational modelling shows mutation rate switching accelerates immune escape

The work so far provides evidence that frameshift mutations and reversions at the *MSH6* and *MSH3* homopolymers allow these sites to act as molecular switches, controlling mutability during tumour growth. The dynamics of frameshift mutation and reversion at these loci is likely to regulate immune selection and adaptation in these tumours. To investigate the dynamics of mutation rate switching, we next developed a computational simulation to model mutation rate switching during tumour growth. The model extends previous work by Lakatos et al., 2020, where early tumour growth from 100 to 100,000 cells is followed using a stochastic birth-death process. During each cell division, tumours cells can either die or proliferate according to their fitness value, and also gain new mutations which may affect their fitness value (see figure 4A). Fitness of tumour cells depends on both the burden of accumulated neoantigens and the prevailing immune selection pressure (*S)*. At high immune selection strength neoantigenic mutations result in an increased likelihood of cell death. Tumour cells can also gain immune independent lethal mutations represented in the model by P_{lethal}. Mutation rate μ can be either at the baseline MMRd hypermutated state (average 6 mutations per cell division) or ultra-hypermutated state (average 120 mutations per cell division) representing gain of secondary MMR frameshifts. The probability of switching between the two mutation rate states is determined by the switch rate *ß*, where *ß*=0 indicates lack of switching and *ß*= 0.01 indicates frequent switching to or from the ultra-hypermutated state. The parameters used for the model were similar to those used in previous published work (Lakatos et al., 2020) but with some important adaptations. The baseline mutation rate (*µ)* of 6 mutations per cell division was based on previous work showing that this would approximately correspond to the mutation burden in whole exome sequenced MMRd tumours (See supplementary note, Lakatos et al., 2020). Since our hypothesis is *MSH6/MSH3* frameshifts create a transient increase in the mutation rate later during tumour evolution (after initial loss of MMR), the mutation rate for this ultra-hypermutated state was set 20-fold higher at 120 mutations per cell division. We modelled mutation rate switching (β) over a range of values, to explore the impact of differing switch rates. For the purposes of the model we wanted to test a realistic range of switch rate values, since at a population level, switch rate will also depend on the number of dividing cells present in the tumour. Details of model parameters are provided in table 5.1 below.

Model parameter	
Mutation rate switch probability (per cell division, β)	0, 0.0005, 0.001, 0.002, 0.005, 0.01, 0.02
Mutation rate (mutations per cell division, μ)	6, 120
Immune selection strength	-0.1, -0.8, -2
Mutation probability (per cell division):	
Lethal mutation	5×10^{-4}
Immune escape mutation	1×10^{-6}
Neoantigen mutation	0.1

Table 5.1: Parameters used in mutation rate switch simulation model

The parameter values are listed for mutation rate switch probability, mutation rate and immune selection strength (s), The probabilities of each type of mutation per cell division (lethal mutation, immune escape and neoantigen mutation are listed.

To test the model, we first investigated how mutation rate switching influences population diversity. We studied the change in read length of a single microsatellite during tumour growth and calculated the Shannon diversity at this locus for each surviving detectable tumour. This showed that a higher mutation rate was associated with increased microsatellite diversity (figure 5.5B). This was in keeping with our previous data from patient tumours (see Chapter 4 figure 4.7A) and provided validation that the model was operating as expected. Mutation rate switching also increased or decreased microsatellite diversity depending on the mutation rate of the initial cell population, in keeping with expectations.

We then investigated the effect of mutation rate switching on tumour growth. Since ultrahypermutated cells have a higher chance of gaining an immune escape mutation we predicted switching to a higher mutation rate would decrease the time to immune escape and the rate of lineage elimination. To assess this, we followed the number of lineages that were eliminated until 10 lineages survived using different switch rates (figure 5.6). At higher mutation rate switching there was a reduced number of lineages eliminated consistent with a growth advantage. However, mutation rate switching had no effect at low levels of immune selection, as in this scenario lineages could survive without developing immune escape.

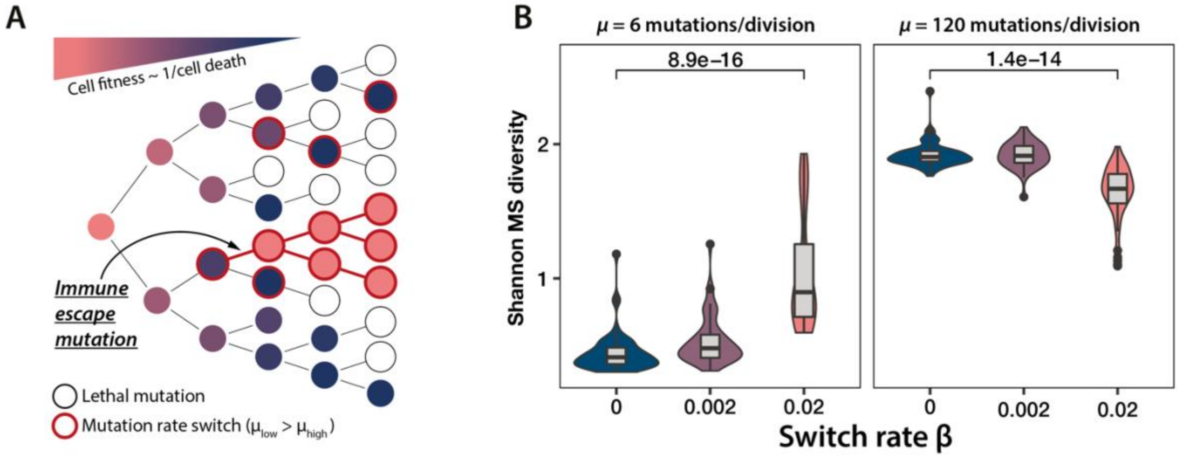


Figure 5.5: Mathematical model of mutation rate switching and impact on microsatellite diversity

A) Cartoon of tumour growth model incorporating mutation rate switching. Shade of circles represents each cells fitness value and outline colour the mutation rate. B) Shannon microsatellite diversity at a single homopolymer locus at different mutation rate switching frequency, both at baseline hypermutated starting mutation rate (left plot) and at ultrahypermutated starting mutation rate (right plot).

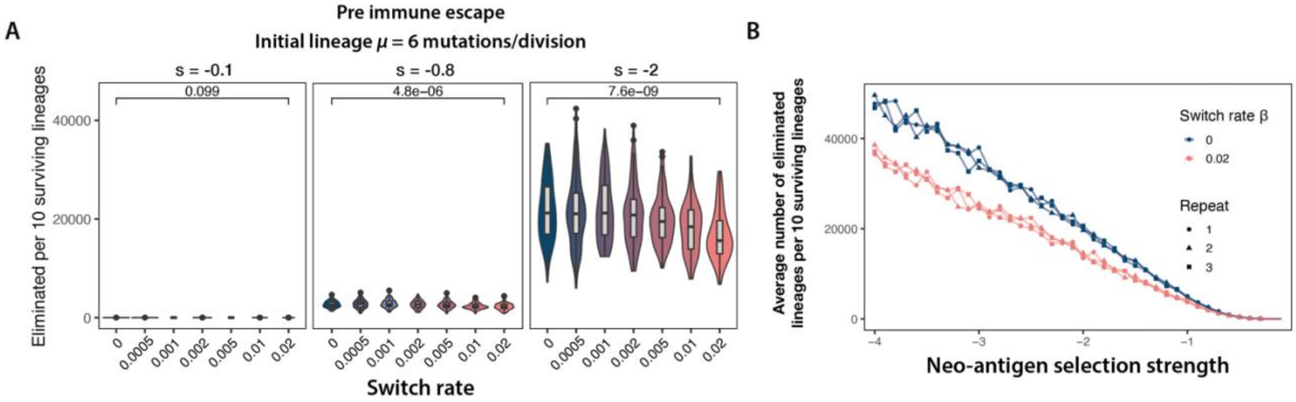


Figure 5.6: Modelling impact of mutation rate switching on tumour growth pre-immune escape

A) Number of eliminated lineages per 10 surviving at varying mutation switch rates and immune selection strength (s). B) Number of eliminated lineages per 10 surviving lineages versus varying immune selection strength. Results plot scenario without mutation rate switching (blue line) and with frequent switching (pink line).

After immune escape tumours can grow unimpeded by neoantigen accumulation. However other deleterious (P_{lethal}) mutations independent of the immune system may still reduce cellular fitness. We hypothesised that since ultrahypermutated tumours accumulate such deleterious mutations more rapidly, that switching back down to a lower mutation rate would provide a growth advantage in this scenario. We explored growth time between complete immune escape and the tumour reaching a clinically detectable size. At higher mutation rate switching there was a shorter growth time, except when immune selection strength was low (figure 5.6 and 5.7).

Overall, the simulation results provide confirmation that mutation rate switching provides a growth advantage in hypermutated cancers. Before immune escape, switching to a higher mutation rate accelerates the time to immune escape. Following immune escape, ongoing deleterious mutations reduce fitness providing selection pressure to switch back down to a lower mutation rate. Whilst for the purposes of the model immune escape is simplified as a binary process, in an evolving tumour there could be multiple cycles of switching to and from an increased mutation rate, adapting mutability to match immune selection pressure.

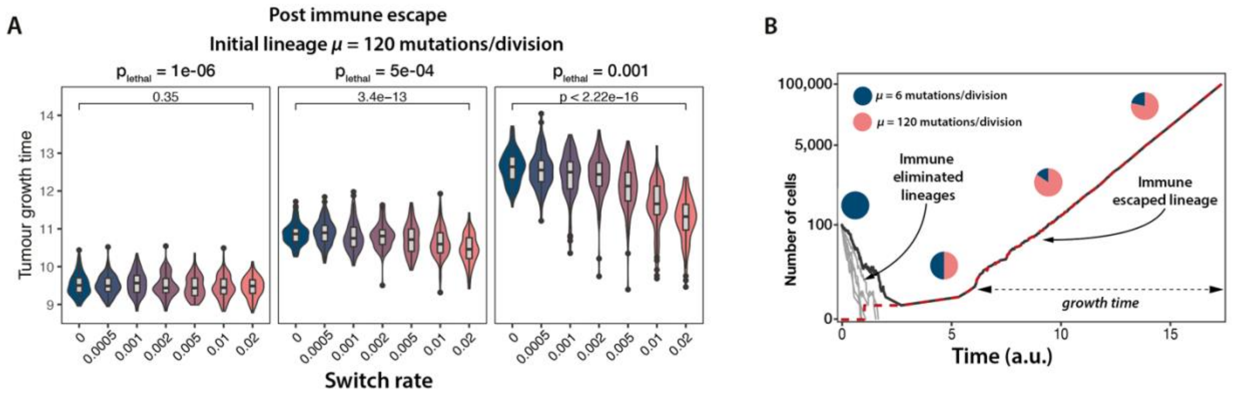


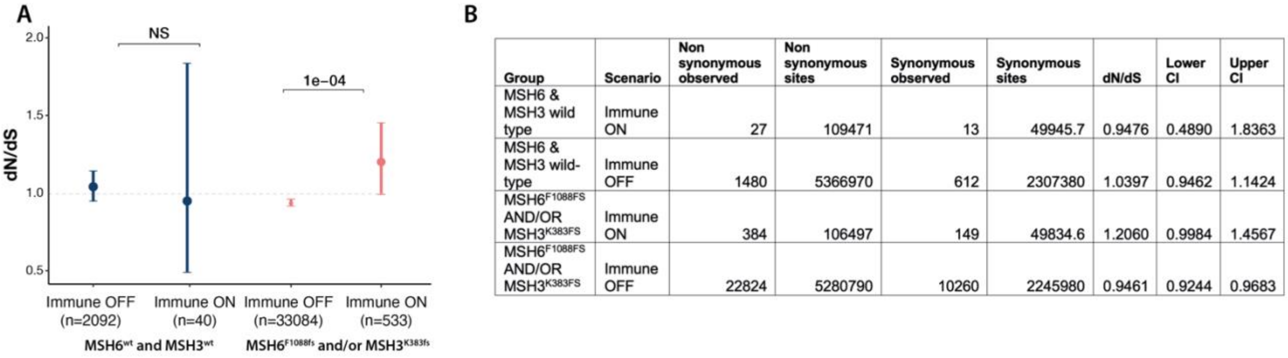
Figure 5.7: Modelling impact of mutation rate switching on tumour growth post-immune escape

A) Impact of varying frequency of mutation rate switching and lethal (P_{lethal}) mutations on tumour growth time B) Dynamics of mutation rate switching during growth of 6 simulated tumours. Grey lines indicate eliminated lineages. One lineage survives shown in black with the number of immune escaped cells shown overlapped in red. Pie charts display the proportion of tumour cells at baseline and increased mutation rate.

5.6 Immune dN/dS analysis

This work has uncovered a mechanism allowing MMRd tumours to fluctuate mutation rate using homopolymers contained within *MSH6* and *MSH3* as molecular switches. *MSH6* and *MSH3* frameshifts may not be directly selected but instead may hitchhike to prominence with the adaptive variants that they generate. This second order selection is well described in bacteria(Tenaillon et al., 2001). For instance, experimental data has shown that MMR deficient E. coli which had adapted to growth in the mouse intestine, outcompeted an intestinal non-mutator population even after restoration of the MMR defect, indicating selection of mutator alleles is indirect and dispensable following adaptation (Giraud et al., 2001).

To determine whether indirect selection of *MSH6* and *MSH3* frameshifts also occurs in MMRd tumours, immune dN/dS analysis was performed (see methods section 7.3). Immune dN/dS measures the ratio of nonsynonymous to synonymous mutations at genomic loci that are predicted to be exposed to the immune system (Immune ON)(Zapata et al., 2020). In this analysis we calculated dN/dS across genomic regions that bind to HLA-A0201, the most common HLA class I allele in the Caucasian population allowing comparison between patients in our UCL WES cohort. Immune ON dN/dS values were compared to dN/dS values outside regions exposed to the immune system (immune OFF) as a control. The results show that samples in the group with *MSH6* and/or *MSH3* frameshifts have a significantly increased immune ON dN/dS compared to immune OFF dN/dS (figure 7). Meanwhile in *MSH6/MSH3* wild-type samples there was no significant difference in immune ON dN/dS values compared to immune OFF dN/dS values. These results show that in the presence *MSH6* and *MSH3* frameshifts, there is enrichment of non-synonymous mutations in genomic regions exposed to the immune system. This is in keeping with previous work by Zapata et al showing that increased immune dN/dS values are associated with higher rates of immune escape (Zapata et al., 2020). This supports our understanding that *MSH6* and *MSH3* frameshifts are indirectly selected through linkage with immune escape variants.



Group	Scenario	Non synonymous observed	Non synonymous sites	Synonymous observed	Synonymous sites	dN/dS	Lower CI	Upper CI
MSH6 & MSH3 wild type	Immune ON	27	109471	13	49945.7	0.9476	0.4890	1.8363
MSH6 & MSH3 wild-type	Immune OFF	1480	5366970	612	2307380	1.0397	0.9462	1.1424
MSH6^{F1088FS} AND/OR MSH3^{K383FS}	Immune ON	384	106497	149	49834.6	1.2060	0.9984	1.4567
MSH6^{F1088FS} AND/OR MSH3^{K383FS}	Immune OFF	22824	5280790	10260	2245980	0.9461	0.9244	0.9683

Figure 2 Immune dN/dS analysis in samples according to MSH6 and MSH3 frameshift mutation status.

A) dN/dS values for immune ON and immune OFF regions in samples according to MSH6^{F1088fs} and MSH3^{K383fs} grouping. Total number of mutations under each category are reported in brackets. B) Table showing data used to calculate dN/dS scores for immune ON and OFF regions of exome.

5.7 Estimating mutation rates from the tail of neutral mutations

Measuring in-vivo mutation rates in tumours from single time point biopsies is challenging. In any given sample, the tumour age and number of divisions since the most recent common ancestor are not readily available. A recently published computational method for subclonal reconstruction proposes a method to estimate mutation rates from the tail of neutral mutations within the allele frequency of a given sample (Caravagna et al., 2020). The approach, named MOBSTER (MOdel Based cluSTering in cancER), combines population genetics approaches with machine learning to perform more accurate tumour subclonal reconstruction. We applied MOBSTER's population genetics approach to estimate sample-specific mutation rates μ from the fit of neutral tails. Analyses were performed on our exome sequencing dataset, normalised to regions confirmed as diploid (see Methods section 7.4). A limitation of this analysis was that often samples lacked good quality subclonal tails, likely due to poor coverage of low frequency variants. Following Qc checks, we found that samples for patient UCL-1014 (figure 5.9C) showed an order of magnitude increase in mutation rate going from the MSH6 proficient to deficient region of this tumour. In this tumour μ=2.11^{-7} for the MSH6-proficient sample and μ=1.51^{-6} for the MSH6-deficient sample. These results further support our data that mutation rate increases following secondary frameshifts in *MSH6* and *MSH3*. Future work using whole genome sequencing data at high sequencing depth will allow more robust mutation rate analysis from multi-region data.

