INVESTIGATING THE CLINICOPATHOLOGICAL SPECTRUM AND ASSOCIATED GENETICS OF COLORECTAL CARCINOMA IN YOUNG (<60 YEARS OF AGE) PATIENTS IN THE WESTERN CAPE PROVINCE

ALESSANDRO PIETRO ALDERA

ALDALE001

Thesis Presented for the Degree of
DOCTOR OF PHILOSOPHY

Division of Human Genetics
Department of Pathology
UNIVERSITY OF CAPE TOWN

February 2025

Primary Supervisor: Professor Raj Ramesar
Co-supervisors: Professor Komala Pillay, Associate Professor Adam Boutall

PLAGIARISM DECLARATION

I, **Dr Alessandro Pietro Aldera**, hereby declare that the work on which this dissertation is based is my original work (except where acknowledgements indicate otherwise) and that neither the whole work nor any part of it has been, is being, or is to be submitted for another degree in this or any other university.

I empower the university to reproduce for the purpose of research either the whole or any portion of the contents in any manner whatsoever.

This thesis/dissertation has been submitted to the Turnitin module (similarity and originality checking software), and I confirm that my supervisor has seen my report and any concerns revealed by such have been resolved with my supervisor.

Name of student: Alessandro Pietro Aldera

Student number: ALDALE001

Signature : Signed by candidate

Date : 10/02/2025

LIST OF PUBLICATIONS

"I confirm that I have been granted permission by the University of Cape Town's Doctoral Degrees Board to include the following publication(s) in my PhD thesis, and where co-authorships are involved, my co-authors have agreed that I may include the publication(s)":

1. **Aldera AP**, Pillay K, Robertson B, Boutall A, Ramesar R. Genomic landscape of colorectal carcinoma in sub-Saharan Africa. *J Clin Pathol*. 2023. 76(1):5-10.

2. **Aldera AP**, van der Westhuizen J, Tsai WJ, Krause MJ, Yildiz S, Pillay K, Boutall A, Ramesar R. Investigating somatic variants and pathways in mismatch repair-deficient (dMMR) colorectal carcinoma in South Africa. *J Clin Pathol*. 2024 May 15.

3. **Aldera AP**, Veldhuizen GP, Cifci D, Tsai WJ, Pillay K, Boutall A, Kather J, Ramesar R. Deep Learning based on H&E histopathological slides Predicts MSI Status in Ethnically Heterogeneous Population of South Africa. *J Clin Pathol.* 2025 (under review).

4. **Aldera AP**, Owusu D, Biral L, Pillay K, Boutall A, Dave S, Ramesar R. Molecular Pathology of Colorectal Cancer in African Populations: Insights from Somatic Whole Exome and Transcriptome Sequencing. *J Pathol Clin Res*. 2025 (under review).

DEDICATION

This work is dedicated to my wife Dr Cassandra Bruce-Brand, my inspiration and true north. In the time that I have been holed up producing this work, you have climbed your own personal mountain, of which you should be equally proud. May we continue to share our passion for histopathology, explore the world, and grow our (animal) family.

Sienna, Dante, Apollo, Lupo and Alex, thank you for your patience, I promise that there will be more walks and couch time now.

To my parents Dushanka and Marco for their unwavering support and love, and my parents-in-law Pam and Alick for their support and encouragement.

ACKNOWLEDGEMENTS

To Prof. Raj Ramesar for his guidance, encouragement, contagious enthusiasm, and mentorship.

To Dr Roger James for his encouragement, support and tolerance of this endeavour.

To Prof. Komala Pillay and A/Prof. Adam Boutall for their co-supervision.

To Jenny Molde, Kayla Benjamin, Subash Govender, and the other JDW Pathology Inc. techs for re-cutting and staining the tissue sections for the artificial intelligence project.

To Subash Govender for her assistance with DNA extraction for the project.

To Dr Debbie Tsai for her assistance with collecting the study material and keeping me company while evaluating the slides.

To my collaborators May Krause and Safiye Yildiz, and the sequencing wizard Akshay Vanmali for their assistance with wet bench experiments in the Ramesar laboratory.

To Gregory Veldhuizen, Didem Cifci and Prof. Jakob Kather for their assistance and guidance with the artificial intelligence work.

To Dennis Owusu, Leo Biral and Sandeep Dave for their invaluable assistance with the exome sequencing and bioinformatics analysis.

PREFACE

Research Conceptualisation

The candidate undertook a literature review and wrote a research proposal under the supervision of Prof. Raj Ramesar. This was approved by the Department of Pathology Research Committee. This work was supported by independent funding obtained by the candidate through the National Health Laboratory Service Research Trust and research funding from the South African Medical Research Council obtained by Prof. Raj Ramesar as the principal investigator. The candidate conceived the study and all experiments in collaboration with the primary supervisor, Prof Raj Ramesar, and co-supervisors Prof. Komala Pillay and A/Prof. Adam Boutall.

Ethics Approval

Ethical approval was obtained from the University of Cape Town Human Research Ethics Committee (HREC) reference number: 202/2002. The principles of the Declaration of Helsinki and subsequent amendments were adhered to.

Experiments and Analysis

DNA extraction was performed in the Molecular Laboratory at JDW Pathology. The panel-based NGS experiment was conducted in the Ramesar Laboratory at the University of Cape Town. The Artificial Intelligence experiment was performed at the Kather Laboratory, University of Dresden, Germany. Whole-exome sequencing was conducted in the Dave Laboratory in the Department of Medicine, Duke University, North Carolina.

PLAN OF THIS THESIS

This thesis is presented in six chapters. **Chapter 1** provides a comprehensive background of colorectal carcinoma, including the molecular pathogenesis. This includes a state-of-the-art literature review (**paper 1**) summarising the molecular classification of colorectal carcinoma, with a particular focus on reviewing all the available literature pertinent to sub-Saharan Africa. The aims and objectives are also presented here. **Chapter 2** comprehensively details the clinicopathological characteristics of the entire research cohort.

Chapters 3-5 represent the original research (papers 2, 3 and 4) and are linked to achieve the objectives of this degree.

Chapter 3 (paper 2) investigates somatic mutations and pathways with panel sequencing, in a selected group of 32 dMMR CRC where no germline variants were detected diagnostically. In **Chapter 4 (paper 3)** whole slide scanned digital images of the entire research cohort were subjected to a machine learning platform to assess its effectiveness in predicting dMMR CRC in an ethnically heterogeneous South African cohort. **Chapter 5 (paper 4)** describes the somatic mutational landscape of the pMMR CRC subset utilising whole-exome sequencing technology.

The last chapter provides a general discussion of the most important findings, a review of the various studies' strengths and weaknesses, and a brief overview of perspectives for future research.

ABSTRACT

The incidence of colorectal carcinoma (CRC) in young patients is rising in sub-Saharan Africa and is set to become a major public health problem within the next decade. Despite this, there is a paucity of large-scale genomic studies in the subregion. To investigate driver genes, oncogenic signalling pathways and the spectrum of pathogenic variants, we retrospectively identified 197 CRC cases over a 5-year period.

Thirty-two mismatch repair-deficient (dMMR) cases, without known germline variants, were investigated with amplicon-based panel next-generation sequencing (NGS). Pathogenic or likely pathogenic variants were detected in the corresponding MMR gene in 14 of 18 (78%) MLH1/PMS2-deficient tumours, 5 of 8 (63%) MSH2/MSH6-deficient tumours, 1 of 4 (25%) tumours with isolated MSH6 loss, and 0 of 2 tumours with isolated PMS2 loss. Cases with a variant allele frequency suggesting a germline mutation were identified in *MLH1* (eight), *MSH2* (two) and *MSH6* (one). NGS-based strategies for Lynch syndrome screening are advised to detect the broad spectrum of disease-causing MMR gene variants in our population.

Resource constraints prohibit the rollout of universal MMR screening in sub-Saharan Africa. We sought to determine the performance of a deep learning model in our ethnically heterogeneous cohort. Our model yielded an AUROC of 0.91 (±0.02). Calibrating the classification threshold to 0.15, the overall sensitivity achieved in our cohort was optimised to 96% (95% CI 90-100) with a specificity of 60% (95% CI 52-82). This model could therefore be employed to accurately pre-screen for dMMR cases, thereby reducing the burden of downstream immunohistochemical and molecular testing in our resource-limited setting.

Whole-exome sequencing was performed on a subset of the research cohort. Eighty-three cases were included in the analysis (77 MSS, 4 MSI, 2 POL). *APC*, *TP53* and *KRAS* were among the most frequently mutated driver genes, although at a lower frequency than described in the literature. *BRAF* V600E mutations were absent. Although there were differences in the frequencies of mutations in the major driver genes, the frequencies of oncogenic pathway alterations were similar. *FAT4* (26%) and *TET2* (15%) have emerged as important novel driver genes in left-sided tumours and are potential therapeutic targets for further investigation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

AI	-	Artificial intelligence
APC	-	Adenomatous polyposis coli
CIN	-	Chromosomal instability
CNV	-	Copy number variant
CRC	-	Colorectal carcinoma
DL	-	Deep learning
DNA	-	Deoxyribose nucleic acid
dMMR	-	Mismatch repair deficient
ED	-	Exonuclease domain
FAP	-	Familial adenomatous polyposis
FFPE	-	Formalin fixed paraffin embedded
HIC	-	High-income country
IHC	-	Immunohistochemistry
IVD	-	In vitro diagnostics
MSI	-	Microsatellite instability
MSS	-	Microsatellite stable
NGS	-	Next-generation sequencing
NSAIDs	-	Nonsteroidal anti-inflammatory drugs
PCR	-	Polymerase chain reaction
pMMR	-	Mismatch repair proficient
NOS	-	Not otherwise specified
POLE	-	Polymerase ε
SCNAs	-	Somatic copy-number alterations
TCGA	-	The cancer genome atlas
TILs	-	Tumour infiltrating lymphocytes
VAF	-	Variant allele frequency
WES	-	Whole-exome sequencing
WNT	-	Wingless-related integration site

1. CHAPTER ONE - INTRODUCTION

1.1. Colorectal Carcinoma

1.1.1. Definition

Colorectal carcinoma (CRC) is a malignant epithelial neoplasm of the large bowel that shows glandular or mucinous differentiation.

1.1.2. Epidemiology

The worldwide incidence of CRC according to GLOBOCAN 2022 was 1 926 118 (third, amongst all cancers), and the absolute mortality was 903 859 (second, amongst all cancers).[1] Males are affected marginally more than females. Australia, Europe, North America and Asia account for most new cases of CRC. There is a decline in the incidence of CRC in many high-income countries (HIC) which likely reflects the trend of population level shifts to a healthier lifestyle and improved access to screening programmes.[2, 3] However, the incidence of CRC is rising in developing countries and is set to do so dramatically over the coming decade.[3]

1.1.3. Aetiology

Although most cases are sporadic, hereditary genetic predisposition is a significant risk factor for CRC; the risk varies depending on the gene and type of mutation (Table 1.1).

Established modifiable risk factors include processed and red meat, alcohol, and excess body fat.[4-6] Consuming a diet high in fibre and dairy products and increased physical activity decrease the risk.[6, 7] Nonsteroidal anti-inflammatory drugs (NSAIDs) also lower the risk.[8] Chronic inflammation of the large bowel, as occurs in inflammatory bowel disease (Crohn disease and ulcerative colitis) is another risk factor. In the tropics, chronic infections with various parasitic and bacterial organisms have been implicated in the development of CRC, presumably related to chronic inflammation.[9, 10]

Table 1.1: Summary of Hereditary Cancer Syndromes in CRC[11]

Syndrome	Genes Involved	Inheritance pattern	CRC risk
Lynch syndrome	Mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*, others)	AD	10–50%[a]
Familial adenomatous polyposis	*APC*	AD	100%[b]
MUTYH-associated polyposis	*MUTYH*	AR	60–70%[b]
NTHL1-associated polyposis	*NTHL1*	AR	Unknown
Polymerase proofreading–associated polyposis	*POLD1*, *POLE*	AD	30–70%[c]
Constitutional mismatch repair deficiency syndrome	Mismatch repair genes (*MLH1*, *MSH2*, *MSH6*, *PMS2*)	AR	Unknown
Hereditary mixed polyposis syndrome	*GREM1*	AD	Unknown
MSH3-associated polyposis	*MSH3*	AR	Unknown
AXIN2-associated polyposis	*AXIN2*	AD	Unknown
Immune deficiency–associated polyposis	Various	Various	Unknown
Serrated polyposis	Unknown	Unknown	Unknown
Juvenile polyposis syndrome	*SMAD4* or *BMPR1A*	AD	68%[d]
Peutz–Jeghers syndrome	*STK11* (*LKB1*)	AD	39%[c]
Cowden syndrome	*PTEN*	AD	9%[b]
Li–Fraumeni syndrome	*TP53*	AD	Unknown

AD, autosomal dominant; AR, autosomal recessive.

[a] Risk at 75 years (varies by gene, age, and sex). [b] Lifetime risk. [c] Risk at 65–70 years.

[d] Risk at 60 years.

1.1.4. Pathogenesis

CRC is thought to develop via one of three pathways. The majority of tumours develop through the adenoma-carcinoma sequence originally described by Vogelstein.[12] The remaining tumours develop via the microsatellite instability (MSI) pathway (hypermutant) or as a result of mutations in the proofreading domain of polymerase ε *(POLE)* (ultramutant). Hypermutant CRC is defined as more than 12 mutations per 10^6 base pairs.[13]

1.1.4.1. Chromosomal Instability

Approximately 84% of CRCs develop from an adenomatous polyp via the chromosomal instability (CIN) pathway (also referred to as microsatellite stable – MSS). Molecularly, these tumours are characterised by high levels of deoxyribonucleic acid (DNA) somatic copy-number alterations (SCNAs) with gains and losses affecting a small group of genes.[14] Biallelic inactivation of the adenomatous polyposis coli (*APC*) tumour suppressor gene is the hallmark of this pathway, occurring as an early event in the formation of an adenomatous polyp (Figure 1.1), and seen in up to 80% of CIN CRC.[12, 15] Somatic inactivation may either be due to a mutation or an epigenetic event. Germline mutations in *APC* lead to a hereditary form of CRC, familial adenomatous polyposis (FAP) which is an autosomal dominant condition. Inactivation of *APC* leads to the dysregulation of the WNT signalling pathway and the cytoplasmic accumulation of β-catenin, which then translocates to the nucleus and activates transcription factors such as *MYC* and *cyclin D1*.[16] This leads to increased cellular proliferation. Additional mutations in proto-oncogenes are required in the adenoma-carcinoma sequence. *KRAS*-activating mutations are the most frequent, occurring in approximately 50% of adenomas larger than 1cm in size.[17] Mutations in other tumour suppressor genes include *SMAD2* and *SMAD4* (effectors of the TGF-β signalling pathway) and *TP53*. Loss of function of *TP53* is a late event in the adenoma-carcinoma sequence and is found in 70-80% of CRC. Loss of function of a tumour suppressor gene is often caused by chromosomal deletions but may also occur due to CpG island hypermethylation and epigenetic silencing.

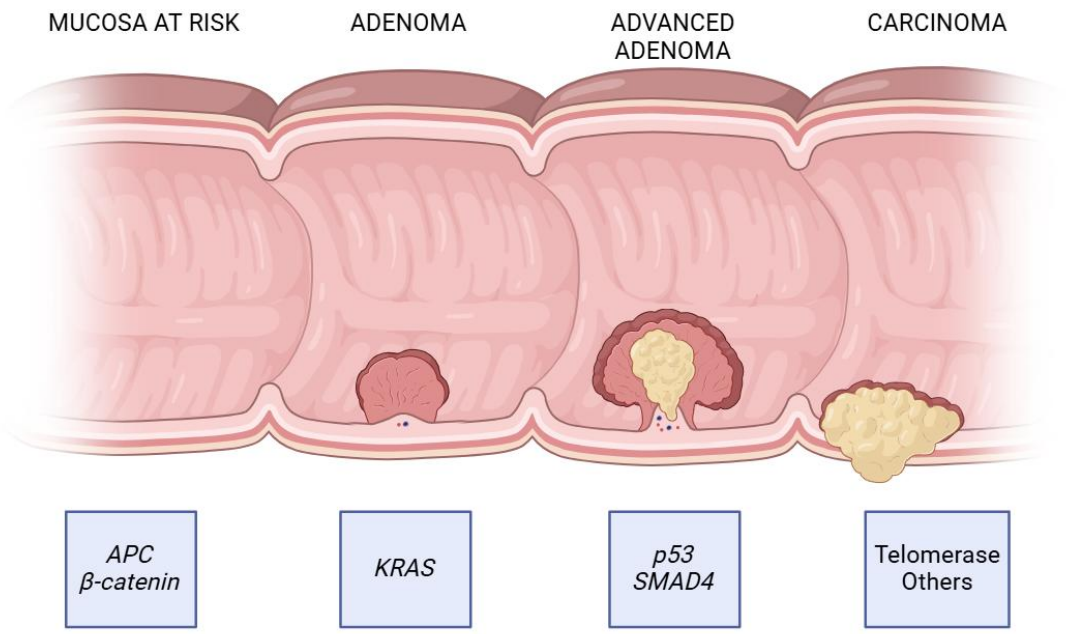


Figure 1.1: Morphological and Molecular Changes in the adenoma-carcinoma Sequence[18]

1.1.4.2. Hypermutated

MSI accounts for 13-16% of CRC and is caused by the loss of function of an MMR gene. Most cases are sporadic, caused by biallelic somatic inactivation, but in up to 20% of cases the patient carries a germline MMR (or rarely *EPCAM*) gene mutation. There are at least 7 MMR genes in humans. *MLH1*, *MSH2*, *MSH6* and *PMS2* are the most important in cancer biology. *MLH1* is the most frequently affected MMR gene, and the typical mechanism of loss of function is hypermethylation of the promoter region of the gene. This is frequently associated with *BRAF* V600E mutations. In patients with dMMR, mutations accumulate in microsatellite repeat regions of the DNA, leading to MSI.

1.1.4.2.1. Mismatch repair mechanisms

Microsatellites are areas of short tandem repeat sequences that typically (but not always) occur in non-coding regions of a gene. DNA damage accumulates over time due to base-base

mismatches, indels and slippage. Where microsatellites are in coding regions, this has the potential to generate mutations that are deleterious to cell function. DNA mismatch repair (MMR) is a cellular mechanism that safeguards the genome from such errors during replication.[19] There are several components involved in DNA MMR (Table 1.2). MMR has been extensively studied in *Escherichia coli* where the proteins involved are MutS and MutL. In humans, these complexes are heterodimers and are referred to as hMutS and hMutL. hMutSα (MSH2-MSH6) and hMutSβ (MSH2-MSH3) are responsible for detecting DNA base pair mismatches.[20] The hMutSα complex recognises mismatched bases of 1-2 nucleotides, whereas the hMutSβ complex recognises larger mismatched base sequences. These complexes bind to the unravelled DNA strand and recruit hMutLα (MLH1-PMS2), hMutLβ (MLH1-PMS1) or hMutLγ (MLH1-MLH3) complexes which are responsible for nicking the DNA strand and initiating DNA excision by EXO1. Once the segment of DNA containing the mismatched bases has been excised, the hMutL complex is responsible for terminating excision. This is followed by DNA resynthesis and nick ligation, which is mediated by Pol δ and DNA ligase 1.

Table 1.2: Major Components in Human DNA Mismatch Repair

Component	**Function**
hMutSα (MSH2-MSH6) hMutSβ (MSH2-MSH3)	DNA mismatch recognition
hMutLα (MLH1-PMS2) hMutLβ (MLH1-PMS1) hMutLγ (MLH1-MLH3)	Molecular matchmaker, endonuclease, termination of mismatch-provoked excision
Exo1	DNA excision
Pol δ	DNA re-synthesis
DNA ligase 1	Nick ligation

1.1.4.2.2. Microsatellite Instability Pathway to Colorectal Cancer

Most DNA mismatch mutations are silent as microsatellites typically occur in the non-coding regions of genes. However, some microsatellites occur in the coding or promoter regions of genes. In particular, *TGF-βRII* and the pro-apoptotic protein BAX are involved (Figure 1.2).[21] Under normal conditions, *TGF-β* inhibits colonic epithelial cell proliferation. Mutations in *TGF-βRII*, therefore, may lead to increased cell proliferation.

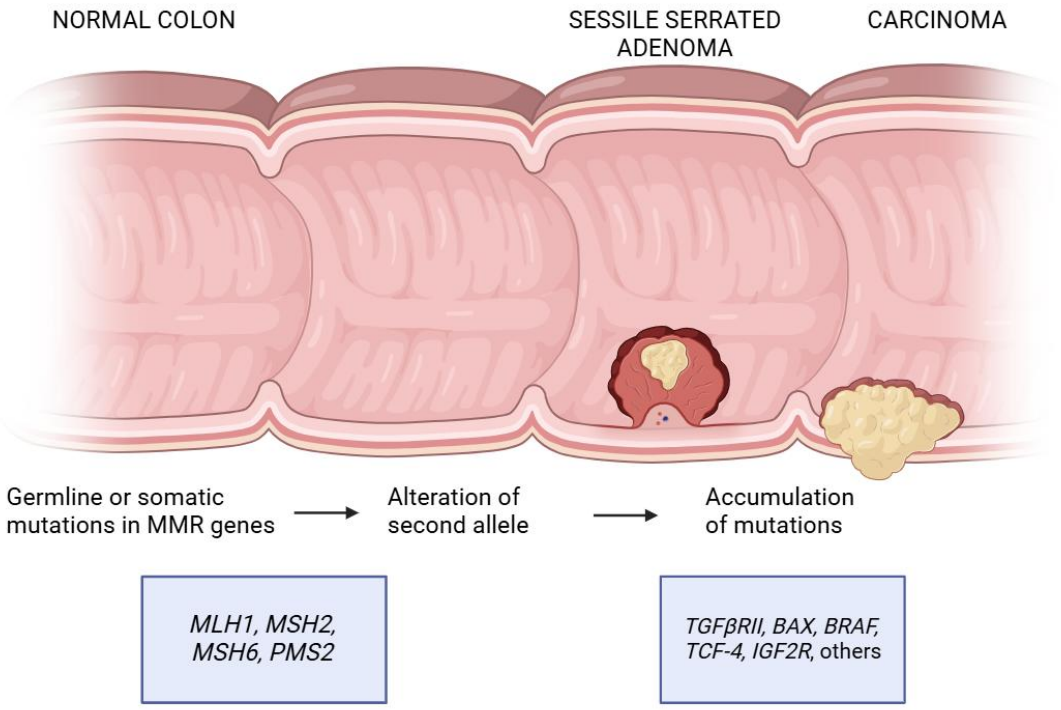


Figure 1.2: Morphological and Molecular Changes in the Mismatch Repair Pathway of CRC[18]

1.1.4.3. Ultramutated

The Cancer Genome Atlas (TCGA) identified a subset of ultramutated but MSS CRC in 2012 by analysing exome sequencing data from 224 sporadic CRC cases.[13] These tumours harboured recurrent inactivating somatic mutations within the exonuclease domain (ED) of *POLE*. Ultramutated CRC accounts for 1-3% of CRC, and features a characteristic nucleotide base

change spectrum with increased C-to-A transversions and an extremely high tumour mutation burden.[17] Most of these mutations are passenger mutations.

Both Pol ε and Pol δ comprise four subunits in humans. The largest subunit contains the catalytic and proofreading exonuclease active sites encoded by *POLE* and *POLD1,* respectively.[22] The most frequent *POLE* mutations occur at codons 286, 411, and 459 (Figure 1.3).[13, 23] Nine somatic and six germline pathogenic *POLE* ED mutations have been described in human cancer.[24] Although pathogenic *POLD1* mutations are described in other cancer types, none have been identified in CRC to date.

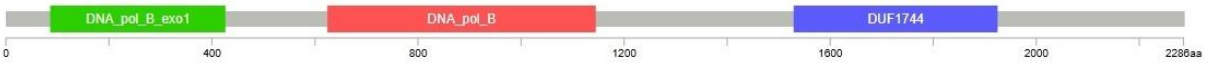


Figure 1.3: Lolliplot of the *POLE* gene, highlighting the position of the exonuclease domain at codon 86-426[25, 26]

1.1.5. Pathology

1.1.5.1. Histological Features that Suggest Microsatellite Instability

Most CRCs are histologically classified as adenocarcinoma, not otherwise specified (NOS). There are several well-defined histological subtypes that have strict diagnostic criteria and are associated with MSI and Lynch syndrome.[11] Mucinous adenocarcinoma is the most common subtype (5-20%) and is defined as having >50% pools of extracellular mucin, containing malignant epithelium arranged in clumps or as single cells. Signet ring cell adenocarcinoma is diagnosed when >50% of the tumour comprises cells with a globule of intracytoplasmic mucin, which peripherally displaces the nucleus. Medullary carcinoma contains sheets of malignant epithelial cells with abundant eosinophilic cytoplasm, prominent nucleoli, and a prominent infiltrate of lymphocytes.

Tumour-infiltrating lymphocytes (TILs) and a Crohn-like response are independently associated with MSI and a more favourable prognosis.[27] Several different methods for quantitatively

assessing TILs have been described. One such system is the Klintrup-Makinen Score, wherein the immune infiltrate is scored from 0 to 3 (Table 1.3).[28]

Table 1.3: Klintrup-Makinen Score for Assessment of Tumour Infiltrating Lymphocytes

Score	**Description**
0	no increase in inflammatory cells
1	patchy increase of inflammatory cells at the invasive margin
2	band-like infiltrate at the invasive margin with some destruction of cancer cell islets
3	very prominent inflammatory reaction, and frequent destruction of cancer cell islets

1.1.5.2. MMR Protein Immunohistochemistry

dMMR is assessed using IHC by staining for the four MMR proteins (MLH1, PMS2, MSH2, MSH6). This method is technically straightforward to perform and has a faster turnaround time compared to PCR-based methods. Interpretation of the IHC results is performed by a histopathologist and is relatively straightforward (Table 1.4). The tissue section should contain a built-in positive control, such as native colonic epithelium or lymphocytes. Loss of staining in tumour cells is interpreted as a significant finding, and the combination of staining loss can be used to identify the MMR gene, which is likely harbouring a mutation. MSH2 (and MLH1) can form heterodimers with MMR proteins other than MSH6 (and PMS2), and therefore, expression of MSH2 (and MLH1) is usually not lost when an MSH6 (and PMS2) mutation is present.

Table 1.4: Common Patterns of MMR Protein Loss with Immunohistochemistry

Pattern	**Interpretation**
MLH1 and PMS2	*MLH1* mutation (or inactivation)
MSH2 and MSH6	*MSH2* mutation *EPCAM* mutation (less common)
Isolated MSH6	*MSH6* mutation
Isolated PMS2	*PMS2* mutation

1.1.5.3. Molecular Tests for Microsatellite Instability

PCR has been widely used historically to detect MSI and remains the gold standard. This test, however, is more costly than IHC, has a longer turnaround time, and does not indicate which MMR gene is likely to be affected. The original National Cancer Institute Bethesda panel assessed five microsatellite repeats: two mononucleotide repeats (BAT25 and BAT26) and three dinucleotide repeats (D2S123, D5S346, and D17S250). Newer panels, however, assess mononucleotide repeats only (BAT25, BAT26, NR21, NR24 and NR27) as they are more sensitive and specific. Loss of 2 or more loci is considered MSI-H, loss of 1 locus MSI-L, and no shifted microsatellites is considered MSS.

Recently, bioinformatics tools have emerged that can be applied to data generated from large panel NGS or WES to infer MSI status. The most widely used tools are MSIsensor, MANTIS, and mSINGS. Most recently, MSIsensor-pro was developed, and it does not require a matched normal sample.[29] These tests are currently predominantly utilised in the research environment, although once diagnostic NGS becomes more widespread, this may become a more attractive option.

1.2. Literature Review (Paper 1)

Title: The Genomic Landscape of Colorectal Carcinoma in sub-Saharan Africa

Authors: Alessandro Pietro Aldera (1,4), Komala Pillay (1), Barbara Robertson (2), Adam Boutall (3), Raj Ramesar (4)

Journal: Journal of Clinical Pathology

Citation: Aldera AP, Pillay K, Robertson B, Boutall A, Ramesar R. Genomic landscape of colorectal carcinoma in sub-Saharan Africa. J Clin Pathol. 2023 Jan 1;76(1):5-10.

1. Division of Anatomical Pathology, National Health Laboratory Services / University of Cape Town, Cape Town, South Africa

2. Division of Radiation Oncology, Groote Schuur Hospital and University of Cape Town

3. Division of Surgical Gastroenterology, Groote Schuur Hospital and University of Cape Town

4. UCT MRC Genomic and Precision Medicine Research Unit, Division of Human genetics, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences and University of Cape Town

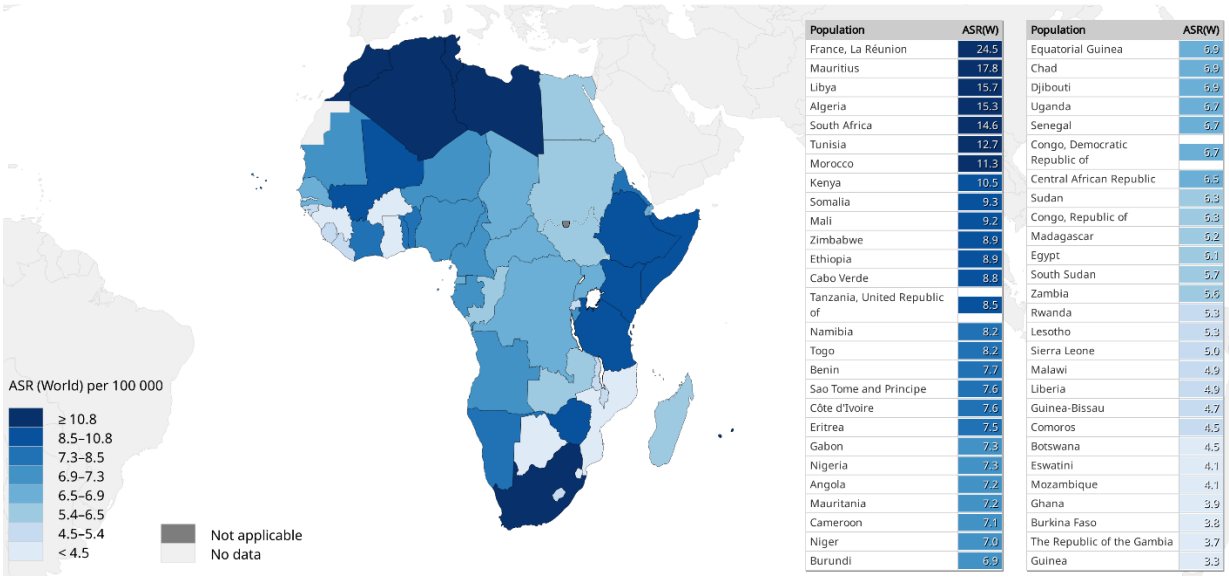
Key words: Colorectal carcinoma; molecular pathology; microsatellite instability

1.2.1. Abstract

Our understanding of the molecular classification of colorectal carcinoma (CRC) has evolved significantly over the past two decades. Tumours can be broadly categorised as microsatellite stable (MSS), microsatellite instability (MSI) or CpG island methylator phenotype (CIMP). Prognostic and predictive information is provided by these categories. The overwhelming majority of the data on which these categories are based have originated from Europe and North America. There is a dearth of information represented from Africa and indigenous African patients. However, some small studies and preliminary data have shown significant differences in all of these groups. The prevalence of MSI in Africa is consistently reported as almost double that of European and North American data. Interestingly, *BRAF* V600E mutations and *MLH1* promotor hypermethylation seem to be uncommon in Africa. The high proportion of MSI tumours is only partly accounted for by germline mutations in mismatch repair genes (Lynch syndrome) suggesting that there are likely to be other mechanisms at play. Within the MSS group, preliminary data suggest that the typical molecular pathways (WNT pathway activation) may not be as dominant in Africa. The purpose of this review it to summarize the current state of the molecular genetic landscape of CRC in Africa and provide insights into areas for further study.

Introduction

Colorectal carcinoma (CRC) is the third most common cancer worldwide (10%), and is the second leading cause of cancer-related deaths, accounting for 935 173 deaths in 2020.[1] The incidence of CRC is set to rise dramatically over the coming decade, and most of this growth is predicted to occur in low- and middle- income countries.[2] The International Agency for Research on Cancer (IARC), through the GLOBOCAN estimates, provides up to date statistics on CRC incidence and mortality rates.[3] In Africa, age-standardised incidence rates (ASR) are highest in the Indian Ocean islands of Mauritius (17.8) and Le Reunion (24.5), South Africa (14.6) and North Africa (11.3-15.7) (Figure 1.4). The rest of the continent shows a relatively low burden of disease, although this may be confounded by the lack of well-utilised centralised cancer registries. These figures are in stark contrast to the relatively high age-standardised incidence rates from Eastern Europe (45.3), Japan (38.5) and North America (31.2).



Population	ASR(W)
France, La Réunion	24.5
Mauritius	17.8
Libya	15.7
Algeria	15.3
South Africa	14.6
Tunisia	12.7
Morocco	11.3
Kenya	10.5
Somalia	9.3
Mali	9.2
Zimbabwe	8.9
Ethiopia	8.9
Cabo Verde	8.8
Tanzania, United Republic of	8.5
Namibia	8.2
Togo	8.2
Benin	7.7
Sao Tome and Principe	7.6
Côte d'Ivoire	7.6
Eritrea	7.5
Gabon	7.3
Nigeria	7.3
Angola	7.2
Mauritania	7.2
Cameroon	7.1
Niger	7.0
Burundi	6.9

Population	ASR(W)
Equatorial Guinea	6.9
Chad	6.9
Djibouti	6.9
Uganda	6.7
Senegal	6.7
Congo, Democratic Republic of	6.7
Central African Republic	6.5
Sudan	6.3
Congo, Republic of	6.3
Madagascar	6.2
Egypt	6.1
South Sudan	5.7
Zambia	5.6
Rwanda	5.3
Lesotho	5.3
Sierra Leone	5.0
Malawi	4.9
Liberia	4.9
Guinea-Bissau	4.7
Comoros	4.5
Botswana	4.5
Eswatini	4.1
Mozambique	4.1
Ghana	3.9
Burkina Faso	3.8
The Republic of the Gambia	3.7
Guinea	3.3

Figure 1.4: Estimated Age-Standardised Incidence Rates (ASR) of Colorectal Carcinoma in Africa in 2020. Data from GLOBOCAN 2020.[3]

Several authors have noted differences in the epidemiology and clinicopathological features of CRC in cohorts from countries in Africa, as compared to high-income countries (HIC).[4-11] These include a younger age of onset, more advanced disease at presentation, increased mucinous differentiation and increased frequency of microsatellite instability (MSI). More recently, larger studies have shown an increased frequency of left-sided tumours amongst young indigenous Africans.[12, 13] This is in contrast to data from Europe, Japan and the USA which shows decreasing frequency of left-sided disease.[14-16]

Despite significant advances in molecular pathology and increasingly sophisticated molecular-based cancer classification systems, very little data has been contributed by African studies. This review attempts to comprehensively summarise the available molecular data from all of the relevant studies originating from sub-Saharan Africa.

1.2.3. Current Concepts in the Molecular Classification of Colorectal Carcinoma

Traditionally, two molecular pathways have been recognised in CRC, namely, MSI and chromosomal instability (CIN).[17, 18] The majority of cases in both categories are sporadic (70-80%), while the remaining have a hereditary component.

CIN (also referred to as microsatellite stable – MSS) CRC develops via the classic adenoma-carcinoma sequence and is characterised by high frequency somatic copy number alterations (SCNA).[19] This pathway accounts for the majority (65-75%) of CRC cases. Loss of the *adenomatous polyposis coli* (*APC*) tumour suppressor gene is an early event.[17, 20] These tumours are also associated with *KRAS* activation and *TP53* inactivation.[21]

MSI CRC occurs in the setting of mutations (which may be somatic or germline) in the DNA mismatch repair (MMR) system. This pathway accounts for 13-16% of CRC globally. There are at least seven MMR genes in humans, the most significant of which, in cancer biology, include *MLH1, MSH2, MSH6* and *PMS2*. Other MMR genes which are less frequently mutated in CRC are *PMS1*, *MLH3* and *MSH3*.[22-24] Somatic-type variation and modification of DNA accounts for the majority of MSI CRC cases and results from epigenetic silencing (methylation) of the promotor sequence of *MLH1* and are associated with *BRAF* V600E point mutations.[25] Germline mutations (Lynch syndrome) occur as a result of mutations mostly in one of the four MMR genes or, less commonly, the *epithelial cell adhesion molecule* (*EPCAM*) gene. Germline-*EPCAM* contiguous deletions lead to transcriptional silencing of *MSH2* and manifestation of Lynch syndrome.[26] CpG island-methylator phenotype (CIMP) CRC shows hypermethylation of several key promotor CpG islands leading to epigenetic inactivation of tumour-suppressor and tumour-related genes.[27] CIMP CRC and sporadic MSI CRC show significant overlap in clinicopathological features and both develop via the serrated neoplasia pathway.

The molecular classification of CRC has evolved over the past decade to include novel molecular and transcriptomic data, with proposed classification systems provided by The Cancer Genome

Atlas (TCGA) project and Colorectal Cancer Subtyping Consortium (CRCSC), dominating recent literature.[28, 29]

The TCGA project investigated 276 cases of CRC by studying exome sequence, DNA copy number, promoter methylation and messenger RNA (mRNA) and microRNA expression. Based on this data, 13% of cases could be classified as hypermutated (MSI), 3% as ultramutated (MSI coupled with *POLE* or *POLD1* mutations) and the majority (84%) as CIN. TCGA also identified potentially therapeutic targets in *ERBB2* and *IGF2* amplifications.

CRCSC aggregated transcriptomic data on RNA expression from several studies and identified four major consensus molecular subtype (CMS) groups, namely CMS1 (MSI-immune), CMS2 (canonical), CMS3 (metabolic) and CMS4 (mesenchymal) (Figure 1.5). Almost all of the hypermutated MSI cancers fell into the CMS1 group (MSI-immune, 14%). These were associated with *MLH1* silencing, CIMP, *BRAF* mutations and a low number of SCNAs. In addition, these tumours were found to show evidence of immune activation with prominent CD8+ tumour-infiltrating lymphocytes. The significance of recognising this group is their potential response to immune checkpoint inhibitors.[21] The remaining MSS cases were categorised as CMS2 (canonical, 37%), CMS3 (metabolic, 13%) or CMS4 (mesenchymal, 23%). There was a residual group showing mixed features which represented 13% of cases. CMS2 was associated with a high number of SCNAs, and WNT and *MYC* activation. CMS3 showed a low number of SCNAs, CIMP-low, frequent *KRAS* mutations, metabolic dysregulation, and an epithelial signature. CMS4 showed a high number of SCNAs, stromal infiltration, *TGF-β* activation, epithelial-to-mesenchymal (EMT) transition, angiogenesis, and matrix remodelling.

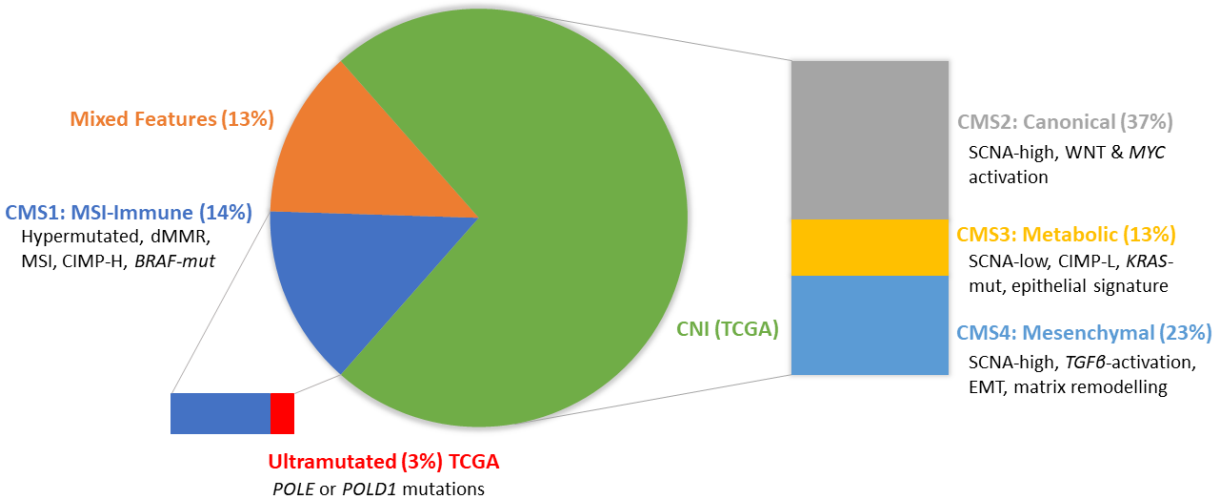


Figure 1.5: Molecular Classification of Colorectal Carcinoma Incorporating the Cancer Genome Atlas (TCGA) and Colorectal Cancer Subtyping Consortium (CRCSC) Schemas. CMS consensus molecular subtype; dMMR deficient mismatch repair; MSI microsatellite instability; CIMP CpG island methylator phenotype, SCNA somatic copy number alteration, EMT epithelial-mesenchymal transition.

Liu *et al*. further analysed TCGA data in 2018, which included 459 colorectal adenocarcinoma cases, and identified a group of genomically stable tumours which lacked hypermethylation and aneuploidy.[30] This molecular subgroup comprised 12.6% (58/459) of the cohort and enriched for earlier stage disease, later mean age at diagnosis, and *SOX9* and *PCBP1* mutations.

1.2.4. Colorectal Carcinoma in Africa

Although our understanding of the underlying genomics and pathogenesis of CRC has evolved significantly over the past two decades, most data have originated from groups in North America and Europe. Studies have suggested a unique molecular profile and disproportionate tendency for early age of onset among African Americans.[31-33] Several centres in Sub-Saharan Africa

have published work on CRC. Most of these studies are descriptive and have focused on the clinicopathological characteristics, and to a lesser extent the prevalence of MSI tumours. To our knowledge, there has only been one study in sub-Saharan Africa that has extensively interrogated the genetic basis of CRC in a cohort of unselected Nigerian patients.[13] This represents an obvious knowledge gap.

1.2.4.1. Interrogating the MSI Pathway

Studies from Southern, West and East Africa have consistently demonstrated an increased frequency of MSI CRC in their respective populations. There have been a variety of techniques used to demonstrate MSI, and these are not necessarily comprehensive or comparable between studies. Most work appears to have originated from South Africa, with studies in Gauteng, Mpumalanga, North-West, Western Cape and Northern Cape provinces.[4, 34, 35] These studies provide the clearest overview of the MMR protein expression status as determined by immunohistochemistry (IHC) in sub-Saharan Africa (Table 1.5). The findings of the Northern Cape study, however, are likely attributed to a unique founder mutation prevalent in the region.[36] Holla *et al*. recently examined a cohort of 27 indigenous African patients with CRC and found that MMR-proficient tumours occurred at a significantly younger age (44 vs. 35 years) and showed a predominance for the distal colon (73%).[37] Other studies with larger sample sizes have, however, found an older age of onset for MMR-proficient tumours.[13, 34] McCabe *et al*. described a greater proportion of dMSH2/6 (49%) in black patients compared to other ethnic groups who predominantly showed a dMLH1/PMS2 pattern.[34] Cronje *et al*. published data in 2009 on a cohort from Gauteng, Mpumalanga and North-West provinces, and found the rate of MSI to be 67% in that population group.[4] This study showed a very high proportion of tumours with isolated loss of expression of MSH6 (75/192) which has not been demonstrated in other studies in the same region and should probably be interpreted cautiously. These authors also did not perform PMS2 immunohistochemistry on their cohort.

Table 1.5: Distribution of Loss of Mismatch Repair Protein Immunohistochemistry Expression in All Available Sub-Saharan African Studies with Complete Data.

	Province (South Africa)	**n**	**MLH1 & PMS2**	**MSH2 & MSH6**	**Isolated MSH6**	**Isolated PMS2**	**dMMR**
Cronje *et al*. (2009)	Gauteng, Mpumalanga, North West	192	37	17	75	ND	**129 (67%)**
Vergouwe *et al*. (2013)	Northern Cape	78	11	6	ND	ND	**17 (22%)**
McCabe *et al*. (2019)	Gauteng	439	31	22	1	2	**56 (13%)**
Katsidzira *et al*. (2019)	Zimbabwe	89	4	2	0	0	**6 (7%)**
Holla *et al*. (2022)	Western Cape	27	6	3	0	1	**10 (37%)**
		825	**89 (11%)**	**50 (6%)**	**76 (9%)**	**3 (0.4%)**	**218 (26%)**

pMMR proficient mismatch repair; dMMR deficient mismatch repair; ND not done

Data from a histopathological study in Uganda showed that 27.3% of tumours met the Revised Bethesda guidelines for MSI testing.[10] These tumours were, unfortunately, not subjected to MSI testing.

Interestingly, recent data from a small Zimbabwean cohort showed a prevalence of MSI tumours (7%) similar to that described in the international literature.[38] These authors proceeded to germline testing with a multigene next generation sequencing (NGS) panel and found the frequency of Lynch syndrome in their cohort to be 3.3-5.6%. This is within the global prevalence range of Lynch syndrome (2-4%).[39] Some authors have suggested that the prevalence of germline variants in young patients with CRC is underestimated.[40, 41]

Studies using IHC and polymerase chain reaction (PCR)-based methods to detect MSI in Nigeria reported the frequency of MSI to be 23-43%.[7, 42, 43] Alatise *et al*. recently published work on a Nigerian cohort which likely represents the most sophisticated study to date in sub-Saharan Africa.[13] They used a hybrid approach to determine MSI status including MSIsensor (NGS data) and mismatch repair (MMR) IHC and found the rate of MSI to be 28.1% (18/64) and 21.3% (20/94), respectively. In Ghana, Raskin *et al*. showed a frequency of MSI-H in 29/71 (41%) of sequenced tumours.[44] Limited MMR IHC data is available from this study, but loss of MLH1/PMS2 was found in 3 tumours and isolated loss of MLH1 and PMS2 in one tumour each.

The 5 studies identified in sub-Saharan Africa which utilised PCR-based strategies, have shown the frequency of MSI-H tumours to be 11-43% (Table 1.6).[7, 34, 38, 44, 45] Irabor *et al*., Raskin *et al*. and Cronje *et al*. used the National Cancer Institute (NCI) 5 marker panel (BAT25, BAT26, D5S346 (APC), D17S250 (Mfd15CA) and D2S123) and McCabe used the pentaplex mononucleotide panel (BAT-25, BAT-26, NR-21, NR-24, and NR-27). Polymorphisms in the mononucleotide repeats BAT25 and BAT26 are well known to occur at increased frequency (18.4% and 12.6% of these alleles respectively) in the African American population.[46, 47] The use of a 5 marker panel and inclusion of sampled non-tumour tissue as a control should prevent overcalling MSI-H in this population group.

Table 1.6: Comparison of Molecular and Immunohistochemical Methods for Determining MSI in All Available Studies in Sub-Saharan Africa with Molecular Data

			MSI	
	n	**Country**	**MSI-H (PCR)**	**dMMR (IHC)**
Alatise *et al*. (2021)	157	Nigeria	28% (18/64)*	21% (20/94)
McCabe *et al*. (2019)	439	South Africa (Gauteng)	12% (31/267)	13% (56/439)
Katsidzira *et al*. (2019)	89	Zimbabwe	11% (6/53)	7% (6/89)
Raskin *et al*. (2013)	71	Ghana	41% (29/70)	**
Irabor *et al*. (2017)	83	Nigeria	43% (15/35)	ND
Cronje *et al*. (2009)	44	South Africa (Gauteng, Mpumalanga, North West)	25% (11/44)	34% (15/44)

MSI microsatellite instability; PCR polymerase chain reaction; dMMR deficient mismatch repair; IHC immunohistochemistry; ND not done; *NGS data using MSIsensor; **cannot be extracted accurately

Only the more recent studies have demonstrated comprehensive interrogation of MSI status with a full panel of MMR protein immunohistochemical stains or NGS-based methods (MSIsensor). The picture on MSI in Africa is still incomplete and requires more comprehensive studies comprising larger cohorts and utilising consistent standardised techniques.

1.2.4.2. Further Dissecting the MSI group – MLH1 hypermethylation, CIMP and BRAF V600E

Only three studies in sub-Saharan Africa have attempted to further differentiate between subgroups within the overarching category of MSI tumours. Alatise *et al*. found the rate of MSI in their Nigerian cohort to be more than double that reported in HIC.[13] However, germline testing on a subset of the cases found pathogenic mutations to be present in only 17% of the MSI-H cases which is similar to internationally reported rates (11-18.5%).[39, 48] In addition, less than a third of selected MSI-H cases showed *MLH1* promotor hypermethylation, only 8% were CIMP-H and *BRAF* V600E mutations were absent. These findings suggest an alternative molecular pathway leading to MSI-H in African patients.

In South Africa, Cronje *et al*. showed a high frequency of MSI-H and a low frequency of promotor methylation and *BRAF* V600E point mutations in their young indigenous African cohort.[45] McCabe *et al*. performed *BRAF* PCR on the 34 MSI cases in their study and found mutations in only 4/34 (12%).[34] All of these mutations were found in the "other ethnic group" category and all indigenous Africans with MSI were found to be *BRAF* V600E wildtype. Other groups have also shown a low frequency of *BRAF* V600E mutations in sporadic CRC in sub-Saharan Africa.[38, 44]

Interestingly, McCabe *et al*. found that 5/31 (16%) of patients with MSI-H on PCR were MMR proficient on IHC suggesting that aberrant (missense) staining patterns or other mismatch repair genes (*MSH3, MLH1, PMS1*) may be in play.[31] Discrepancies in the results from molecular-based techniques (PCR and NGS) and IHC have been highlighted by 4 studies in sub-Saharan Africa (Table 2).[13, 34, 38, 45] Interpretation of MMR protein immunohistochemistry is limited by several factors including tissue fixation, intratumoural heterogeneity, tumour infiltrating lymphocytes, and the presence of MMR gene missense mutations which may result in retained protein expression.[49] Hechtman *et al*. have recently shown that MMR protein expression may be retained in as many as 6% of MSI-H tumours with missense mutations in MMR genes.[50]

1.2.4.3. Preliminary Data on MSS tumours in Africa

The molecular pathogenesis of MSS CRC is dominated by activation of the WNT pathway in HIC.[28] This is typically characterised by inactivation mutations in *APC* and/or gain-of-function mutations in *CTNNB1*. Alatise *et al*. recently published the only study to comprehensively investigate somatic alterations in the molecular pathways associated with MSS CRC in sub-Saharan Africa.[13] As part of this study, the authors subjected 46 tumours from their cohort of 380 indigenous African patients to NGS using the 341 gene Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT) assay.[51] The findings were compared to a cohort of 1040 MSK patients with available genomic data. Significant differences were found in the frequency of *APC* mutations (39.1% Nigeria vs. 76.0% MSKCC, $P < 0.01$) and WNT pathway alterations (47.8% Nigeria vs. 81.9% MSKCC, $P < 0.001$). *CTNNB1* gene alterations were similar between the two groups (4.4 vs. 3.7%, $P = 0.68$). *RAS* pathway alterations were also more frequent in the Nigerian cohort (76.1 vs. 59.6% MSKCC, $P = 0.03$). *TP53* (71.7% Nigeria vs. 77.5% MSKCC, $P = 0.370$) and *KRAS* (56.5% Nigeria vs. 44.1% MSKCC, $P = 0.13$) somatic mutation frequencies were not different between the cohorts. Raskin *et al*. used Sanger sequencing to investigate *KRAS* exons 2 and 3 in 76 tumours from a Ghanaian cohort, and found activating mutations in 23/76 (30%) cases.[44]

Although sample numbers were limited in these studies, these findings represent a significant departure from the traditional molecular pathways in MSS in HIE.

1.2.5. Bridging the Genomic Rift

There is a somewhat fragmented, but significant and growing body of literature demonstrating an increased prevalence of MSI CRC in sub-Saharan Africa. The isolated studies which have examined paired germline material have not shown there to be a significantly increased prevalence of Lynch syndrome amongst these cases.[13, 38] Investigators attempting to tease out the group of sporadic MSI tumours with *BRAF* V600E and *MLH1* promotor methylation studies have failed to demonstrate that any significant proportion of these tumours falls into the CIMP-H category.[13, 34, 45] This has led authors to conclude that other mechanisms of *MLH1*

silencing may be at play. This remains uninvestigated. Interestingly, no studies in Africa have utilised *BRAF* V600E IHC to investigate their dMLH1/PMS2 tumours. This is an easily accessible, reliable, and relatively inexpensive alternative to PCR-based strategies.

In addition, the discordance between MSI status assigned by PCR (and more recently in some instances by NGS-based models) and by the routinely used MMR protein IHC markers, MLH1, MSH2, MSH6 and PMS2 have led some to suggest that other, largely uninvestigated, MMR genes (*MSH3*, *MLH3*, *PMS1*) may be at play. Larger studies based on well collected and preserved samples, well-populated databases and utilising modern high-throughput technologies are required from different centres to gain a clear picture of the MSI tumour landscape in Africa.

The MSS CRC group remains even more poorly characterised in Africa. Preliminary data suggests that there may be vast differences between the activated molecular pathways in this group of tumours in Africa compared to Europe and North America.

Gaining a clear understanding of the molecular pathways underpinning and driving the development of CRC is critical in a resource limited setting like Africa. This will result in better allocation of available resources for screening hereditary and sporadic CRC and precursor lesions, as well as inform which treatment options are likely to be most effective in the local context.

1.2.6. Author Statements

Contributor Statement: Conceptualisation (AA and RR), Literature review and drafting of manuscript (AA), Proof reading and editing (KP, AB, BR, RR).

1.2.7. References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: **Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA: a cancer journal for clinicians* 2021, **71**(3):209-249.

2. Karsa L, Lignini T, Patnick J, Lambert R, Sauvaget C: **The dimensions of the CRC problem.** *Best Pract Res Clin Gastroenterol* 2010, **24**(4):381-396.

3. Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, Bray F: **Cancer statistics for the year 2020: An overview.** *Int J Cancer 2021*, **149**(4):778-789.

4. Cronje L, Paterson A, Becker P: **Colorectal cancer in South Africa: a heritable cause suspected in many young black patients.** *South African Med J* 2009, **99**(2):103-106.

5. Saluja S, Alatise OI, Adewale A, Misholy J, Chou J, Gonen M, Weiser M, Kingham TP: **A comparison of colorectal cancer in Nigerian and North American patients: is the cancer biology different?** *Surgery* 2014, **156**(2):305-310.

6. Irabor D, Adedeji O: **Colorectal cancer in Nigeria: 40 years on. A review.** *Eur J Cancer Care* 2009, **18**(2):110-115.