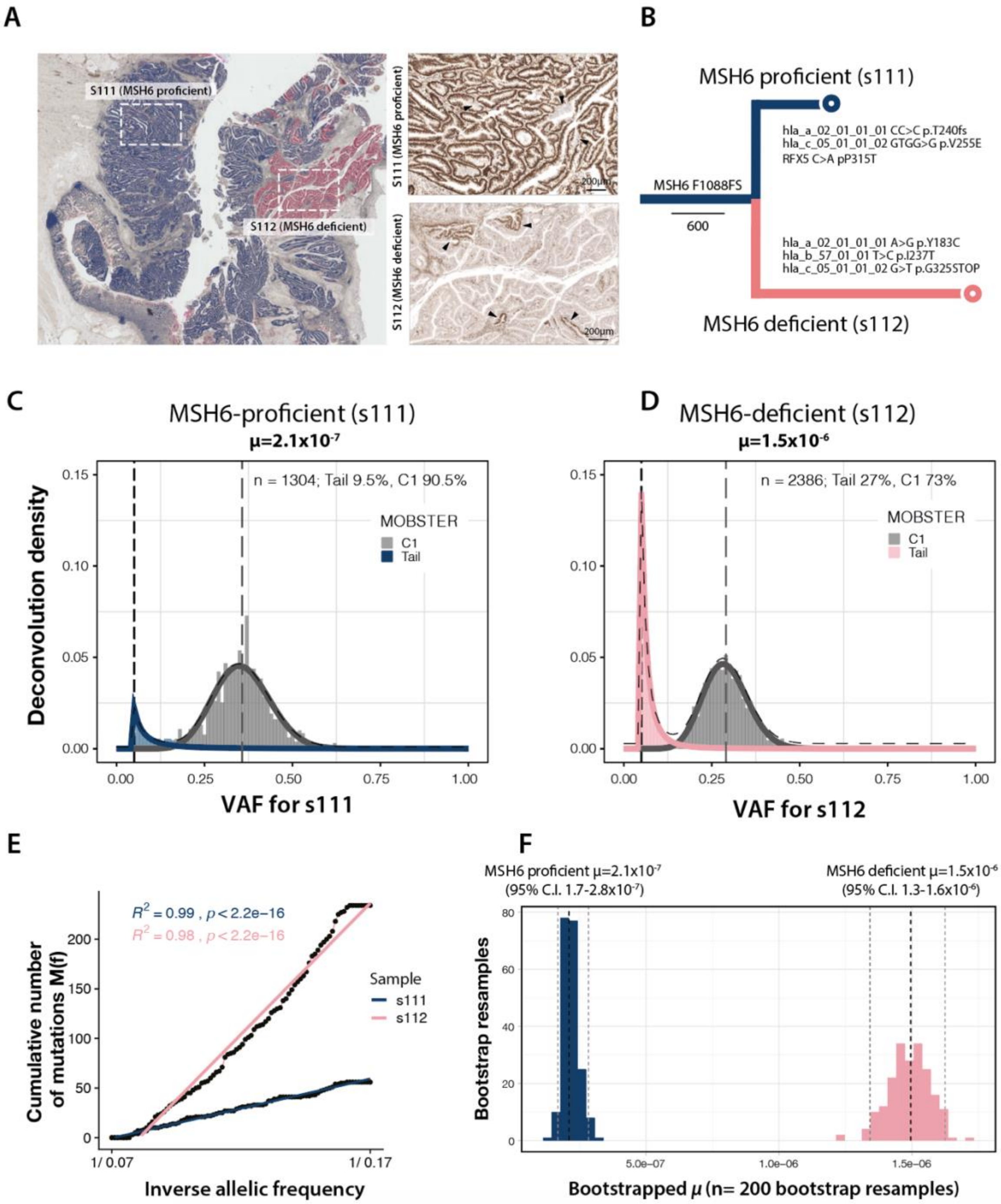


Figure 5.9: Results of MOBSTER mutation rate analysis in samples for tumour UCL_1014

A) MSH6 IHC labelling with overview and high-power images displayed. Arrowheads show small reverter subclones. B) Phylogenetic tree for tumour UCL_1014. C-D) Subclonal reconstruction of tumour UCL_1014 for sample s111 and s112 respectively. E) Cumulative frequency distribution of neutral tail mutations in s111 and s112. The point estimate of the mutation rate μ is obtained from the gradient of the cumulative frequency distribution. F) Details of bootstrapped percentile confidence interval for the mutation rate point estimate in panel E.

5.8 Discussion

The work presented in this chapter has helped characterise the dynamics of secondary *MSH6* and *MSH3* frameshifts during MMRd tumour evolution. In the previous chapters we found that *MSH6* and *MSH3* frameshifts increase mutability and broaden mutation repertoire. We also recognised that the resulting increased genetic diversity creates a trade-off between beneficial immune escape variants and deleterious neoantigens. In this chapter I show that MMRd tumours exploit the short-term adaptive benefits of increased mutability whilst also limiting the long-term costs by utilising reversion mutations to restore reading frame and expression of MSH6 and MSH3 following adaptation. This results in a novel mechanism of gene regulation, whereby the *MSH6*/*MSH3* homopolymer tracts act as a stochastic on-off switch, akin to the action of contingency loci in bacterial evolution(Moxon et al., 2006). Computational modelling work further confirmed that mutation rate switching provides a growth advantage, particularly where immune selection strength is high, as is expected to be the case in tumours with background MLH1/PMS2 loss. This results in a model whereby MMRd tumours undergo repeated cycles of mutation rate switching during their evolution, accelerating immune escape in response to changing immune selection pressure, followed by decreased mutation rate once adaptation is achieved. *MSH6* and *MSH3* frameshifts increase the evolvability of MMRd tumours and are particularly useful where multiple beneficial mutations are required for successful adaptation. *MSH6* and *MSH3* frameshifts are therefore not directly selected but are instead indirectly selected through the adaptive variants that they generate.

Key findings:

- Tumour phylogenetic trees reveal evidence for frameshift reversion mutations at the *MSH6* and *MSH3* homopolymers
- Computational modelling confirms mutation rate switching provides a growth advantage by accelerating immune escape followed by return to baseline mutation rate after adaptation.
- Immune dN/dS analysis shows enrichment of non-synonymous mutations in exonic regions exposed to the immune system in samples with *MSH6* and *MSH3* frameshifts reflecting indirect selection through hitchhiking with immune escape variants.

- Mutation rate estimation using the MOBSTER clonal deconvolution tool shows evidence of intra-tumour increase in mutation rate going from a MSH6 proficient to deficient subclone.

Chapter 6. Final discussion and conclusion

This work highlights the importance of considering the evolutionary basis for molecular findings in cancer. Here I started with a simple question; are there secondary regulators of mutation rate in MMRd cancers, that allow these tumours to manage the long-term genotoxic costs associated with hypermutation. Based on the knowledge that MMRd tumours progress through frameshift mutations at microsatellite sequences, I was able to show that frameshifts in *MSH6* and *MSH3* associated with a significant increase in mutation burden, whilst frameshifts in other common MSI targets such as *TGFBR2* and *BAX* showed no significant correlation. This unexpected finding suggested that *MSH6* and *MSH3* frameshifts were causal to increased mutation burden rather than simply the consequence of mutation rate variation between tumours. Visualising MSH6 expression using immunohistochemistry allowed me to observe that MSH6 deficient regions frequently contain nested positive clones ranging from isolated cells to large patches, suggesting frequent on/off switching of gene expression. Sanger sequencing confirmed that the expression of MSH6 was controlled by successive expansions and contractions at the coding homopolymer contained within the gene and tumour phylogenies later confirmed cases of reversion mutations in MSH6 proficient subclones within negative regions. Using laser capture microdissection to enrich for a clonally pure population of MSH6 deficient cells, I found a significant increase in mutation burden together with a shift in mutation bias, further confirming that secondary MMR frameshifts leave their footprint on the genome. Computational modelling confirmed that mutation rate switching provides a growth advantage when immune selection strength is high as is the case in tumours with mismatch repair deficiency.

Overall, this work has identified an elegant mechanism regulating the growth dynamics of MMRd cancers. Microsatellite instability unmasks reversible secondary frameshifts in *MSH6* and *MSH3*, which in turn further increase mutation rate and bias. Clones with *MSH6* and *MSH3* disruption obtain an indirect growth advantage as they acquire the necessary combination of immune escape mutations more quickly. Following immune adaptation, the ongoing increased mutation rate becomes a disadvantage due to the accumulation of deleterious mutations. This drives the selection of reversion mutations in *MSH6* and *MSH3* to restore mutation rate to the baseline MMRd level. Tumours may undergo multiple cycles of mutation rate switching in response to changing immune selection during growth. This proposed model is summarised in the cartoon in figure 6.1 below.

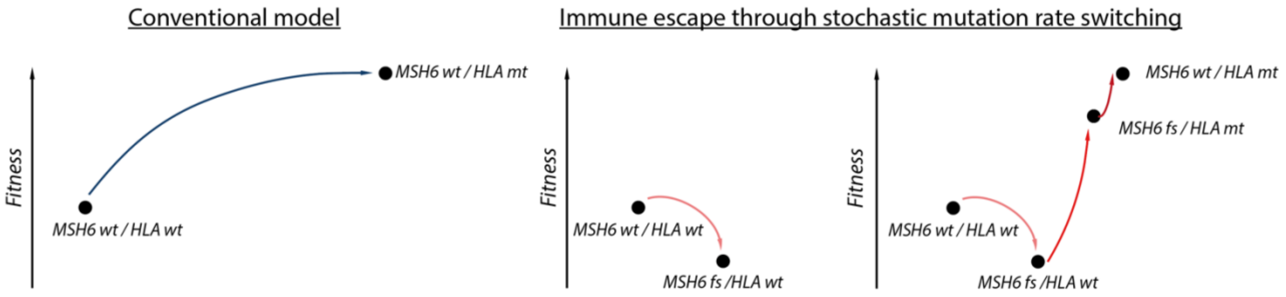


Figure 6.1: Model illustrating the fitness impact of mutation rate switching

In the conventional model (left) mutation rate remains constant and immune escape takes relatively longer. In the model with mutation rate switching (right), there is an initial decrease in fitness followed by accelerated gain of immune escape variants decreasing the overall time for fitness peak to be reached.

A question that arises from this work is why secondary *MSH6* and *MSH3* frameshifts, increase mutation rate in tumours that already have background loss of MLH1/PMS2 (MutLα). This can be addressed by considering the role of MutS and MutLα in the mismatch repair pathway. To recap, MSH6 and MSH3 partner with MSH2 to form the MutSα and MutSβ heterodimers respectively. Disruption of both MSH6 and MSH3 thus completely disrupts MutS function. Data from large cancer clinical cohorts (Salem et al., 2020) and cell lines (Zou et al., 2021) show that primary loss of MSH2 results in a higher mutation burden compared to loss of MLH1 or PMS2, indicating disruption of MutS leads to a more severe hypermutator phenotype compared to MutLα loss. A likely explanation for this is that the MutS mismatch recognition heterodimer operates earlier in the MMR pathway and has overlapping functions with other DNA repair pathways. Cross talk between MutS proteins and the base excision repair (BER) pathway have been reported (Lai et al., 2016). More specifically experimental data in cell lines has found that MutSα together with the BER glycosylase OGG1 both contribute to repair of 8-oxo-guanine damage (Mazurek et al., 2002), preventing the formation of 8-oxo-G:A mismatches which contribute to C>A/G>T mutations. More recently, a study of normal tissue from patients with constitutional mismatch repair deficiency found that MutSα loss (either MSH2 or MSH6) specifically had a major role in the repair of spontaneous methyl-cytosine deamination to thymine at CpG sites (Sanders et al., 2021). This activity was found to be independent of DNA replication and it was suggested that MutSα may partner with the thymine DNA glycosylase

MBD4 to perform this activity. Thus, there is mounting evidence for a non-canonical role of the MutS heterodimer in DNA repair, with potential cross-talk with the BER pathway.

The mismatch repair system is known to preferentially repair mutations occurring at early replicating regions of the genome and in particular at exonic regions (Frigola et al., 2017). In the recent study of patients with constitutional mismatch repair deficiency it was found that MMR loss due to MutSα (MSH2 or MSH6) as compared to MutLα loss (MLH1 or PMS2) resulted in significantly increased mutation burden at early replicating coding regions of the genome (Sanders et al., 2021). This finding is relevant because it also implies that secondary *MSH6* and *MSH3* frameshifts, by disrupting the MutS heterodimer, may shift mutation bias towards early replicating, functionally important genomic regions. Future work using whole genome sequencing data could confirm this in the context of secondary *MSH6* and *MSH3* frameshifts.

This work also contributes to an improved understanding of how mismatch repair deficient tumours evolve. Mutations in cancer are generally considered to be fixed and irreversible. Here we find evidence that in MMRd tumours, frameshift and reversion events at the *MSH6* and *MSH3* homopolymers allows these loci to act as molecular on/off switches regulating mutation rate. Since microsatellites are hypermutable loci in MMRd tumours other microsatellites may also show similar frameshift switching activity and identifying these would be a valuable area of future investigation. One can speculate occurrence of reversion mutations at sites with frameshift neoantigens or other deleterious loss of function frameshifts. Frameshift reversion mutations potentially provide a mechanism for MMRd tumours to mitigate the impact of Muller's ratchet (Muller, 1964), which states that in an asexually reproducing population progressive deleterious mutations lead to loss of cellular fitness over time.

I am also interested to consider the clinical applications of this work in patients. Patients with MMRd cancers are candidates for immune checkpoint inhibitor therapy (Marcus et al., 2019), however not all benefit. For example, in metastatic MMRd colorectal cancer approximately 40% of patients do not get long-term benefit from single agent anti-PD1 immunotherapy (Overman et al., 2017). It remains to be seen how subclonal *MSH6* and *MSH3* frameshifts impact on immunotherapy response. However, this work provides a rationale for earlier use of immunotherapy in the treatment pathway such that less time may have elapsed for immune escape mechanisms to establish. There may also be an application to therapeutically block MSH6 and MSH3 such that the mutation rate switch is permanently turned on. This could increase neoantigen generation and potentially drive improved immunotherapy response (Germano et al., 2017). However, caution is needed with such an approach. For instance, *MSH6* mutations often emerge following alkylating agent therapy in glioblastoma (GBM) where they act as a mechanism of resistance to temozolomide in MGMT methylated tumours (Yip et al., 2009). Recent data however shows that GBM tumours with *MSH6* mutations following temozolomide, did not respond to immunotherapy (Touat et al., 2020). The reasons for this are unclear, but may possibly reflect insufficient clonal neoantigens or that MSH6 loss mainly drives increased SNVs rather than high quality frameshift neoantigens.

In summary this work identifies a mechanism for adaptive mutability in mismatch repair deficient cancers, revealing that secondary reversible frameshifts in *MSH6* and *MSH3* toggle mutation rate and bias during tumour evolution to accelerate adaptation in response to immune selection. This process bears strong parallels with bacterial resistance evolution, where genetic diversification in response to environmental stress facilitates adaptation. Understanding the evolutionary trajectory of mismatch repair deficient cancers may provide opportunities to improve immunotherapy response in patients.

Chapter 7: Methods

7.1 Contribution statement

I performed all wetlab experiment. Variant calling for whole exome sequencing data was performed by Kevin Litchfield. Downstream bioinformatic analysis was performed by myself. Panagiotis Barmpoutis performed image analysis of multiplex immunofluorescence data. Luis Zapata performed immune dN/dS analysis. Eszter Lakatos developed the mathematical model of mutation rate switching. Guilio Caravagna performed mutation rate analysis using the MOBSTER tool. William Cross identified MSI tumours in the Genomics England colorectal cancer cohort. All other analyses were performed by myself.

7.2 UCL MMRd Colorectal Cancer Cohort

All samples were processed as per protocols approved by the UCL/UCLH Biobank of health and disease ethical review committee (Project Reference Number NC21.18). A database search of the biobank was performed to identify MMRd colorectal cancers diagnosed between 2014 to 2018. 88 of 546 (16%) cancers tested by immunohistochemistry (IHC) showed evidence of mismatch repair protein loss. FFPE tumour blocks were retrieved and MSH6 IHC performed on cut sections using an established protocol. Antibody details and immunohistochemistry conditions are provided in Table 7.1 below. Immunohistochemitry was completed using a Leica Bond autostainer.

Epitope	Clone	Company	Dilution	Antigen retrieval (mins)	Primary incubation (mins)	Post-primary (mins)	HRP polymer (mins)	DAB (mins)
MSH6	EP49	Agilent	1:400	ER2 40	30	20	20	5
MLH1	ES05	Agilent	1:200	ER2 40	30	20	20	5
PMS2	A16-4	BD Biosciences	1:300	ER2 40	15	8	8	5
MSH2	FE11	Agilent	1:100	ER2 40	30	20	20	5
MSH3	611390	BD Biosciences	1:100	ER1 30	30	8	8	5

Table 7.1: Antibody conditions for mismatch repair immunohistochemistry.

Eleven (n=11/40, 28%) tumours with subclonal MSH6 loss in at least one tumour block were identified. A stage and age-matched cohort of 11 MLH1/PMS2 deficient MMRd tumours without immunohistochemical MSH6 loss were also selected as the comparison group. For each tumour, a normal block from the resection margin was also retrieved to be used as the germline sample. MSH6 labelled slides for each tumour were scanned using a slide scanner (Hamamatsu NanoZoomer).

Laser Capture Microdissection (LCM)

Tumours with MSH6 deficient subclones (n=11) and those in the MSH6 proficient comparison group (n=11) underwent laser capture microdissection (LCM). Where available multi-region samples from more than one tumour block were taken. IHC labelling results in loss of DNA yield. Therefore, a bespoke protocol using IHC labelled sections to guide microdissection of thicker adjacent unlabelled sections was developed. Each tumour block underwent serial sectioning as follows: one 3µm thick section on to a glass slide, five 10µm thick sections onto poly-ethylene naphtholate (PEN) membrane slides (Zeiss Ag) and one 3µm thick section on to a glass slide. The 3µm thick sections underwent IHC against MSH6 and were used to guide microdissection of the thicker intervening sections. Membrane slides were pre-treated with ultraviolet light and 0.01% poly-l-lysine to improve tissue adherence. Mounted sections were baked at 50^{o}C for 4 hours in an oven. Next haematoxylin staining of the 10µm thick sections was performed using the following steps:

xylene (10 minutes, two changes), 100% ethanol (1 minute, two changes), 90% ethanol (1 minute, one change), rinse in deionized water, Gill’s haematoxylin (1 minute, one change), rinse gently in running water, 90% ethanol (1 minute, two changes), 100% ethanol (1 minute, two changes), xylene (1 minute, two changes).

Laser capture microdissection (LCM) was performed to obtain samples from MSH6 proficient and deficient regions. Using the Palm Microbeam microscope (Zeiss Ag), specific MSH6 deficient and proficient tumour regions approximately 2-3mm^2 in area were individually microdissected and collected in 500µL AdhesiveCap tubes (Zeiss Ag). Tissue originating from the same location was pooled across serial sections and processed as one sample. For each tumour a separate sample was also taken from the normal mucosa to be used as the germline sample. A sample number was allocated to each unique microdissected region and the location recorded for future reference.

DNA extraction of microdissected tissue samples

The manufacturer's instructions were followed. Briefly, 6ul of proteinase K and 200ul of lysis buffer (Perkin Elmer Inc.) was added to micro-dissected tissue samples and incubated overnight at 56^{o}C followed by 1 hour at 70^{o}C to reverse formaldehyde crosslinks. DNA extraction was completed with the Chemagic Prepito automated instrument (Perkin Elmer Inc) which uses a magnetic particle separation technique. Extracted DNA was quantified with a Qubit fluorometer (Thermo Fisher) as per the manufacturer's instructions.

Sanger Sequencing of the MSH6 C8 homopolymer

To detect presence of frameshift mutation in the C8 coding homopolymer of *MSH6,* PCR followed by BigDye terminator Sanger sequencing was performed.
PCR was performed using the following reagents:
Forward primer TTTTAACAGATGTTTTACTGTGC
Reverse primer TCATTAGGAATAAAATCATCTCC
Q5 polymerase mastermix (New England Biolabs)
10ng of genomic DNA
PCR was performed as follows: 35 cycles of denaturation at 95^{o}C for 30 seconds, followed by primer annealing at 60^{o}C for 1 minute, followed by extension at 72^{o}C for 30 seconds.