7. Irabor DO, Oluwasola OA, Ogunbiyi OJ, Ogun OG, Okolo CA, Melas M, Gruber SB, Shi C, Raskin L**: Microsatellite instability is common in colorectal cancer in native Nigerians.** *Anticancer Res* 2017, **37**(5):2649-2654.

8. Wentink M, Räkers M, Stupart D, Algar U, Ramesar R, Goldberg P: **Incidence and histological features of colorectal cancer in the Northern Cape Province, South Africa.** *S Afr J Surg* 2010, **48**(4):109-113.

9. Chalya PL, Mchembe MD, Mabula JB, Rambau PF, Jaka H, Koy M, Mkongo E, Masalu N: **Clinicopathological patterns and challenges of management of colorectal cancer in a resource-limited setting: a Tanzanian experience.** *World J Surg Oncol* 2013, **11**(1):1-9.

10. Dijxhoorn DN, Boutall A, Mulder C, Ssebuufu R, Mall A, Kalungi S, Baigrie C, Goldberg P: **Colorectal cancer in patients from Uganda: a histopathological study.** *East Cent Afr J Surg* 2014, **19**(1):112-119.

11. Katsidzira L, Chokunonga E, Gangaidzo IT, Rusakaniko S, Borok M, Matsena-Zingoni Z, Thomson S, Ramesar R, Matenga JA: **The incidence and histo-pathological characteristics of colorectal cancer in a population based cancer registry in Zimbabwe.** *Cancer epidemiol* 2016, **44**:96-100.

12. Madiba T, Moodley Y, Sartorius B, Sartorius K, Aldous C, Naidoo M, Govindasamy V, Bhadree S, Stopforth L, Ning Y: **Clinicopathological spectrum of colorectal cancer among the population of the KwaZulu-Natal Province in South Africa.** *Pan Afr Med J* 2020, **37**(74).

13. Alatise OI, Knapp GC, Sharma A, Chatila WK, Arowolo OA, Olasehinde O, Famurewa OC, Omisore AD, Komolafe AO, Olaofe OO: **Molecular and phenotypic profiling of colorectal cancer patients in West Africa reveals biological insights.** *Nat Commun* 2021, **12**(1):1-8.

14. Cheng L, Eng C, Nieman LZ, Kapadia AS, Du XL: **Trends in colorectal cancer incidence by anatomic site and disease stage in the United States from 1976 to 2005.** *Am J Clin Oncol* 2011, **34**(6):573-580.

15. Chauvenet M, Cottet V, Lepage C, Jooste V, Faivre J, Bouvier A-M: **Trends in colorectal cancer incidence: a period and birth-cohort analysis in a well-defined French population.** *BMC cancer* 2011, **11**(1):1-6.

16. Takada H, Ohsawa T, Iwamoto S, Yoshida R, Nakano M, Imada S, Yoshioka K, Okuno M, Masuya Y, Hasegawa K: **Changing site distribution of colorectal cancer in Japan.** *Dis Colon Rectum* 2002, **45**(9):1249-1254.

17. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis.** *Cell* 1990, **61**(5):759-767.

18. Kinzler KW, Vogelstein B: **Lessons from hereditary colorectal cancer.** *Cell* 1996, **87**(2):159-170.

19. Poulogiannis G, Ichimura K, Hamoudi RA, Luo F, Leung SY, Yuen ST, Harrison DJ, Wyllie AH, Arends MJ: **Prognostic relevance of DNA copy number changes in colorectal cancer.** *J Pathol* 2010, **220**(3):338-347.

20. Arends MJ: **Pathways of colorectal carcinogenesis.** *Appl Immunohistochem Mol Morphol* 2013, **21**(2):97-102.

21. Müller MF, Ibrahim AE, Arends MJ: **Molecular pathological classification of colorectal cancer.** *Virchows Archiv* 2016, **469**(2):125-134.

22. Nicolaides NC, Papadopoulos N, Liu B, Weit Y-F, Carter KC, Ruben SM, Rosen CA, Haseltine WA, Fleischmann RD, Fraser CM: **Mutations of two P/WS homologues in hereditary nonpolyposis colon cancer.** *Nature* 1994, **371**(6492):75-80.

23. Wu Y, Berends MJ, Sijmons RH, Mensink RG, Verlind E, Kooi KA, van der Sluis T, Kempinga C, van der Zee AG, Hollema H: **A role for MLH3 in hereditary nonpolyposis colorectal cancer.** *Nat Genet* 2001, **29**(2):137-138.

24. Huang J, Kuismanen SA, Liu T, Chadwick RB, Johnson CK, Stevens MW, Richards SK, Meek JE, Gao X, Wright FA: **MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer.** *Cancer Res* 2001, **61**(4):1619-1623.

25. Poulogiannis G, Frayling IM, Arends MJ: **DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome.** *Histopathol* 2010, **56**(2):167-179.

26. Ligtenberg MJ, Kuiper RP, Chan TL, Goossens M, Hebeda KM, Voorendt M, Lee TY, Bodmer D, Hoenselaar E, Hendriks-Cornelissen SJ: **Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1.** *Nat Genet* 2009, **41**(1):112-117.

27. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa J-PJ: **CpG island methylator phenotype in colorectal cancer.** *Proc Natl Acad Sci* 1999, **96**(15):8681-8686.

28. Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**(7407):330-337.

29. Guinney J, Dienstmann R, Wang X, De Reynies A, Schlicker A, Soneson C, Marisa L, Roepman P, Nyamundanda G, Angelino P: **The consensus molecular subtypes of colorectal cancer**. *Nat Med* 2015, **21**(11):1350-1356.

30. Liu Y, Sethi NS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, Seoane JA, Farshidfar F, Bowlby R, Islam M: **Comparative molecular analysis of gastrointestinal adenocarcinomas.** *Cancer cell* 2018, **33**(4):721-735. e728.

31. Ashktorab H, Azimi H, Varma S, Lee EL, Laiyemo AO, Nickerson ML, Brim H: **Driver genes exome sequencing reveals distinct variants in African Americans with colorectal neoplasia.** *Oncotarget* 2019, **10**(27):2607.

32. Guda K, Veigl ML, Varadan V, Nosrati A, Ravi L, Lutterbaugh J, Beard L, Willson JK, Sedwick WD, Wang ZJ: **Novel recurrently mutated genes in African American colon cancers.** *Proceedings of the National Academy of Sciences* 2015, **112**(4):1149-1154.

33. Ashktorab H, Daremipouran M, Devaney J, Varma S, Rahi H, Lee E, Shokrani B, Schwartz R, Nickerson ML, Brim H: **Identification of novel mutations by exome sequencing in African American colorectal cancer patients.** *Cancer* 2015, **121**(1):34-42.

34. McCabe M, Perner Y, Magobo R, Magangane P, Mirza S, Penny C: **Microsatellite Instability assessment in Black South African Colorectal Cancer patients reveal an increased incidence of suspected Lynch syndrome.** *Sci Rep* 2019, **9**(1):1-10.

35. Vergouwe F, Boutall A, Stupart D, Algar U, Govender D, Van der Linde G, Mall A, Ramesar R, Goldberg P: **Mismatch repair deficiency in colorectal cancer patients in a low-incidence area.** *S Afr J Surg* 2013, **51**(1):16-21.

36. Stupart DA, Goldberg P, Algar U, Ramesar R: **Surveillance colonoscopy improves survival in a cohort of subjects with a single mismatch repair gene mutation.** *Colorectal Dis* 2009, **11**(2):126-130.

37. Holla R, Vorster A, Locketz M, De Haas M, Oke O, Govender D, Ramesar R, Goldberg P: **Immunohistochemical determination of mismatch repair gene product in colorectal carcinomas in a young indigenous African cohort.** *S Afr J Surg* 2022, **60**(1):28-33.

38. Katsidzira L, Vorster A, Gangaidzo IT, Makunike-Mutasa R, Govender D, Rusakaniko S, Thomson S, Matenga JA, Ramesar R: **Investigation on the hereditary basis of colorectal cancers in an African population with frequent early onset cases.** *PLOS one* 2019, **14**(10):e0224023.

39. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, Clendenning M, Sotamaa K, Prior T, Westman JA**: Feasibility of screening for Lynch syndrome among patients with colorectal cancer.** *J Clin Oncol* 2008, **26**(35):5783.

40. Stoffel EM, Koeppe E, Everett J, Ulintz P, Kiel M, Osborne J, Williams L, Hanson K, Gruber SB, Rozek LS: **Germline genetic features of young individuals with colorectal cancer.** *Gastroenterol* 2018, **154**(4):897-905. e891.

41. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, Bacher J, Bigley C, Nelsen L, Goodfellow PJ: **Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer.** *JAMA oncol* 2017, **3**(4):464-471.

42. Adegoke O, Komolafe A, Ojo O: **Microsatellite instability statuses and clinicopathological characteristics of colorectal carcinomas in a Sub-Saharan African population.** *Gastroenterol & Hepatol Int J* 2017, **2**(2):000119.

43. Duduyemi B, Akang E, Adegboyega P, Thomas J: **Significance of DNA mismatch repair genes and microsatellite instability in colorectal carcinoma in Ibadan, Nigeria.** *Am J Med Biol Res* 2013, **1**:145-148.

44. Raskin L, Dakubo JC, Palaski N, Greenson JK, Gruber SB: **Distinct molecular features of colorectal cancer in Ghana.** *Cancer Epidemiol* 2013, **37**(5):556-561.

45. Cronjé L, Becker P, Paterson A, Ramsay M: **Hereditary non-polyposis colorectal cancer is predicted to contribute towards colorectal cancer in young South African blacks.** *S Afr J Sci* 2009, **105**(1):68-72.

46. Pyatt R, Chadwick RB, Johnson CK, Adebamowo C, de la Chapelle A, Prior TW: **Polymorphic variation at the BAT-25 and BAT-26 loci in individuals of African origin: implications for microsatellite instability testing.** *Am J Pathol* 1999, **155**(2):349-353.

47. Samowitz WS, Slattery ML, Potter JD, Leppert MF: **BAT-26 and BAT-40 instability in colorectal adenomas and carcinomas and germline polymorphisms.** *Am J Pathol* 1999, **154**(6):1637-1641.

48. Bessa X, Ballesté B, Andreu M, Castells A, Bellosillo B, Balaguer F, Castellví–bel S, Paya A, Jover R, Alenda C: **A prospective, multicenter, population-based study of BRAF mutational analysis for Lynch syndrome screening.** *Clin Gastroenterol Hepatol* 2008, **6**(2):206-214.

49. Bateman AC: **DNA mismatch repair proteins: scientific update and practical guide.** *J Clin Pathol* 2021, **74**(4):264-268.

50. Hechtman JF, Rana S, Middha S, Stadler ZK, Latham A, Benayed R, Soslow R, Ladanyi M, Yaeger R, Zehir A: **Retained mismatch repair protein expression occurs in approximately 6% of microsatellite instability-high cancers and is associated with missense mutations in mismatch repair genes.** *Mod pathol* 2020, **33**(5):871-879.

51. Cheng DT, Mitchell TN, Zehir A, Shah RH, Benayed R, Syed A, Chandramohan R, Liu ZY, Won HH, Scott SN: **Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): a hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology.** *J Mol Diagn* 2015, **17**(3):251-264.

1.3. Specific Aims and Objectives

1.3.1. Primary Aim

To characterise the somatic molecular landscape of colorectal carcinoma in a cohort of relatively young patients (<60 years of age) in Cape Town, South Africa.

1.3.2. Specific Objectives

(i) To investigate and characterise the cohort of dMMR tumours with amplicon-based panel next-generation sequencing (NGS) to identify the spectrum of somatic variants in MMR genes and associated driver genes.

(ii) To evaluate the performance of a deep learning (DL) artificial intelligence (AI) model as a pre-screening test to identify dMMR tumours in our unique cohort of whole-slide scanned images from sub-Saharan Africa.

(iii) To comprehensively characterise the somatic molecular genetics of the cohort of pMMR cases by whole exome sequencing, with a particular focus on driver genes and oncogenic pathway activation.

1.4. Rationale

CRC is one of the most common cancers globally, in both men and women, and contributes significantly to cancer-associated deaths.[1] Currently, most of this burden of disease is reported in developed countries. However, over the course of the next decade, the incidence of CRC is set to rise drastically in developing regions, such as sub-Saharan Africa.[30]

As described in detail in the introductory section above, CRC can be partitioned into dMMR and pMMR categories. Historically, the purpose of recognising dMMR tumours was to identify patients who might have Lynch syndrome for further testing. Recently however, it has become important to identify these patients as they may be candidates for immune checkpoint inhibitor therapy.[17] Universal screening of all CRC cases for dMMR by immunohistochemistry (IHC) or PCR is currently recommended by several international advisory groups.[31] However, this

has added cost implications and is neither standardised nor practiced routinely in South Africa. As a result, patients who would benefit from referral to a genetic counselling and testing service, are either not identified in the first instance, or are lost to follow up. This is a major barrier to providing a comprehensive Lynch syndrome screening program.

Recently, it has been shown that pre-screening whole slide scanned images of CRC resection samples with deep-learning (DL) models can significantly reduce the burden of downstream IHC or PCR testing.[32-34] These models have been trained and validated on large international cohorts. However, there has been little representation of tumours from diverse populations, and notably, sub-Saharan Africa. This is a potential shortcoming of these studies, and further validation is required before these models can be applied diagnostically in sub-Saharan Africa. The cost saving potential of this pre-screening test (by reducing the overall number of cases which require wet bench testing) would allow more patients to have access to the more costly confirmatory IHC or PCR tests.

pMMR CRC represents the majority of tumours encountered in clinical practice requiring therapy. These tumours develop via the well described adenoma-carcinoma sequence and are characterised by alterations in the WNT signalling pathway, *KRAS* mutations and *TP53* inactivation. Few studies have investigated the tumour biology of pMMR CRC in sub-Saharan Africa. The preliminary work which has been done shows that alterations in these classical pathways are not as frequent as described in developed or high-income countries.[35] This is an intriguing observation and warrants further investigation. Understanding if there are alternative driver genes or altered oncogenic signalling pathways is essential to developing modern treatment strategies for these tumours.

Three main experiments were conducted to achieve the study objectives. The dMMR cohort, without known germline pathogenic variants, was subjected to amplicon-based panel sequencing in the Ramesar laboratory (Chapter 3). The main purpose of this was to identify and characterise the spectrum of MMR gene variants which led to loss of MMR protein expression in this cohort.

Secondary aims were to identify any MMR genes with a variant allele frequency (VAF) suggestive of a germline variant, and to identify any *POLE* variants, as well as describe the frequency of established driver variants in this cohort. The second experiment (Chapter 4) sought to determine the performance of a DL model on detecting dMMR in our novel cohort. As universal MMR protein immunohistochemistry is not performed in South Africa uniformly, the goal of this study was to investigate whether this AI model could be used as a pre-screening test to decrease the burden (and associated cost) of downstream diagnostic IHC or PCR testing. In the final experiment (Chapter 5) we sought to comprehensively characterise and profile the molecular landscape of the pMMR cohort with whole exome sequencing (WES) in the Dave laboratory (Duke University).

2. CHAPTER TWO

2.1. Overview of Sample Collection and Methodology

CRC resection specimens reported in the Division of Anatomical Pathology at the University of Cape Town / Groote Schuur Hospital were identified by searching the laboratory information system, Trakcare. The study period included cases received over 5 years, between 1 January 2016 and 31 December 2020. Patients under the age of 60 years at the time of resection were included in this study. Early-onset CRC is defined as diagnosis before 50 years of age. The reason for choosing an age cutoff of 60 for this study was twofold. In the first instance, local cost-saving protocols mandated that only patients younger than 60 were screened for dMMR with IHC. Therefore, this was the cohort with complete IHC results. Second, it is being increasingly recognised that hereditary CRC (including Lynch syndrome) may present over the age of 50 years, and so we wanted to include as many of these patients as possible.[36] Demographic and clinicopathological data were collected from the pathology reports which were accessed on Trackcare. In cases where there was incomplete pathological data, the glass slides were reviewed, and the missing variables were recorded.

Glass slides and formalin fixed paraffin embedded (FFPE) tissue wax blocks were retrieved from the Division of Anatomical Pathology archives. Glass slides were reviewed by a consultant histopathologist to select appropriate representative tumour blocks. Criteria of at least 1000 viable tumour cells in a section, and more than 50% tumour volume were used to guide block selection. If there was insufficient residual tumour tissue (i.e. rectal adenocarcinoma which was resected post neoadjuvant therapy), these cases were excluded. Once the FFPE tissue wax blocks were retrieved, they were assessed for adequacy. The amount of residual tissue and the tissue preservation were visually assessed by a consultant histopathologist. In cases where the selected block was deemed inadequate, we attempted to retrieve and assess an alternate block. The results of the MMR protein IHC (MLH1, MSH2, MSH6, PMS2) were reviewed by two histopathologists. Only cases with complete IHC results were included in the study.

dMMR cases were cross referenced against the Division of Human Genetics CRC database to identify patients who were appropriately referred for further germline testing. In cases where germline testing was performed, these results were recorded. Where a known germline pathogenic variant was detected, these cases were excluded from further sequencing studies. Cases with no germline pathogenic variant detected (in the limited panel tested) were retained for further downstream sequencing. Cases which were not referred for testing, but where a family member was known with a germline pathogenic variant were also retained for sequencing studies.

2.2. Summary of Overall Clinicopathological Data and Samples

Two hundred and twenty-two (222) CRC resection samples were retrieved from the Division of Anatomical Pathology archives. Overall, 197 samples with sufficient material for further molecular experiments were identified. The sample workflow is summarised below (Figure 2.1).

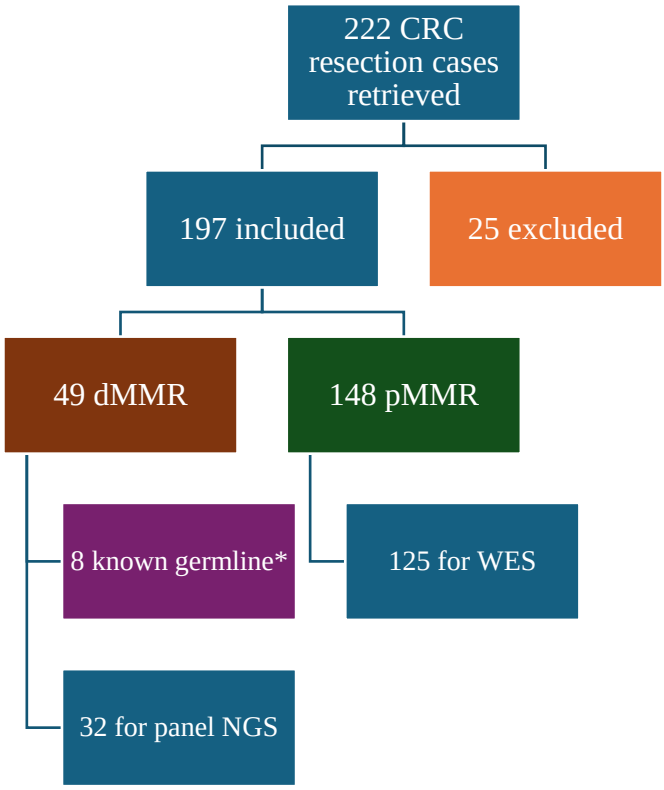


Figure 2.1: Workflow of Samples Collected

* 8 cases had known germline pathogenic variants (these were all *MLH1* 1528C>T)

dMMR mismatch repair deficient; pMMR mismatch repair proficient, WES whole-exome sequencing, NGS next-generation sequencing

Forty-nine dMMR (20.2%) and 148 pMMR (79.8%) cases were included in the overall cohort. The clinicopathological characteristics of the entire cohort are shown in Table 2.1. The frequency of loss of IHC staining patterns in dMMR cases is shown below (Figure 2.2). The group of 6 dMMR cases with "other" unconventional loss of staining patterns included isolated loss of MLH1 (1), isolated loss of MSH2 (2), loss of MSH6 and PMS2 (1), loss of MLH1 and MSH6 (2) and loss of MLH1, PMS2 and MSH6 (1).

Table 2.1 Demographic and Clinicopathological Characteristics of the Overall Cohort (n=197)

	dMMR (%)	**pMMR** (%)
Total Number	49 (24.9)	148 (75.1)
Age (years)	48 (40.5,54)*	52 (47,56)*
Gender		
Female	19 (38.8)	71 (48.0)
Male	30 (61.2)	77 (52.0)
Tumour Side		
Right	30 (61.2)	42 (28.4)
Left	19 (38.8)	106 (71.6)
Tumour Site		
Caecum	14 (28.6)	12 (8.1)
Ascending	8 (16.3)	15 (10.1)
Hepatic flexure	3 (6.1)	4 (2.7)
Transverse	5 (10.2)	11 (7.4)
Splenic flexure	5 (10.2)	5 (3.4)
Descending	1 (2.0)	11 (7.4)
Sigmoid	9 (18.4)	42 (28.4)
Rectum	4 (8.2)	48 (32.4)
Histological Type		
Adenocarcinoma, NOS	37 (75.5)	138 (91.9)
Mucinous	12 (24.5)	12 (8.1)
Histological Grade		
1	4 (8.2)	9 (6.1)
2	32 (65.3)	134 (90.5)
3	13 (26.5)	5 (3.4)
Pathological Tumour Stage		
1	0 (0)	3 (2.0)
2	3 (6.1)	9 (6.1)
3	25 (51.0)	99 (66.9)
4a	13 (26.5)	28 (18.9)
4b	8 (16.3)	9 (6.1)
Tumour infiltrating lymphocytes		
0	14 (28.6)	97 (65.5)
1	17 (34,7)	40 (27.0)
2	8 (16.3)	7 (4.7)
3	10 (20.4)	4 (2.7)
Peritumoral (Crohn-like) lymphocytic response		
0	21 (42.9)	88 (59.5)
1	19 (38.8)	43 (29.1)
2	3 (6.1)	9 (6.1)
3	6 (12.2)	8

* median (Q1,Q3); NOS not otherwise specified

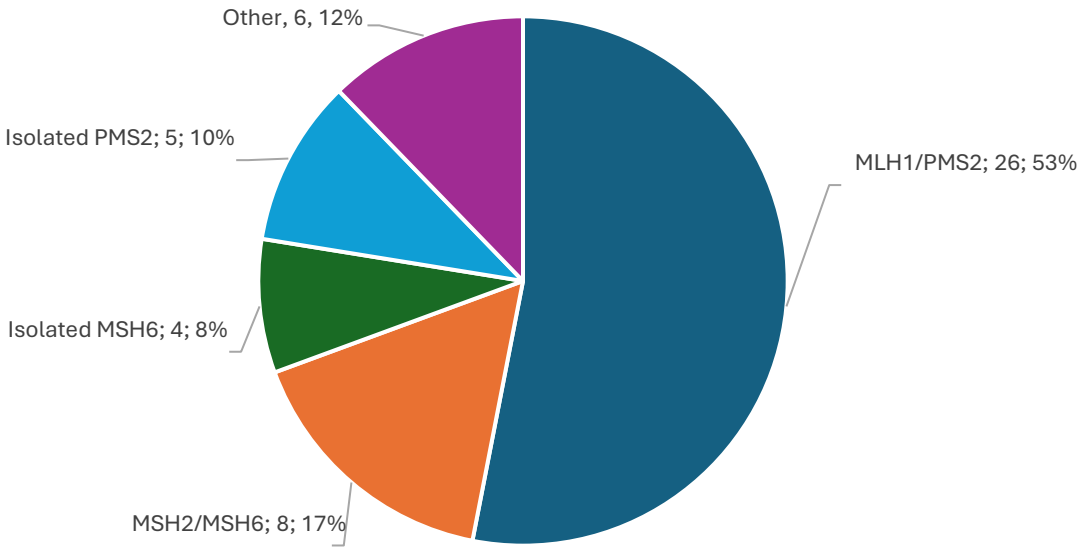


Figure 2.2: Loss of MMR Protein Expression Patterns in CRC Study Cohort by Immunohistochemistry

Of the 49 dMMR cases, 20 were referred and had already undergone germline testing in the Division of Human Genetics, University of Cape Town (Table 2.2). Of these, 8 were found to have an *MLH1* 1528C>T pathogenic variant. Four patients tested negative for the "common 5" panel. This is an inhouse PCR panel which detects the 5 most frequent germline variants in the local mixed ancestry population (*MLH1* c.1528C>T, *MSH2* c.387_388delTC, *MSH2* c.1221_1222delCT, *MSH2* c.1340_1341insGG and *MSH2* c.1046C>G). Two patients belonged to families known to the Division of Human Genetics to harbour a germline pathogenic variant, but do not seem to have undergone testing themselves. Six were registered but did not have a test result recorded in the database.

Table 2.2 Results of Patients Referred to the Division of Human Genetics for Germline Testing

Number	**Germline Test Result**
8 (40%)	*MLH1* c.1528C>T
1 (5%)	*MSH2* c.731_734del (family)
1 (5%)	*MSH2* c.731_733del (family)
4 (20%)	Negative for “common 5” panel
6 (30%)	No test results found

Thirty-two dMMR cases, without a known germline mutation and with adequate DNA quantity and quality, were selected randomly with an even split between MLH1/PMS2 loss and other patterns of MMR loss for NGS. We were limited to 125 cases for WES and prioritized the youngest patients for this experiment.

3. CHAPTER THREE (Paper 2)

Title: Investigating Somatic Variants and Pathways in Mismatch Repair Deficient (dMMR) Colorectal Carcinoma in South Africa

Authors: Alessandro Pietro Aldera(1,2,3), Jana van der Westhuizen(2), Wan-Jung Tsai(1), May J Krause(2), Safiye Yildiz(2), Komala Pillay(1), Adam Boutall(4), Raj Ramesar(2)

Journal: Journal of Clinical Pathology

Citation: Aldera AP, van der Westhuizen J, Tsai WJ, Krause MJ, Yildiz S, Pillay K, Boutall A, Ramesar R. Investigating somatic variants and pathways in mismatch repair-deficient (dMMR) colorectal carcinoma in South Africa. J Clin Pathol. 2024 May 15: Epub ahead of print.

1. Division of Anatomical Pathology, National Health Laboratory Services / University of Cape Town, Cape Town, South Africa

2. Division of Human Genetics, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences and University of Cape Town

3. JDW Pathology Inc, Cape Town, South Africa

4. Division of General Surgery, Groote Schuur Hospital and University of Cape Town

Key words: Colorectal carcinoma, mismatch repair deficiency, Lynch syndrome

3.1. Abstract

Background: Colorectal carcinoma (CRC) is a common cause of morbidity and mortality worldwide, and an emerging public health problem in sub-Saharan Africa. Several authors have described an increased frequency of mismatch repair deficient (dMMR) CRC in sub-Saharan Africa, but these tumours remain poorly characterised molecularly. We sought to interrogate the somatic molecular genetic landscape of dMMR CRC in a cohort of young patients to better inform Lynch syndrome (LS) screening strategies and personalized medicine approaches in our setting.

Methods: Thirty-two patients (age < 60 years) were identified with dMMR CRC. Known Lynch syndrome patients were excluded. DNA was extracted from selected formalin fixed paraffin embedded (FFPE) tissue resection samples and subjected to amplicon-based Next Generation Sequencing (NGS).

Results: Pathogenic or likely pathogenic variants were detected in the corresponding MMR gene in 14/18 (78%) MLH1/PMS2 deficient tumours, 5/8 (63%) MSH2/MSH6 deficient tumours, 1/4 (25%) tumours with isolated MSH6 loss, and 0/2 tumours with isolated PMS2 loss. Previously unreported variants were identified in *MLH1* (3), and *MSH2* (1). Cases with a variant allele frequency suggesting a germline mutation were identified in *MLH1* (8), *MSH2* (2) and *MSH6* (1). Only one MMR gene variant was detected in more than one patient (*MLH1* p.Q510*). Four *POLE*/*POLD1* exonuclease domain variants were identified, one of which was previously unreported.

Conclusion: The spectrum of disease-causing MMR gene variants in our population necessitates NGS testing for LS screening. This study also highlights the role of somatic testing on readily available FFPE samples to generate data on the epidemiology of CRC in different settings.

3.2. Key Messages

- Colorectal carcinoma (CRC) is an emerging public health epidemic in sub-Saharan Africa, and preliminary work has shown an increased frequency of mismatch repair deficient (dMMR) cases in this population.

- This study demonstrates mixed correlation between MMR immunohistochemistry (IHC) results and MMR gene mutations, detected by panel-based next generation sequencing (NGS), with a wide range of different MMR gene variants identified. Many of these mutations harboured variant allele frequencies which would suggest a germline variant (i.e. Lynch syndrome).
- Further high throughput genomic work (exomes and RNA-Seq) is required to better characterise the cases with discordant IHC and NGS findings. Given the spectrum of variants encountered in our population, screening for Lynch syndrome in clinical practice should be done with panel-based NGS.

3.3. Introduction

Colorectal carcinoma (CRC) is the third most common cancer worldwide, and contributes significantly to cancer-associated morbidity and mortality.[1] The incidence of CRC in sub-Saharan Africa is set to rise dramatically over the next decade.[2] Despite this, the molecular genetics and tumour biology of CRC in sub-Saharan Africa remains poorly characterised.[3] In an effort to disentangle its molecular heterogeneity, CRC may be broadly classified into mismatch repair proficient (pMMR) and mismatch repair deficient (dMMR) groups based on the expression of MMR proteins by immunohistochemistry (IHC). The majority of dMMR tumours are sporadic and associated with *MLH1* promotor hypermethylation, but a proportion (3-4% of CRC overall) are associated with a germline loss-of-function mutation in one of the MMR genes (Lynch syndrome; LS).[4] DNA mismatch repair is a highly conserved mechanism which improves fidelity during DNA replication (S phase) by identifying mismatched nucleotide base pairs. Four major MMR proteins function as heterodimers. MSH2 pairs with MSH6 (hMutL), recognises the mismatched nucleotide base pairs and initiates repair, while MLH1 heterodimerises with PMS2 (hMutS) and functions as a molecular matchmaker, acts as an endonuclease and terminates mismatch provoked excision.[5] These proteins are encoded by their corresponding genes *MLH1*, *PMS2*, *MSH2* and *MSH6*. Downstream from this, EXO1 excises the mismatched segment and polymerase epsilon (POLE) is recruited to resynthesise the correct nucleotide sequence. Structural abnormalities in *EPCAM*, which encodes a cell adhesion molecule, may also cause LS because it is adjacent to *MSH2*.[6]

Over the past two decades, several authors have noted an increased proportion of dMMR CRC in sub-Saharan Africa (mean 26%; range 7-67%) [7-11] compared to the reported global average of 12%. The few studies examining paired germline material have however, not shown an increased prevalence of LS.[11, 12] Recognising dMMR CRC in pathology specimens has prognostic and predictive importance and is used to triage patients to LS screening programmes, for whom universal screening and management has shown significant reduction in morbidity and mortality. To our knowledge, no published sub-Saharan African studies have sought to characterise the spectrum of somatic mutations in the MMR genes in dMMR CRC. This is of importance to correlate the MMR gene protein expression with somatic mutations, identify any frequently occurring possible founder mutations, and provide real world locally relevant data for designing germline screening panels and algorithms.

The aim of this study is to describe the somatic mutations occurring in the MMR genes in a cohort of young South African patients with dMMR CRC.

3.4. Methods

3.4.1. Study Design and Data Collection

Archived dMMR CRC resection samples, reported in the Division of Anatomical Pathology at Groote Schuur Hospital (Cape Town, South Africa) between 2016-2020, were identified by searching the National Health Laboratory Services (NHLS) laboratory information system. IHC was performed to determine MMR status during routine clinical practice, and for this study these slides were reviewed by a consultant histopathologist. Four stains (MLH1, PMS2, MSH2, MSH6) were assessed on whole slide resection specimens. In the presence of adequate internal and external controls, staining in any proportion of tumour cell nuclei of any intensity was regarded as retained expression. Slides and formalin fixed paraffin embedded (FFPE) tissue wax blocks were reviewed by a histopathologist for representativeness and adequacy. Patients known to the University of Cape Town Division of Human Genetics with an established MMR gene pathogenic variant were excluded from this study. Samples from 32 patients were randomly

selected with an even split between cases showing immunohistochemical loss of MLH1/PMS2 and those showing other combinations of loss of MMR gene protein expression.

3.4.2. Molecular Profiling

Three 10µm scrolls were taken from each representative FFPE tissue wax block. DNA was extracted using the QIAamp DNA FFPE Advanced UNG Kit (Qiagen, Maryland, USA) according to the manufacturer's protocol. Assessment of the extracted DNA quantity, integrity and purity, was performed utilising microfluidic capillary electrophoresis, spectrophotometry, and fluorometry. The NGS panel used consisted of the 24 genes in the commercially available Oncomine™ Colorectal and Pancreatic panel (Thermo Fisher Scientific, Massachusetts, USA) *APC*, *ARID1A*, *BRAF*, *CDKN2A*, *CTNNB1*, *ERBB2*, *ERBB3*, *FBXW7*, *GNAS*, *KRAS*, *MLH1*, *MSH2*, *MSH6*, *MYC*, *PIK3CA*, *PMS2*, *POLE*, *PTEN*, *RNF43*, *SMAD4*, *SOX9*, *TCF7L2*, *TP53*, *NRAS*, and an additional three spiked genes *EPCAM*, *POLD1* and *MUTYH*. Libraries were constructed using the Ion AmpliSeq™ Kit for Chef DL8 (Thermo Fisher Scientific, Massachusetts, USA) on the Ion Chef™ System (Thermo Fisher Scientific, Massachusetts, USA). The amplified libraries were quantified using the Ion Library TaqMan Quantitation Kit (QuantStudioTM 3 Real-Time PCR System). Diluted libraries were pooled and templated using the Ion 540™ Chip Kit (Thermo Fisher Scientific, Massachusetts, USA), and sequenced on the Ion GeneStudio™ S5 Prime (Thermo Fisher Scientific, Massachusetts, USA).

3.4.3. Variant Calling and Annotation

The Oncomine™ Colorectal and Pancreatic 540 w4.3 workflow (https://ionreporter.thermofisher.com/ir) was used to align reads to the human genome (hg19) and to generate variant call files (VCF), which were uploaded to Franklin by Genoox (https://franklin.genoox.com/clinical-db/home). Somatic alterations including single-nucleotide variants and small insertions/deletions with variant allele frequency (VAF) > 5% and depth of coverage > 500 were filtered and manually inspected. OncoKB knowledgebase, ClinVar (https://www.ncbi.nlm.nih.gov/clinvar) and American College of Medical Genetics and

Genomics (ACMG) guidelines were used to assign variant pathogenicity.[13] Identified pathogenic variants were visually inspected on Integrative Genomics Viewer.[14]

3.4.4. Statistics

Descriptive statistics, median and range, were used for continuous variables. Count and percentage were used for categorical variables, with comparisons made using Fisher's exact test. IBM SPSS Statistics version 29 (IBM, Armonk, NY) was used for all analyses. All tests were two-sided, with $P < 0.05$ considered significant.

3.4.5. Ethical Approval

Approval for access to patient demographic data, pathology reports and the use of archival FFPE tissue samples for IHC and molecular studies was granted by the University of Cape Town Human Research Ethics Committee (HREC 202/2002).

3.5. Results

3.5.1. Clinicopathological Findings

Thirty-two dMMR cases were selected for NGS (Table 3.1). There were 18 males (56%) and 14 females (44%). The median age at diagnosis was 46 years (range 27-59). The majority of tumours were located in the right colon (n=21; 66%), while 11 occurred in the left colon (n=11; 34%). One patient (T19) was found to have synchronous tumours in the right colon (hepatic flexure and transverse colon). MLH1-deficient tumours were more likely to occur in the right colon compared to non-MLH1 deficient tumours (p=0.027). By IHC, 18 tumours (56%) showed loss of MLH1 and PMS2 expression, two (6%) showed isolated PMS2 loss, eight (25%) showed loss of MSH2 and MSH6 expression, and four (13%) showed isolated loss of MSH6 expression (Figure 3.1). One tumour showed a combination pattern with loss of MLH1, PMS2 and MSH6 staining (T6). No *BRAF* V600E mutation was detected by IHC or PCR in any of the 18 tumours showing MLH1 and PMS2 loss of staining. The majority of tumours were classified histologically as adenocarcinoma not otherwise specified (NOS), with only 25% of tumours

meeting the criteria for mucinous adenocarcinoma. Overall, most tumours were locally advanced (pT3 or pT4) and presented with involvement of 0 (pN0) or 1-3 (pN1) regional lymph nodes.

Table 3.1: Descriptive Analysis of Clinicopathological Variables, Stratified by MLH1 versus non-MLH1 Immunohistochemistry Protein Loss.

	Overall (N=32)		**MLH1 (N=18)**		**non-MLH1 (N=14)**		**P value**
Age at diagnosis, years							
Median (range)	46,5 (27-59)		46 (27-59)		48,5 (30-58)		
Sex							
Female	14	43,8%	10	55,6%	4	28,6%	0.165
Male	18	56,3%	8	44,4%	10	71,4%	
Anatomical Site							
Caecum	11	34,4%	8	44,4%	3	21,4%	
Ascending	5	15,6%	3	16,7%	2	14,3%	
Hepatic Flexure	2	6,3%	1	5,6%	1	7,1%	
Transverse	3	9,4%	3	16,7%	0	0,0%	
Splenic Flexure	2	6,3%	1	5,6%	1	7,1%	
Descending	0	0,0%	0	0,0%	0	0,0%	
Sigmoid	8	25,0%	2	11,1%	6	42,9%	
Rectum	1	3,1%	0	0,0%	1	7,1%	
L/R							
Right	21	65,6%	15	83,3%	6	42,9%	**0.027**
Left	11	34,4%	3	16,7%	8	57,1%	
Histological Type							
Adenocarcinoma NOS	24	75,0%	12	66,7%	12	85,7%	0.412
Mucinous adenocarcinoma	8	25,0%	6	33,3%	2	14,3%	

Primary Tumour							
pT1	1	3,1%	0	0,0%	1	7,1%	0.867
pT2	1	3,1%	1	5,6%	0	0,0%	
pT3	17	53,1%	11	61,1%	6	42,9%	
pT4a	10	31,3%	5	27,8%	5	35,7%	
pT4b	3	9,4%	1	5,6%	2	14,3%	
Regional Lymph Nodes							
pN0	14	43,8%	8	44,4%	6	42,9%	0,763
pN1a	9	28,1%	5	27,8%	4	28,6%	
pN1b	3	9,4%	2	11,1%	1	7,1%	
pN1c	2	6,3%	0	0,0%	2	14,3%	
pN2a	3	9,4%	2	11,1%	1	7,1%	
pN2b	1	3,1%	1	5,6%	0	0,0%	

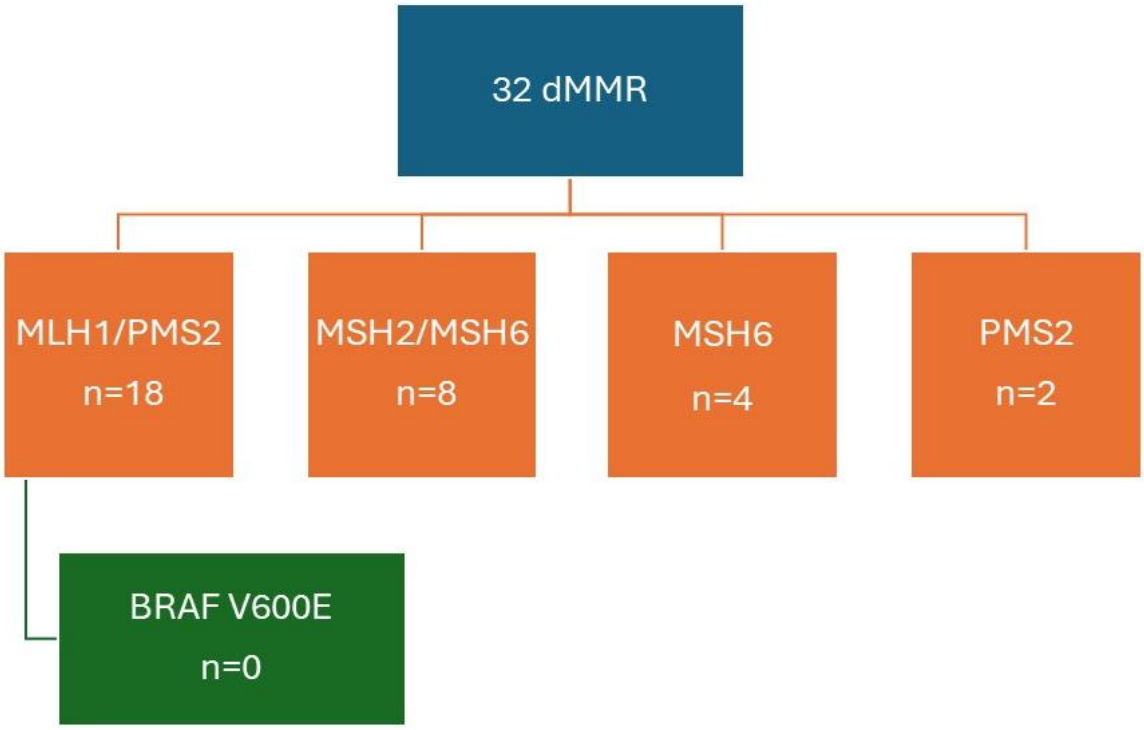


Figure 3.1: Summary of Immunohistochemical Staining Results. The second tier indicates the IHC loss of staining pattern in these mismatch repair deficient (dMMR) tumours.

3.5.2. Mismatch Repair Gene Mutations

Fourteen of the 18 MLH1/PMS2 deficient tumours (78%) harboured pathogenic or likely pathogenic *MLH1* variants (Table 3.2). In the residual four cases, one showed a previously unreported variant of uncertain significance (VUS) in a splice region of *MLH1* (c.1761+5G>T; AF=36,36%), two showed previously reported VUS in *MSH6*, and one harboured no detectable pathogenic MMR gene variants. Eight of the 14 (57%) *MLH1* pathogenic or likely pathogenic variants had an allele frequency suggestive of a germline variant. One of these cases without a *MLH1* variant (T6) was the case showing loss of MLH1, PMS2 and MSH6 on IHC. This case was found to have a likely germline VUS in *MSH6*. Three previously unreported *MLH1* variants were discovered in our cohort. The c.1731+5G>T splice variant already described above, a missense mutation c.83C>G (pP28R) and a frameshift mutation c.2017_2038del (p.F673Afs*103). The previously described c1528C>T (p.Q510*) truncating mutation, which

represents a large founder effect in the northwest region of South Africa, was the only variant to be present in more than one sample (n=2).

Table 3.2: Identified Mismatch Repair Gene Variants and their Likely Significance

	Gene	Coding DNA*	Protein Change	VAF	ClinVar	ACMG	Reported before
T1	*MLH1*	c.1731G>A	p.Ser577=	77,33%	Pathogenic	Pathogenic	Yes
T2	*MLH1*	c.1731+5G>T	Splice region	36,36%	Not reported	VUS	**No**
T3	*MLH1*	c.1975C>T	p.R659*	9,10%	Pathogenic	Pathogenic	Yes
T4	*MLH1*	c.1029C>A	p.Y343*	73,04%	Pathogenic	Pathogenic	Yes
T5	*MLH1*	c.117-1G>A	Splice acceptor	53,18%	Pathogenic	Pathogenic	Yes
T6	*MSH6*	c.3936_3951del	p.K1315R	49,10%	VUS	VUS	Yes
T8	*MLH1*	c.1852_1854del	p.K618del	86,45%	Pathogenic	Pathogenic	Yes
T9	*MLH1*	c.1410-2A>G	Splice acceptor	32,16%	Likely Pathogenic	Pathogenic	Yes
T10	*MLH1*	c.2084C>A	p.S695*	22,94%	Pathogenic	Pathogenic	Yes
T11	*MLH1*	c.83C>G	p.P28R	47,07%	Not reported	Likely Pathogenic	**No**
T12	*MLH1*	c.1528C>T	p.Q510*	51,91%	Pathogenic	Pathogenic	Yes
T13	*MLH1*	c.2017_2038del	p.F673Afs*103	10,64%	Not reported	Likely Pathogenic	**No**
T14	*MLH1*	c.350C>T	p.T117M	8,90%	Pathogenic	Pathogenic	Yes
T15	*MLH1*	c.1528C>T	p.Q510*	64,38%	Pathogenic	Pathogenic	Yes
T16	*MLH1*	c.790+1G>A	Splice donor	25,94%	Pathogenic	Pathogenic	Yes
T17	*MLH1*	c.2048T>C	p.F683S	68,10%	Likely Pathogenic	Pathogenic	Yes
T18	*MSH6*	c.1292A>C	p.K431T	16,85%	VUS	Likely Pathogenic	Yes
T19	*MUTYH*	c.452A>G	p.Y151C	99,60%	Pathogenic	Pathogenic	Yes
T20	*MSH2*	c.1165C>T	p.R389*	5,16%	Pathogenic	Pathogenic	Yes

T22	*PMS2*	c.556C>T	p.Q186*	29,45%	Pathogenic	Pathogenic	Yes
T23	*MSH6*	c.2062_2063del	p.V688Lfs*9	48,81%	Pathogenic	Pathogenic	Yes
T25	*MSH2*	c.1906dup	p.A636Gfs*8	31,62%	Not reported	Likely Pathogenic	**No**
T26	*PMS2*	c.1882C>T	p.R628*	16,57%	Pathogenic	Pathogenic	Yes
T27	*MSH2*	c.2131C>T	p.R711*	55,56%	Pathogenic	Pathogenic	Yes
T28	*MSH6*	c.2722G>T	p.E908*	7,76%	Pathogenic	Pathogenic	Yes
T29	*MSH2*	c.1221_1222del	p.Y408Sfs*8	48,29%	Pathogenic	Pathogenic	Yes
T30	*MSH2*	c2132G>A	R711Q	10,25%	VUS	Likely Pathogenic	Yes
T31	*MSH2*	c.2377C>T	p.Q793*	78,43%	Pathogenic	Pathogenic	Yes
T32	*MSH6*	c.1691C>A	p.S564*	49,17%	Pathogenic	Pathogenic	Yes

* Variants in red indicate likely germline, based on VAF.

Variant Allele Frequency (VAF), American College of Medical Genetics and Genomics (ACMG)

Five of the eight MSH2/MSH6 deficient tumours (63%) harboured pathogenic or likely pathogenic variants in *MSH2*, with two of these (40%) showing an allele frequency suggesting a germline variant. One of these variants (*MSH2* c.1906dup; p.A636Gfs*8) was a previously unreported frameshift mutation. Of the three MSH2/MSH6 cases where no *MSH2* variant was called, one harboured a pathogenic truncating variant in *PMS2* (c1882C>T; p.R628*), one showed a pathogenic variant in *MSH6* (c.2722G>T; p.E908*) and one had no detectable MMR gene variants.

Only one of the four cases (25%) with isolated MSH6 loss showed a pathogenic variant in *MSH6*, with a likely germline allele frequency. Two of the residual three cases showed variants in other MMR genes; namely *PMS2* (c.556C>T; p.Q186*) and *MSH2* (c.2377C>T; p.Q793*). The latter being a likely germline variant. The remaining case showed a likely germline pathogenic variant in *MUTYH*.

No *PMS2* variants were called in the two cases showing isolated loss of PMS2 on IHC. One of these cases demonstrated a likely germline pathogenic variant in *MSH6* (c.2062_2063del; p.V688Lfs*9). The other case had no detectable pathogenic MMR gene variants.

3.5.3. POLE and POLD1 Mutational Status

Variants in the exonuclease domains of *POLE* (268-471) and *POLD1* (304-517) were detected in four of the 32 cases (13%). One of these variants had previously been reported as pathogenic (Table 3.3). Two of the other three variants (*POLD1* c.1511C>T and c.1055G>A) were previously reported on ClinVar as VUS and one (*POLE* c.1588G>A) was unreported.

Table 3.3: POLE and POLD1 Exonuclease Domain Variants with Corresponding Case Data

	IHC	NGS		Coding DNA	Protein Change	ClinVar	ACMG
T1	MLH1/PMS2	*MLH1*	*POLD1*	c.1511C>T	p.T504I	VUS	VUS
T31	MSH6	*MSH6*	*POLD1*	c.1055G>A	p.R352H	VUS	VUS
T18	MLH1/PMS2	None	*POLE*	c.857C>G	p.P286R	Likely Pathogenic	Likely Pathogenic
T28	MSH2/MSH6	*MSH6*	*POLE*	c.1588G>A	p.D530N	Not reported	VUS

Polymerase epsilon (POLE), polymerase D1 (POLD1), immunohistochemistry (IHC), next generation sequencing (NGS), American College of Genetics and Genomics (ACMG), variant of uncertain significance (VUS).

3.5.4. Pathway Activation

Commonly mutated genes in four oncogenic pathways were assessed utilising this NGS panel (Table 3.4). The genes in these pathways were mapped against the MMR gene and *POLE/D1* gene status of each case (Figure 3.2).[15, 16] PIK3K (*PTEN*, *PIK3CA*, *ARID1A*) and WNT (*APC*, *CTNNB1*) pathway activation was the most frequent, followed by RTK-RAS (*KRAS*, *NRAS*, *BRAF*) activation and alteration in the cell cycle (*TP53*, *FBXW7*).

Table 3.4: Pathway Activation Summary Data, Stratified by MLH1 versus non-MLH1 Immunohistochemistry Protein Loss.

	Overall	n=28	**MLH1**	n=15	**non-MLH1**	n=13	**p = 0.843**
WNT	**18**	**64,29%**	11	73,33%	7	53,85%	
APC	14	50,00%	8	53,33%	6	46,15%	
CTNNB1	6	21,43%	5	33,33%	1	7,69%	
PI3K	**18**	**64,29%**	9	60,00%	9	69,23%	
PTEN	8	28,57%	5	33,33%	3	23,08%	
PIK3CA	12	42,86%	5	33,33%	7	53,85%	
ARID1A	5	17,86%	1	6,67%	4	30,77%	
RTK-RAS	**11**	**39,29%**	5	33,33%	6	46,15%	
KRAS	7	25,00%	4	26,67%	3	23,08%	
NRAS	4	14,29%	1	6,67%	3	23,08%	
BRAF	2	7,14%	1	6,67%	1	7,69%	
Cell Cycle	**9**	**32,14%**	6	40,00%	3	23,08%	
TP53	5	17,86%	3	20,00%	2	15,38%	
FBXW7	5	17,86%	3	20,00%	2	15,38%	
***TGF*-b**	**3**	**10,71%**	1	6,67%	2	15,38%	
***POLE*/D1**	**4**	**14,29%**	1	6,67%	3	23,08%	

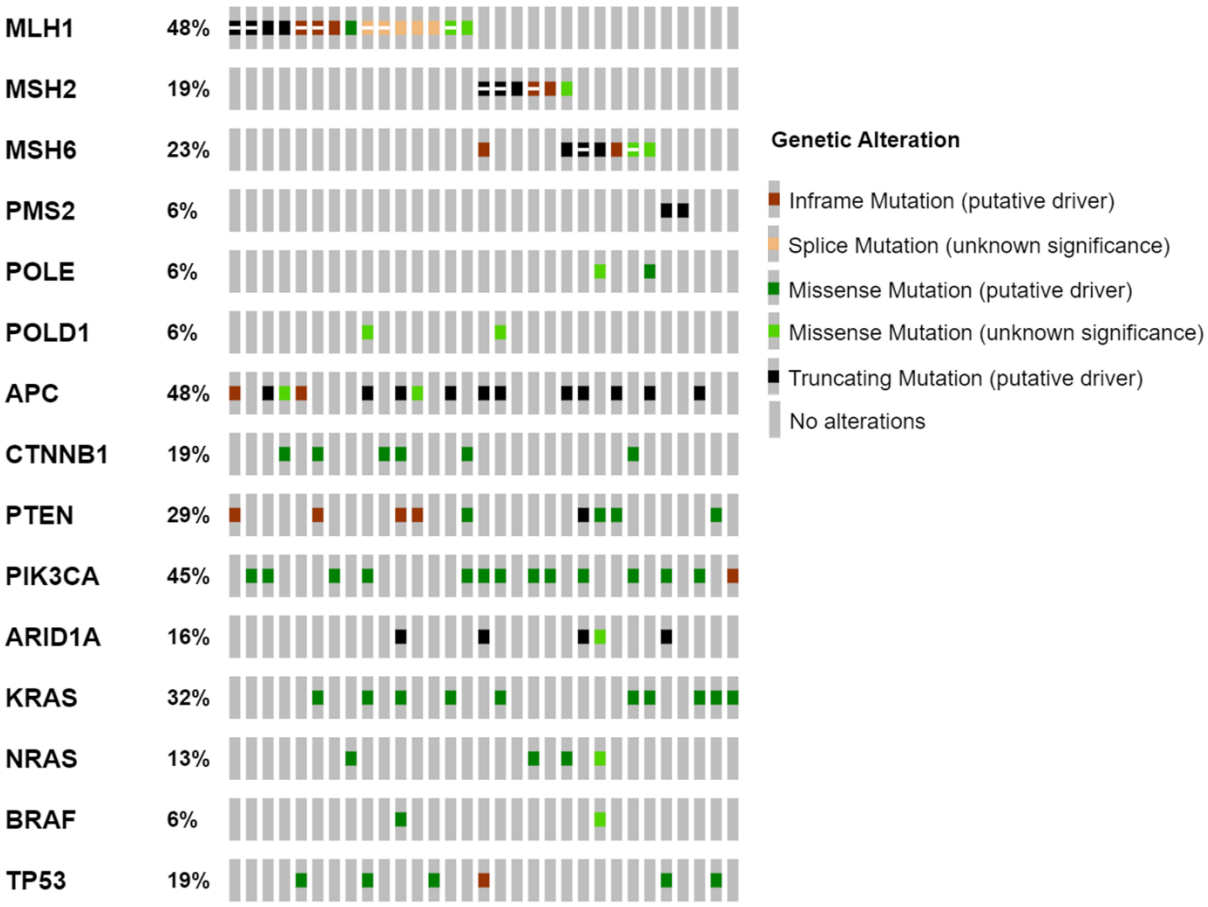


Figure 3.2: Summary of Pathogenic and Likely Pathogenic Variants by Case

Variants with likely germline VAF are denoted by a white bar in the oncoprint.

3.6. Discussion

Left-sided tumours were found in a significant proportion of cases (n=11; 32%) and these tumours were more likely to show loss of MSH2/MSH6, MSH6 or PMS2 protein expression (p=0.027). Conventionally, dMMR tumours are thought to be associated with a right-sided tumour location.[17, 18] This may reflect a MLH1-deficient phenotype which is more common overall, and, in larger studies with older patients, mostly as a result of *MLH1* promotor hypermethylation. McCabe *et al*. also found a high prevalence of left-sided microsatellite instability (MSI) tumours (36%) in their cohort of indigenous African patients.[8] It should be

noted, for the purposes of LS screening programs in South Africa, that up to a third of dMMR tumours occur in the left colon.