Sample processing for Whole Exome Sequencing

DNA acoustic fragmentation was performed with a Covaris E220 device. 125ng of sample DNA was inserted into snap-cap microtubes (Covaris) at a total volume of 50uL. The manufacturer's guidelines for the Covaris device were followed and the following settings were used: Duty factor=10%, peak incident power (W)=175, cycles per burst=200, time (seconds)=300. Afterwards, fragmented DNA samples were transferred to 1.5ml Eppendorf tubes. Samples were next subjected to FFPE repair to minimise artefacts related to formalin fixation. A validated FFPE repair kit (M6630L, New England Biolabs) was used as per the manufacturer's protocol. Briefly 48ul of fragmented DNA sample was mixed with 3.5ul of FFPE DNA repair buffer, 3.5ul of end-prep buffer and 2ul of FFPE DNA repair mix and the mixture incubated at 20^{o}C for 30 minutes.

DNA library preparation and exome capture

DNA Library preparation was performed using the NebNext Ultra II kit (New England Biolabs) as per the manufacturer's protocol. Samples underwent end repair and A-tailing followed by adapter ligation by adding 30ul of ligation master mix, 1ul of ligation enhancer and 2.5ul of sequencing adapters. The mixture was incubated for 15 minutes at 20^{o}C. Sequencing adapters were diluted 10x as per manufacturer's protocol. 0.9x (87ul) magnetic bead clean-up of adapter ligated libraries was performed using Ampure XP beads (Beckman Coulter) followed by ethanol washes and elution in 17ul of 10mM Tris-HCl. Adapter ligated libraries (15ul) were added to 25ul of Q5 master mix, 10ul of index primers and amplified using 10 cycles of PCR. NEBNext Multiplex Oligos (#E7335) were used for sample indexing and the index number used for each sample recorded. A Tapestation (Agilent Technologies) device was used to analyse the resulting library fragment size using the high sensitivity screentape. Library quantification was also performed using a Qubit fluorometer (Thermo Fisher).

Exome capture was performed using a commercial kit called SeqCap EZ kit (Roche Sequencing Solutions). As per the manufacturer's guidelines, 250ng of library samples from the previous step were pooled in groups of four to give a total mass of 1 microgram. The resulting multiplexed library pool was hybridized with exome capture probes for 16 hours at 47^{o}C. Following this, unbound probes were washed away and the hybridized DNA was amplified with 14 cycles of PCR. This was followed by 1x Ampure XP bead clean up and finally eluted in 33ul of 0.1x TE solution. The resulting captured amplified library was then quantified by qPCR using a commercial kit (NEBNext Library Quant kit for Illumina, New England Biolabs).

Next Generation Sequencing of exome libraries

Sample libraries were diluted to 2nM and sequenced in batches of 12 samples on a NovaSeq instrument (Illumina). An S1 flowcell with 100bp paired end reads was used in accordance with the manufacturer's instructions.

Aligment and Variant Calling of WES data

Generated FastQ sequencing files were aligned to the human Hg19 reference genome using BWA-mem (version 0.7.7). Aligned FastQ files were converted to BAM files and sorted and indexed using Samtools. Sequencing duplicates were marked using Picard Markduplicates and the GATK (version 2.8) workflow was used for local indel realignment. Initial Qc metrics were produced using PicardTools, GATK (version 2.8) and FastQC. Bases with a Phred score of less than 20 and reads with a mapping quality of less than 20 were omitted. SNV variant calling was performed using MuTect (version 1.1.4). Only SNVs where the variant allele frequency (VAF) was ≥ 5% and the total number of reads in the tumour and germline at that position was ≥20 were kept. For insertion/deletions (INDELs), two separate callers, VarScan2 and Scalpel were used and the consensus of high confidence calls between these callers used to avoid the known high level of artefacts often observed with indels. Variants were annotated using Annovar (version 2016Feb01).

Purity, ploidy and Copy Number (CN) estimation

The Sequenza package was used to derive copy number estimates for each sample. Using the package instructions, tumour purity and ploidy estimates were obtained by using the probabilistic parameter search. A quality control step was included, where the SNV allele frequency distribution in each sample was manually reviewed. Following correction for tumour purity, predicted copy number states were in keeping with expected allele frequency shifts (i.e. peaks at 0.33 and 0.67 in trisomy regions and 0.5 and 1 in copy neutral LOH).

Clonality assessment

The Palimpsest package was used to classify mutations as clonal or subclonal. For each SNV, the cancer cell fraction (CCF) was calculated using the variant allele fraction, tumour purity, copy number of the locus in tumour and normal as per (Letouzé et al., 2017). CCF was calculated within the package using the following formula:

$$CCF = VAF \text{ x} \frac{pN_t + (1-p)N_n}{pn_{chr}}$$

Where p is the tumour purity, N_t and N_n are the local copy number in the tumour and normal cells and n_{chr} is the number of chromosomal copies carrying the mutation in tumour cells.

The package determines the 95% confidence interval of the VAF using a binomial test and provides a 95% confidence interval for the CCF. A mutation is considered subclonal if the upper boundary of the 95% confidence interval was <0.95, and clonal otherwise.

Extraction of read length distribution of exonic microsatellites

The MSIsensor package (version 0.6) using the ‘msi scoring’ command was used to obtain the read length distribution of all length 6 to 11 exonic homopolymers in each sample. Tumour and matching normal bam files were supplied as input. MSIsensor read length distribution files were then tabulated ready for downstream analysis.

Accurate calling of *MSH6*$^{\mathbf{F1088}}$ and *MSH3*$^{\mathbf{K383}}$ frameshift mutation

Frameshift mutations are known to be challenging to call in NGS data. To accurately call homopolymer frameshifts within *MSH6* and *MSH3* homopolymers we used the results of MSIsensor and Scalpel/Varscan calling to reach a consensus. Any discrepancies between these methods were manually checked using IGV (Integrated Genomics Viewer v2.3) software. Mutations were called where a minimum of 5% of reads showed instability and with a minimum of 50 reads present.

Linear mixed effect model

A linear mixed effect model was created to account for the non-independence of multiple sampling per patient tumour. This model assessed the relationship between frameshift mutation status of *MSH6* and *MSH3* on total mutation burden. For random effects the model used individual variation in mutation burden between tumours. For fixed effects the model used the presence of mutation in *MSH6* and/or *MSH3* homopolymers, age at diagnosis and tumour purity. P-values were obtained by likelihood ratio tests of the full model with the effect of *MSH6/MSH3* mutation status against the null model without the effect MSH6/MSH3 status. The model was defined in the R package LME4 as follows:

lmer(MT_burden ~ MSH6_MSH3_status + age + tumour_purity+(1|Tumour_ID)
Results of the linear mixed effects model are provided in supplementary table 1.

Mutation signature analysis

The Sigprofiler (version 3.1) package was used to perform mutation signature analysis. Samples were grouped into 3 categories according to MSH6/MSH3 mutation status as follows: wild-type for both *MSH6* and *MSH3*, presence of either *MSH6*^{F1088fs} or *MSH3*^{K383fs} and presence of both *MSH6*^{F1088fs} and *MSH3*^{K383fs}. SNV and indel data for each sample was combined to create 3 meta-files according to these groups. Three de-novo SBS signatures were identified and their 96-channel trinucleotide distribution plotted. The percentage contribution of each signature according to *MSH6/3* grouping was further plotted. Indel and double base signatures were analysed in a similar manner. COSMIC signatures were also extracted using the Sigprofiler tool. As with de-novo signatures, mutation calls from samples were pooled into three groups according to *MSH6/MSH3* frameshift mutation status. Signature extraction was performed with the following settings: number of non-negative matrix factorisation replicates=500, minimum signatures to be extracted=1 and maximum signatures to be extracted=6. An upper limit of 6 maximum signatures was used to avoid overcalling low prevalence signatures.

Shannon microsatellite diversity

The Shannon entropy of all exonic length 6 to 11 microsatellites in each sample was calculated using the formula:

$$Shannon\ diversity = -\sum_{i=1}^{R} \quad [p_i ln\ (p_i)\ ]$$

Where p_i = the proportion of total reads represented by the ith microsatellite length
R= total number of read lengths present at a microsatellite

For each sample an average is then taken across all Shannon diversity scores for microsatellites of the same length. This means that each sample has individual Shannon microsatellite diversity scores for microsatellites in the length range of 6-11.

Phylogenetic trees

In tumours where more than 3 samples had been sequenced, phylogenies were derived using the maximum parsimony method. Only SNV calls were used to infer phylogenies given the inherent polymorphic nature of indels.

The Paup package (http://phylosolutions.com/paup-test/) was used using parameters as previously described(Cross et al., 2018).

The SNV calls were converted into a binary 0/1 matrix where 0 represents absence and 1 represents presence of a mutation. Each sample is represented by a row and each variant is represented by a column. Phylogenetic trees are derived as follows: 1) Phylogenies are rooted to the normal sample using the root function, 2) the hsearch function was used to perform a heuristic search of available trees and 1,000 of the shortest trees were output, 3) the bootstrap function was used to randomly resample the data 10,000 times with replacement, with the proportion of each branch instance reported. The most parsimonious tree was reported for each tumour. The Figtree software was used to view the resultant trees and (http://tree.bio.ed.ac.uk/software/figtree/) and converted into PDF files for storage. In the case of tumours with only 2 or 3 samples parsimony trees are not possible. Instead, simple inferences about clonality based on shared and private mutations were made. Mutations shared in all samples were allocated to the trunk whilst private mutations were allocated to branches. In tumours with 3 samples available, sample pairs with the most shared mutations were allocated to the same clade and mutations unique to each sample allocated to the terminal branches.

HLA class I genotyping and mutation calling

Using the Polysolver package (version 4) HLA class I haplotyping and mutation calling was performed for HLA-A, HLA-B and HLA-C alleles. Germline and tumour sequencing data was supplied in the form of BAM files.

Neoantigen prediction

A pipeline called neopredpipe was used for neoantigen prediction. This uses the patient’s specific HLA haplotype from Polysolver and the NetMHCpan prediction workflow (Schenck et al., 2019).

7.3 Immune dN/dS analysis

Immune dN/dS measures enrichment of nonsynonmous mutations in the portion of the genome exposed to immune recognition. It was calculated using a previously published method called SOPRANO(Zapata et al., 2020). It involves measuring dN/dS values in a target region exposed to the immune system (ON-target) and in the rest of the proteome (OFF-target). In this work for the ON region, we used genomic regions that translate to peptides that bind HLA-A0201 allele, since this is the most common HLA type in the Caucasian population. The file used as target region is available at github.com/luisgls/SOPRANO.

7.4 MOBSTER clonal deconvolution and mutation rate analysis

Mutation rate analysis using the pipeline MOBSTER (MOdel Based cluSTering in cancER) was performed by Caravagna as per his previous publication (Caravagna et al., 2020). Briefly this workflow can perform subclonal deconvolution using population genetics and machine learning approaches. In addition MOBSTER is able to estimate the mutation rate (μ) in a tumour sample from the tail of neutral mutations in the variant allele frequency plot. The analysis was restricted to SNVs since the VAF distribution of indels was found to be less reliable in MMRd tumours (likely due to frequent bi-allelic mutations). Somatic SNVs were mapped to corresponding copy number segments, which confirmed that the majority of the tumour exome was in the heterozygous diploid state as expected for tumours with microsatellite instability. Only SNVs mapping to diploid segments were retained as they were associated with less noise. Subclonal deconvolution was then performed with raw VAFs using MOBSTER. The MOBSTER tool was used to search for up to 2 subclones (k=2) in each tumour and an optional neutral tail using previously developed workflows. Parameters of the fit for the Power Law Type-I tail were then used to retrieve the tumour mutation rate µ. The mutation rate here was expressed as time units of tumour cell doublings. To make it comparable across multiple samples of the same patient, and to account for the fact that we used whole-exome data, we normalised by the size of the diploid exome. To obtain a Confidence Interval (CI) for μ, a non-parametric bootstrap procedure was used. We bootstrapped with repetitions from the mutations available in each sample and built n=200 datasets per patient; then we re-ran the MOBSTER analysis conditioned on retrieving the expected monoclonal architecture (k=1) identified in the main run and re-computed the normalised values for the bootstrap estimate of μ. With the distribution of bootstrapped μ values, we built a percentile CI corresponding to an α-level of 5% by taking the 2.5% and 97.5% empirical quantiles.

7.5 Multiplex Immunofluorescence

Multiplex immunofluorescence was performed using a commercial kit (Opal multiplex automation kit by Akoya). This system allows visualisation of 6 markers of interest. It uses HRP (horse radish peroxidase) conjugated secondary antibodies to covalently link fluorophores to the target of interest. A panel of markers consisting of MSH6, CD20, FOXP3, CD4, PANCK and CD8 was optimised for the multiplex immunofluorescence assay. Antibody details are provided in table 7.2 below. The manufacturer's protocol was followed and immuno-labelling performed using the Leica Bond RX autostainer (Leica Biosystems). A brief description of the optimisation process is described below.

Opal fluorophore	Opal fluorophore dilution	Primary epitope	Clone	Company	Primary dilution	Primary incubation (mins)	Antigen retrieval
520	1:150	MSH6	EP49	Agilent	1:100	60	ER2 (40)
540	1:150	CD20	L26	Agilent	1:150	15	ER1(20)
570	1:150	FOXP3	D608R	Cell Signalling Technology	1:200	30	ER2(20)
620	1:150	CD4	4B12	Agilent	1:50	30	ER1(20)
650	1:150	PANCK	AE1/3	Agilent	1:700	15	ER1(20)
690	1:150	CD8	4B11	Agilent	1:100	15	ER1(20)

Table 7.2: Antibody details and conditions used for multiplex immunofluorescence experiment.

Monoplex labelling optimization:

Initially monoplex slides were created to optimize labelling of each marker individually. Each primary antibody was assigned an opal fluorophore. Monoplex slides were processed with appropriate number of antibody stripping steps before and after staining reflecting the eventual multiplex sequence. Following staining, monoplex slides were scanned using the Vectra 3.0 fluorescence microscope and signal counts measured using the Inform software.

Autofluorescence (AF) slide:

To correct for background autofluorescence a representative tumour section was labelled with Pan-CK primary antibody and without secondary antibody or fluorophore. This was scanned and used to substract the AF signal in subsequent experiments.

Library development

A spectral unmixing library was created by labelling slides with the most abundant marker (Pan-CK) and each opal fluorophore individually, resulting in 6 library slides. Library slides were scanned on the Vectra 3.0 microscope using all 5 epi-fluorescence filters (DAPI, FITC, Cy3, Texas red, Cy5). The spectral unmixing library was developed using the Inform software and saved for subsequent experiments.

Multiplex immunofluorescence assay

Tissue sections were cut for the cohort of MMRd colorectal tumours. Multiplex immunofluorescence labelling was performed using the Leica BondRx autostainer. Using the previously optimised conditions, the multiplex assay was run using the conditions detailed in table 7.2. The following steps were followed for each staining run:

1. Deparaffinization using Bond dewax solution
2. Antigen retrieval solution using Bond ER1 or ER2 solution
3. Blocking buffer
4. Primary antibody incubation
5. Opal polymer HRP incubation
6. Opal fluorophore incubation
7. Stripping of antibody complexes using Bond ER1 or ER2 solution
8. Repeat steps 2-7 until all primary antibodies applied
9. DAPI counterstain

Following immunolabelling the slides were cover-slipped using Diamond antifade mountant (Invitrogen) and scanned using the Vectra 3.0 fluorescence microscope.

ORION fluorescence cell segmentation workflow

Image analysis was performed using a bespoke workflow named ORION (FluORescence cell segmentatION) developed by Panagiotis Barmpoutis. Initially this workflow was developed on a training dataset using a subset of the data. Following training ORION was run on the full dataset of 194 multispectral images tiles from 27 tumours. The workflow involves spectral unmixing, separation of cells based on an ellipsoidal model and Bayesian classification. Since both tumour cells and immune cells may express MSH6, colocalization with panCK was used to correctly classify tumour cells as MSH6 proficient or deficient. Each imaged tile was classified as MSH6 proficient or deficient based on the expression status tumour cells. Neighbourhood analysis was performed to quantify the number of immune cells of each subtype within a 100um radius of each MSH6 proficient and deficient tumour cell in an imaged tile. Image tiles were classified as MSH6 proficient or deficient according to the MSH6 expression status of the majority of tumour cells. For each tile we reported the sum total of immune cells of each subtype identified from neighbourhood analysis.

7.6 Genomics England (GEL) CRC cohort

Whole genome sequencing data were generated through a standardised workflow as part of the Genomics England 100,000 genomes project with samples sequenced on the Illumina HiSeq platform. 992 colorectal cancers were identified from the v8 release of the 100,000 Genomes dataset. Reads were aligned to the GrCh38 version of the reference human genome using the Illumina iSAAC aligner and variant calling performed using Strelka. Variants achieving the Strelka ‘PASS’ filter and also with a minimum of 50 tumour reads and with a variant allele frequency of greater than 5% were included in the downstream analysis. The Sequenza package was used to derive tumour purity, ploidy and copy number estimates for each sample as detailed above.

Identification of MSI cancers in GEL cohort

MSIsensor (version 0.6) was used to identify samples with microsatellite instability. Further validation was also performed using two steps. First, mutation signature data was used to confirm that there was significant mismatch repair SBS mutation signature present in identified MSI cases. Second, where available loss of mismatch repair protein expression by immunohistochemistry was confirmed (IHC validation dataset, n = 101, 98% classification accuracy). Tumours with known pathogenic POLE or POLD1 exonuclease domain mutations were also excluded. This resulted in a cohort of 217 cases for subsequent analysis and referred hereafter as the GEL CRC MSI cohort.

Identification of primary cause of MMR deficiency in GEL cohort

Germline and somatic mutations in the MMR genes (*MLH1, PMS2, MSH2, MSH6* and *MSH3*) were identified by searching the Genomics England main programme tiering data for tier 1 pathogenic mutations. To identify tumours with *MLH1* promoter methylation, the presence of somatic *BRAF*^{V600E} mutation was used as a surrogate marker. This approach is in keeping with current clinical guidelines where it is recognized that amongst MMRd colorectal tumours the presence of somatic *BRAF* mutation associates strongly with MLH1 promoter methylation (Durno et al., 2017; National institute for health and care excellence (NICE), 2017).

Measuring for enrichment of frameshifts at the *MSH6* and *MSH3* homopolymers

This analysis compared the frequency of frameshifts at the *MSH6* C8 and *MSH3* A8 homopolymers against the frequency of frameshifts at other coding homopolymers of the same length and nucleotide base composition. This allowed measurement of enrichment of frameshifts at *MSH6* and *MSH3* compared to other homopolymers. Genomic coordinates of all length 8 exonic homopolymers was obtained using the SciRoKo package. Using these coordinates the mutation status of all length 8 coding homopolymers and the base affected was extracted from the variant call files. The percentage of cases with frameshift mutation in C:G or A:T microsatellites were calculated separately and compared to the mutation frequency observed at the *MSH6* (C8) and *MSH3* (A8) microsatellites respectively. Comparisons were between groups were then made using the Chi squared test.