The absence of *BRAF* V600E in any of our samples is in keeping with a growing body of evidence that this point mutation is not commonly found in CRC in indigenous Africans.[8, 12] T6 showed an unusual combination of MLH1, PMS2 and MSH6 protein expression loss, and was found to have a likely germline (VAF=49.10%) *MSH6* VUS (c.3936_3951del). Secondary mutations in MMR genes are described in the literature, resulting in unusual immunohistochemical protein expression patterns.[19] Typically, these are found in *MSH6*. The loss of MLH1/PMS2 in this case may be explained by *MLH1* promotor hypermethylation occurring independently of the *MSH6* variant. Up to 50% of such *MLH1* epigenetic events are associated with *BRAF* V600E mutations, but a minority of cases that are *BRAF* V600 wildtype are associated with *KRAS* variants or *NTRK* fusion events.[20] T6 was shown to have a *KRAS* pathogenic variant. Ultimately, *MLH1* promotor methylation studies would be required to resolve this case.

Concordance between immunohistochemical loss of MLH1/PMS2 and pathogenic variants in the *MLH1* gene was high (n=14; 78%). Of the four discordant cases, one harboured a *MLH1* VUS, two showed a *MSH6* VUS, and one showed no detectable pathogenic variants in the MMR genes. The latter harboured no *BRAF* or *KRAS* variants, but the possibility of *MLH1* epigenetic silencing cannot be excluded without promotor methylation studies. More than half (n=8; 57%) of the 14 *MLH1* pathogenic or likely pathogenic variants detected in this study demonstrated a VAF approximating 50%, suggesting a possible germline mutation. Although this is not recommended as the sole criterion for inferring a likely germline variant, pathogenic variants in somatic genomic profiling with a VAF approximating 50% (accepted range 30-70%) are commonly accepted as representative of a heterozygous germline variant.[21, 22] Tumour heterogeneity and structural aberrations such as copy number variations may, however, introduce uncertainty and these results should be interpreted cautiously.[23] Ultimately, sequencing of paired germline samples would be required to confidently resolve this. We described 12 different pathogenic / likely pathogenic *MLH1* variants, with only one variant, c1528C>T (p.Q510*), which is a well described founder (disease-causing truncation) mutation in our setting, occurring

in more than one patient in the cohort.[24] There were two previously undescribed *MLH1* likely pathogenic variants and one VUS.

There was less agreement between the immunohistochemical staining results and the detected pathogenic variants in cases showing loss of both MSH2/MSH6, isolated loss of MSH6 and isolated loss PMS2 protein expression. Only six (43%) of these cases showed a pathogenic or likely pathogenic variant in the expected gene. Overall however, 11 of the 14 cases (79%) showed a pathogenic or likely pathogenic variant in a non-MLH1 MMR gene. Interestingly, there were no pathogenic *MLH1* variants in any of these cases. Five of these 11 cases (45%) harboured a likely germline variant, and one novel *MSH2* frameshift variant was detected (c.1906dup).

T19 showed isolated MSH6 loss on IHC, but a likely germline pathogenic *MUTYH* variant (with biallelic loss range frequency 99.6%). Aberrant immunohistochemical expression of MMR proteins mimicking LS has been described in isolated patients with *MUTYH*-associated polyposis.[25-28] To our knowledge, this is the index report of isolated loss of MSH6 protein expression in this setting. This finding highlights the importance of including non-MMR genes implicated in familial gastrointestinal cancer when screening for LS. The spectrum of variants encountered in our study population supports the recommendation of other authors to advocate for NGS screening for hereditary CRC and LS.[29]

The frequency of *POLE* and *POLD1* exonuclease variants was 13% (n=4) in our cohort. This proportion is significantly lower than that reported by Brim *et al*. in a small African American cohort, although their cases were unselected for MSI status.[30] The *POLE* P286R variant is well described in the literature. However, the *POLE* D530N variant is previously unreported. Leon-Castillo *et al*. devised a *POLE* score calculator for endometrial carcinoma to address the clinical difficulty in calling pathogenicity in these variants.[31] This schema included a score for tumour mutational burden, nucleotide base change frequencies and MSI status. To our knowledge, such a framework does not exist for calling *POLE* variants in CRC.

WNT and RTK-RAS pathway alterations were less frequent in our dMMR cohort (64% and 39%) compared to those reported by Alatise *et al*. in their Nigerian MSI-H cohort (93% and 82%) and MSKCC African American cohort (100% and 85%).[12] This is likely due to the limited panel of genes included in our panel, as the Alatise study included whole exome sequencing data.

This study has several limitations. Whole exome sequencing would have provided more comprehensive coverage of the MMR genes and would also have more thoroughly interrogated the major pathways in CRC. Despite the negative *BRAF* IHC and PCR in the cases showing MLH1/PMS2 loss of staining, only up to 50% of cases with *MLH1* epigenetic silencing harbour *BRAF* V600E mutations. *MLH1* promotor hypermethylation studies would more definitively classify this mechanism of *MLH1* inactivation. Comparing our tumour genomic sequencing results with paired germline samples (either peripheral blood or non-tumour FFPE tissue) would definitively resolve cases where a germline variant is suspected. This should be the focus of future work. The small sample size also limited the power of this study. Further research involving a larger cohort with more comprehensive molecular profiling is recommended to better characterize this disease in our setting.

3.7. Conclusion

This study has provided preliminary data on the somatic genetic landscape of dMMR CRC in sub-Saharan African with protein expression correlates. In addition, this study highlights the value of utilising readily available archived FFPE samples to generate data regarding the epidemiology of CRC, and tailor downstream diagnostic genetic tests in the local population. The spectrum of mutations encountered in this population and the high frequency of likely germline MMR gene variants supports more widespread screening for hereditary CRC and LS with panel based NGS.

3.8. References

1. Sung H, Ferlay J, Siegel RL, et al. **Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA: a cancer journal for clinicians.* 2021;**71**(3):209-249.

2. Karsa L, Lignini T, Patnick J, et al. **The dimensions of the CRC problem**. *Best Pract Res Clin Gastroenterol.* 2010;**24**(4):381-396.

3. Aldera AP, Pillay K, Robertson B, et al. **Genomic landscape of colorectal carcinoma in sub-Saharan Africa.** *J Clin Pathol.* 2023;**76**(1):5-10.

4. Poulogiannis G, Frayling IM, Arends MJ. **DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome.** *Histopathol*. 2010;**56**(2):167-179.

5. Tamura K, Kaneda M, Futagawa M, et al. **Genetic and genomic basis of the mismatch repair system involved in Lynch syndrome.** *Int J Clin Oncol.* 2019;**24**:999-1011.

6. Ligtenberg MJ, Kuiper RP, Chan TL, et al**. Heritable somatic methylation and inactivation of MSH2 in families with Lynch syndrome due to deletion of the 3′ exons of TACSTD1.** *Nat Genet.* 2009;**41**(1):112-117.

7. Cronje L, Paterson A, Becker P. **Colorectal cancer in South Africa: a heritable cause suspected in many young black patients.** *SAMJ*. 2009;**99**(2):103-106.

8. McCabe M, Perner Y, Magobo R, et al. **Microsatellite Instability assessment in Black South African Colorectal Cancer patients reveal an increased incidence of suspected Lynch syndrome.** *Sci Rep.* 2019;**9**(1):1-10.

9. Vergouwe F, Boutall A, Stupart D, et al. **Mismatch repair deficiency in colorectal cancer patients in a low-incidence area.** *S Afr J Surg.* 2013;**51**(1):16-21.

10. Holla R, Vorster A, Locketz M, et al. **Immunohistochemical determination of mismatch repair gene product in colorectal carcinomas in a young indigenous African cohort.** *S Afr J Surg.* 2022;**60**(1):28-33.

11. Katsidzira L, Vorster A, Gangaidzo IT, et al**. Investigation on the hereditary basis of colorectal cancers in an African population with frequent early onset cases.** *PLOS One.* 2019;**14**(10):e0224023.

12. Alatise OI, Knapp GC, Sharma A, et al. **Molecular and phenotypic profiling of colorectal cancer patients in West Africa reveals biological insights.** *Nat Commun.* 2021;**12**(1):1-8.

13. Richards S, Aziz N, Bale S, et al. **Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology**. *Genet Med.* 2015;**17**(5):405-423.

14. Robinson JT, Thorvaldsdóttir H, Turner D, et al. **igv. js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV).** *Bioinformatics*. 2023;**39**(1):btac830.

15. Gao J, Aksoy BA, Dogrusoz U, et al. **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.** *Sci Signal.* 2013;**6**(269):pl1-pl1.

16. Cerami E, Gao J, Dogrusoz U, et al**. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data.** *Cancer Discov.* 2012;**2**(5):401-404.

17. Okamoto K, Kitamura S, Kimura T, et al**. Clinicopathological characteristics of serrated polyps as precursors to colorectal cancer: current status and management.** *J Gastroenterol Hepatol.* 2017;**32**(2):358-367.

18. Inamura K. **Colorectal cancers: an update on their molecular pathology.** *Cancers*. 2018;**10**(1):26.

19. Shia J, Zhang L, Shike M, et al. **Secondary mutation in a coding mononucleotide tract in MSH6 causes loss of immunoexpression of MSH6 in colorectal carcinomas with MLH1/PMS2 deficiency.** *Mod Pathol.* 2013;**26**(1):131-138.

20. Cocco E, Benhamida J, Middha S, et al. **Colorectal carcinomas containing hypermethylated MLH1 promoter and wild-type BRAF/KRAS are enriched for targetable kinase fusions.** *Cancer Res.* 2019;**79**(6):1047-1053.

21. DeLeonardis K, Hogan L, Cannistra SA, et al. **When should tumor genomic profiling prompt consideration of germline testing?** *J Oncol Prac.* 2019;**15**(9):465-473.

22. Raymond VM, Gray SW, Roychowdhury S, et al**. Germline findings in tumor-only sequencing: points to consider for clinicians and laboratories.** *JNCI.* 2016;**108**(4):djv351.

23. Li MM, Datto M, Duncavage EJ, et al. **Standards and guidelines for the interpretation and reporting of sequence variants in cancer: a joint consensus recommendation of the Association for Molecular Pathology, American Society of Clinical Oncology, and College of American Pathologists.** *J Mol Diagn*. 2017;**19**(1):4-23.

24. Stupart DA, Goldberg P, Algar U, et al. **Surveillance colonoscopy improves survival in a cohort of subjects with a single mismatch repair gene mutation.** *Colorectal Dis.* 2009;**11**(2):126-130.

25. Morak M, Heidenreich B, Keller G, et al. **Biallelic MUTYH mutations can mimic Lynch syndrome.** *Eur J Hum Genet.* 2014;**22**(11):1334-1337.

26. Lefevre JH, Colas C, Coulet F, et al. **MYH biallelic mutation can inactivate the two genetic pathways of colorectal cancer by APC or MLH1 transversions.** *Fam Cancer.* 2010;**9**:589-594.

27. Colebatch A, Hitchins M, Williams R, et al. **The role of MYH and microsatellite instability in the development of sporadic colorectal cancer.** *Br J Cancer*. 2006;**95**(9):1239-1243.

28. Cleary SP, Cotterchio M, Jenkins MA, et al. **Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study.** *Gastroenterol.* 2009;**136**(4):1251-1260.

29. Yildiz S, Musarurwa TN, Algar U, et al. **Genetic insights: High germline variant rate in an indigenous African cohort with early-onset colorectal cancer.** *Front Oncol.* 2023;**13**.

30. Brim H, Shokrani B, Azimi H, et al. **POLE gene mutations in African Americans colorectal cancer.** *Cancer Res*. 2021;**81**(13_Supplement):2413-2413.

31. León-Castillo A, Britton H, McConechy MK, et al. **Interpretation of somatic POLE mutations in endometrial carcinoma.** *J Pathol*. 2020;**250**(3):323-335.

4. CHAPTER FOUR (Paper 3)

Title: Deep Learning Predicts Microsatellite Instability Status in Colorectal Carcinoma in an Ethnically Heterogeneous Population in South Africa

Authors: Alessandro Pietro Aldera* (1, 2, 3), Gregory Patrick Veldhuizen* (4), Didem Cifci (4), Wan-Jung Tsai (1, 5), Komala Pillay (1, 5), Adam Boutall (6), Jakob Nikolas Kather (4, 7, 8), Raj Ramesar (2, 5)

* Contributed equally

Journal: Submitted to Journal of Clinical Pathology (January 2025)

1. Division of Anatomical Pathology, Department of Pathology, University of Cape Town, Cape Town, South Africa

2. UCT MRC Genomic and Precision Medicine Research Unit, Division of Human Genetics, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences and University of Cape Town

3. JDW Pathology Inc, Cape Town, South Africa

4. Else Kroener Fresenius Center for Digital Health, Technical University Dresden, Dresden, Germany

5. National Health Laboratory Services, Groote Schuur Hospital, Cape Town, South Africa

6. Division of General Surgery, Groote Schuur Hospital and University of Cape Town

7. Department of Medicine I, University Hospital Dresden, Dresden, Germany

8. Medical Oncology, National Center for Tumour Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany

Key words: Colorectal carcinoma, microsatellite instability, deep learning, artificial intelligence

4.1. Abstract

Background: Deep learning (DL) models are effective pre-screening tools for detecting mismatch repair deficiency (dMMR) in colorectal carcinoma (CRC). These models have been trained and validated on large cohorts from the Northern Hemisphere, without representation of African samples. We sought to determine the performance of a DL model in an ethnically heterogeneous cohort of patients from South Africa.

Methods: Our cohort comprised 197 CRC resection specimens, with scanned whole slide images (WSI) tessellated and inputted into a transformer-based DL model trained on large international cohorts. Model performance was evaluated using AUROC, sensitivity and specificity. The maximal Youden's J index was calculated to determine the optimal cutoff threshold for model prediction score.

Results: Our model yielded an AUROC of 0.91 (±0.02). Using a prediction score threshold of 0.324 produced an overall sensitivity of 84% (95% CI 71-91) and specificity of 82% (95% CI 76-88). The false negative cases were predominantly left-sided (75%) and did not show the typical dMMR/MSI-H histological phenotype. Sensitivity was lower (50-75%) in cases showing isolated PMS2 or MSH6 loss of staining. Calibrating the classification threshold to 0.15, the sensitivity was optimised to 96% (95% CI 90-100) with a specificity of 60% (95% CI 52-82). This would have resulted in excluding 89 cases (45%) from downstream immunohistochemical (IHC) or molecular testing.

Conclusions: Following appropriate region-specific calibration, we have shown that this model could be employed to accurately pre-screen for dMMR in CRC, thereby reducing the burden of downstream IHC and molecular testing in a resource limited setting.

4.2. Key Messages

- Transformer-based deep learning models can be used to significantly reduce the burden of downstream dMMR testing in CRC in an ethnically heterogeneous resource constrained environment
- Model performance is not transferable between geographic regions and requires local calibration of prediction score threshold

- Reduced sensitivity was found in left-sided CRC, tumours without the typical dMMR histological phenotype, and those with isolated loss of PMS2 or MSH6 on immunohistochemistry

4.3. Introduction

Colorectal carcinoma (CRC) is the third most common cancer worldwide, and contributes significantly to cancer-associated morbidity and mortality.[1] Broadly, CRC is separated into mismatch repair proficient (pMMR) and mismatch repair deficient (dMMR) groups based on the expression of MMR proteins by immunohistochemistry (IHC). dMMR cases may also be identified by testing for microsatellite instability (MSI) with polymerase chain reaction (PCR) or by applying bioinformatics tools to data generated from next generation sequencing (NGS). The accuracy of these techniques is comparable, but IHC is favoured in routine diagnostic practice as it is relatively inexpensive, technically simpler to perform, and has a rapid turnaround time. Identification of dMMR/MSI-H tumours is important for Lynch syndrome (LS) screening programs, has prognostic and predictive therapeutic value, and in the age of immunotherapy, is becoming increasingly relevant.[2, 3]

The incidence of CRC in sub-Saharan Africa is set to rise dramatically over the next decade.[4] Over the past two decades, several authors have noted an increased proportion of dMMR CRC in sub-Saharan Africa (mean 26%; range 7-67%) compared to the reported global average of 12%.[5-9] Despite this, the molecular genetics and tumour biology of CRC in sub-Saharan Africa remains poorly characterised.[10] The ethnic heterogeneity of the South African population, encompassing a diverse mix of African, European, Asian, and mixed ancestries, is important in the context of MSI prediction. Unique genetic and molecular features arising from this diversity may influence the performance of diagnostic models developed in more ethnically homogeneous populations. Several international professional bodies recommend universal screening of MMR/MSI status in all index CRC cases.[11] This recommendation has not been adopted widely, if at all, in sub-Saharan Africa, due to prohibitive costs and lack of ready access to testing facilities.

Deep learning (DL) models, applied to histopathology whole slide images (WSI), have emerged as a reliable and accurate tool to reduce the need for downstream IHC and molecular testing.[12-17] However, these models have predominantly been trained on populations from North America and Western Europe, raising concerns about their generalizability to other populations. Applying models across different populations can be challenging due to potential genetic and molecular differences that may affect model performance.[13] Known genetic variations, such as differences in MMR gene mutations and MSI patterns specific to certain ethnic groups, could impact the histopathological features that DL models rely on for accurate prediction. This underscores the need to evaluate and possibly recalibrate these models for use in diverse populations like that of South Africa.

The primary aim of this study was to determine whether a DL model can be effectively used as a pre-screening tool in the ethnically heterogeneous population of South Africa, to decrease the downstream burden of MMR/MSI testing. Secondary aims were to describe the clinicopathological features of false negative (FN) cases which may reveal biological insights.

4.4. Methods

4.4.1. Study Design and Data Collection

This was a retrospective cross sectional study. Cases were identified from CRC resection specimens received between 2016 and 2020 in the Division of Anatomical Pathology, University of Cape Town and National Health Laboratory Services, Groote Schuur Hospital, Cape Town, South Africa. Patients who were 60 years old or younger at the time of resection were included in the study. This age restriction was chosen to focus on a population more likely to have hereditary or early-onset CRC, which is often associated with different genetic and molecular profiles compared to CRC in older patients. Resection specimens with insufficient residual tumour or a complete response to neoadjuvant therapy were excluded. Clinicopathological data were obtained from the histopathology reports. The MMR IHC slides (MLH1, MSH2, MSH6, PMS2) for each case were reviewed by two pathologists to confirm the MMR status (AA, WJT).

4.4.2. Slide Selection and Image Acquisition

Representative formalin fixed paraffin embedded (FFPE) tissue blocks, with adequate viable tumour, were selected by a pathologist for each case. These blocks were recut at 4μm, and the tissue sections were stained with Haematoxylin and Eosin (H&E) to produce the slides. Slides were scanned on an Aperio GT450 DX slide scanner (Leica Biosystems, Nussloch, Germany) at 40x magnification. The scanned slides were manually inspected for quality prior to downstream analysis.

4.4.3. WSI Preprocessing and Feature Extraction Using the STAMP Protocol

To predict the MSI status from whole-slide images (WSIs) of CRC tissue samples, we employed the Solid Tumor Associative Modeling in Pathology (STAMP) protocol, an open-source, standardized workflow designed for end-to-end DL in computational pathology.[18] The STAMP protocol facilitates the extraction of phenotypic features from routine histopathology slides and their integration with molecular data to predict clinically relevant biomarkers. Initially, each WSI was divided into non-overlapping image tiles to manage the high resolution and large size of the slides. These tiles were then processed to remove background regions and low-quality areas using an adaptive thresholding method, ensuring that only informative regions of the tissue were retained for subsequent analysis.

4.4.4. Model Development and Deployment

A weakly-supervised DL approach was employed, wherein each WSI was labelled with the corresponding MSI status. A state-of-the-art pre-trained self-supervised learning model was used to extract features from each tile, which were then aggregated to generate a patient-level feature vector. The extracted features were input into a transformer-based model, which was trained to predict MSI status. The model was optimized using cross-entropy loss. We based our study on the model trained by Wagner *et al*. previously and merely retrained it with our updated pipeline, which was published separately as a STAMP.[16] For model validation we used our novel CAPE cohort. All source codes for preprocessing are available at

https://github.com/KatherLab/preProcessing and all source codes for DL are available as part of the 'Histology Image Analysis' (HIA) package at https://github.com/KatherLab/HIA.

4.4.5. Statistics

The primary method of performance evaluation was based on the analysis of the receiver operating characteristic (ROC) curve, with a focus on the area under the ROC curve (AUROC). To determine the optimal screening threshold for the model, we employed Youden's J Index, which balances sensitivity and specificity. The maximal Youden's J index was calculated as 0.67 which corresponded to a prediction threshold of 0.324. Additionally, we conducted subgroup analyses to assess the model's performance across various demographic and clinical subgroups, including sex, age, tumour sidedness, histological grade, histological type, tumour infiltrating lymphocytes, and immunohistochemical (IHC) staining for mismatch repair (dMMR) genes. This comprehensive analysis allowed us to explore the robustness and consistency of the model's predictive capabilities across different patient subpopulations.

4.5. Results

197 scanned slides with corresponding complete clinicopathological data were included in the analysis (Table 4.1). There were 49 (25%) dMMR cases and 148 (75%) pMMR cases in this cohort. Of the dMMR cases, 26 (53%) showed MLH1/PMS2 loss, 8 (16%) showed MSH2/MSH6 loss, 4 (8%) showed isolated PMS2 loss, and 4 (8%) showed isolated MSH6 loss of staining on IHC. Seven (14%) dMMR cases showed an unconventional combination of loss of staining and were excluded from the subgroup analysis. Our model achieved an AUROC of 0.91 (±0.02) on the CAPE cohort (Figure 4.1). Youden's J index was used to select the optimal cutoff threshold of 0.324 (Table 4.2). This yielded an overall sensitivity of 84% (95% CI 71-91) and specificity of 82% (95% CI 76-88) for predicting dMMR (Table 4.3). Representative H&E stained tissue sections are shown with their corresponding heatmaps at different model prediction scores (Figure 4.2).

Table 4.1: Clinicopathological Characteristics of the CAPE Cohort

	n	dMMR		pMMR	
Gender					
Male	107	30	28%	77	72%
Female	90	19	21%	71	79%
Age					
<40	27	8	30%	19	70%
40-49	54	22	41%	32	59%
50-60	116	19	16%	97	84%
Side					
Right	72	30	42%	42	58%
Left	125	19	15%	106	85%
Anatomical Site					
Caecum	26	14	54%	12	46%
Ascending	30	11	37%	19	63%
Transverse	16	5	31%	11	69%
Descending	22	6	27%	16	73%
Sigmoid	51	9	18%	42	82%
Rectum	52	4	8%	48	92%
Histological Grade					
1	13	4	31%	9	69%
2	166	32	19%	134	81%
3	29	13	45%	16	55%
Histological Type					
Adenocarcinoma, NOS	173	37	21%	136	79%
Mucinous	24	12	50%	12	50%
TILs					
0	111	11	10%	100	90%
1	57	17	30%	40	70%
2	15	8	53%	7	47%
3	14	10	71%	4	29%
Overall	197	49	25%	148	75%

dMMR mismatch repair deficient, pMMR mismatch repair proficient**,** NOS not otherwise specified, TILs tumour infiltrating lymphocytes

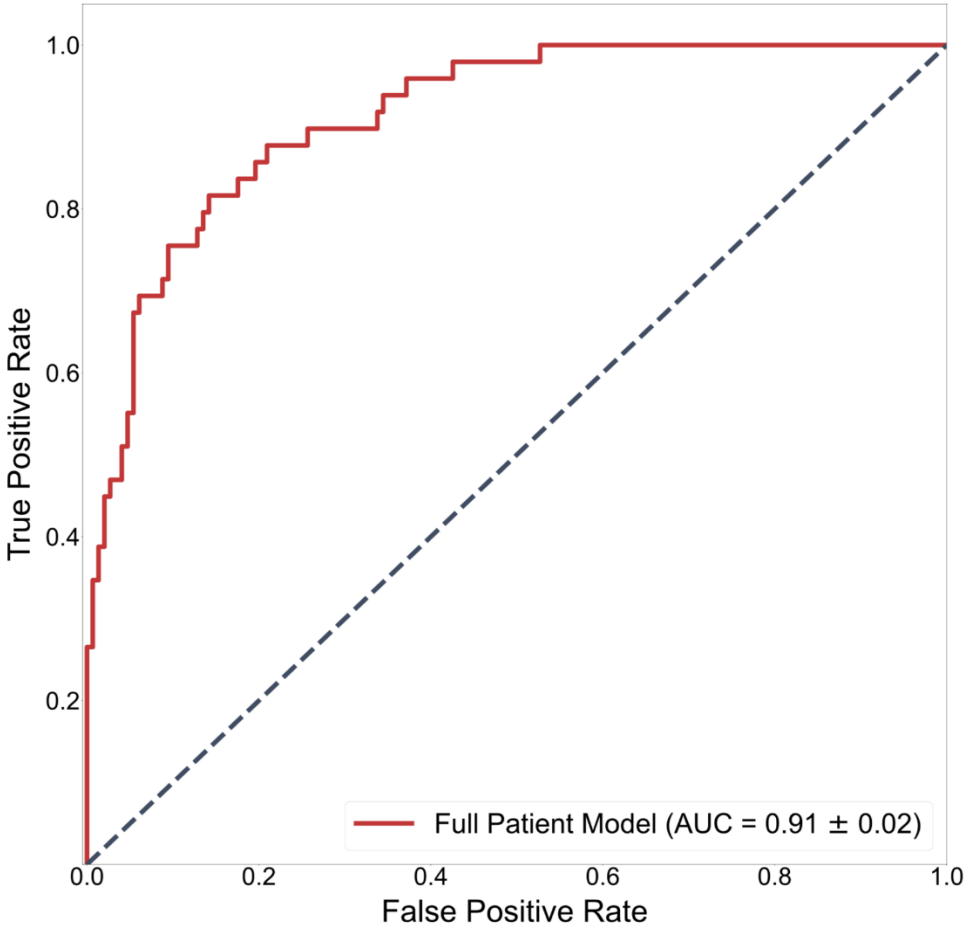


Figure 4.1: ROC Curve for South African CAPE Cohort (AUROC=0.91)

Table 4.2: Evaluation of Model Performance in the CAPE Cohort at Different Cut-off Thresholds

	Model Calibration Threshold						
	0,624	**0,524**	**0,424**	**0,324**	**0,224**	**0,174**	**0,15**
TP	33	35	40	43	44	46	47
TN	140	134	126	117	105	97	89
FP	8	14	22	31	43	51	59
FN	16	14	9	6	5	3	2
Sensitivity	0,67	0,71	0,82	0,88	0,9	0,94	0,96
Specificity	0,95	0,91	0,85	0,79	0,71	0,66	0,6
Youden's J index	0,62	0,62	0,67	**0,67**	0,61	0,6	0,56

TR true positive, TN true negative, FP false positive, FN false negative

Table 4.3: Sensitivity and Specificity Subgroup Analysis by Key Clinicopathological Variables

	Yoden's J Index Threshold (0.324)				Sensitivity Optimised Threshold (0.15)			
	Sensitivity % (95% CI)		Specificity % (95% CI)		Sensitivity % (95% CI)		Specificity % (95% CI)	
Overall	84	(71-91)	82	(76-88)	95.9	(90.4-101.5)	60.1	(52.2-68.0)
Age								
<40	88	(53-98)	63	(41-81)	100	(100-100)	57.9	(35.7-80.1)
40-49	86	(67-95)	75	(58-87)	95.5	(86.8-104.6)	50	(32.7-67.3)
50-60	79	(57-91)	89	(81-94)	94.7	(84.7-104.8)	63.9	(54.4-73.5)
Side								
Right	93	(79-98)	76	(61-87)	100	(100-100)	54.8	(39.7-69.8)
Left	68	(46-85)	85	(77-90)	89.5	(75.7-103.3)	62.3	(53.0-71.5)
Histological Grade								
1	100	(51-100)	78	(45-94)	100	(100-100)	66.7	(35.9-97.5)
2	75	(58-87)	85	(78-90)	93.8	(85.4-102.1)	61.2	(52.9-69.4)
3	100	(77-100)	20	(4-62)	100	(100-100)	20.0	(-.15-55)
Histological Type								
Adenocarcinoma, NOS	78	(63-89)	84	(77-89)	94.6	(87.3-101.9)	61	(52.8-69.2)
Mucinous	100	(76-100)	67	(39-86)	100	(100-100)	50	(21.7-78.3)
TILs								
0	71	(45-88)	82	(74-89)	92.9	(79.4-106.3)	60.8	(51.1-70.5)
1	88	(66-97)	82	(68-91)	100	(100-100)	62.5	(47.5-77.5)
2	75	(41-93)	86	(49-97)	87.5	(64.6-110.4)	57.1	(20.5-93.8)
3	100	(72-100)	75	(30-95)	100	(100-100)	25.0	(-17.4-67.4)

NOS not otherwise specified, TILs tumour infiltrating lymphocytes

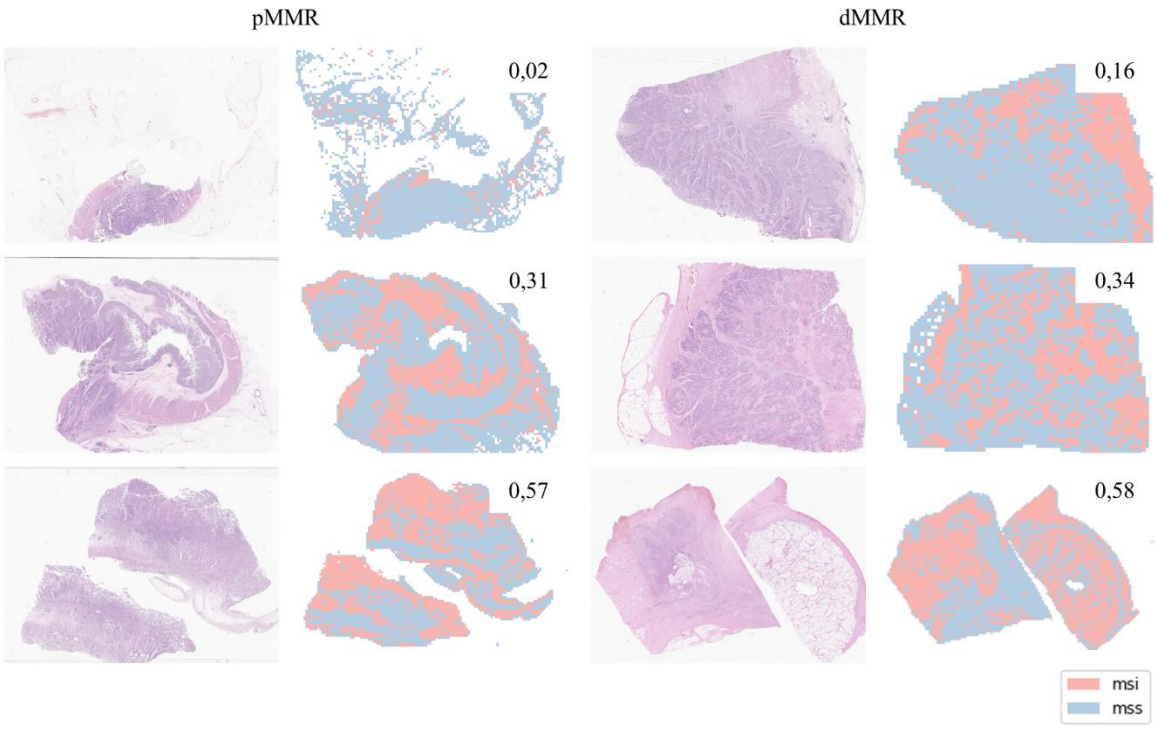


Figure 4.2: Representative Haematoxylin-and-Eosin (H&E) Stained Sections with Corresponding Heatmaps Stratified by MMR Status and Model Prediction Scores

4.5.1. Performance Stratified by Key Clinicopathological Variables

Patient age at the time of resection was stratified into 3 tiers; <40 years of age showed 88% sensitivity and 63% specificity, 40-49 years 86% sensitivity and 75% specificity, and 50-60 years 79% sensitivity and 89% specificity. The overall specificity in right-sided tumours (caecum, ascending colon, hepatic flexure and transverse colon) was 76%, as compared to 85% in left-sided tumours (splenic flexure, descending colon, sigmoid colon and rectum). The sensitivity was lower for left-sided tumours (68%) compared to right-sided tumours (93%). Tumours which were histologically classified as mucinous (>50% pools of extracellular mucin containing clumps of malignant epithelium) showed a sensitivity of 100%, compared to 78% in conventional adenocarcinoma not otherwise specified (NOS). The specificity in mucinous tumours was 67%, and 84% in adenocarcinoma NOS. The sensitivity was 100% in grade 1 tumours, 75% in grade 2 tumours, and 100% in grade 3 tumours. The sensitivity at different TILs scores was 71% (score 0), 88% (score 1), 75% (score 2) and 100% (score 3). The sensitivity in dMMR tumours

showing MLH1/PMS2 loss on IHC was 92%, MSH2/MSH6 loss 68%, isolated PMS2 loss 50% and isolated MSH6 loss 75%.

4.5.2. Clinicopathological Characteristics of Misclassified Cases

The clinicopathological and molecular characteristics of the misclassified cases, specifically those assigned MSS scores which were dMMR by IHC (false negative; FN), were further investigated (Table 4.4). Photomicrographs from representative H&E stained sections with corresponding model prediction scores are shown (Figure 4.3). Four of the 49 dMMR cases were assigned MSI probability scores of <0.324. Among these 4 cases, 3 (75%) occurred in the left colon, 4 (100%) did not have a significant mucinous component, and 3 (75%) had a TILs score of 0. Two (50%) cases showed IHC loss of MLH1/PMS2, 1 (25%) isolated PMS2 loss, and 1 (25%) isolated MSH6 loss. One case had targeted germline sequencing, which demonstrated a pathogenic *MLH1* C1528T variant (Lynch syndrome).

Table 4.4: Analysis of Initial False Negative and False Positive Cases Using Yoden's J Index Determined Threshold (0.324) Stratified by Clinicopathological Variables

	False Negative		False Positive	
Gender				
Male	3	75%	20	65%
Female	1	25%	11	35%
Age				
<40	1	25%	7	23%
40-49	1	25%	11	35%
50-60	2	50%	13	42%
Side				
Right	1	25%	12	39%
Left	3	75%	19	61%
Histological Grade				
1	0	0%	2	6%
2	4	100%	25	81%
3	0	0%	4	13%
Histological Type				
Adenocarcinoma, NOS	4	100%	27	87%
Mucinous	0	0%	4	13%
TILs				
0	3	75%	20	65%
1	1	25%	9	29%
2	0	0%	1	3%
3	0	0%	1	3%
Immunohistochemistry				
MLH1/PMS2	2	50%		
MSH2/MSH6	0	0%		
PMS2	1	25%		
MSH6	1	25%		
Overall	4		31	

NOS not otherwise specified, TILs tumour infiltrating lymphocytes

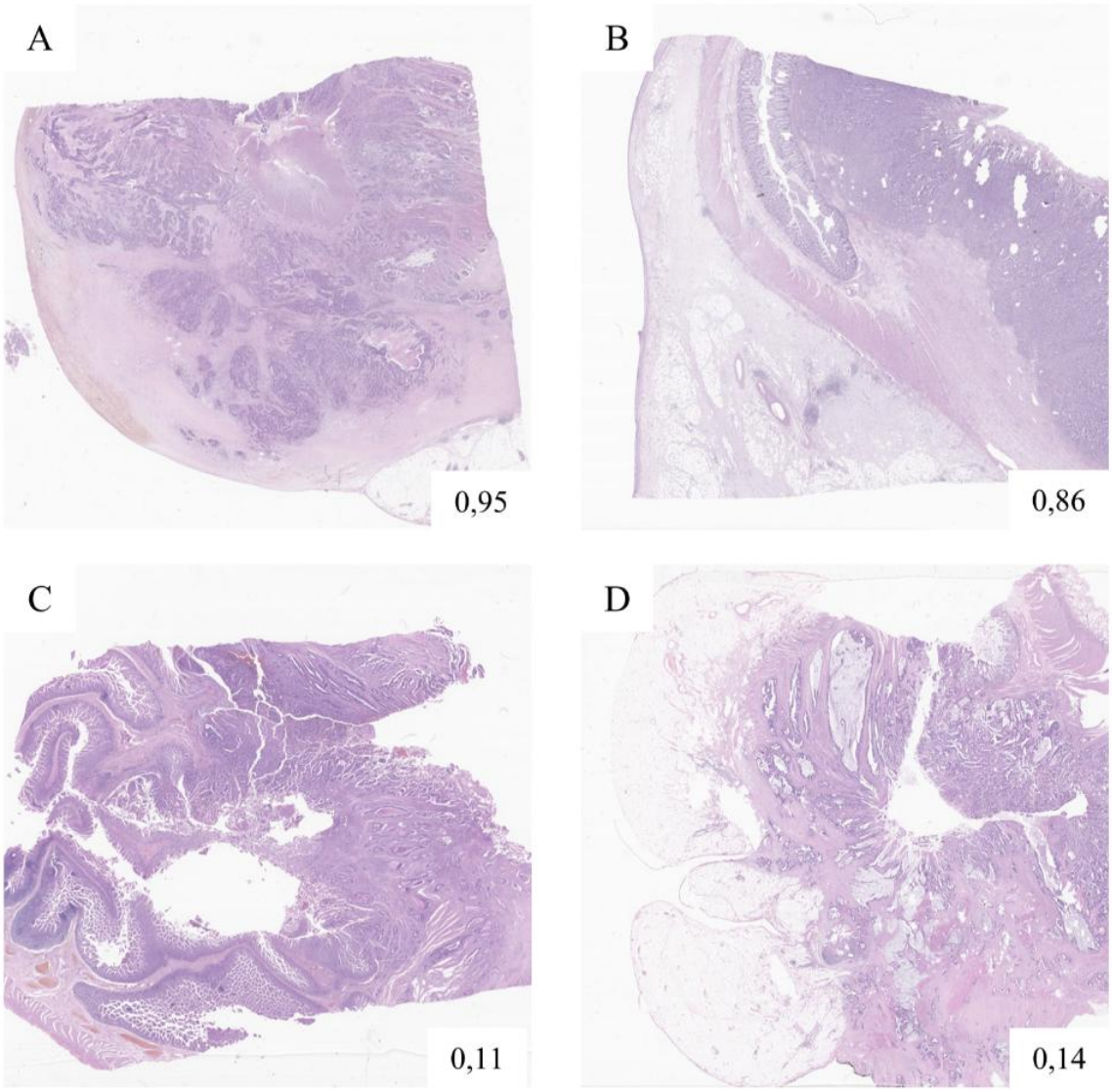


Figure 4.3: Haematoxylin-and-Eosin (H&E) Stained Sections of Selected Misclassified Cases with Corresponding Model Prediction Scores. **A&B** pMMR Tumours with High MSI Prediction Scores. **C&D** dMMR Tumours with Low MSI Prediction Scores.

4.5.3. Adjusting the Cutoff Threshold to Achieve Optimal Sensitivity

Using this model at a prediction score cutoff threshold of 0.324 with a sensitivity of 84% would have resulted in 4 FN cases. This is not ideal for a screening test, which prompted us to examine model performance at different cutoff thresholds and attempt to identify the optimal threshold to

achieve clinical grade sensitivity of >95% (Table 2). Using a cutoff of 0.15, we showed a sensitivity of 96% with a specificity of 60%. This would result in only 2 FN cases and reduce the number of cases needing downstream diagnostic tests by 89 (45%). These 2 FN cases both occurred in the left colon, were histologically classified as adenocarcinoma NOS, and both showed a less common dMMR pattern of staining (one isolated PMS2 loss, and the other isolated MSH6 loss).

4.6. Discussion

In this study we investigated the efficacy of a transformer-based DL model for predicting dMMR/MSI-H in CRC, in a novel ethnically heterogeneous study population from South Africa. The overall performance (AUROC 0.91) in our cohort is lower than what has been shown in previous studies.[12, 13, 15, 17] Using the prediction threshold of 0.324, we achieved an overall sensitivity of 84% (95% CI 71-91) and specificity of 82% (95% CI 76-88). This sensitivity is low for a screening test, and risks missing dMMR cases. This might impact patient outcomes, particularly in resource-limited settings where the adoption of such models is intended to alleviate the burden of more expensive or less accessible diagnostic methods. The low sensitivity may be due to selection bias, as our cohort was enriched for younger patients, and as a result, the proportion of dMMR cases was 25%, which is more than twice that reported in the general population. There was also a significant proportion of cases in our cohort which did not demonstrate the typical dMMR/MSI-H morphological phenotype (mucinous differentiation, high TILs scores and a brisk peritumoral lymphocytic response), and this may have impacted the ability of the DL model to accurately classify these cases.[14, 19, 20]

Stratifying the performance of the model by clinicopathological variables revealed several interesting observations. Younger age at resection (<40 years) and poorly differentiated (grade 3) tumours were associated with a lower specificity. The performance for excluding dMMR/MSH-H in poorly differentiated (grade 3) tumours was particularly poor (20%). Poorly differentiated tumours are often associated with dMMR/MSI-H, and this may have led the model to assign a higher score to these cases. The sensitivity was higher for detecting dMMR in

tumours showing loss of MSH2/MSH6 (100%) or MLH1/PMS2 (96%), compared to those with isolated PMS2 (50%) or isolated MSH6 (75%) loss. This finding could be related to the small number of dMMR cases in this cohort with isolated PMS2 or MSH6 loss, as these are infrequent patterns of dMMR. However, this may be important to note in population groups where less common mechanisms of dMMR are more frequent. Most of the original work done with DL models and dMMR/MSI-H prediction in CRC used MSI PCR as the ground truth and so the performance variation with different MMR protein expression patterns is not fully known.

Amongst the 4 dMMR cases which were initially missed by our model, the majority (75%) of tumours were left-sided and did not exhibit the typical morphological phenotype associated with dMMR/MSI-H. It is important to highlight that 1 of these cases had confirmed LS by germline testing. Echle *et al*. suggested a variety of technical factors (scanning artifacts and little viable tumour tissue) as possible explanations for their FN cases.[13] After review of our FN slides, we did not find such technical factors, and propose that the DL prediction model has a lower accuracy in left-sided tumours and dMMR tumours which do not exhibit the typical morphological phenotype. This is of particular relevance in sub-Saharan Africa, where a larger proportion of left-sided tumours are reported to occur.[21-23]

In order to identify the optimal cut-off threshold for maximizing sensitivity, we explored the effect of incrementally reducing the threshold on sensitivity and specificity in the CAPE cohort. To achieve clinical grade sensitivity of >95%, a threshold of 0.15 was required. With this calibration, the specificity was reduced to 60%, which was still able to reduce the number of cases which would need downstream diagnostic testing by 45%. Determining the optimal sensitivity and specificity for a screening test likely requires at least matching the current gold standard in terms of sensitivity, so as to minimize FN cases. The small sample size limits the power of this study. Future work should be done on larger cohorts from sub-Saharan Africa, with specific focus on developing region-specific cutoff thresholds to account for the unique genetic characteristics of the population. Additionally, integrating other data types, such as genomics or transcriptomics, could potentially enhance model accuracy and ensure better clinical outcomes.

In summary, we have shown good performance of this DL model in our ethnically heterogeneous cohort (AUROC=0.91) and that sensitivity can be optimized to clinical grade by adjusting the prediction score cutoff threshold in a systematic manner. Once calibrated to the regional population, this model could be employed routinely as a pre-screening tool to reduce the need for downstream diagnostic tests by at least 45%.

4.7. Author Statements

Ethical Approval: This study was performed in accordance with the Declaration of Helsinki. This study is a retrospective analysis of digital images of anonymized archival tissue samples of multiple cohorts of colorectal cancer patients. The overall analysis was approved by the Ethics board at University Hospital Carl Gustav Carus, Dresden, Germany. Approval for access to patient demographic data, pathology reports and the use of archival FFPE tissue was granted by the University of Cape Town Human Research Ethics Committee (HREC 202/2002).

Funding: JNK is supported by the German Cancer Aid (DECADE, 70115166), the German Federal Ministry of Education and Research (PEARL, 01KD2104C; CAMINO, 01EO2101; SWAG, 01KD2215A; TRANSFORM LIVER, 031L0312A; TANGERINE, 01KT2302 through ERA-NET Transcan; Come2Data, 16DKZ2044A; DEEP-HCC, 031L0315A), the German Academic Exchange Service (SECAI, 57616814), the German Federal Joint Committee (TransplantKI, 01VSF21048) the European Union's Horizon Europe and innovation programme (ODELIA, 101057091; GENIAL, 101096312), the European Research Council (ERC; NADIR, 101114631), the National Institutes of Health (EPICO, R01 CA263318) and the National Institute for Health and Care Research (NIHR, NIHR203331) Leeds Biomedical Research Centre.

The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Contributions:

Alessandro Pietro Aldera* (conceptualisation, writing manuscript, data analysis)

Gregory Patrick Veldhuizen* (conceptualisation, writing manuscript, data analysis)

Didem Cifci (data analysis, editing manuscript)

Wan-Jung Tsai (collection of sample material, editing manuscript)

Komala Pillay (supervision, editing manuscript)

Adam Boutall (supervision, editing manuscript)

Jakob Nikolas Kather (conceptualisation, supervision, editing manuscript)

Raj Ramesar (conceptualisation, supervision, editing manuscript)

* Contributed equally

4.8. References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, Bray F: **Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA: a cancer journal for clinicians* 2021, **71**(3):209-249.

2. Kather JN, Halama N, Jaeger D: **Genomics and emerging biomarkers for immunotherapy of colorectal cancer.** *In: Seminars in cancer biology:* 2018: Elsevier; 2018, **52**:89-197.

3. André T, Shiu K-K, Kim TW, Jensen BV, Jensen LH, Punt C, Smith D, Garcia-Carbonero R, Benavides M, Gibbs P: **Pembrolizumab in microsatellite-instability–high advanced colorectal cancer.** *NEJM* 2020, **383**(23):2207-2218.

4. Karsa L, Lignini T, Patnick J, Lambert R, Sauvaget C: **The dimensions of the CRC problem**. *Best Pract Res Clin Gastroenterol* 2010, **24**(4):381-396.

5. Cronje L, Paterson A, Becker P: Colorectal cancer in South Africa: a heritable cause suspected in many young black patients. *S Afri Med J* 2009, **99**(2):103-106.

6. McCabe M, Perner Y, Magobo R, Magangane P, Mirza S, Penny C: **Microsatellite Instability assessment in Black South African Colorectal Cancer patients reveal an increased incidence of suspected Lynch syndrome.** *Sci Rep* 2019, **9**(1):1-10.

7. Vergouwe F, Boutall A, Stupart D, Algar U, Govender D, Van der Linde G, Mall A, Ramesar R, Goldberg P: **Mismatch repair deficiency in colorectal cancer patients in a low-incidence area.** *S Afr J Surg* 2013, **51**(1):16-21.

8. Holla R, Vorster A, Locketz M, De Haas M, Oke O, Govender D, Ramesar R, Goldberg P: **Immunohistochemical determination of mismatch repair gene product in colorectal carcinomas in a young indigenous African cohort**. *S Afr J Surg* 2022, **60**(1):28-33.

9. Katsidzira L, Vorster A, Gangaidzo IT, Makunike-Mutasa R, Govender D, Rusakaniko S, Thomson S, Matenga JA, Ramesar R: **Investigation on the hereditary basis of colorectal cancers in an African population with frequent early onset cases**. *PLOS one* 2019, **14**(10):e0224023.

10. Aldera AP, Pillay K, Robertson B, Boutall A, Ramesar R: **Genomic landscape of colorectal carcinoma in sub-Saharan Africa.** *J Clin Pathol* 2023, **76**(1):5-10.

11. Marks K, West N: **Molecular assessment of colorectal cancer through Lynch syndrome screening.** *Diagn Histopathol* 2020, **26**(1):47-50.

12. Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, Carpentier S, Reyre J, Enea D, von Loga K: **Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides.** *Nat Commun* 2023, **14**(1):6695.

13. Echle A, Laleh NG, Quirke P, Grabsch H, Muti H, Saldanha O, Brockmoeller S, van den Brandt P, Hutchins G, Richman S: **Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application.** *ESMO Open* 2022, **7**(2):100400.

14. Echle A, Laleh NG, Schrammen PL, West NP, Trautwein C, Brinker TJ, Gruber SB, Buelow RD, Boor P, Grabsch HI: **Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review**. *ImmunoInformatics* 2021, **3**:100008.

15. Kather JN, Pearson AT, Halama N, Jäger D, Krause J, Loosen SH, Marx A, Boor P, Tacke F, Neumann UP: **Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer.** *Nat Med* 2019, **25**(7):1054-1056.

16. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, Veldhuizen GP, Quirke P, Grabsch HI, van den Brandt PA: **Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study.** *Cancer Cell* 2023, **41**(9):1650-1661. e1654.

17. Yamashita R, Long J, Longacre T, Peng L, Berry G, Martin B, Higgins J, Rubin DL, Shen J: **Deep learning model for the prediction of microsatellite instability in colorectal cancer: a diagnostic study**. *Lancet Oncol* 2021, **22**(1):132-141.

18. El Nahhas OS, van Treeck M, Wölflein G, Unger M, Ligero M, Lenz T, Wagner SJ, Hewitt KJ, Khader F, Foersch S: **From whole-slide image to biomarker prediction: a protocol for end-to-end deep learning in computational pathology**. *arXiv preprint* arXiv:231210944 2023.

19. Greenson JK, Bonner JD, Ben-Yzhak O, Cohen HI, Miselevich I, Resnick MB, Trougouboff P, Tomsho LD, Kim E, Low M: **Phenotype of microsatellite unstable colorectal carcinomas: well-differentiated and focally mucinous tumors and the absence of dirty necrosis correlate with microsatellite instability.** *Am J Surg Pathol* 2003, **27**(5):563-570.

20. Jenkins MA, Hayashi S, O'shea A-M, Burgart LJ, Smyrk TC, Shimizu D, Waring PM, Ruszkiewicz AR, Pollett AF, Redston M: **Pathology features in Bethesda guidelines predict colorectal cancer microsatellite instability: a population-based study.** *Gastroenterol* 2007, **133**(1):48-56.

21. Alatise OI, Knapp GC, Sharma A, Chatila WK, Arowolo OA, Olasehinde O, Famurewa OC, Omisore AD, Komolafe AO, Olaofe OO: **Molecular and phenotypic profiling of colorectal cancer patients in West Africa reveals biological insights.** *Nat Commun* 2021, **12**(1):1-8.

22. Madiba T, Moodley Y, Sartorius B, Sartorius K, Aldous C, Naidoo M, Govindasamy V, Bhadree S, Stopforth L, Ning Y: **Clinicopathological spectrum of colorectal cancer among the population of the KwaZulu-Natal Province in South Africa.** *Pan Afr Med J* 2020, **37**(74).

23. McCabe M, Penny C, Magangane P, Mirza S, Perner Y: **Left-sided colorectal cancer distinct in indigenous African patients compared to other ethnic groups in South Africa**. *BMC cancer* 2022, **22**(1):1089.

5. CHAPTER FIVE (Paper 4)

Title: Somatic Whole Exome Sequencing of Colorectal Carcinoma in Young Patients from sub-Saharan Africa Reveals Novel Insights

Authors: Alessandro Pietro Aldera (1, 2, 3), Dennis Owusu (4), Leonardo Biral (4), Komala Pillay (1, 5), Adam Boutall (6), Sandeep Dave (4), Raj Ramesar (2, 5)

Journal: Submitted to The Journal of Pathology: Clinical Research (January 2025)

1. Division of Anatomical Pathology, Department of Pathology, University of Cape Town, Cape Town, South Africa

2. UCT MRC Genomic and Precision Medicine Research Unit, Division of Human Genetics, Department of Pathology, Institute of Infectious Diseases and Molecular Medicine, Faculty of Health Sciences and University of Cape Town

3. JDW Pathology Inc, Cape Town, South Africa

4. Center for Genomic and Computational Biology and Department of Medicine, Duke University, Durham, NC, USA

5. National Health Laboratory Services, Groote Schuur Hospital, Cape Town, South Africa

6. Division of General Surgery, Groote Schuur Hospital and University of Cape Town

Key words: Colorectal cancer, sub-Saharan Africa, whole exome sequencing

5.1. Abstract

Colorectal cancer (CRC) is a frequent cause of morbidity and mortality in sub-Saharan Africa. The incidence of early-onset, microsatellite stable (MSS) CRC is on the rise, and the tumour biology of these lesions is poorly categorised. Preliminary data from one centre in Nigeria found differences in the frequencies of mutations in driver genes and altered signalling pathways. We sought to investigate potential alternative driver genes and signalling pathways by whole exome sequencing (WES). Eighty-three cases passed quality control filters and were included in the analysis (77 MSS, 4 MSI, 2 POL). *APC*, *TP53* and *KRAS* were among the most frequently mutated driver genes, although at a lower frequency than expected. *BRAF* V600E mutations were absent in our cohort. Although there were differences in the frequencies of mutations in the major driver genes, the frequencies of oncogenic pathway alterations were found to be similar. *FAT4* (26%) and *TET2* (15%) emerged as important mutated driver genes and potential therapeutic targets for further investigation. We have highlighted distinct differences in driver gene mutations in our cohort of young CRC from sub-Saharan Africa and identified *FAT4* and *TET2* as potential drivers that are more common and are potential therapeutic targets.

5.2. Introduction

Colorectal carcinoma (CRC) is the fourth most common cancer and the fifth leading cause of cancer-associated mortality in sub-Saharan Africa.[1] The incidence is predicted to increase significantly over the next decade, particularly in low- and middle-income countries.[2] Earlier work has suggested that a significant proportion of CRC in sub-Saharan Africa has a clinicopathological phenotype that differs from what is seen in developed countries.[3-9] Specifically, there appears to be an increased incidence of early-onset left-sided CRC in the indigenous African population.[8-11] Previous studies have investigated the reported increased incidence of microsatellite instability high (MSI-H) CRC in the indigenous African and African American population.[12] The molecular profile of microsatellite stable (MSS) tumours remains poorly categorised in sub-Saharan Africa. Alatise *et al*. are the only authors to have performed panel sequencing on a cohort of 64 CRC tumour specimens from Nigeria.[9] Their findings highlighted distinct differences in the major CRC signalling pathways and driver genes, particularly in MSS tumours.

Recent large-scale high-throughput sequencing studies have confirmed that *APC*, *TP53* and *KRAS* are the dominant mutated driver genes in CRC.[13, 14] Several authors have shown a higher frequency of tumours with wild-type *BRAF* and *KRAS* in sub-Saharan Africa.[4, 9] Recently, Alatise *et al*. showed that the frequency of mutations in these driver genes was significantly different in their Nigerian cohort, though they did not identify any novel potential drivers.[9] Identifying the dominant molecular signalling pathways in these tumours is critical to understanding the tumour biology and guiding screening and treatment strategies in sub-Saharan Africa.