Multiple linear regression model

This analysis tested whether frameshift disruption of any MSI target genes associated with a significant change in overall mutation burden. A multiple linear regression model was created to test the relationship between frameshifts in coding homopolymers and total mutation burden. A total of 23 coding homopolymers, including the genes MSH6 and MSH3 were used as independent variables in the model. This list of genes were those reported as recurrently mutated in MMRd colorectal cancer (Cortes-Ciriano et al., 2017) and consisted of *AIM2, ACVR2A, RFX5, MBD4, DOCK3, TGFBR2, GLYR1, OR51E, CLOCK, CASP, JAK1, TAF1B, BAX, MYH11, HPS1, SLAMF1, HNF1A, RGS12, ELAVL3, SMAP1, SLC22A9*. Age and tumour purity were also included in the model to account for potential confounding. The model was created in R using the lm function. Results were plotted as a volcano plot with regression coefficients of contribution to mutation burden versus -$\log_{10}$ p value of the t-statistic. The model was also run using only *MSH6* and *MSH3* frameshifts as independent variables to obtain estimates of the contribution of these frameshifts individually and combined on total mutation burden. Tabulated results of the model are provided in supplementary table 2.

Identification of *MSH6* and *MSH3* coding mutations

To identify all MSH6 and MSH3 coding mutations variant call files were queried for all coding *MSH6* and *MSH3* mutations. Data were extracted and tabulated by the frequency and type of mutation (see chapter 2, figure 2.3D).

Mutation burden analysis

SNV, indel and total mutation burden for each sample was obtained from variant call files. Both synonymous and non-synonymous SNVs were included. Violin and waterfall plots were generated using the ggplot package in R.

MLH1/PMS2 (MutL) deficient subset

To verify that the differences in the underlying primary cause mismatch repair loss in these tumours was not confounding the results, the analysis was performed on a subset of cases with known MutLα (MLH1/PMS2) loss. In MSI colorectal cancer, the presence of *BRAF*^{V600E} mutation is known to correspond with MLH1 promoter hypermethylation (Kambara et al., 2004). We therefore restricted our analysis to MSI cases with *BRAF*^{V600E} mutation or samples with tier 1 pathogenic germline *MLH1* or *PMS2* mutations. We then repeated the analysis generating violin plots for SNV, indel and total mutation burden.

7.7 TCGA dataset

The case ID of MSI tumours from the TCGA dataset was obtained from a previous study (Cortes-Ciriano et al., 2017). Variant call data for these tumours was downloaded from the National Cancer Institute Genomics Data Commons portal (https://portal.gdc.cancer.gov/repository). Cases with frameshift mutation in the *MSH6* and *MSH3* coding microsatellites were identified from supplemental data provided in (Cortes-Ciriano et al., 2017). Tumours with known *POLE* or *POLD* exonuclease mutations were excluded (Temko et al., 2018). The final MSI cohort consisted of the following tumour types, colorectal (n=48), uterine (n=67), stomach (n=63) and esophageal (n=3). Tumours were grouped according to *MSH6* and *MSH3* frameshift mutation status as previously and violin plots displaying SNV, indel and total mutation burden were generated in R using the ggplot package. Neoantigen data was obtained from Lakatos et al ((Lakatos et al., 2020)) as previously described.

RNA expression data was available for 127 samples (colorectal n=42, uterine n=56, stomach n=29). Raw RNA expression counts were obtained for MSH6 and MSH3 and converted to FPKM (fragments per kilobase million) and then to TPM (transcripts per kilobase million) values. The following conversion formula was used:

TPM = fpkm/sum(fpkm) * 10^6

7.8 Mathematical model of mutation rate switching on tumour growth

The model of mutation rate switching was developed by Ezster Lakatos who provided the notes below on its development. It is an extension of previous work (Lakatos et al., 2020) and uses a stochastic branching birth/death process to model tumour growth and neoantigen accumulation. Under the model each cell can either (i) die with a probability inversely proportional to their fitness or (ii) divide into two daughter cells that accumulate novel mutations according to their respective mutation rate. Cells in hyper-mutated and ultra-hyper-mutated state gain a number of mutations with each division sampled from a Poisson distribution with parameter (mutation rate, μ) 6 and 120, respectively. New mutations are either (i) neutral with no effect on cell fitness; (ii) antigenic, decreasing the cell's fitness; (iii) immune escape mutations that eliminate immune predation and therefore nullify antigen-induced fitness decrease; (iv) lethal, irreversibly decreasing cell fitness regardless of immune escape. The probability of a given mutation being non-neutral is defined by p(antigen), p(escape), p(lethal), respectively. Note that these mutation types are non-exclusive and a mutation can be for example both antigenic and lethal (though with only a small probability). In addition, at each division the daughter cells may undergo mutation rate switching with probability β. β=0 corresponds to no switching (mutation rates remain constant) while β>1/100 represent frequent switches to or from ultra-hyper-mutated state. Each tumour was initiated with a homogeneous population of 100 tumour cells all in either hyper-mutated or ultra-hyper-mutated state; and simulated until elimination (no tumour cells left) or until reached detectable size (>100,000 cells).

Microsatellite diversity

We encoded the mutation status of the microsatellite locus as an integer: 0 represented a wild-type allele, -1/+1 a single deletion/insertion, and so on. Upon division, daughter cells inherited the mutation status of their ancestor. Every new mutation had a probability, p(ms), to affect the microsatellite: if they did, the state of the locus was changed from N to N+1 or N-1 with equal probability. At the end of the simulation, the mutation status of all cells was read out and the total Shannon diversity of the population was computed in R (using the package entropy).

Growth time

We defined the start of the “growth period” as the last time-point when the population count went below 20 (immune escaped) cells. The final time was the time-point when the population reached 100,000 cells. Growth time was computed as T(final) – T(growth-start). We chose this measure over T(final) as the latter had a very high uncertainty due to the variable time lineages spent before probabilistically acquiring immune escape and initiating unimpeded growth.

Parameter values

The following default parameter values were used in all simulations, unless indicated otherwise (e.g. a range of p(lethal) values in Fig. 4F): Neoantigen probability (p(antigen)): 0.1; immune escape probability (p(escape)): 10^{-6}; lethal mutation probability (p(lethal)): $5*10^{-4}$; microsatellite-shifting rate (p(ms)): 10^{-3}; immune-related selection coefficient (s): -0.8 (representing moderate selection).

Supplementary data

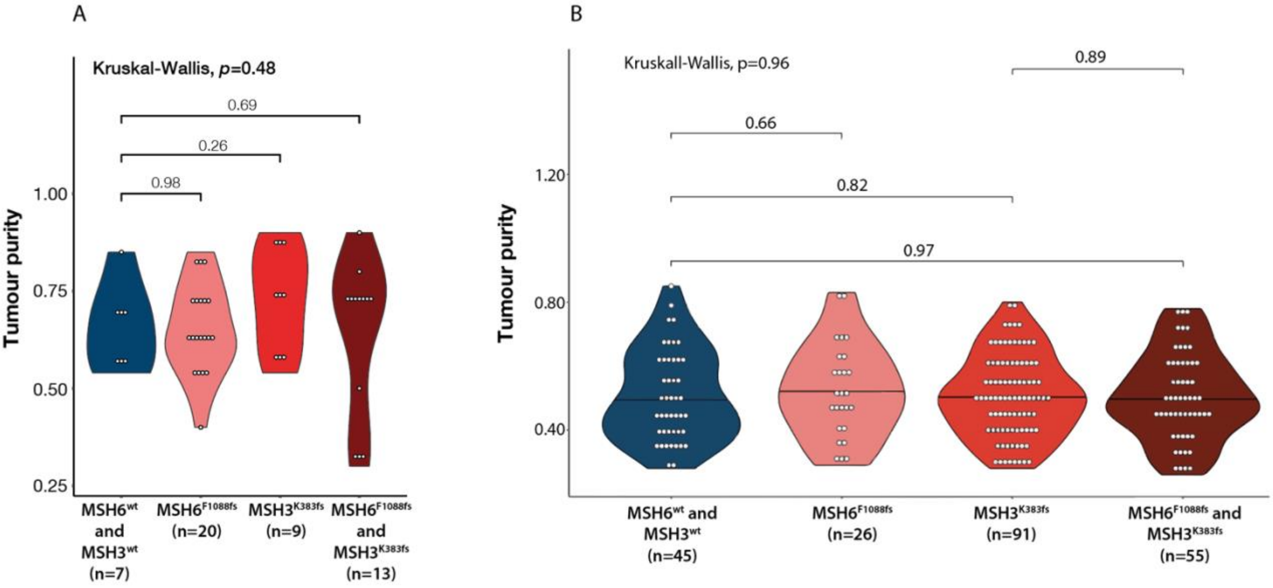


Figure S1: Analysis of tumour purity in UCH and GEL cohorts.

Violin plot showing tumour purity in samples grouped according to MSH3 and MSH6 microsatellite frameshift status in A) UCH cohort and B) GEL MSI colorectal cancer cohorts.

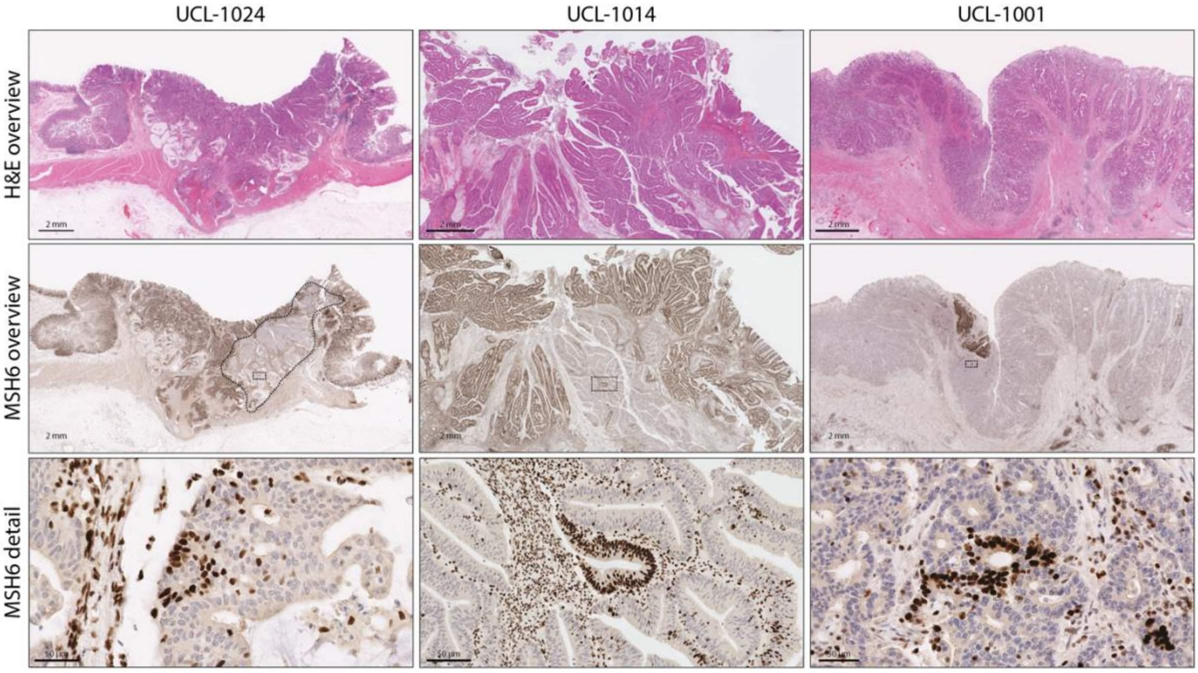


Figure S2: Additional examples of tumours with subclonal MSH6 loss in the context of background MLH1/PMS2 deficiency.

Images show for three tumours, H&E overview image, MSH6 IHC overview and high-power micrographs of nested proficient subclones within deficient regions.

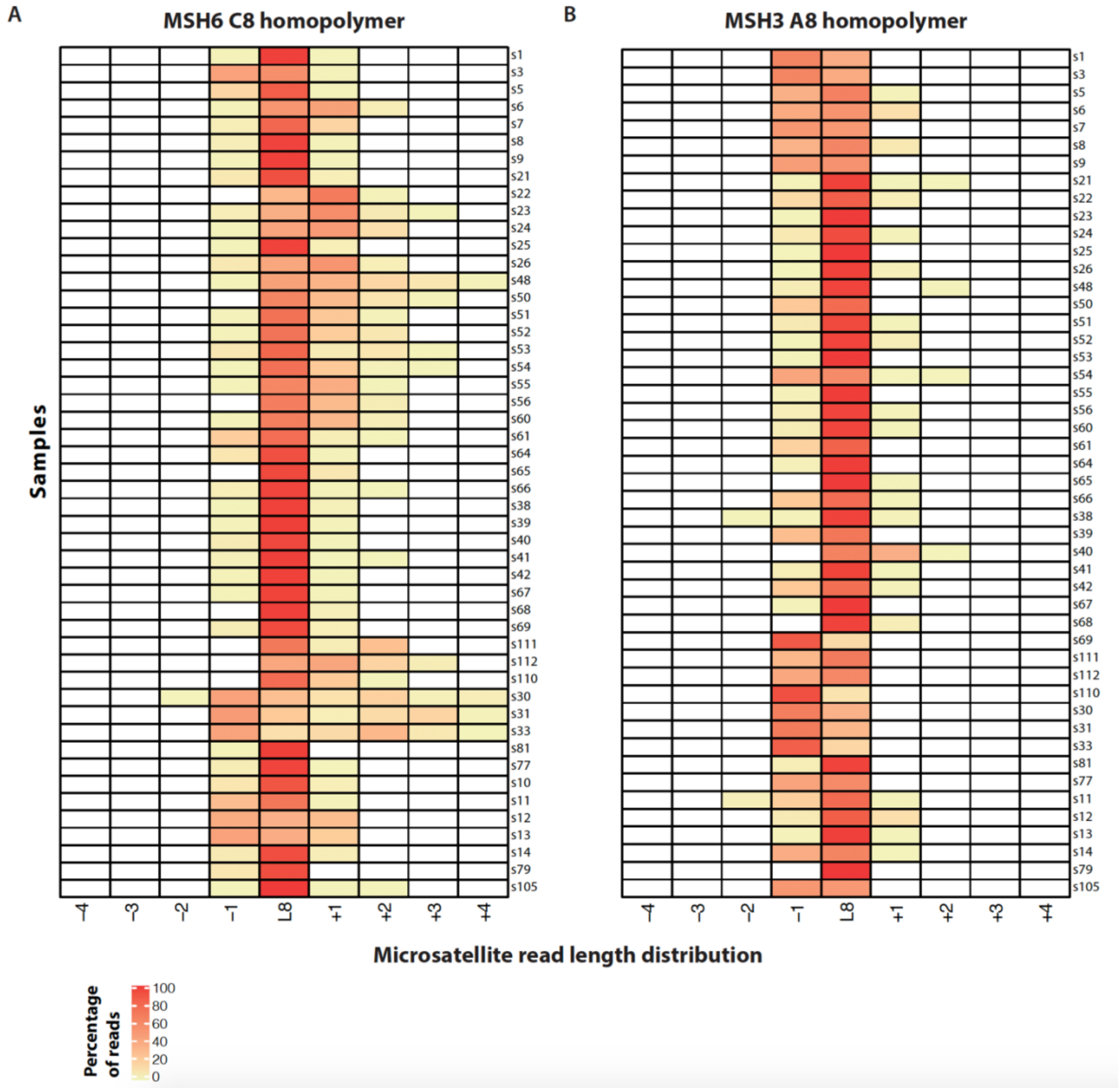
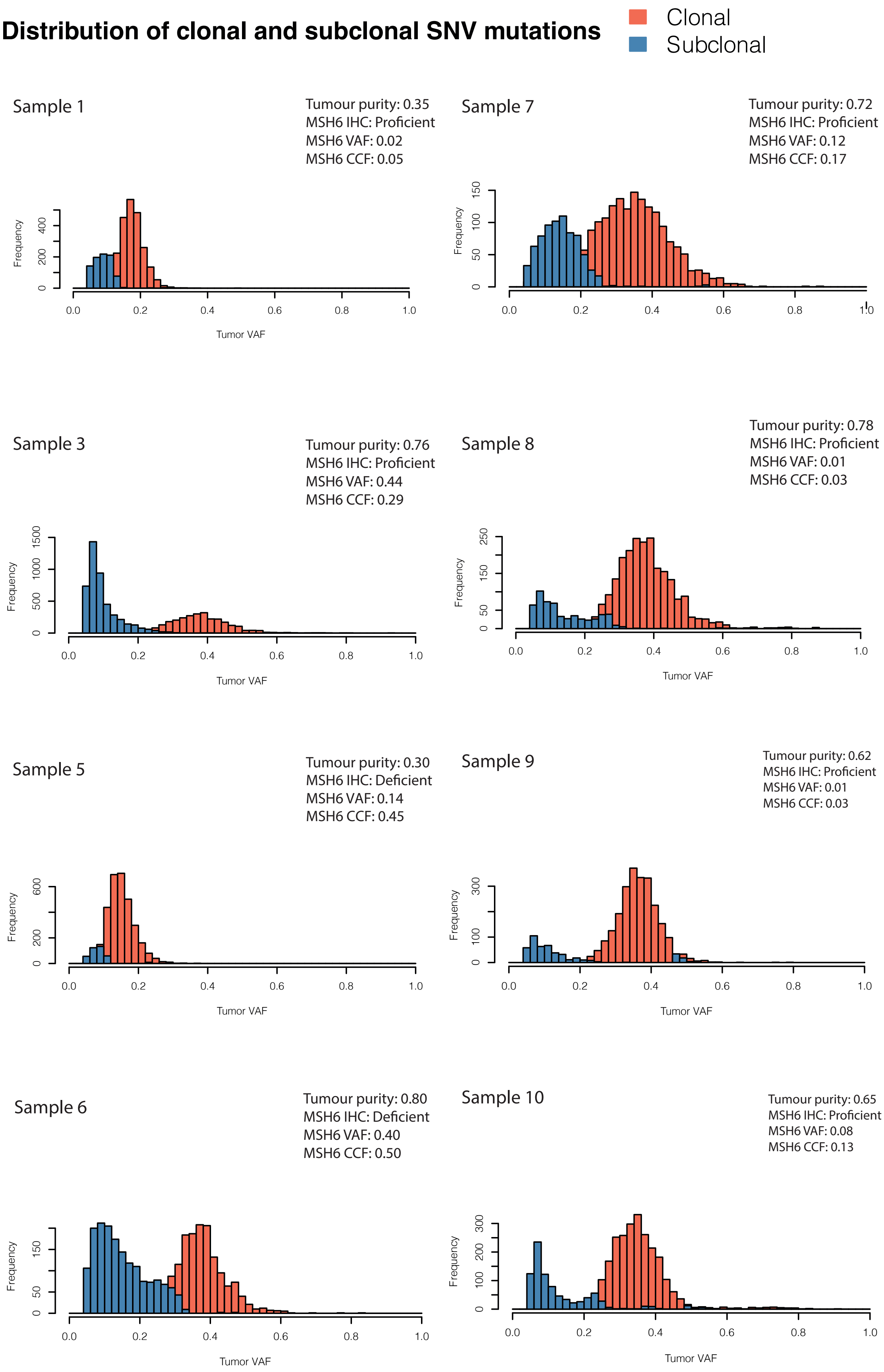


Figure S3: Analysis of MSH6 and MSH3 homopolymer read length distribution in the UCH cohort.

Read length distribution of A) MSH6 C8 and B) MSH3 A8 homopolymers. Each row is a sample.



Distribution of clonal and subclonal SNV mutations

🟥 Clonal
🟦 Subclonal

Sample 1
Tumour purity: 0.35
MSH6 IHC: Proficient
MSH6 VAF: 0.02
MSH6 CCF: 0.05

Sample 7
Tumour purity: 0.72
MSH6 IHC: Proficient
MSH6 VAF: 0.12
MSH6 CCF: 0.17

Sample 3
Tumour purity: 0.76
MSH6 IHC: Proficient
MSH6 VAF: 0.44
MSH6 CCF: 0.29

Sample 8
Tumour purity: 0.78
MSH6 IHC: Proficient
MSH6 VAF: 0.01
MSH6 CCF: 0.03

Sample 5
Tumour purity: 0.30
MSH6 IHC: Deficient
MSH6 VAF: 0.14
MSH6 CCF: 0.45

Sample 9
Tumour purity: 0.62
MSH6 IHC: Proficient
MSH6 VAF: 0.01
MSH6 CCF: 0.03

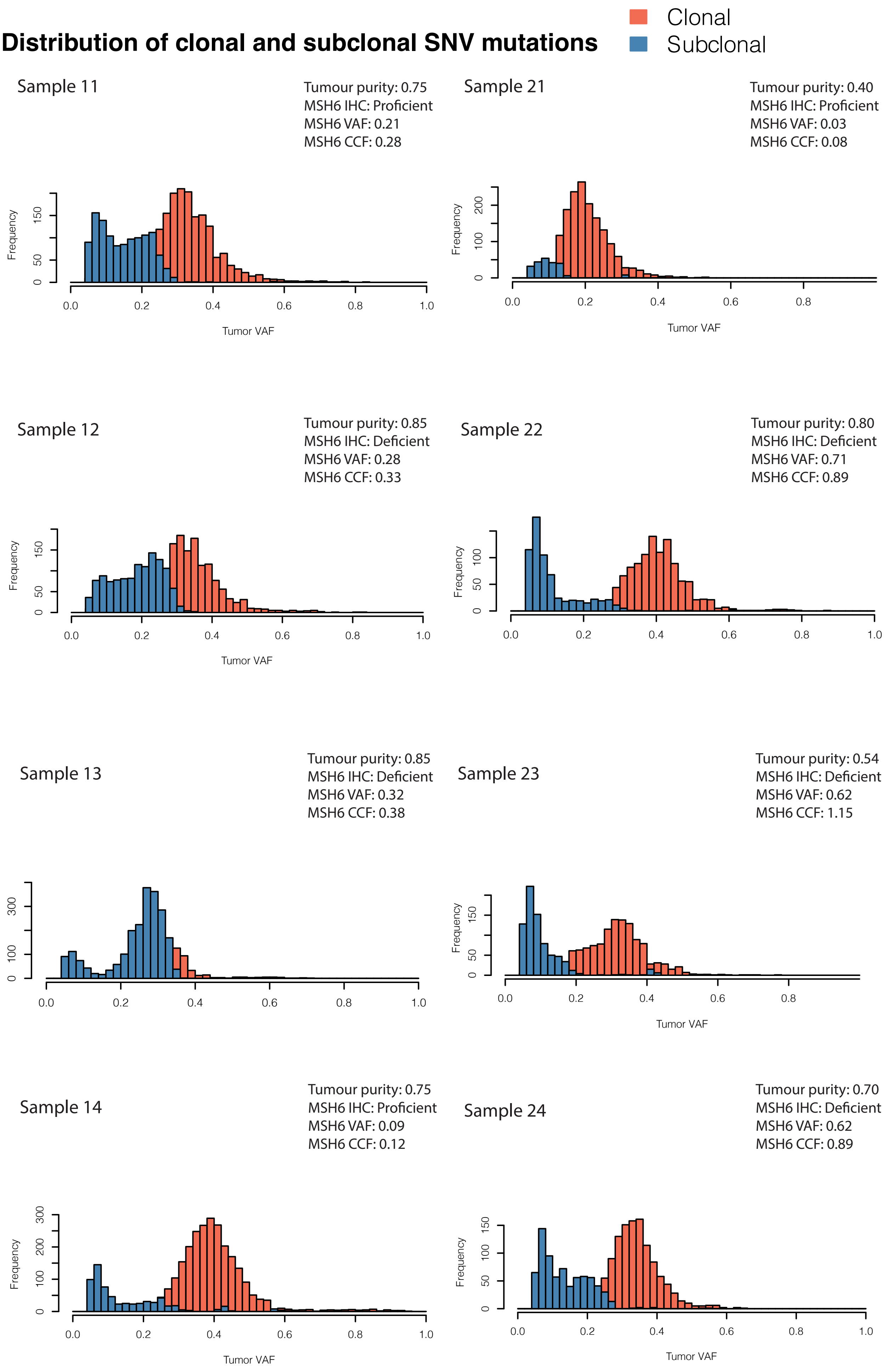
Sample 6
Tumour purity: 0.80
MSH6 IHC: Deficient
MSH6 VAF: 0.40
MSH6 CCF: 0.50

Sample 10
Tumour purity: 0.65
MSH6 IHC: Proficient
MSH6 VAF: 0.08
MSH6 CCF: 0.13

Figure S3

Distribution of clonal and subclonal SNV mutations

🟥 Clonal
🟦 Subclonal



Sample 11

Tumour purity: 0.75
MSH6 IHC: Proficient
MSH6 VAF: 0.21
MSH6 CCF: 0.28

Sample 21

Tumour purity: 0.40
MSH6 IHC: Proficient
MSH6 VAF: 0.03
MSH6 CCF: 0.08

Sample 12

Tumour purity: 0.85
MSH6 IHC: Deficient
MSH6 VAF: 0.28
MSH6 CCF: 0.33

Sample 22

Tumour purity: 0.80
MSH6 IHC: Deficient
MSH6 VAF: 0.71
MSH6 CCF: 0.89

Sample 13

Tumour purity: 0.85
MSH6 IHC: Deficient
MSH6 VAF: 0.32
MSH6 CCF: 0.38

Sample 23

Tumour purity: 0.54
MSH6 IHC: Deficient
MSH6 VAF: 0.62
MSH6 CCF: 1.15

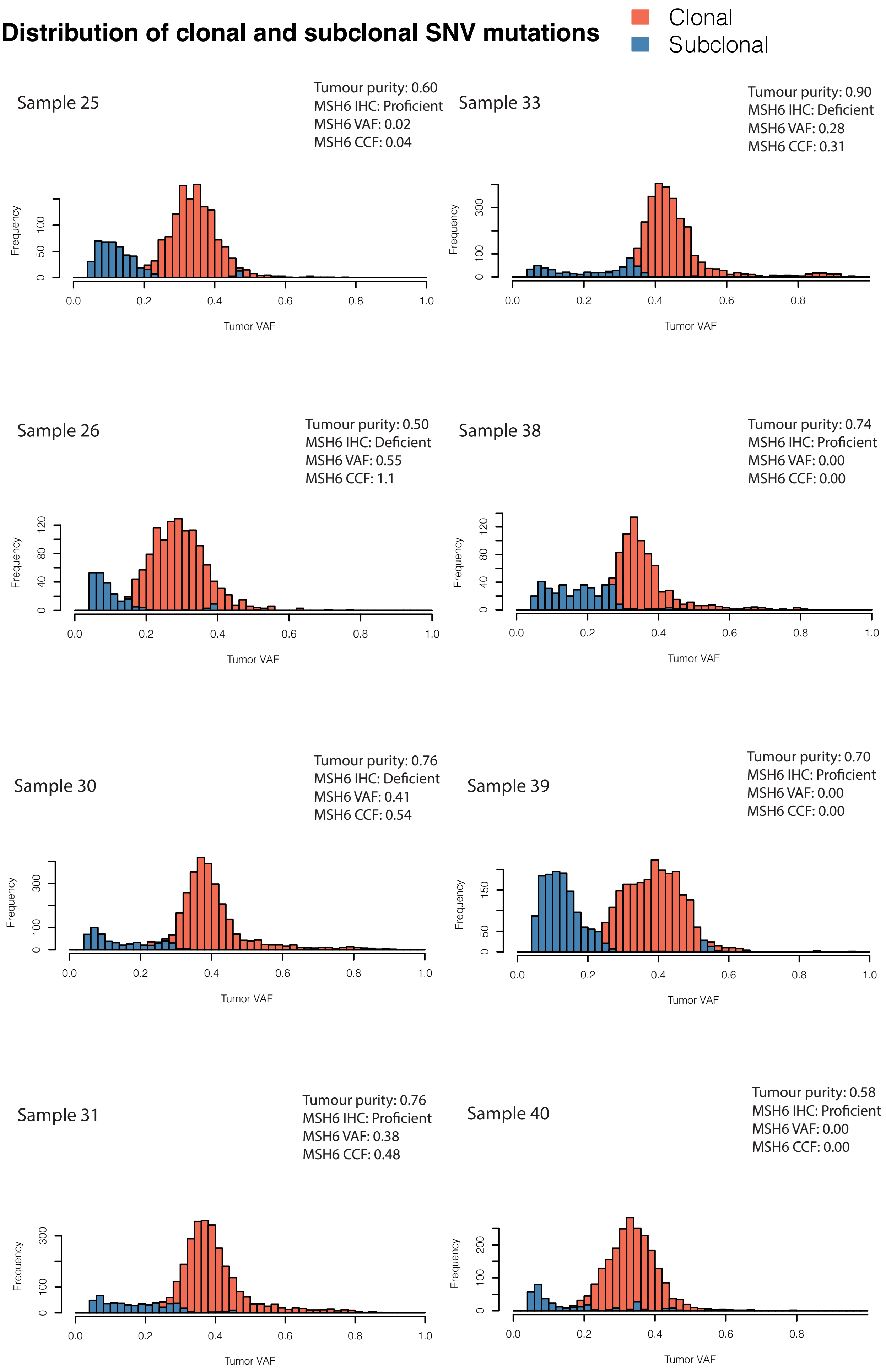
Sample 14

Tumour purity: 0.75
MSH6 IHC: Proficient
MSH6 VAF: 0.09
MSH6 CCF: 0.12

Sample 24

Tumour purity: 0.70
MSH6 IHC: Deficient
MSH6 VAF: 0.62
MSH6 CCF: 0.89

Figure S3



Distribution of clonal and subclonal SNV mutations

🟥 Clonal

🟦 Subclonal

Sample 25

Tumour purity: 0.60
MSH6 IHC: Proficient
MSH6 VAF: 0.02
MSH6 CCF: 0.04

Sample 33

Tumour purity: 0.90
MSH6 IHC: Deficient
MSH6 VAF: 0.28
MSH6 CCF: 0.31

Sample 26

Tumour purity: 0.50
MSH6 IHC: Deficient
MSH6 VAF: 0.55
MSH6 CCF: 1.1

Sample 38

Tumour purity: 0.74
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 30

Tumour purity: 0.76
MSH6 IHC: Deficient
MSH6 VAF: 0.41
MSH6 CCF: 0.54

Sample 39

Tumour purity: 0.70
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 31

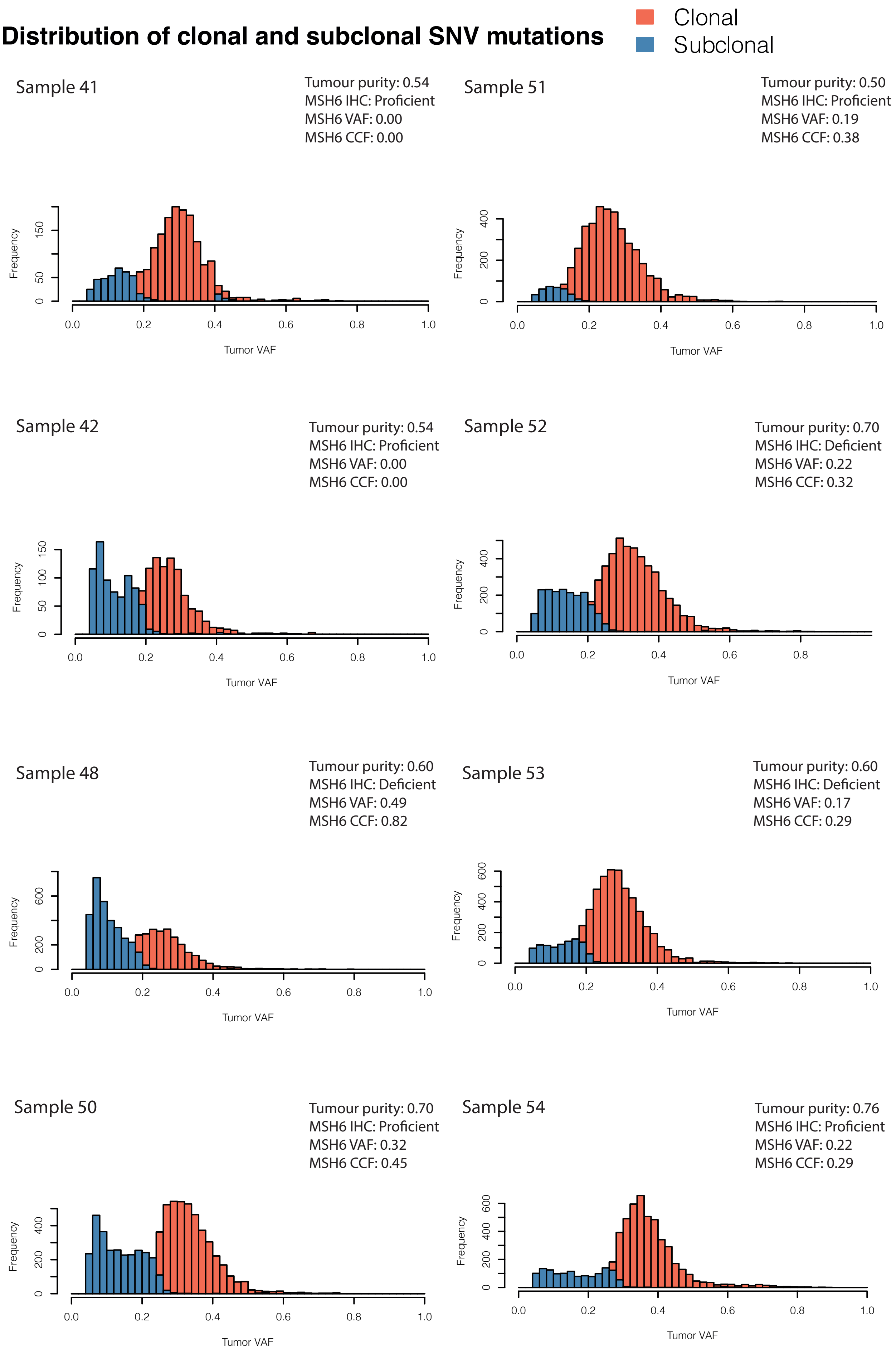
Tumour purity: 0.76
MSH6 IHC: Proficient
MSH6 VAF: 0.38
MSH6 CCF: 0.48

Sample 40

Tumour purity: 0.58
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Figure S3