In this study, we sought to describe the driver gene mutations and dominant oncogenic signalling pathways with somatic whole exome sequencing (WES).

5.3. Methods

5.3.1. Study Design and Data Collection

CRC resection samples, reported in the Division of Anatomical Pathology at Groote Schuur Hospital (Cape Town, South Africa) between 2016-2020, were identified retrospectively by searching the National Health Laboratory Service (NHLS) laboratory information system. Representativeness and adequacy of archived Haematoxylin and Eosin (H&E) stained glass slides and formalin-fixed paraffin-embedded (FFPE) tissue wax blocks were assessed by a consultant histopathologist. One hundred and thirty-eight cases with complete clinicopathological information and age under 60 years at the time of resection were selected. Cases that were mismatch repair proficient (pMMR) by immunohistochemistry were preferentially selected for WES, although some young unexplained dMMR cases were included. Rectal tumours with insufficient residual viable tumour cells post-neoadjuvant chemotherapy were excluded.

5.3.2. Sample collection and DNA Extraction

Ten 10μm scrolls were taken from a representative FFPE tissue wax block for each case and sent to Duke University for molecular analysis. Following the manufacturer's protocol, nucleic acid extraction and sequencing library preparation were conducted using the Duoseq Research Kit (EPXv3, Data Driven Bioscience, Durham, NC).

5.3.3. WES and Alignment

Sequencing was performed on the Illumina platform, adhering to the manufacturer's guidelines for data generation and analysis. Of the samples submitted, 137 were successfully extracted and underwent WES. The median exonic coverage for these cases was 26.8X. FASTQ files containing DNA sequencing reads were processed using Trimmomatic (v0.39) in paired-end mode to remove adapter sequences and low-quality reads.[15, 16] DNA reads were then aligned to the human reference genome (GRCh38.p12, with a PAR mask on chary) using Sentieon BWAmem (v201911) with the default parameters.[17] PCR duplicate reads were identified and marked using Picard (v2.8.1) (http://broadinstitute.github.io/picard). Quality control metrics were generated using Picard, FASTQC (v0.11.8) (www.bioinformatics.babraham.ac.uk/projects/fastqc), and samtools (v1.13).[18]

5.3.4. Single-nucleotide Variant Calling

Variant calling analysis was performed using DNA reads from the exome-sequenced samples, leveraging a pipeline that uses the union of variants called by each of Strelka2 (v2.9.10), DeepVariant (v1.1.0), and Sentieon Haplotyper (v201911) to call variants.[17, 19] Of the 137 analyzed samples, 83 met the quality criteria for variant calling. Synonymous single nucleotide variants (SNVs), and variants with a population frequency of greater than 0.01 reported in the gnomAD8 databases were excluded. CRC drivers were identified as outlined in a recent major comprehensive CRC genomics profiling study.[13] All identified variants underwent manual review using Integrative Genomics Viewer (IGV).[20]

5.3.5. Copy Number Variant Calling

Copy number variants were identified using CNVkit and GISTIC.[21, 22] Fourteen samples were excluded from the CNV analysis because they failed to pass the CNV quality metrics. Recurrently mutated cytobands were plotted using the GenVisR R package.[23]

5.3.6. Ancestry Analysis

Global ancestry inference was conducted using LD-pruned variants (R^2 cutoff of 0.2) across all chromosomes, analyzed with ADMIXTURE (v1.3).[24] We included reference populations from the HapMap3 project, representing a diverse set of ancestries: Yoruba in Ibadan, Nigeria (YRI); Maasai in Kinyawa, Kenya (MKK); Luhya in Webuye, Kenya (LWK); Toscani in Italia (TSI); Utah residents with Northern and Western European ancestry (CEU); Han Chinese in Beijing, China (CHB); Chinese in Denver, USA (CHD); and Japanese in Tokyo, Japan (JPT). Meta-population ancestry proportions were calculated by summing proportions of related populations: African ancestry (YRI, MKK, LWK), European ancestry (TSI, CEU), and Asian ancestry (CHB, CHD, JPT). Individuals were categorized into ancestry groups based on heuristic thresholds, previously developed, evaluated, and established using a large external dataset. These heuristic thresholds are as follows: Black if the African meta-population proportion was ≥50%, White if the European meta-population proportion was ≥50%, Asian if the Asian meta-population proportion was ≥50%, and Mixed if no single meta-population proportion was ≥50%.

5.3.7. Microsatellite Instability Identification

Microsatellite instability-high (MSI-H) tumors were identified using MSIsensor, a validated computational algorithm for assessing microsatellite instability (MSI) and mismatch repair (MMR) status.[25, 26] Tumors with an MSI sensor score ≥10 were classified as MSI-H, while those with a score <10 were classified as microsatellite stable (MSS), based on thresholds established in prior colorectal cancer studies.[25]

5.3.8. Identification of Pathogenic POLE Variants

Tumours with pathogenic somatic variants in *POLE* exonuclease domain (ED) were identified based on the 22 known pathogenic variants which were previously reported.[27, 28] High mutational burden tumors were either classified as microsatellite instability-high (MSI-H) or found to carry a known pathogenic *POLE* variant. Tumors with pathogenic *POLE* variants were confirmed to have elevated activity of SBS10a and SBS10b mutational signatures, which are well-established markers of *POLE* ED mutations.[29]

5.3.9. SBS Mutational Signatures

Single base substitution (SBS) signatures were extracted de novo using SigProfilerExtractor, referencing known COSMIC mutational signatures (v3.4).[29] The analysis was performed assuming 1 and 30 SBS signatures (minimum and maximum signature parameters, respectively), while default settings were applied for all other parameters, as described in prior studies.

5.3.10. Statistical Analysis

Descriptive statistics were used to summarize data: median and range for continuous variables, with comparisons conducted using the Wilcoxon rank-sum test, and count and percentage for categorical variables, with comparisons made using two-sided Fisher's exact tests. The frequency of oncogenic alterations between groups was compared using two-sided Fisher's exact tests. Pathway analysis was performed using templates from Sanchez-Vega *et al*.[30] Genomic results were considered hypothesis-generating and were not corrected for multiple testing due to the limited sample size. All statistical analyses were conducted using R software, with a significance threshold of $p < 0.05$ for two-sided tests.

5.4. Results

Eighty-three (out of the original 138) cases passed quality control checks and were included for whole exome sequencing (WES) analysis. Salient demographic and clinicopathological features are summarised in Table 5.1. The majority of cases were left-sided (63.9%), locally advanced tumours (pT3 or pT4), and histologically classified as moderately differentiated adenocarcinoma NOS. Ancestry analysis was performed on these cases which confirmed that the majority were of African (56.6%) or mixed (26.5%) ancestry descent (Figure 5.1). The majority of the cases (n=77, 92.8%) were classified as MSS by MSIsensor, while 4 MSI cases (4.8%) were included. Two cases (2.4%) with pathogenic *POLE* ED variants were identified (P286R) which represented the POLE ultramutated group. This classification correlated well with the mean SNV count per sample (Figure 5.2).

Table 5.1: Demographic and Clinicopathological Characteristics of a Cohort of Colorectal Cancer Patients (n=83)

Age at Diagnosis	52 years (45, 56)*
Gender	
Female	38 (45.8%)
Male	45 (54.2%)
Tumour Site	
Caecum	9 (10.8%)
Ascending	8 (9.6%)
Hepatic flexure	3 (3.6%)
Transverse	10 (12.0%)
Splenic flexure	5 (6.0%)
Descending	8 (9.6%)
Sigmoid	19 (22.9%)
Rectum	21 (25.3%)
Tumour Side	
Left	53 (63.9%)
Right	30 (36.1%)
Pathological Tumour Stage (pT)	
1	1 (1.2%)
2	3 (3.6%)
3	52 (62.7%)
4a	18 (21.7%)
4b	9 (10.8%)
Histological Grade	
1	3 (3.6%)
2	77 (92.8%)
3	3 (3.6%)
Histological Type:	
Adenocarcinoma NOS	81 (97.6%)
Mucinous	2 (2.4%)
MSI Status	
MSI	4 (4.8%)
MSS	77 (92.8%)
POLE ultramutated	2 (2.4%)
Calculated Ethnicity	
African	47 (56.6%)
Mixed	22 (26.5%)
White	13 (15.7%)
Unknown	1 (1.2%)

* Median (Q1, Q3)

This cohort was enriched for pMMR cases with only select unexplained dMMR cases included. NOS not otherwise specified, MSI microsatellite instability, MSS microsatellite stable.

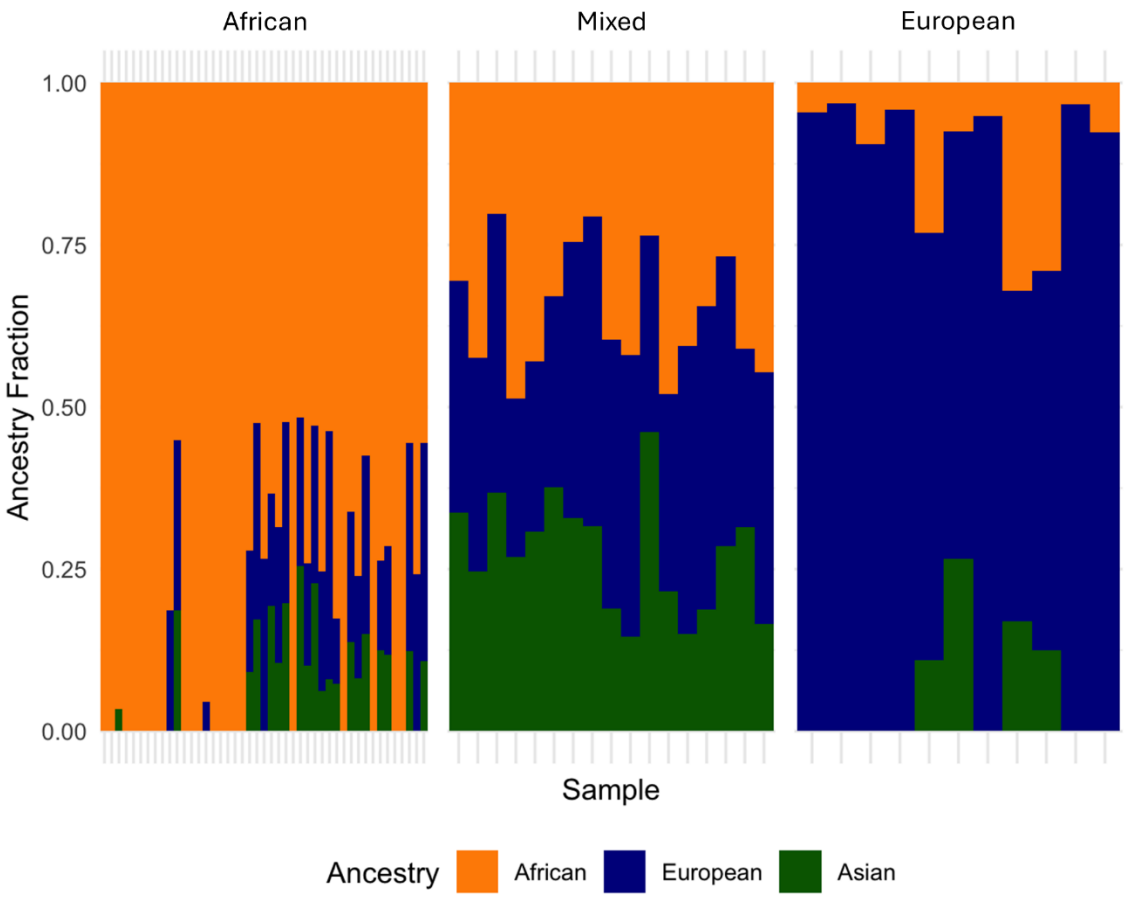


Figure 5.1: Ethnicity Determination by Ancestry Analysis. The majority of samples are shown to originate form African and mixed ethnic groups.

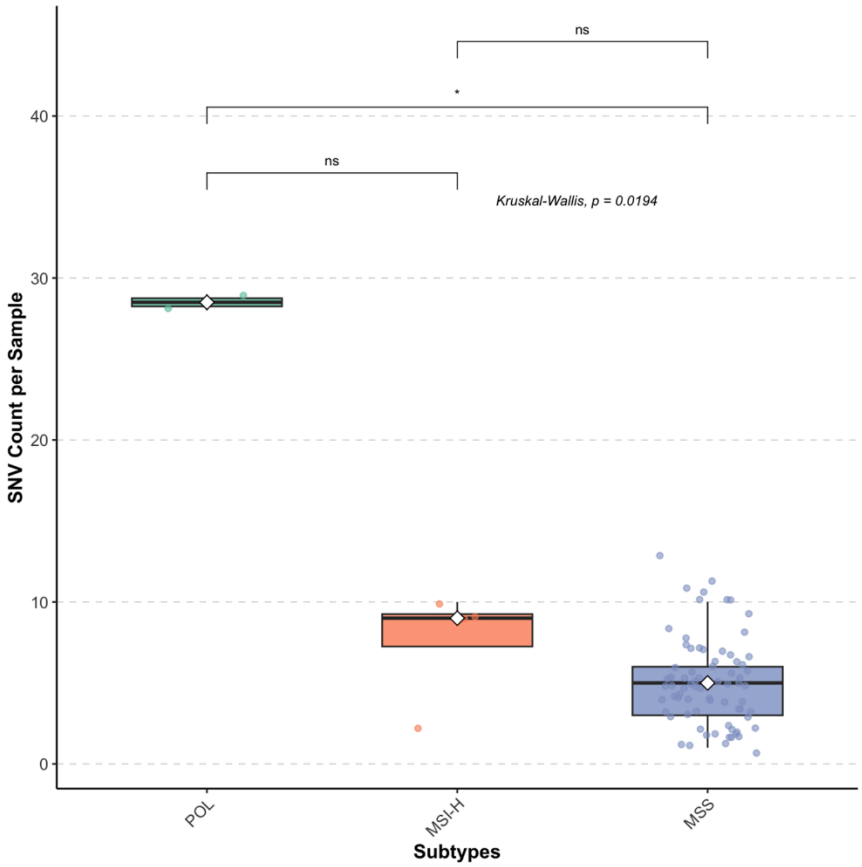


Figure 5.2: Distribution of SNV Counts per Sample across CRC Molecular Subtypes.

Driver gene identification at a base-pair level was performed separately in MSS, MSI and POLE ultramutated CRC (Figure 5.3). As expected, *APC* (49%) and *TP53* (35%) were the most frequently mutated genes in CRC. The frequency of *KRAS* (18%) and *PIK3CA* (7%) mutations was lower than expected. Only one *BRAF* mutation (1.3%) was detected, this was a K253Q variant occurring in an MSS tumour. *FAT4*, *BIRC6*, *TET2*, *CSMD3*, *ATM* and *EPHA3* were among the top 10 mutated genes across all molecular subtypes. Interestingly, an association between co-occurring *FAT4* and *TET2* driver mutations was found (p=0.0093).

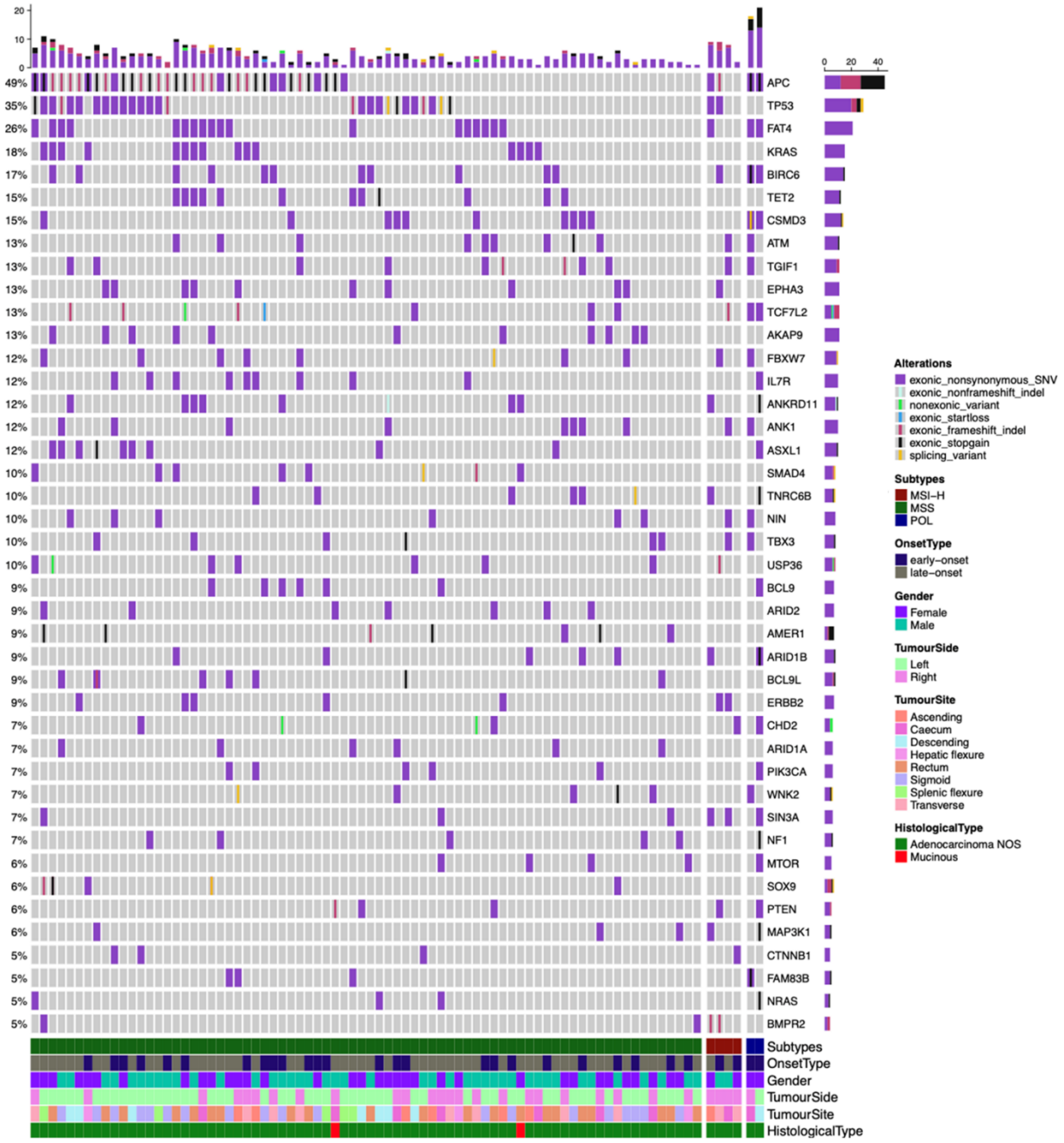


Figure 5.3: Oncoprint. The top 5% of mutated driver genes based on SNV analysis are shown in this Oncoprint. MSS, MSI and POL ultramutated cases are shown in separate vertical panels. SNV counts per sample are shown in the upper panel. Early-onset age at diagnosis <50.

SNV counts per sample in MSS CRC were analysed by salient clinicopathological variables. The mean SNV count was similar for early- and late-onset MSS CRC (Figure 5.4a). Left-sided tumours, African ancestry and females showed a higher mean SNV count, though these differences were not statistically significant (Figure 5.4b-d).

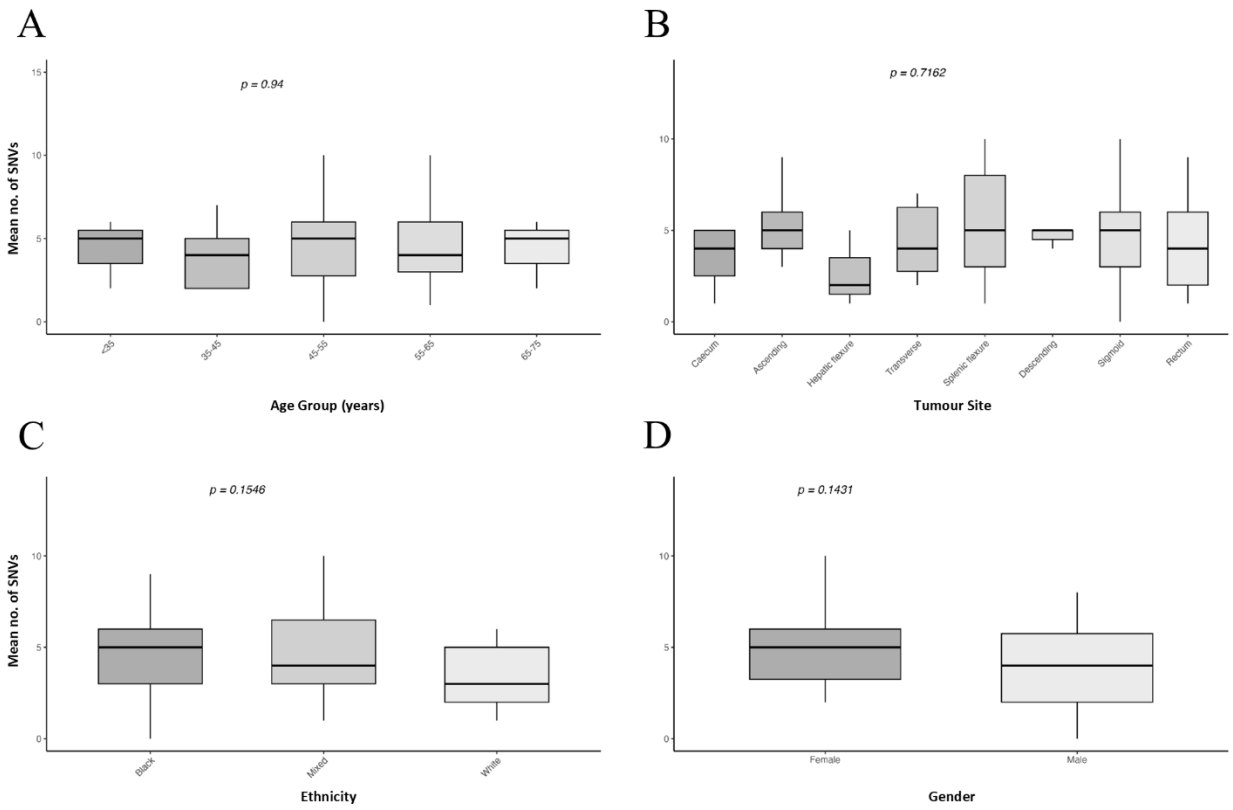


Figure 5.4: SNV Counts in MSS CRC Stratified by Salient Clinicopathological Features. **A**, age group in 10 year bands. **B**, anatomical location by tumour site. **C**, ethnicity determined by ancestry analysis. **D**, gender.

Further analysis of the top five mutated driver genes was performed based on anatomical location and age group. *APC*, *TP53*, *FAT4* and *TET2* driver mutations occurred at a higher frequency in left-sided tumours, while *KRAS* mutations were enriched for in right-sided tumours (Figure 5.5a-b). *APC* drivers were more frequent in early-onset CRC, while *FAT4*, *KRAS*, TET2, and *TP53* drivers occurred more often in later-onset CRC (Figure 5.5c-d). *FAT4* driver mutations

were only found in mixed and African ancestry groups (Figure 5.6c). *ARID1B*, *BCL9*, *CHD2* and *PTEN* drivers were found exclusively in tumours from African patients in out cohort.

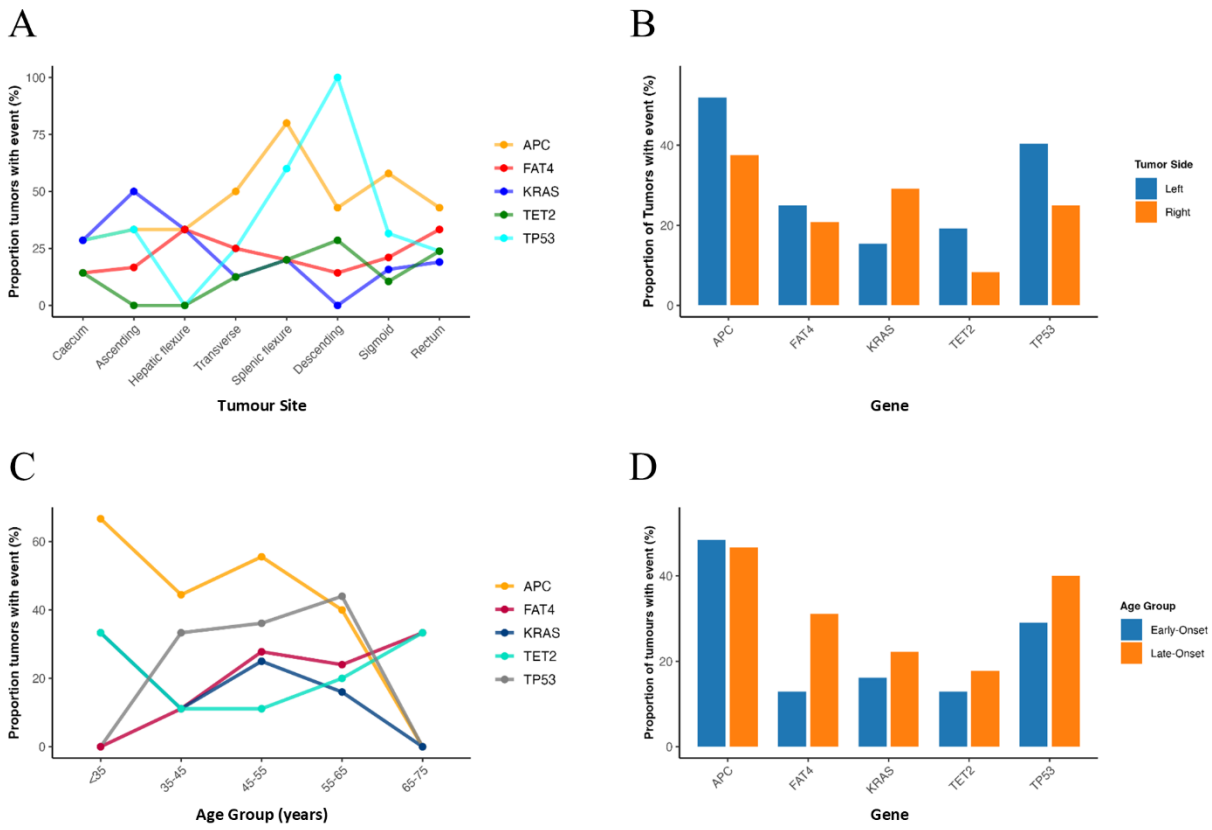


Figure 5.5: Top 5 Mutated Driver Genes in MSS CRC Stratified by Salient Clinicopathological Features. **A**, anatomical location by tumour site. **B**, anatomical location classified as left vs. right. **C**, age group in 10 year bands. **D**, age group classified as early-onset (<50 years) vs. late-onset.