Distribution of clonal and subclonal SNV mutations



Sample 41

Tumour purity: 0.54
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 51

Tumour purity: 0.50
MSH6 IHC: Proficient
MSH6 VAF: 0.19
MSH6 CCF: 0.38

Sample 42

Tumour purity: 0.54
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 52

Tumour purity: 0.70
MSH6 IHC: Deficient
MSH6 VAF: 0.22
MSH6 CCF: 0.32

Sample 48

Tumour purity: 0.60
MSH6 IHC: Deficient
MSH6 VAF: 0.49
MSH6 CCF: 0.82

Sample 53

Tumour purity: 0.60
MSH6 IHC: Deficient
MSH6 VAF: 0.17
MSH6 CCF: 0.29

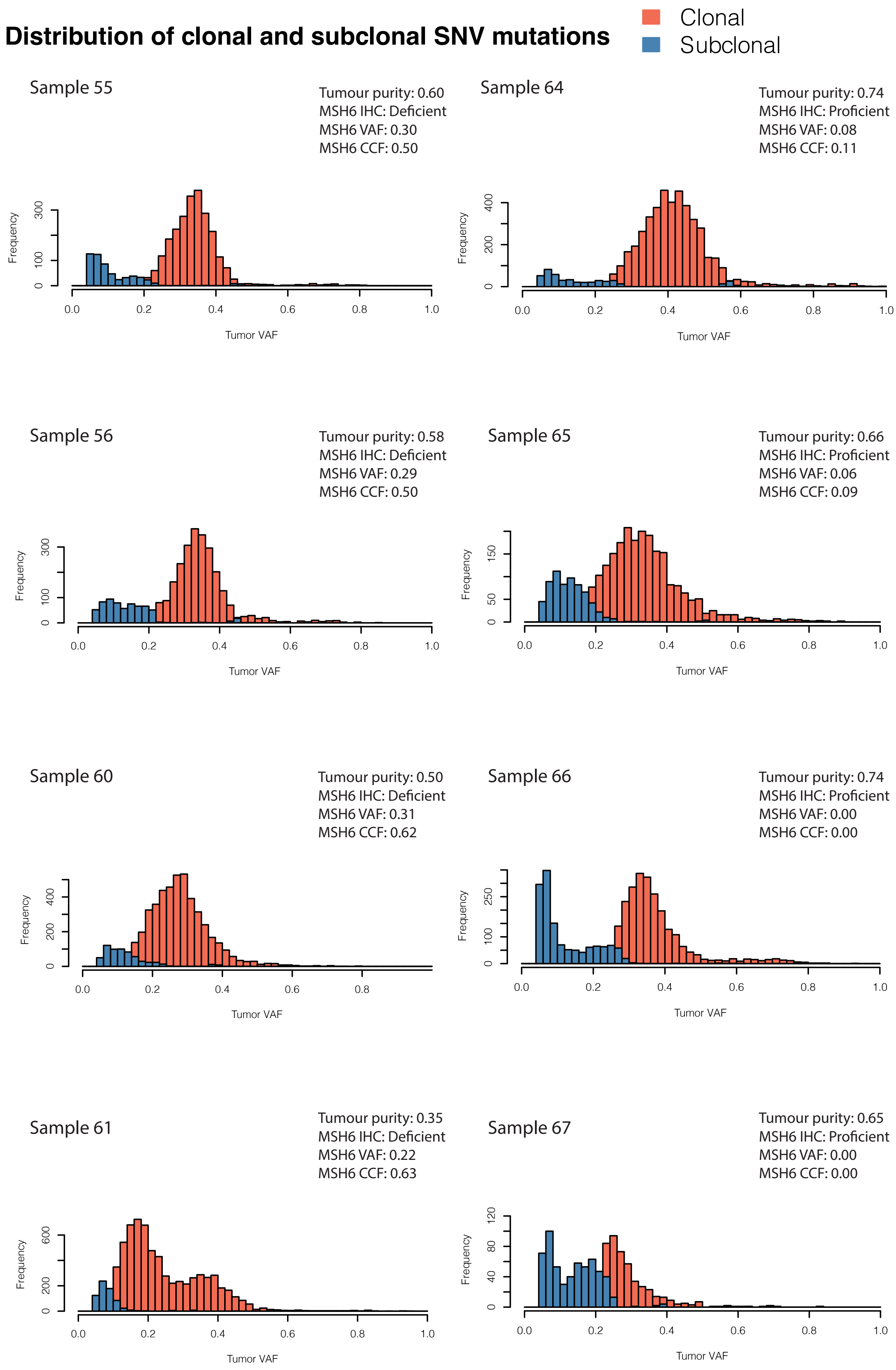
Sample 50

Tumour purity: 0.70
MSH6 IHC: Proficient
MSH6 VAF: 0.32
MSH6 CCF: 0.45

Sample 54

Tumour purity: 0.76
MSH6 IHC: Proficient
MSH6 VAF: 0.22
MSH6 CCF: 0.29

Figure S3



Distribution of clonal and subclonal SNV mutations

🟥 Clonal

🟦 Subclonal

Sample 55

Tumour purity: 0.60
MSH6 IHC: Deficient
MSH6 VAF: 0.30
MSH6 CCF: 0.50

Sample 64

Tumour purity: 0.74
MSH6 IHC: Proficient
MSH6 VAF: 0.08
MSH6 CCF: 0.11

Sample 56

Tumour purity: 0.58
MSH6 IHC: Deficient
MSH6 VAF: 0.29
MSH6 CCF: 0.50

Sample 65

Tumour purity: 0.66
MSH6 IHC: Proficient
MSH6 VAF: 0.06
MSH6 CCF: 0.09

Sample 60

Tumour purity: 0.50
MSH6 IHC: Deficient
MSH6 VAF: 0.31
MSH6 CCF: 0.62

Sample 66

Tumour purity: 0.74
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

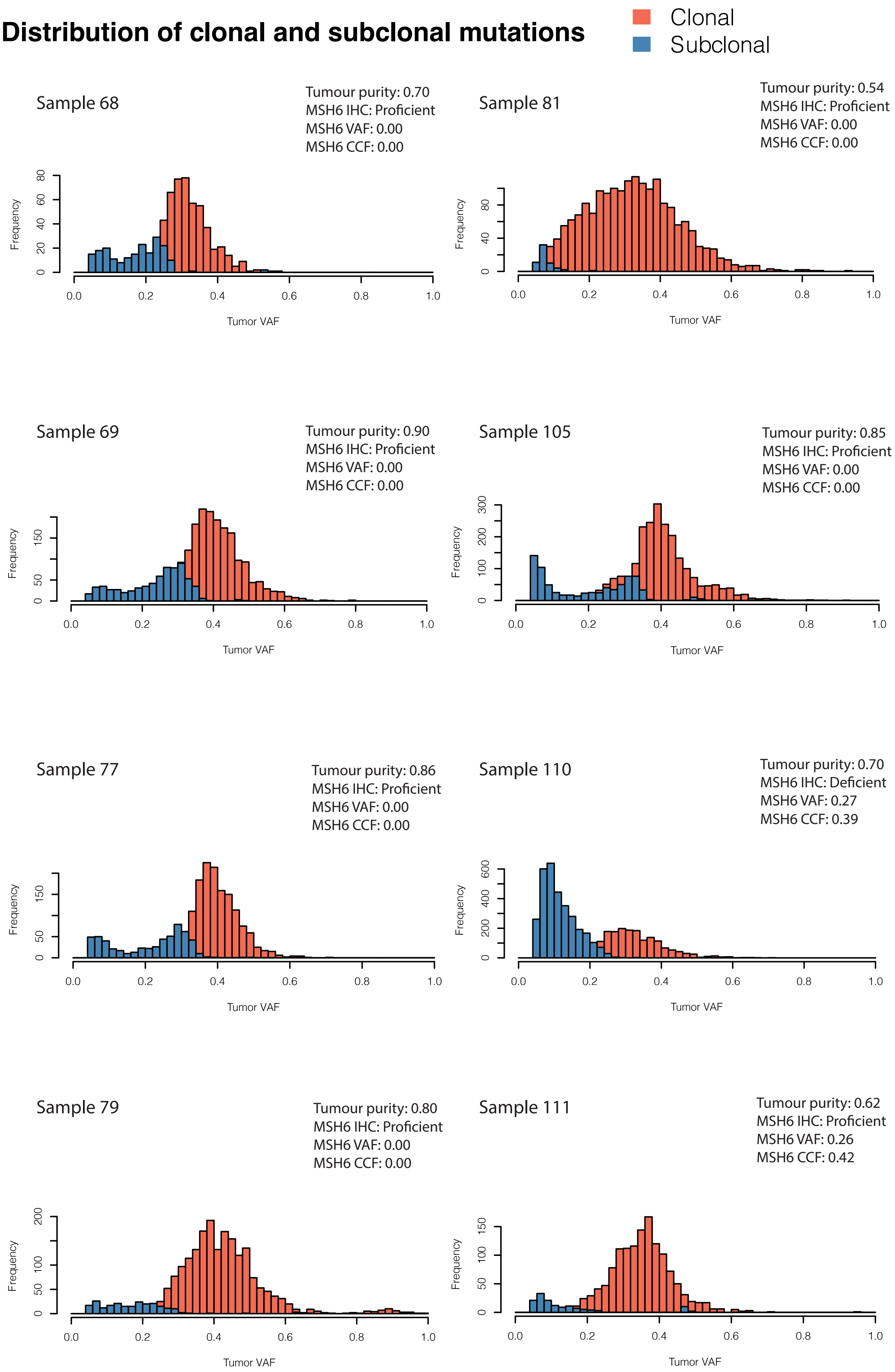
Sample 61

Tumour purity: 0.35
MSH6 IHC: Deficient
MSH6 VAF: 0.22
MSH6 CCF: 0.63

Sample 67

Tumour purity: 0.65
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Figure S3



Distribution of clonal and subclonal mutations

🟥 Clonal
🟦 Subclonal

Sample 68
Tumour purity: 0.70
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 81
Tumour purity: 0.54
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 69
Tumour purity: 0.90
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 105
Tumour purity: 0.85
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 77
Tumour purity: 0.86
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 110
Tumour purity: 0.70
MSH6 IHC: Deficient
MSH6 VAF: 0.27
MSH6 CCF: 0.39

Sample 79
Tumour purity: 0.80
MSH6 IHC: Proficient
MSH6 VAF: 0.00
MSH6 CCF: 0.00

Sample 111
Tumour purity: 0.62
MSH6 IHC: Proficient
MSH6 VAF: 0.26
MSH6 CCF: 0.42

Figure S3

Distribution of clonal and subclonal mutations

🟥 Clonal
🟦 Subclonal

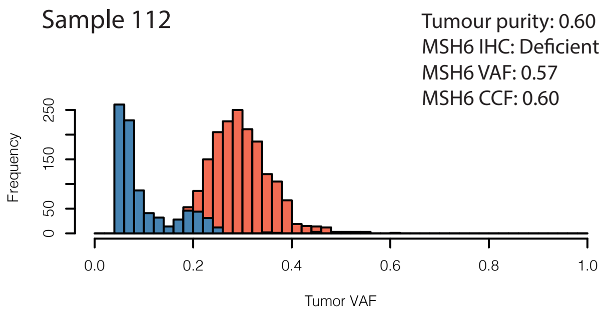


Figure S3

Linear mixed effect model

Full model:
Formula: MT_burden ~ MSH6_MSH3_status + purity + age + (1 | Tumours)
Data: cohort

AIC	BIC	logLik	deviance	df.resid
881.9	893.2	-434.9	869.9	43

Random effects:

Groups	Name	Std.Dev.
Tumours	(Intercept)	2209
Residual		1061

Number of obs: 49, groups: Tumours, 22

Fixed Effects:

(Intercept)	MSH6_MSH3_status	purity	age
1509.53	612.23	1580.30	29.67

Null model:
Formula: MT_burden ~ purity + age + (1 | Tumours)
Data: cohort

AIC	BIC	logLik	deviance	df.resid
884.5813	894.0404	437.2906	874.5813	44

Random effects:

Groups	Name	Std.Dev.
Tumours	(Intercept)	2428
Residual		1080

Number of obs: 49, groups: Tumours, 22

Fixed Effects:

(Intercept)	MSH6_MSH3_status	purity	age
	3065.10	29.91	1491.92

```
> anova(full_model,null_model)
```
Data: cohort
Models:
null_model: MT_burden ~ age + purity + (1 | Tumours)
full_model: MT_burden ~ MSH6_MSH3_status + purity + age + (1 | Tumours)

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
null_model	5	884.58	894.04	-437.29	874.58			
full_model	6	881.85	893.20	-434.93	869.85	4.7309	1	0.02963 *

```
#remove each fixed effect
> lme4:::drop1.merMod(full_model, test = "Chisq")
```
Single term deletions
Model: MT_burden ~ MSH6_MSH3_status + purity + age + (1 | Tumours)

	npar	AIC	LRT	Pr(Chi)
<none>		881.85		
MSH6_MSH3_status	1	884.58	4.7309	0.02963 *
purity	1	880.80	0.9470	0.33048
age	1	881.17	1.3232	0.25001

Supplementary table 1: Linear mixed effect modelling

Linear mixed effect model assessing random effect of individual variation between tumours and fixed effects of age, MSH6/MSH3 mutation status and tumour purity on total mutation burden.

(A)

Independent variables	Homopolymer	Cases mutated	Beta coefficient	Standard error	t value	P value
MSH3	A8	143 (66%)	57294	16962.2	3.378	0.000886
MSH6	C8	81 (37%)	47190.3	16039.6	2.942	0.003663
RFX5	G7	29 (13%)	-6426.2	22979.3	-0.28	0.780049
MBD4	T10	100 (46%)	-226.7	15291	-0.015	0.988185
AIM2	T10	184 (85%)	217.7	21976.5	0.01	0.992106
ACVR2A	A8	212 (98%)	-60886.3	56695.4	-1.074	0.284214
DOCK3	C7	128 (59%)	18511.4	16435.9	1.126	0.261462
TGFBR2	A10	205 (94%)	58377.2	37461	1.558	0.120807
GLYR1	C8	125 (58%)	9028.4	15735.2	0.574	0.566801
OR51E	A8	66 (30%)	-1741.2	16822.1	-0.104	0.917668
CLOCK	A9	140 (65%)	14815.2	17093.2	0.867	0.387177
CASP5	T8	15 (7%)	-39338.4	30024.3	-1.31	0.191695
JAK1	T8	21 (10%)	21234.3	26056.1	0.815	0.416118
TAF1B	A11	208 (96%)	34930.8	42449.4	0.823	0.411602
BAX	G8	115 (53%)	24170.5	15973.8	1.513	0.131898
MYH11	G8	98 (45%)	-15812.1	15776.1	-1.002	0.317476
HPS1	G8	82 (38%)	21541.7	15546.3	1.386	0.167471
SLAMF1	T9	87 (40%)	14910.3	15689	0.95	0.343125
HNF1A	C8	22 (10%)	4553.4	26119.2	0.174	0.861788
RGS12	A9	72 (33%)	12400.9	16625.4	0.746	0.456646
ELAVL3	C9	85 (39%)	-6014.3	15507.7	-0.388	0.698578
SMAP1	A10	183 (84%)	38820.9	22473.8	1.727	0.085715
SLC22A9	A11	206 (95%)	62795.6	37207.8	1.688	0.093101
Tumour purity	-	-	178452.1	58599.6	3.045	0.002653
Patient age	-	-	792.4	608.5	1.302	0.1944
			R^2 (adjusted)			0.225
			P value			4.20×10^{-7}

(B)

Independent variables	Beta coefficient	Standard error	t value	P value
MSH3 K383fs	88038.2	20078.9	4.385	0.0000183
MSH6 F1088fs	63675.6	27247.8	2.337	0.0204
MSH6 F1088fs & MSH3 K383fs	139338.8	22163.6	6.287	1.83x10-9
tumour purity	145461.6	56173.8	2.589	0.0103
patient age	874	591.4	1.478	0.141
R^2 (adjusted)				0.17
p-value				1.87×10^{-8}

Supplementary table 2: Multiple linear regression analysis for association between frameshifts in MSI target genes and total mutation burden.

A) Full model with 23 MSI target genes. B) Model using MSH6 and MSH3 frameshifts individually and combined.

Independent variables	Homopolymer	Cases mutated	Beta coefficient	Standard error	t value	P value
MLL3	9T	99 (46%)	44477.1	15781.7	2.818	0.00542
MSH6	8C	81 (37%)	43313.8	16517.1	2.622	0.00955
OR52N5	10A	87 (40%)	37020	15513.3	2.386	0.01816
MSH3	8A	146 (67%)	39296.3	17872	2.199	0.02929
UBR5	8T	125 (58%)	36827.4	17074.2	2.157	0.03247
ACVR2A	8A	212 (98%)	-40563.3	58673.9	-0.691	0.49033
KIAA2018	11T	117 (54%)	21657.5	15196.8	1.425	0.15602
SLC22A9	11A	206 (95%)	40013.9	37691	1.062	0.28996
ASTE1	11T	121 (56%)	3632.8	15648.3	0.232	0.81671
TGFBR2	10A	205 (94%)	15591.6	38694	0.403	0.68751
NDUFC2	9A	162 (75%)	12053.8	18211.8	0.662	0.50899
SEC31A	9T	150 (69%)	21463.8	17875.9	1.201	0.2316
LTN1	11T	205 (94%)	-36439.9	37060.2	-0.983	0.32693
C18orf34	10T	146 (67%)	-5434.7	17631.9	-0.308	0.7583
AIM2	10T	184 (85%)	-11597.6	22431.5	-0.517	0.60584
RPL22	6T	6 (3%)	-31242.4	49152.2	-0.636	0.52591
OR7E24	11T	134 (62%)	1377.4	16393.3	0.084	0.93314
CCDC150	11A	121 (56%)	-14546	15922.8	-0.914	0.3623
RNF43	7C	152 (70%)	27933.3	18312.3	1.525	0.12909
CASP5	10T	15 (7%)	-35481.9	30725.3	-1.155	0.24985
MIS18BP1	11T	165 (76%)	-2820.2	19156.8	-0.147	0.88314
PHACTR4	10A	118 (54%)	4320.1	15893.4	0.272	0.7861
SLC35F5	9A	217 (100%)	NA	NA	NA	NA
CASP5.1	10T	177 (82%)	-20420.5	21181	-0.964	0.33642
SRPR	8T	116 (53%)	-3784.4	15620.5	-0.242	0.80887
LMAN1	9T	121 (56%)	2457.7	16943.8	0.145	0.88485
RBM27	9A	117 (54%)	-16576.8	15376.2	-1.078	0.28258
SMAP1	10A	183 (84%)	19154.1	17033	1.125	0.26243
SLAMF1	9T	87 (40%)	7528.7	15485.2	0.486	0.62748
DDX27	8A	141 (65%)	17720.7	16694.7	1.061	0.29004
TMEM22	9A	117 (54%)	22802.4	15535.1	1.468	0.14407
TEAD2	8G	110 (51%)	26780.9	16321.8	1.641	0.10275
MNS1	10T	90 (41%)	11519.2	15903.8	0.724	0.46991
PRDM2	9A	74 (34%)	-6140.6	16134.9	-0.381	0.70401
SEC63	10T	122 (56%)	-11300.1	16450.7	-0.687	0.49311
CLOCK	9A	140 (65%)	18403.4	17992	1.023	0.30788
TMEM60	9T	115 (53%)	11993.9	16051.1	0.747	0.45599
PRRG1	8C	43 (20%)	30340.3	19133.2	1.586	0.11472
AASDH	10A	122 (56%)	25448.9	15966	1.594	0.11287
XPOT	9T	79 (36%)	2425.6	15852.1	0.153	0.87858
UPF3A	9A	161 (74%)	27587.7	18824.7	1.466	0.1447

C15orf40	14A	192 (88%)	14684	24331	0.604	0.547
FAM111B	10A	119 (55%)	12412.6	15768.7	0.787	0.43232
CCDC168	9T	34 (16%)	4674	21572.6	0.217	0.82874
KIAA1919	9T	126 (58%)	18022	15609.2	1.155	0.24994
COBLL1	9A	124 (57%)	3787.5	16087.2	0.235	0.81416
MIS18BP1.1	9T	94 (43%)	10500.1	15665.8	0.67	0.50364
FGFBP1	9T	108 (50%)	28995.9	15941.8	1.819	0.07076
DOCK3	7C	128 (59%)	4880.9	15950.7	0.306	0.75999
SPINK5	10A	118 (54%)	11686.4	15285.6	0.765	0.44564
JPH4	9C	130 (60%)	7470.2	16407.3	0.455	0.6495
purity	-	-	146928.1	58518.2	2.511	0.01301
age	-	-	-139.3	653.3	-0.213	0.83135
			R^2 (adjusted)			0.318
			P value			1.12×10^{-7}

Supplementary table 3: Multiple linear regression analysis for association between frameshift mutations in top 50 recurrently mutated microsatellites and total mutation burden.

Top 50 recurrently mutated microsatellites are as per genes listed in figure 2A of publication by Cortes-Ciriano et al (Cortes-Ciriano et al., 2017).

References:

Abdel-Rahman, W. M., Georgiades, I. B., Curtis, L. J., Arends, M. J., & Wyllie, A. H. (1999). Role of BAX mutations in mismatch repair-deficient colorectal carcinogenesis. *Oncogene*, *18*(12), 2139–2142. https://doi.org/10.1038/sj.onc.1202589

Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., Boot, A., Covington, K. R., Gordenin, D. A., Bergstrom, E. N., Islam, S. M. A., Lopez-Bigas, N., Klimczak, L. J., McPherson, J. R., Morganella, S., Sabarinathan, R., Wheeler, D. A., Mustonen, V., Alexandrov, L. B., … Consortium, P. (2020). The repertoire of mutational signatures in human cancer. *Nature*, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

André, T., Shiu, K.-K., Kim, T. W., Jensen, B. V., Jensen, L. H., Punt, C., Smith, D., Garcia-Carbonero, R., Benavides, M., Gibbs, P., de la Fouchardiere, C., Rivera, F., Elez, E., Bendell, J., Le, D. T., Yoshino, T., Van Cutsem, E., Yang, P., Farooqui, M. Z. H., … Diaz, L. A. (2020). Pembrolizumab in Microsatellite-Instability–High Advanced Colorectal Cancer. *New England Journal of Medicine*, *383*(23), 2207–2218. https://doi.org/10.1056/NEJMoa2017699

Bader, S., Walker, M., Hendrich, B., Bird, A., Bird, C., Hooper, M., & Wyllie, A. (1999). Somatic frameshift mutations in the MBD4 gene of sporadic colon cancers with mismatch repair deficiency. *Oncogene*, *18*(56), 8044–8047. https://doi.org/10.1038/sj.onc.1203229

Baranovskaya, S., Soto, J. L., Perucho, M., & Malkhosyan, S. R. (2001). Functional significance of concomitant inactivation of MLH1 and MSH6 in tumor cells of the microsatellite mutator phenotype. *Proceedings of the National Academy of Sciences*, *98*(26), 15107 LP – 15112. https://doi.org/10.1073/pnas.251234498

Bayliss, C. D., Field, D., & Moxon, E. R. (2001). The simple sequence contingency loci of Haemophilus influenzae and Neisseria meningitidis. *The Journal of Clinical Investigation*, *107*(6), 657–662. https://doi.org/10.1172/JCI12557

Bendell, J., Ciardiello, F., Tabernero, J., Tebbutt, N., Eng, C., Di Bartolomeo, M., Falcone, A., Fakih, M., Kozloff, M., Segal, N., Sobrero, A., Shi, Y., Roberts, L., Yan, Y., Chang, I., Uyei, A., & Kim, T. (2018). Efficacy and safety results from IMblaze370, a randomised Phase III study comparing atezolizumab+cobimetinib and atezolizumab monotherapy vs regorafenib in chemotherapy-refractory metastatic colorectal cancer. *Annals of Oncology*, *29*, v123. https://doi.org/10.1093/annonc/mdy208.003

Biller, L. H., & Schrag, D. (2021). Diagnosis and Treatment of Metastatic Colorectal Cancer: A Review. *JAMA*, *325*(7), 669–685. https://doi.org/10.1001/jama.2021.0106

Bjedov, I., Tenaillon, O., Gérard, B., Souza, V., Denamur, E., Radman, M., Taddei, F., & Matic, I. (2003). Stress-Induced Mutagenesis in Bacteria. *Science*, *300*(5624), 1404 LP – 1409. https://doi.org/10.1126/science.1082240

Caravagna, G., Heide, T., Williams, M. J., Zapata, L., Nichol, D., Chkhaidze, K., Cross, W., Cresswell, G. D., Werner, B., Acar, A., Chesler, L., Barnes, C. P., Sanguinetti, G., Graham, T. A., & Sottoriva, A. (2020). Subclonal reconstruction of tumors by using machine learning and population genetics. *Nature Genetics*, *52*(9), 898–907.

https://doi.org/10.1038/s41588-020-0675-5

Chan, T.-L., Wong, C. W., Chan, A. S., Guo, D. L., Ho, J. W., Yuen, S. T., & Leung, S. Y. (2004). BRAF and KRAS mutations in mismatch repair deficient colorectal cancer. *Cancer Research*, *64*(7 Supplement), 744 LP – 745. http://cancerres.aacrjournals.org/content/64/7_Supplement/744.5.abstract

Chang, A., Liu, L., Ashby, J. M., Wu, D., Chen, Y., O&#039;Neill, S. S., Huang, S., Wang, J., Wang, G., Cheng, D., Tan, X., Petty, W. J., Pasche, B. C., Xiang, R., Zhang, W., & Sun, P. (2021). Recruitment of KMT2C/MLL3 to DNA Damage Sites Mediates DNA Damage Responses and Regulates PARP Inhibitor Sensitivity in Cancer. *Cancer Research*, *81*(12), 3358 LP – 3373. https://doi.org/10.1158/0008-5472.CAN-21-0688

Chen, L., Liu, P., Evans, T. C. J., & Ettwiller, L. M. (2017). DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science (New York, N.Y.)*, *355*(6326), 752–756. https://doi.org/10.1126/science.aai8690

Cipponi, A., Goode, D. L., Bedo, J., McCabe, M. J., Pajic, M., Croucher, D. R., Rajal, A. G., Junankar, S. R., Saunders, D. N., Lobachevsky, P., Papenfuss, A. T., Nessem, D., Nobis, M., Warren, S. C., Timpson, P., Cowley, M., Vargas, A. C., Qiu, M. R., Generali, D. G., … Thomas, D. M. (2020). MTOR signaling orchestrates stress-induced mutagenesis, facilitating adaptive evolution in cancer. *Science*, *368*(6495), 1127 LP – 1131. https://doi.org/10.1126/science.aau8768

Colas, C., Coulet, F., Svrcek, M., Collura, A., Fléjou, J.-F., Duval, A., & Hamelin, R. (2012). Lynch or not Lynch? Is that always a question? *Advances in Cancer Research*, *113*, 121–166. https://doi.org/10.1016/B978-0-12-394280-7.00004-X

Cortes-Ciriano, I., Lee, S., Park, W.-Y., Kim, T.-M., & Park, P. J. (2017). A molecular portrait of microsatellite instability across multiple cancers. *Nature Communications*, *8*(1), 15180. https://doi.org/10.1038/ncomms15180

Cross, W., Kovac, M., Mustonen, V., Temko, D., Davis, H., Baker, A.-M., Biswas, S., Arnold, R., Chegwidden, L., Gatenbee, C., Anderson, A. R., Koelzer, V. H., Martinez, P., Jiang, X., Domingo, E., Woodcock, D. J., Feng, Y., Kovacova, M., Maughan, T., … Consortium, T. S. (2018). The evolutionary landscape of colorectal tumorigenesis. *Nature Ecology & Evolution*, *2*(10), 1661–1672. https://doi.org/10.1038/s41559-018-0642-z

CRUK. (2021). *CRUK cancer mortality statistics by cancer type*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/mortality/common-cancers-compared#heading-Zero

de Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., Yates, L., Jamal-Hanjani, M., Shafi, S., Murugaesu, N., Rowan, A. J., Grönroos, E., Muhammad, M. A., Horswell, S., Gerlinger, M., Varela, I., Jones, D., Marshall, J., Voet, T., Van Loo, P., … Swanton, C. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science (New York, N.Y.)*, *346*(6206), 251–256. https://doi.org/10.1126/science.1253462

Deng, G., Bell, I., Crawley, S., Gum, J., Terdiman, J. P., Allen, B. A., Truta, B., Sleisenger, M. H., & Kim, Y. S. (2004). BRAF mutation is frequently present in sporadic colorectal cancer with methylated hMLH1, but not in hereditary nonpolyposis colorectal cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *10*(1 Pt 1), 191–195. https://doi.org/10.1158/1078-0432.ccr-1118-3

Dentro, S. C., Wedge, D. C., & Van Loo, P. (2017). Principles of Reconstructing the Subclonal Architecture of Cancers. *Cold Spring Harbor Perspectives in Medicine*, *7*(8). https://doi.org/10.1101/cshperspect.a026625

Do, H., & Dobrovic, A. (2015). Sequence artifacts in DNA from formalin-fixed tissues: causes and strategies for minimization. *Clinical Chemistry*, *61*(1), 64–71. https://doi.org/10.1373/clinchem.2014.223040

Dominguez-Valentin, M., Sampson, J. R., Seppälä, T. T., ten Broeke, S. W., Plazzer, J.-P., Nakken, S., Engel, C., Aretz, S., Jenkins, M. A., Sunde, L., Bernstein, I., Capella, G., Balaguer, F., Thomas, H., Evans, D. G., Burn, J., Greenblatt, M., Hovig, E., de Vos tot Nederveen Cappel, W. H., … Møller, P. (2020). Cancer risks by gene, age, and gender in 6350 carriers ofpathogenic mismatch repair variants: findings from the Prospective Lynch SyndromeDatabase. *Genetics in Medicine*, *22*(1), 15–25. https://doi.org/10.1038/s41436-019-0596-9

Durno, C., Boland, C. R., Cohen, S., Dominitz, J. A., Giardiello, F. M., Johnson, D. A., Kaltenbach, T., Levin, T. R., Lieberman, D., Robertson, D. J., & Rex, D. K. (2017). Recommendations on Surveillance and Management of Biallelic Mismatch Repair Deficiency (BMMRD) Syndrome: A Consensus Statement by the US Multi-Society Task Force on Colorectal Cancer. *Gastroenterology*, *152*(6), 1605–1614. https://doi.org/https://doi.org/10.1053/j.gastro.2017.02.011

Duval, A., & Hamelin, R. (2002). Mutations at Coding Repeat Sequences in Mismatch Repair-deficient Human Cancers. *Cancer Research*, *62*(9), 2447 LP – 2454. http://cancerres.aacrjournals.org/content/62/9/2447.abstract

Engel, C., Ahadova, A., Seppälä, T. T., Aretz, S., Bigirwamungu-Bargeman, M., Bläker, H., Bucksch, K., Büttner, R., de Vos Tot Nederveen Cappel, W. T., Endris, V., Holinski-Feder, E., Holzapfel, S., Hüneburg, R., Jacobs, M. A. J. M., Koornstra, J. J., Langers, A. M., Lepistö, A., Morak, M., Möslein, G., … Vasen, H. F. (2020). Associations of Pathogenic Variants in MLH1, MSH2, and MSH6 With Risk of Colorectal Adenomas and Tumors and With Somatic Mutations in Patients With Lynch Syndrome. *Gastroenterology*, *158*(5), 1326–1333. https://doi.org/10.1053/j.gastro.2019.12.032

Farchoukh, L., Kuan, S.-F., Dudley, B., Brand, R., Nikiforova, M., & Pai, R. K. (2016). MLH1-deficient Colorectal Carcinoma With Wild-type BRAF and MLH1 Promoter Hypermethylation Harbor KRAS Mutations and Arise From Conventional Adenomas. *The American Journal of Surgical Pathology*, *40*(10), 1390–1399. https://doi.org/10.1097/PAS.0000000000000695

FDA. (2017). *FDA grants accelerated approval to pembrolizumab for first tissue/site agnostic indication*. Case Medical Research. https://doi.org/10.31525/fda1-ucm560040.htm

Ford, D. J., & Dingwall, A. K. (2015). The cancer COMPASS: navigating the functions of MLL complexes in cancer. *Cancer Genetics*, *208*(5), 178–191. https://doi.org/https://doi.org/10.1016/j.cancergen.2015.01.005

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature Genetics*, *49*(12), 1684–1692. https://doi.org/10.1038/ng.3991

Gandhi, L., Rodríguez-Abreu, D., Gadgeel, S., Esteban, E., Felip, E., De Angelis, F., Domine, M., Clingan, P., Hochmair, M. J., Powell, S. F., Cheng, S. Y.-S., Bischoff, H. G., Peled, N., Grossi, F., Jennens, R. R., Reck, M., Hui, R., Garon, E. B., Boyer, M., … Garassino, M. C. (2018). Pembrolizumab plus Chemotherapy in Metastatic Non-Small-Cell Lung Cancer. *The New England Journal of Medicine*, *378*(22), 2078–2092. https://doi.org/10.1056/NEJMoa1801005

Germano, G., Lamba, S., Rospo, G., Barault, L., Magrì, A., Maione, F., Russo, M., Crisafulli, G., Bartolini, A., Lerda, G., Siravegna, G., Mussolin, B., Frapolli, R., Montone, M., Morano, F., de Braud, F., Amirouchene-Angelozzi, N., Marsoni, S., D'Incalci, M., … Bardelli, A. (2017). Inactivation of DNA repair triggers neoantigen generation and impairs tumour growth. *Nature*, *552*(7683), 116–120. https://doi.org/10.1038/nature24673

Germano, G., Lu, S., Rospo, G., Lamba, S., Rousseau, B., Fanelli, S., Stenech, D., Le, D. T., Hays, J., Totaro, M. G., Amodio, V., Chila, R., Mondino, A., Diaz, L. A., Di Nicolantonio, F., & Bardelli, A. (2021). CD4 T cell dependent rejection of beta 2 microglobulin null mismatch repair deficient tumors. *Cancer Discovery*, candisc.0987.2020. https://doi.org/10.1158/2159-8290.CD-20-0987

Giardiello, F. M., Allen, J. I., Axilbund, J. E., Boland, C. R., Burke, C. A., Burt, R. W., Church, J. M., Dominitz, J. A., Johnson, D. A., Kaltenbach, T., Levin, T. R., Lieberman, D. A., Robertson, D. J., Syngal, S., & Rex, D. K. (2014). Guidelines on genetic evaluation and management of Lynch syndrome: a consensus statement by the US Multi-Society Task Force on colorectal cancer. *Gastroenterology*, *147*(2), 502–526. https://doi.org/10.1053/j.gastro.2014.04.001

Giraud, A., Matic, I., Tenaillon, O., Clara, A., Radman, M., Fons, M., & Taddei, F. (2001). Costs and benefits of high mutation rates: adaptive evolution of bacteria in the mouse gut. *Science (New York, N.Y.)*, *291*(5513), 2606–2608. https://doi.org/10.1126/science.1056421

Grasso, C. S., Giannakis, M., Wells, D. K., Hamada, T., Mu, X. J., Quist, M., Nowak, J. A., Nishihara, R., Qian, Z. R., Inamura, K., Morikawa, T., Nosho, K., Abril-Rodriguez, G., Connolly, C., Escuin-Ordinas, H., Geybels, M. S., Grady, W. M., Hsu, L., Hu-Lieskovan, S., … Peters, U. (2018). Genetic Mechanisms of Immune Evasion in Colorectal Cancer. *Cancer Discovery*, *8*(6), 730 LP – 749. https://doi.org/10.1158/2159-8290.CD-17-1327

Grin, I., & Ishchenko, A. A. (2016). An interplay of the base excision repair and mismatch repair pathways in active DNA demethylation. *Nucleic Acids Research*, *44*(8), 3713–3727. https://doi.org/10.1093/nar/gkw059

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. *The New England Journal of Medicine*, *342*(2), 69–77. https://doi.org/10.1056/NEJM200001133420201

Gu, Y., Parker, A., Wilson, T. M., Bai, H., Chang, D.-Y., & Lu, A.-L. (2002). Human MutY homolog, a DNA glycosylase involved in base excision repair, physically and functionally interacts with mismatch repair proteins human MutS homolog 2/human MutS homolog 6. *The Journal of Biological Chemistry*, *277*(13), 11135–11142. https://doi.org/10.1074/jbc.M108618200

Guo, Q., Lakatos, E., Al Bakir, I., Curtius, K., Graham, T. A., & Mustonen, V. (2021). The mutational signatures of formalin fixation on the human genome. *BioRxiv*, 2021.03.11.434918. https://doi.org/10.1101/2021.03.11.434918

Hause, R. J., Pritchard, C. C., Shendure, J., & Salipante, S. J. (2016). Classification and characterization of microsatellite instability across 18 cancer types. *Nature Medicine*, *22*(11), 1342–1350. https://doi.org/10.1038/nm.4191

Hegan, D. C., Narayanan, L., Jirik, F. R., Edelmann, W., Liskay, R. M., & Glazer, P. M. (2006). Differing patterns of genetic instability in mice deficient in the mismatch repair genes Pms2, Mlh1, Msh2, Msh3 and Msh6. *Carcinogenesis*, *27*(12), 2402–2408. https://doi.org/10.1093/carcin/bgl079

Hsieh, P., & Zhang, Y. (2017). The Devil is in the details for DNA mismatch repair. *Proceedings of the National Academy of Sciences*, *114*(14), 3552 LP – 3554.

https://doi.org/10.1073/pnas.1702747114

Huth, C., Kloor, M., Voigt, A. Y., Bozukova, G., Evers, C., Gaspar, H., Tariverdian, M., Schirmacher, P., von Knebel Doeberitz, M., & Bläker, H. (2012). The molecular basis of EPCAM expression loss in Lynch syndrome-associated tumors. *Modern Pathology : An Official Journal of the United States and Canadian Academy of Pathology, Inc*, *25*(6), 911–916. https://doi.org/10.1038/modpathol.2012.30

Kambara, T., Simms, L. A., Whitehall, V. L. J., Spring, K. J., Wynter, C. V. A., Walsh, M. D., Barker, M. A., Arnold, S., McGivern, A., Matsubara, N., Tanaka, N., Higuchi, T., Young, J., Jass, J. R., & Leggett, B. A. (2004). BRAF mutation is associated with DNA methylation in serrated polyps and cancers of the colorectum. *Gut*, *53*(8), 1137 LP – 1144. https://doi.org/10.1136/gut.2003.037671

Kim, T.-M., Laird, P. W., & Park, P. J. (2013). The landscape of microsatellite instability in colorectal and endometrial cancer genomes. *Cell*, *155*(4), 858–868. https://doi.org/10.1016/j.cell.2013.10.015

Kloor, M., & von Knebel Doeberitz, M. (2016). The Immune Biology of Microsatellite-Unstable Cancer. *Trends in Cancer*, *2*(3), 121–133. https://doi.org/10.1016/j.trecan.2016.02.004

Lai, Y., Budworth, H., Beaver, J. M., Chan, N. L. S., Zhang, Z., McMurray, C. T., & Liu, Y. (2016). Crosstalk between MSH2–MSH3 and polβ promotes trinucleotide repeat expansion during base excision repair. *Nature Communications*, *7*(1), 12465. https://doi.org/10.1038/ncomms12465

Lakatos, E., Williams, M. J., Schenck, R. O., Cross, W. C. H., Househam, J., Zapata, L., Werner, B., Gatenbee, C., Robertson-Tessi, M., Barnes, C. P., Anderson, A. R. A., Sottoriva, A., & Graham, T. A. (2020). Evolutionary dynamics of neoantigens in growing tumors. *Nature Genetics*, *52*(10), 1057–1066. https://doi.org/10.1038/s41588-020-0687-1

Lanza, G., Gafà, R., Santini, A., Maestri, I., Guerzoni, L., & Cavazzini, L. (2006). Immunohistochemical test for MLH1 and MSH2 expression predicts clinical outcome in stage II and III colorectal cancer patients. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *24*(15), 2359–2367. https://doi.org/10.1200/JCO.2005.03.2433

Larkin, J., Chiarion-Sileni, V., Gonzalez, R., Grob, J. J., Cowey, C. L., Lao, C. D., Schadendorf, D., Dummer, R., Smylie, M., Rutkowski, P., Ferrucci, P. F., Hill, A., Wagstaff, J., Carlino, M. S., Haanen, J. B., Maio, M., Marquez-Rodas, I., McArthur, G. A., Ascierto, P. A., … Wolchok, J. D. (2015). Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. *The New England Journal of Medicine*, *373*(1), 23–34. https://doi.org/10.1056/NEJMoa1504030

Le, D. T., Durham, J. N., Smith, K. N., Wang, H., Bartlett, B. R., Aulakh, L. K., Lu, S., Kemberling, H., Wilt, C., Luber, B. S., Wong, F., Azad, N. S., Rucki, A. A., Laheru, D., Donehower, R., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., … Diaz, L. A. J. (2017). Mismatch repair deficiency predicts response of solid tumors to PD-1 blockade. *Science (New York, N.Y.)*, *357*(6349), 409–413. https://doi.org/10.1126/science.aan6733

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., … Diaz, L. A. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, *372*(26), 2509–2520. https://doi.org/10.1056/NEJMoa1500596

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., Bioulac-Sage, P., Prévôt, S., Azoulay, D., Paradis, V., Imbeaud, S., Deleuze, J.-F., & Zucman-Rossi, J. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nature Communications*, *8*(1), 1315. https://doi.org/10.1038/s41467-017-01358-x

López, S., Lim, E. L., Horswell, S., Haase, K., Huebner, A., Dietzen, M., Mourikis, T. P., Watkins, T. B. K., Rowan, A., Dewhurst, S. M., Birkbak, N. J., Wilson, G. A., Van Loo, P., Jamal-Hanjani, M., Swanton, C., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Watkins, T. B. K., … Consortium, Tracer. (2020). Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. *Nature Genetics*, *52*(3), 283–293. https://doi.org/10.1038/s41588-020-0584-7

Lynch, H. T., Lynch, P. M., Lanspa, S. J., Snyder, C. L., Lynch, J. F., & Boland, C. R. (2009). Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. *Clinical Genetics*, *76*(1), 1–18. https://doi.org/10.1111/j.1399-0004.2009.01230.x

Maddamsetti, R., & Grant, N. A. (2020). Divergent Evolution of Mutation Rates and Biases in the Long-Term Evolution Experiment with Escherichia coli. *Genome Biology and Evolution*, *12*(9), 1591–1603. https://doi.org/10.1093/gbe/evaa178

Marcus, L., Lemery, S. J., Keegan, P., & Pazdur, R. (2019). FDA Approval Summary: Pembrolizumab for the Treatment of Microsatellite Instability-High Solid Tumors. *Clinical Cancer Research*, *25*(13), 3753 LP – 3758. https://doi.org/10.1158/1078-0432.CCR-18-4070

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R. S., Zborowska, E., Kinzler, K. W., & Vogelstein, B. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science (New York, N.Y.)*, *268*(5215), 1336–1338. https://doi.org/10.1126/science.7761852

Marti, T. M., Kunz, C., & Fleck, O. (2002). DNA mismatch repair and mutation avoidance pathways. *Journal of Cellular Physiology*, *191*(1), 28–41. https://doi.org/10.1002/jcp.10077

Matic, I. (2019). Mutation Rate Heterogeneity Increases Odds of Survival in Unpredictable Environments. *Molecular Cell*, *75*(3), 421–425. https://doi.org/10.1016/j.molcel.2019.06.029

Mazurek, A., Berardini, M., & Fishel, R. (2002). Activation of human MutS homologs by 8-oxo-guanine DNA damage. *The Journal of Biological Chemistry*, *277*(10), 8260–8266. https://doi.org/10.1074/jbc.M111269200

McFarland, C. D., Mirny, L. A., & Korolev, K. S. (2014). Tug-of-war between driver and passenger mutations in cancer and other adaptive processes. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(42), 15138–15143. https://doi.org/10.1073/pnas.1404341111

McGranahan, N., Favero, F., de Bruin, E. C., Birkbak, N. J., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Science Translational Medicine*, *7*(283), 283ra54. https://doi.org/10.1126/scitranslmed.aaa1408

McGranahan, N., Furness, A. J. S., Rosenthal, R., Ramskov, S., Lyngaa, R., Saini, S. K., Jamal-Hanjani, M., Wilson, G. A., Birkbak, N. J., Hiley, C. T., Watkins, T. B. K., Shafi, S., Murugaesu, N., Mitter, R., Akarca, A. U., Linares, J., Marafioti, T., Henry, J. Y., Van Allen, E. M., … Swanton, C. (2016). Clonal neoantigens elicit T cell immunoreactivity and sensitivity to immune checkpoint blockade. *Science*, *351*(6280), 1463 LP – 1469.

https://doi.org/10.1126/science.aaf1490

Mensenkamp, A. R., Vogelaar, I. P., van Zelst-Stams, W. A. G., Goossens, M., Ouchene, H., Hendriks-Cornelissen, S. J. B., Kwint, M. P., Hoogerbrugge, N., Nagtegaal, I. D., & Ligtenberg, M. J. L. (2014). Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. *Gastroenterology*, *146*(3), 643-646.e8. https://doi.org/10.1053/j.gastro.2013.12.002

Miura, S., Vu, T., Deng, J., Buturla, T., Oladeinde, O., Choi, J., & Kumar, S. (2020). Power and pitfalls of computational methods for inferring clone phylogenies and mutation orders from bulk sequencing data. *Scientific Reports*, *10*(1), 3498. https://doi.org/10.1038/s41598-020-59006-2

Moxon, R., Bayliss, C., & Hood, D. (2006). Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annual Review of Genetics*, *40*, 307—333. https://doi.org/10.1146/annurev.genet.40.110405.090442

Muller, H. J. (1964). The relation of recombination to mutational advance. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, *1*(1), 2–9. https://doi.org/https://doi.org/10.1016/0027-5107(64)90047-8

Muzny, D. M., Bainbridge, M. N., Chang, K., Dinh, H. H., Drummond, J. A., Fowler, G., Kovar, C. L., Lewis, L. R., Morgan, M. B., Newsham, I. F., Reid, J. G., Santibanez, J., Shinbrot, E., Trevino, L. R., Wu, Y.-Q., Wang, M., Gunaratne, P., Donehower, L. A., Creighton, C. J., … group, T. source sites and disease working. (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, *487*(7407), 330–337. https://doi.org/10.1038/nature11252

National institute for health and care excellence (NICE). (2017). *NICE guidelines (DG27): Molecular testing strategies for Lynch syndrome in people with colorectal cancer*.

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 (Oxford, England)*, *30*(7), 1015–1016. https://doi.org/10.1093/bioinformatics/btt755

Oh, E., Choi, Y.-L., Kwon, M. J., Kim, R. N., Kim, Y. J., Song, J.-Y., Jung, K. S., & Shin, Y. K. (2015). Comparison of Accuracy of Whole-Exome Sequencing with Formalin-Fixed Paraffin-Embedded and Fresh Frozen Tissue Samples. *PloS One*, *10*(12), e0144162. https://doi.org/10.1371/journal.pone.0144162

Overman, M. J., Lonardi, S., Wong, K. Y. M., Lenz, H.-J., Gelsomino, F., Aglietta, M., Morse, M. A., Van Cutsem, E., McDermott, R., Hill, A., Sawyer, M. B., Hendlisz, A., Neyns, B., Svrcek, M., Moss, R. A., Ledeine, J.-M., Cao, Z. A., Kamble, S., Kopetz, S., & André, T. (2018). Durable Clinical Benefit With Nivolumab Plus Ipilimumab in DNA Mismatch Repair-Deficient/Microsatellite Instability-High Metastatic Colorectal Cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, *36*(8), 773–779. https://doi.org/10.1200/JCO.2017.76.9901

Overman, M. J., McDermott, R., Leach, J. L., Lonardi, S., Lenz, H.-J., Morse, M. A., Desai, J., Hill, A., Axelson, M., Moss, R. A., Goldberg, M. V, Cao, Z. A., Ledeine, J.-M., Maglinte, G. A., Kopetz, S., & André, T. (2017). Nivolumab in patients with metastatic DNA mismatch repair-deficient or microsatellite instability-high colorectal cancer (CheckMate 142): an open-label, multicentre, phase 2 study. *The Lancet Oncology*, *18*(9), 1182–1191. https://doi.org/10.1016/S1470-2045(17)30422-9

Parsons, M. T., Buchanan, D. D., Thompson, B., Young, J. P., & Spurdle, A. B. (2012). Correlation of tumour BRAF mutations and MLH1 methylation with germline mismatch repair (MMR) gene mutation status: a literature review assessing utility of tumour features for MMR variant classification. *Journal of Medical Genetics*, *49*(3), 151 LP – 157. https://doi.org/10.1136/jmedgenet-2011-100714

Plaschke, J., Krüger, S., Jeske, B., Theissig, F., Kreuz, F. R., Pistorius, S., Saeger, H. D., Iaccarino, I., Marra, G., & Schackert, H. K. (2004). Loss of MSH3 protein expression is frequent in MLH1-deficient colorectal cancer and is associated with disease progression. *Cancer Research*, *64*(3), 864–870. https://doi.org/10.1158/0008-5472.can-03-2807

Poynter, J. N., Siegmund, K. D., Weisenberger, D. J., Long, T. I., Thibodeau, S. N., Lindor, N., Young, J., Jenkins, M. A., Hopper, J. L., Baron, J. A., Buchanan, D., Casey, G., Levine, A. J., Le Marchand, L., Gallinger, S., Bapat, B., Potter, J. D., Newcomb, P. A., Haile, R. W., & Laird, P. W. (2008). Molecular characterization of MSI-H colorectal cancer by MLHI promoter methylation, immunohistochemistry, and mismatch repair germline mutation screening. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *17*(11), 3208–3215. https://doi.org/10.1158/1055-9965.EPI-08-0512

Rampias, T., Karagiannis, D., Avgeris, M., Polyzos, A., Kokkalis, A., Kanaki, Z., Kousidou, E., Tzetis, M., Kanavakis, E., Stravodimos, K., Manola, K. N., Pantelias, G. E., Scorilas, A., & Klinakis, A. (2019). The lysine-specific methyltransferase KMT2C/MLL3 regulates DNA repair components in cancer. *EMBO Reports*, *20*(3), e46821. https://doi.org/https://doi.org/10.15252/embr.201846821

Roswell, M., Dushoff, J., & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos*, *130*(3), 321–338. https://doi.org/https://doi.org/10.1111/oik.07202

Russo, M., Crisafulli, G., Sogari, A., Reilly, N. M., Arena, S., Lamba, S., Bartolini, A., Amodio, V., Magrì, A., Novara, L., Sarotto, I., Nagel, Z. D., Piett, C. G., Amatu, A., Sartore-Bianchi, A., Siena, S., Bertotti, A., Trusolino, L., Corigliano, M., … Bardelli, A. (2019). Adaptive mutability of colorectal cancers in response to targeted therapies. *Science*, *366*(6472), 1473 LP – 1480. https://doi.org/10.1126/science.aav4474

Rustgi, A. K. (2013). BRAF: a driver of the serrated pathway in colon cancer. *Cancer Cell*, *24*(1), 1–2. https://doi.org/10.1016/j.ccr.2013.06.008

Salem, M. E., Bodor, J. N., Puccini, A., Xiu, J., Goldberg, R. M., Grothey, A., Korn, W. M., Shields, A. F., Worrilow, W. M., Kim, E. S., Lenz, H.-J., Marshall, J. L., & Hall, M. J. (2020). Relationship between MLH1, PMS2, MSH2 and MSH6 gene-specific alterations and tumor mutational burden in 1057 microsatellite instability-high solid tumors. *International Journal of Cancer*, *147*(10), 2948–2956. https://doi.org/https://doi.org/10.1002/ijc.33115

Sanders, M. A., Vöhringer, H., Forster, V. J., Moore, L., Campbell, B. B., Hooks, Y., Edwards, M., Bianchi, V., Coorens, T. H. H., Butler, T. M., Lee-Six, H., Robinson, P. S., Flensburg, C., Bilardi, R. A., Majewski, I. J., Reschke, A., Cairney, E., Crooks, B., Lindhorst, S., … Campbell, P. J. (2021). Life without mismatch repair. *BioRxiv*, 2021.04.14.437578. https://doi.org/10.1101/2021.04.14.437578

Schaaper, R. M., & Dunn, R. L. (1987). Spectra of spontaneous mutations in Escherichia coli strains defective in mismatch correction: the nature of in vivo DNA replication errors. *Proceedings of the National Academy of Sciences*, *84*(17), 6220 LP – 6224. https://doi.org/10.1073/pnas.84.17.6220

Schenck, R. O., Lakatos, E., Gatenbee, C., Graham, T. A., & Anderson, A. R. A. (2019).

NeoPredPipe: high-throughput neoantigen prediction and recognition potential pipeline. *BMC Bioinformatics*, *20*(1), 264. https://doi.org/10.1186/s12859-019-2876-4

Schrock, A. B., Ouyang, C., Sandhu, J., Sokol, E., Jin, D., Ross, J. S., Miller, V. A., Lim, D., Amanam, I., Chao, J., Catenacci, D., Cho, M., Braiteh, F., Klempner, S. J., Ali, S. M., & Fakih, M. (2019). Tumor mutational burden is predictive of response to immune checkpoint inhibitors in MSI-high metastatic colorectal cancer. *Annals of Oncology*, *30*(7), 1096–1103. https://doi.org/10.1093/annonc/mdz134

Smyrk, T. C., Watson, P., Kaul, K., & Lynch, H. T. (2001). Tumor-infiltrating lymphocytes are a marker for microsatellite instability in colorectal carcinoma. *Cancer*, *91*(12), 2417–2422. https://doi.org/https://doi.org/10.1002/1097-0142(20010615)91:12<2417::AID-CNCR1276>3.0.CO;2-U

Temko, D., Van Gool, I. C., Rayner, E., Glaire, M., Makino, S., Brown, M., Chegwidden, L., Palles, C., Depreeuw, J., Beggs, A., Stathopoulou, C., Mason, J., Baker, A.-M., Williams, M., Cerundolo, V., Rei, M., Taylor, J. C., Schuh, A., Ahmed, A., … Tomlinson, I. (2018). Somatic POLE exonuclease domain mutations are early events in sporadic endometrial and colorectal carcinogenesis, determining driver mutational landscape, clonal neoantigen burden and immune response. *The Journal of Pathology*, *245*(3), 283–296. https://doi.org/10.1002/path.5081

Tenaillon, O., Taddei, F., Radman, M., & Matic, I. (2001). Second-order selection in bacterial evolution: selection acting on mutation and recombination rates in the course of adaptation. *Research in Microbiology*, *152*(1), 11–16. https://doi.org/https://doi.org/10.1016/S0923-2508(00)01163-3

Touat, M., Li, Y. Y., Boynton, A. N., Spurr, L. F., Iorgulescu, J. B., Bohrson, C. L., Cortes-Ciriano, I., Birzu, C., Geduldig, J. E., Pelton, K., Lim-Fat, M. J., Pal, S., Ferrer-Luna, R., Ramkissoon, S. H., Dubois, F., Bellamy, C., Currimjee, N., Bonardi, J., Qian, K., … Ligon, K. L. (2020). Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature*, *580*(7804), 517–523. https://doi.org/10.1038/s41586-020-2209-9

Trigos, A. S., Pearson, R. B., Papenfuss, A. T., & Goode, D. L. (2018). How the evolution of multicellularity set the stage for cancer. *British Journal of Cancer*, *118*(2), 145–152. https://doi.org/10.1038/bjc.2017.398

Turnbull, C., Scott, R. H., Thomas, E., Jones, L., Murugaesu, N., Pretty, F. B., Halai, D., Baple, E., Craig, C., Hamblin, A., Henderson, S., Patch, C., O'Neill, A., Devereau, A., Smith, K., Martin, A. R., Sosinsky, A., McDonagh, E. M., Sultana, R., … Caulfield, M. J. (2018). The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ (Clinical Research Ed.)*, *361*, k1687. https://doi.org/10.1136/bmj.k1687

Tutlewska, K., Lubinski, J., & Kurzawski, G. (2013). Germline deletions in the EPCAM gene as a cause of Lynch syndrome - literature review. *Hereditary Cancer in Clinical Practice*, *11*(1), 9. https://doi.org/10.1186/1897-4287-11-9

Veigl, M. L., Kasturi, L., Olechnowicz, J., Ma, A. H., Lutterbaugh, J. D., Periyasamy, S., Li, G. M., Drummond, J., Modrich, P. L., Sedwick, W. D., & Markowitz, S. D. (1998). Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *Proceedings of the National Academy of Sciences of the United States of America*, *95*(15), 8698–8702. https://doi.org/10.1073/pnas.95.15.8698

Venderbosch, S., Nagtegaal, I. D., Maughan, T. S., Smith, C. G., Cheadle, J. P., Fisher, D., Kaplan, R., Quirke, P., Seymour, M. T., Richman, S. D., Meijer, G. A., Ylstra, B., Heideman, D. A. M., de Haan, A. F. J., Punt, C. J. A., & Koopman, M. (2014). Mismatch repair status and BRAF mutation status in metastatic colorectal cancer patients: a pooled analysis of the CAIRO, CAIRO2, COIN, and FOCUS studies. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *20*(20), 5322–5330. https://doi.org/10.1158/1078-0432.CCR-14-0332

von Loga, K., Woolston, A., Punta, M., Barber, L. J., Griffiths, B., Semiannikova, M., Spain, G., Challoner, B., Fenwick, K., Simon, R., Marx, A., Sauter, G., Lise, S., Matthews, N., & Gerlinger, M. (2020). Extreme intratumour heterogeneity and driver evolution in mismatch repair deficient gastro-oesophageal cancer. *Nature Communications*, *11*(1), 139. https://doi.org/10.1038/s41467-019-13915-7

Wielgoss, S., Barrick, J. E., Tenaillon, O., Wiser, M. J., Dittmar, W. J., Cruveiller, S., Chane-Woon-Ming, B., Médigue, C., Lenski, R. E., & Schneider, D. (2013). Mutation rate dynamics in a bacterial population reflect tension between adaptation and genetic load. *Proceedings of the National Academy of Sciences*, *110*(1), 222 LP – 227. https://doi.org/10.1073/pnas.1219574110

Woerner, S. M., Kloor, M., Schwitalle, Y., Youmans, H., Doeberitz, M. von K., Gebert, J., & Dihlmann, S. (2007). The putative tumor suppressor AIM2 is frequently affected by different genetic alterations in microsatellite unstable colon cancers. *Genes, Chromosomes & Cancer*, *46*(12), 1080–1089. https://doi.org/10.1002/gcc.20493

Yip, S., Miao, J., Cahill, D. P., Iafrate, A. J., Aldape, K., Nutt, C. L., & Louis, D. N. (2009). MSH6 mutations arise in glioblastomas during temozolomide therapy and mediate temozolomide resistance. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *15*(14), 4622–4629. https://doi.org/10.1158/1078-0432.CCR-08-3012

Zapata, L., Caravagna, G., Williams, M. J., Lakatos, E., AbdulJabbar, K., Werner, B., Graham, T. A., & Sottoriva, A. (2020). dN/dS dynamics quantify tumour immunogenicity and predict response to immunotherapy. *BioRxiv*, 2020.07.21.215038. https://doi.org/10.1101/2020.07.21.215038

Zou, X., Koh, G. C. C., Nanda, A. S., Degasperi, A., Urgo, K., Roumeliotis, T. I., Agu, C. A., Badja, C., Momen, S., Young, J., Amarante, T. D., Side, L., Brice, G., Perez-Alonso, V., Rueda, D., Gomez, C., Bushell, W., Harris, R., Choudhary, J. S., … Consortium, G. E. R. (2021). A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer*. https://doi.org/10.1038/s43018-021-00200-0