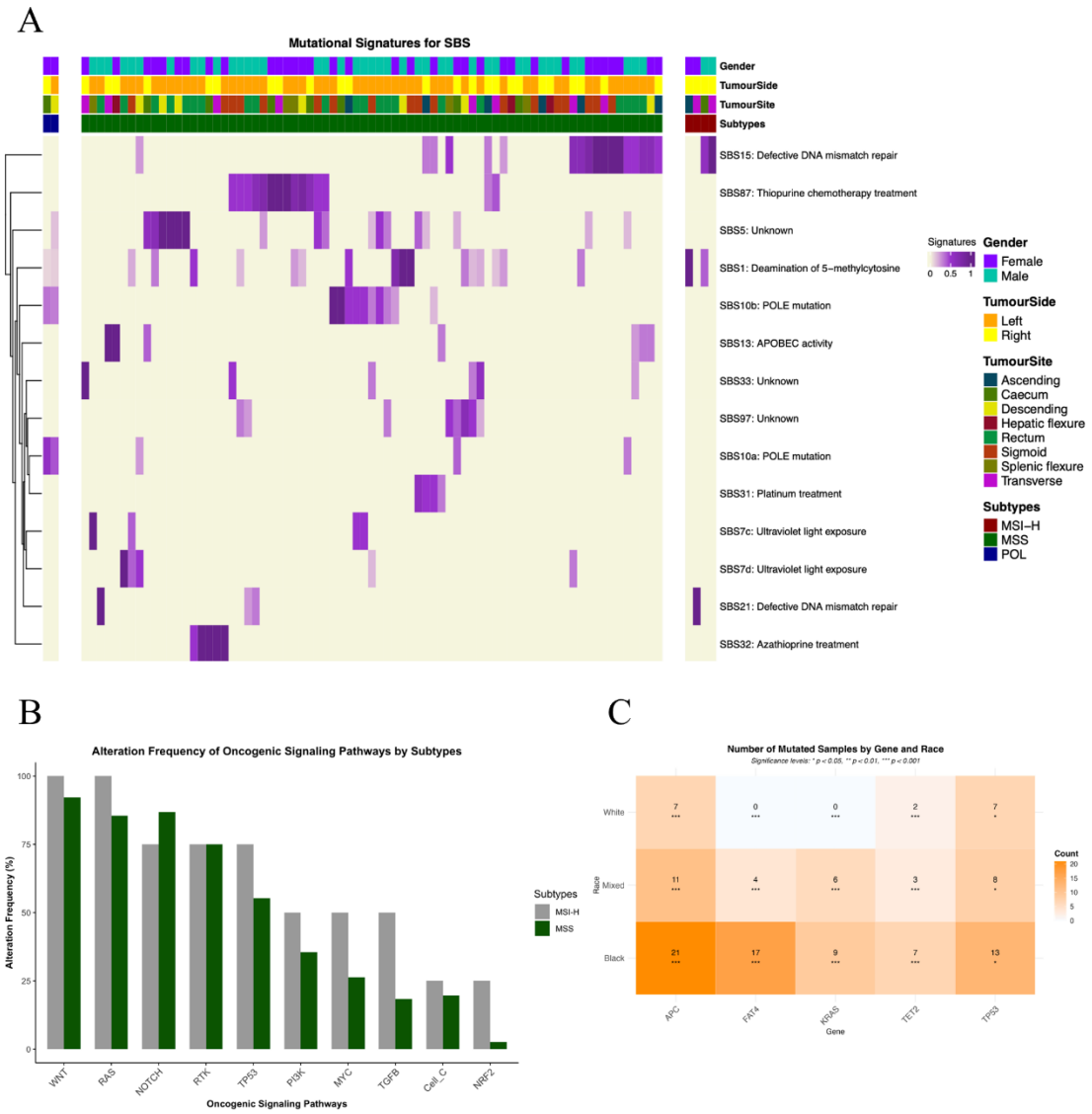


Figure 5.6: Mutational Signatures, Oncogenic Signalling Pathways, and Top 5 Mutated Driver Genes by Ethnicity. **A**, Single base substitution (SBS) mutational signatures focussing on the most important signatures in CRC. **B**, Frequency of alterations in the major signalling pathways in CRC. **C,** Top 5 mutated driver genes stratified by ethnicity.

Single base substitution (SBS) mutational signatures were calculated, and those relevant to CRC are summarised in Figure 5.6a. SBS15, SBS87 and SBS5 were the most frequent in MSS tumours (15.5%, 12.3% and 10.7% respectively). Typically SBS1 and SBS5 have been

described as the most frequent SBS signatures in CRC.[13, 31, 32] The SBS1 signature was seen with a frequency of 9.7% in our cohort (Table 5.2). Interestingly, SBS15 has been previously associated with defective DNA mismatch repair, however this was the most frequent signature in the present MSS group. SBS87 is associated with thiopurine chemotherapy treatment. SBS93 was uncommon (0.01%), in contrast to the findings of Cornish *et al*. (who found this to be present in 40% of MSS cases).[13]

Table 5.2: Established SBS Signatures, Arranged by Frequency

Signature	**Frequency (%)**
SBS15: Defective DNA mismatch repair	15.5
SBS87: Thiopurine chemotherapy treatment	12.3
SBS5: Unknown	10.7
SBS1: Deamination of 5-methylcytosine	9.7
SBS10b: POLE mutation	7.1
SBS32: Azathioprine treatment	5.6
SBS13: APOBEC activity	4.5
SBS97: Unknown	3.9
SBS33: Unknown	3.7
SBS21: Defective DNA mismatch repair	3.2
SBS7c: Ultraviolet light exposure	3.0
SBS7d: Ultraviolet light exposure	3.0
SBS10a: POLE mutation	2.4
SBS31: Platinum treatment	2.4

Alterations in key oncogenic signalling pathways were investigated (Figure 5.6b). WNT, RAS, and PI3K pathway alterations were found with a similar frequency to that reported in the literature in MSS CRC.[9, 14] Alatise *et al*. described a much lower frequency of WNT pathway alteration in their Nigerian MSS cohort.[9] *APC* and *TCF7L2* are well described key genes in the WNT pathway and were also the most frequently mutated in our study. However, we found a significant number (>10% frequency) of mutations in the *CHD8*, *DKK2*, *AMER1*, *CHD4* and *LGR5* genes. These genes were not included in the MSK-IMPACT assay panel that was used by Alatise *et al*. and this may account for the discrepancy. NOTCH and RTK pathways were altered at a higher frequency than previously reported, while *TP53* alteration was found less frequently.[9, 14]

Overall, 69 (83%) samples passed quality control filters for CNV analysis (Figure 5.7). Gains were most significantly found in chromosomes 20q, 13q, and 20p. These findings are similar to Cornish *et al*.[13] Significant gains in chromosome 7q were not detected. Chromosome 6p gains were significantly correlated with African ancestry (p=0.0028). 6p and 7p gains were statistically more frequent in the rectum (p=0.04 & p=0.005).

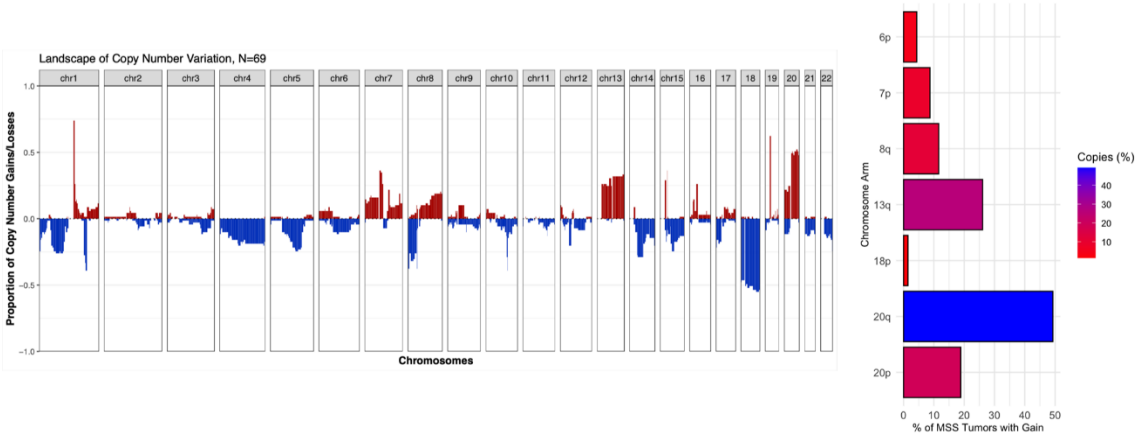


Figure 5.7: Copy Number Variant (CNV) Alterations in MSS CRC. Significant CNV alterations are shown at a chromosome level in the left panel, with the arm location of the most frequent gains and losses highlighted in the right panel.

5.5. Discussion

Although the burden of CRC is increasing in sub-Saharan Africa, there is a paucity of molecular data from African studies. In particular, the rising incidence of left-sided MSS tumours remains poorly categorised at a molecular level.[8] We present the first WES study on a cohort of selected CRC samples from sub-Saharan Africa, as a critical step towards understanding their tumour biology.

APC and *TP53* are well-established as major driver genes in MSS CRC, and were also the two most frequently mutated genes in our study. This finding is similar to that reported recently in a

landmark paper investigating 2033 CRC whole genomes in the UK.[13] However, the frequency of these mutations in our study is approximately half that reported recently by Cornish *et al*. Interestingly, this reduced frequency of *APC* mutations (49%) is similar to that described in Nigeria (37%) by Alatise *et al*.[9] Although factors such as small sample size and quality of sequencing data from FFPE samples may have affected our results, this finding is still noteworthy. The majority of *APC* variants are SNVs, so it is unlikely that other types of variants or epigenetic events were overlooked.[33] Since *APC* mutations are frequent early events in the adenoma-carcinoma sequence in MSS CRC, this finding raises the possibility of alternative early events in CRC pathogenesis. Despite the lower frequency of *APC* mutations, we have shown that the WNT signalling pathway was still altered in the vast majority of tumours. This is in contrast to the findings of Alatise *et al*., and we have highlighted several mutated genes not included in their MSK-IMPACT assay. *KRAS*, another major driver gene in MSS CRC, was also found with less than half the frequency (18% in our study) reported in Nigeria by Alatise *et al*. (57%) and in the UK by Cornish *et al*. (45%).[9, 13] Additionally, *BRAF* V600E mutations were absent in our study, a finding similar to Alatise *et al*.[9] These findings highlight major differences in the frequency of specific driver genes at a base-pair level, between CRC in sub-Saharan Africa, North America, and Europe.

In our study, we identified a high frequency of *FAT4* and *TET2* driver mutations. *FAT4* is a large tumour suppressor gene (translating to a product of 4 983 amino acids) which is thought to inhibit epithelial-to-mesenchymal transition (EMT) via the Wnt/β-catenin pathway, and autophagy via the PI3K/AKT signalling pathway.[34-36] Cornish *et al*. reported a frequency of *FAT4* mutations in 1.0% of MSS tumours.[13] *FAT4* mutations were found in 24% of MSS CRC in our cohort, and exclusively in African and Mixed ancestry groups. This finding may provide an alternative explanation for aberrant Wnt/β-catenin signalling in indigenous African patients where the frequency of *APC* mutations is reduced considerably. In our study, *FAT4* mutations were more frequent in late-onset CRC and tended to occur more frequently in left-sided tumours. Up-regulation of *FAT4* has recently been shown to enhance chemosensitivity of colorectal cancer cells treated by 5-FU.[37] This may have significant therapeutic implications for treating CRC in sub-Saharan Africa.

The ten-eleven translocation (TET) protein acts as a substrate for AMPK and is involved in the conversion of 5-methylcytosine (5-mC) to 5-hydroxymethylcytosine (5-hmC), a key intermediate product in the DNA demethylation reaction.[38] *TET2* mutations are found commonly in haematological malignancies and are also present in solid tumours, including breast cancer.[39] This is a novel pathway in colorectal carcinogenesis and has been linked to obesity-related CRC development.[40] In our study, *TET2* mutations occurred at a frequency of 15%, and more often in late-onset CRC, and in the left colon. *TET2* was reported as a novel gene in CRC by Cornish *et al*., but at a much lower frequency of 0.6% in MSS tumours. Downregulation of *TET2* is an independent poor prognostic factor in CRC.[41] This may be a promising therapeutic target.

Working with FFPE material is known to have limitations in terms of DNA yield and quality. The selective nature of the cohort examined limits direct comparison to other studies. Future work in sub-Saharan Africa should focus on collecting fresh frozen material with paired germline samples. More complete oncology records with follow up data would also add an additional element to this work.

In summary, we have shown differences in oncogenic signalling pathways and driver gene frequencies in MSS CRC in our cohort which was enriched for in patients of African ancestry. *APC*, *TP53*, *KRAS*, and *PIK3CA* appear to only account for drivers in a subset of cases. *FAT4* and *TET2* have emerged as possible alternative driver genes, particularly in late-onset and left-sided tumours. Further studies investigating larger cohorts with paired germline material are required to further dissect these pathways and evaluate these genes as potential therapeutic targets.

5.6. References

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A: **Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.** *CA Cancer J Clin* 2024, **74**(3):229-263.

2. Karsa L, Lignini T, Patnick J, Lambert R, Sauvaget C: **The dimensions of the CRC problem.** *Best Pract Res Clin Gastroenterol* 2010, **24**(4):381-396.

3. Cronje L, Paterson A, Becker P: **Colorectal cancer in South Africa: a heritable cause suspected in many young black patients.** *S Afr Med J* 2009, **99**(2):103-106.

4. McCabe M, Perner Y, Magobo R, Magangane P, Mirza S, Penny C: **Microsatellite Instability assessment in Black South African Colorectal Cancer patients reveal an increased incidence of suspected Lynch syndrome.** *Sci Rep* 2019, **9**(1):1-10.

5. Vergouwe F, Boutall A, Stupart D, Algar U, Govender D, Van der Linde G, Mall A, Ramesar R, Goldberg P: **Mismatch repair deficiency in colorectal cancer patients in a low-incidence area.** *S Afr J Surg* 2013, **51**(1):16-21.

6. Holla R, Vorster A, Locketz M, De Haas M, Oke O, Govender D, Ramesar R, Goldberg P: **Immunohistochemical determination of mismatch repair gene product in colorectal carcinomas in a young indigenous African cohort.** *S Afr J Surg* 2022, **60**(1):28-33.

7. Katsidzira L, Vorster A, Gangaidzo IT, Makunike-Mutasa R, Govender D, Rusakaniko S, Thomson S, Matenga JA, Ramesar R: **Investigation on the hereditary basis of colorectal cancers in an African population with frequent early onset cases.** *PLOS one* 2019, **14**(10):e0224023.

8. McCabe M, Penny C, Magangane P, Mirza S, Perner Y: **Left-sided colorectal cancer distinct in indigenous African patients compared to other ethnic groups in South Africa.** *BMC cancer* 2022, **22**(1):1089.

9. Alatise OI, Knapp GC, Sharma A, Chatila WK, Arowolo OA, Olasehinde O, Famurewa OC, Omisore AD, Komolafe AO, Olaofe OO: **Molecular and phenotypic profiling of colorectal cancer patients in West Africa reveals biological insights.** *Nat Commun* 2021, **12**(1):1-8.

10. Alatise OI, Knapp GC, Bebington B, Ayodeji P, Dare A, Constable J, Olasehinde O, Kingham TP: **Racial differences in the phenotype of colorectal cancer: A prospective comparison between Nigeria and South Africa.** *World J Surg* 2022, **46**(1):47-53.

11. Ashktorab H, Nouraie M, Hosseinkhah F, Lee E, Rotimi C, Smoot D: **A 50-year review of colorectal cancer in African Americans: implications for prevention and treatment.** *Dig Dis Sci* 2009, **54**:1985-1990.

12. Ashktorab H, Smoot DT, Carethers JM, Rahmanian M, Kittles R, Vosganian G, Doura M, Nidhiry E, Naab T, Momen B: **High incidence of microsatellite instability in colorectal cancer from African Americans.** *Clin Cancer Res* 2003, **9**(3):1112-1117.

13. Cornish AJ, Gruber AJ, Kinnersley B, Chubb D, Frangou A, Caravagna G, Noyvert B, Lakatos E, Wood HM, Thorn S: **The genomic landscape of 2,023 colorectal cancers.** *Nature* 2024, **633**(8028):127-136.

14. **Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer.** *Nature* 2012, **487**(7407):330-337.

15. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res* 2010, **38**(6):1767-1771.

16. Bolger AM, Lohse M, Usadel B: **Trimmomatic: a flexible trimmer for Illumina sequence data.** *Bioinformatics* 2014, **30**(15):2114-2120.

17. Saunders CT, Wong WS, Swamy S, Becq J, Murray LJ, Cheetham RK: **Strelka: accurate somatic small-variant calling from sequenced tumor–normal sample pairs.** *Bioinformatics* 2012, **28**(14):1811-1817.

18. Freed D, Aldana R, Weber JA, Edwards JS: **The Sentieon Genomics Tools–A fast and accurate solution to variant calling from next-generation sequence data.** *BioRxiv* 2017.

19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin **R, Subgroup GPDP: The sequence alignment/map format and SAMtools.** *Bioinformatics* 2009, **25**(16):2078-2079.

20. Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP: **igv. js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV)**. *Bioinformatics* 2023, **39**(1):btac830.

21. Talevich E, Shain AH, Botton T, Bastian BC: **CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing.** *PLoS computational biology* 2016, **12**(4):e1004873.

22. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukhim R, Getz G: **GISTIC2. 0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers.** *Genome biol* 2011, **12**:1-14.

23. Skidmore ZL, Wagner AH, Lesurf R, Campbell KM, Kunisaki J, Griffith OL, Griffith M: **GenVisR: genomic visualizations in R.** *Bioinformatics* 2016, **32**(19):3012-3014.

24. Alexander DH, Lange K**: Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.** *BMC bioinformatics* 2011, **12**:1-6.

25. Middha S, Zhang L, Nafa K, Jayakumaran G, Wong D, Kim HR, Sadowska J, Berger MF, Delair DF, Shia J: **Reliable pan-cancer microsatellite instability assessment by using targeted next-generation sequencing data.** *JCO Precis Oncol* 2017, **1**:1-17.

26. Niu B, Ye K, Zhang Q, Lu C, Xie M, McLellan MD, Wendl MC, Ding L**: MSIsensor: microsatellite instability detection using paired tumor-normal sequence data.** *Bioinformatics* 2014, **30**(7):1015-1016.

27. Rayner E, Van Gool IC, Palles C, Kearsey SE, Bosse T, Tomlinson I, Church DN: **A panoply of errors: polymerase proofreading domain mutations in cancer.** *Nat Rev Cancer* 2016, **16**(2):71-81.

28. Briggs S, Tomlinson I: **Germline and somatic polymerase ϵ and δ mutations define a new class of hypermutated colorectal and endometrial cancers.** *J Pathol* 2013, **230**(2):148-153.

29. Degasperi A, Zou X, Dias Amarante T, Martinez-Martinez A, Koh GCC, Dias JM, Heskin L, Chmelova L, Rinaldi G, Wang VYW: **Substitution mutational signatures in whole-genome–sequenced cancers in the UK population.** *Science* 2022, **376**(6591):abl9283.

30. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S: **Oncogenic signalling pathways in the cancer genome atlas.** *Cell* 2018, **173**(2):321-337. e310.

31. Islam SA, Díaz-Gay M, Wu Y, Barnes M, Vangara R, Bergstrom EN, He Y, Vella M, Wang J, Teague JW: **Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor.** *Cell Genom* 2022, **2**(11).

32. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, Boot A, Covington KR, Gordenin DA, Bergstrom EN: **The repertoire of mutational signatures in human cancer.** *Nature* 2020, **578**(7793):94-101.

33. Vogelstein B, Papadopoulos N, Velculescu V, Zhou S, Diaz Jr L, Kinzler K: **Cancer genome landscapes.** *Science* 2013, **339**(6127):1546-1558.

34. Wei R, Xiao Y, Song Y, Yuan H, Luo J, Xu W: **FAT4 regulates the EMT and autophagy in colorectal cancer cells in part via the PI3K-AKT signalling axis.** *J Exp Clin Cancer Res* 2019, **38**:1-14.

35. Mao W, Zhou J, Hu J, Zhao K, Fu Z, Wang J, Mao K: **A pan-cancer analysis of FAT atypical cadherin 4 (FAT4) in human tumors.** *Front Public Health* 2022, **10**:969070.

36. Cai J, Feng D, Hu L, Chen H, Yang G, Cai Q, Gao C, Wei D: **FAT4 functions as a tumour suppressor in gastric cancer by modulating Wnt/β-catenin signalling.** *Br J Cancer* 2015, **113**(12):1720-1729.

37. Li Q, Zhou X, Fang Z, Pan Z, Zhou H: **Up-regulation of FAT4 enhances the chemosensitivity of colorectal cancer cells treated by 5-FU.** *Transl Cancer Res* 2020, **9**(1):309.

38. Kriaucionis S, Heintz N: **The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain.** *Science* 2009, **324**(5929):929-930.

39. Huang Y, Rao A: **Connections between TET proteins and aberrant DNA modification in cancer.** *Trends Genet* 2014, **30**(10):464-474.

40. Kon T, Sasaki Y, Abe Y, Onozato Y, Yagi M, Mizumoto N, Sakai T, Umehara M, Ito M, Nakamura S: **Modulation of AMPK/TET2/5-hmC axis in response to metabolic alterations as a novel pathway for obesity-related colorectal cancer development**. *Sci Rep* 2023, **13**(1):2858.

41. Rawłuszko-Wieczorek AA, Siera A, Horbacka K, Horst N, Krokowicz P, Jagodziński PP: **Clinical significance of DNA methylation mRNA levels of TET family members in colorectal cancer**. *J Cancer Res Clin Oncol* 2015, **141**:1379-1392.

6. CHAPTER SIX – General Discussion, Conclusions, and Future Perspectives

6.1. Routine Testing of MMR/MSI

Universal screening of all CRC tumours for dMMR/MSI is recommended by the UK National Institute for Health and Care Excellence (NICE) guidelines.[31] Screening performed by IHC or PCR is not widely available in sub-Saharan Africa, and the associated additional costs of universal screening are prohibitive. Even in South Africa, where histopathology services are reasonably well developed, testing is performed by arbitrary age cutoffs that vary between regions and facilities. To reduce costs, some authors have suggested IHC screening using only MSH6 and PMS2, with a reflex to the partner stain if either is absent (two-stain method).[37] This practice has limitations and may fail to detect some cases of Lynch syndrome. Although some centres in South Africa have adopted it as a cost-saving measure, it is not widely accepted as a standard of care. During the course of this current work, universal screening with IHC (four gene products) for all CRC has been adopted at Groote Schuur Hospital. It is hoped that this will be adopted nationally in the near future.

Digital pathology is becoming increasingly accessible globally, and several laboratories are incorporating slide scanning into their routine workflow. There are several advantages to digitising histology slides. Other than optimising workflow and turnaround times, there are various AI applications that may further increase productivity, enhance accuracy, and reduce turnaround times. Over the past 5 years, several groups have demonstrated high performance of DL models in detecting dMMR/MSI from scanned whole slide images of CRC resection specimens.[32-34, 38] This has the potential to reduce the burden of downstream testing, thereby saving significant costs.

This could become a viable pre-screening strategy once more South African laboratories have integrated a digital workflow. Most local laboratories have access to a slide scanner, so select cases could be scanned for this purpose even without a completely digital workflow. Commercially available AI models such as MSIntuit are currently available for *in vitro* diagnostic (IVD) use in Europe.[32] Although these models could potentially be used on our

patient population, we have shown that additional calibration would be required. One of the limitations of our study was the selection bias introduced by only including patients <60 years of age at the time of surgery. This artificially increased the proportion of dMMR cases which likely contributed to the reduced performance of the model. Further validation studies with larger patient cohorts are required. However, due to 'haphazard screening strategies' there is a lack of readily available data, and most of the MMR/MSI screening tests would need to be performed and interpreted. This may be easier to accomplish prospectively.

Ultimately, the major challenge in South Africa is the lack of standardised practice for dMMR/MSI screening. There is a need for engagement with the relevant stakeholders (pathologists, clinicians, laboratory groups, patient advocacy groups and funders) to adopt a uniform screening strategy at a national level. Since histopathology in the public sector is facilitated through a single 'National Health Laboratory Service', adopting guidelines at a national level should be attainable. It is hoped that research such as this will raise awareness of current international trends and stimulate discussion between colleagues at a national level.

6.2. *BRAF* Testing in MLH1/PMS2 dMMR CRC

Reflex *BRAF* V600E testing (IHC or PCR) is recommended in all dMMR cases with loss of MLH1/PMS2 to triage patients for further germline testing for Lynch syndrome.[31] The *BRAF* V600E point mutation is associated with *MLH1* promotor hypermethylation in approximately 50% of cases, and suggests a somatic event. Through our work (paper 2), we detected *BRAF* V600E mutations in none of the 18 dMMR cases which showed loss of MLH1 protein expression in any combination. Since only a fraction of these patients are likely to have Lynch syndrome, and only 14/18 were found to have *MLH1* pathogenic variants (implying that 4/18 had epigenetic silencing of *MLH1*), this suggests that *BRAF* V600E may not correlate with *MLH1* promotor hypermethylation in our population.

Several other groups have made similar observations, and it appears that *BRAF* mutations are not a reliable mechanism to differentiate somatic and germline causes of MLH1/PMS2 loss in sub-

Saharan Africa.[35, 39] This would support the use of *MLH1* promotor hypermethylation testing as a means to differentiate somatic from germline *MLH1* inactivation in our local setting. Further work is required to determine whether this finding is transferable to older patients with MLH1/PMS2 loss. In addition, further work is required on our patient cohort on germline (adjacent normal colon) material to clearly delineate germline from somatic cases.

6.3. Appropriate Referral of dMMR Cases for Germline Screening

Guidelines recommend that all cases with MSH2/MSH6, isolated MSH6, and isolated PMS2 loss of staining on IHC should be referred for genetic counselling and germline testing for Lynch syndrome.[31] In addition, cases with MLH1/PMS2 loss and which are *BRAF* V600 wild type, should also undergo germline testing. Only 20 of the 49 dMMR cases (41%) were recorded in the Division of Humen Genetics registry as having been referred for germline testing. Of these, 17 (85%) had an MLH1/PMS2 loss of staining pattern, or an unconventional pattern which included MLH1 loss. We sequenced somatic tissue from 18 of the cases which were not registered for germline testing and found that 9 of them (50%) had an MMR gene mutation with a VAF which indicated a possible germline mutation. This highlights that a significant proportion of likely Lynch syndrome cases are being missed in our system.

Some of these dMMR cases may not have been referred appropriately for genetic counselling and testing. Although the histopathology of these cases was reported at the NHLS laboratory at Groote Schuur Hospital, a significant number of the patients would have been operated at surrounding secondary and district level hospitals (New Somerset Hospital, Victoria Hospital, Mitchel's Plain Hospital, George Hospital). Increasing clinician awareness of interpretation of these results and appropriate referral should increase the number of referred dMMR cases. It would also be useful to include a standardised interpretation comment in the histopathology reports of dMMR cases, advising for germline testing. Another possibility is that patients were appropriately referred for germline testing, but were lost to follow up. Further studies would be needed to investigate this referral pathway more thoroughly and identify gaps which are leading

to less than half of patients in our setting receiving germline testing. An MSc study is currently underway in our laboratory to identify such gaps, and propose means for filling these.

6.4. Adequacy of Current Testing Methods for Germline Screening

There is a focus on targeted testing, via PCR/Sanger sequencing for the founder *MLH1* 1528C>T mutation in the Division of Human Genetics at the University of Cape Town. This founder variant is the most prevalent in the region.[40] The "common 5" panel, which is currently in use, also includes the four most common *MLH2* variants seen in our mixed ancestry population. Our study shows that a wide range of pathogenic or likely pathogenic variants are found in our local environment. Twenty-six different variants were found, and the only variant that occurred in more than one sample was the *MLH1* 1528C>T founder. Thirteen of these 26 variants (48%) occurred at a VAF that suggested a likely germline aetiology. Only 1/13 of these variants (*MSH2* c.1221_1222del) would have been detected with the "common 5" panel currently in use. This highlights the need for NGS, which can be panel-based and optimised to be cost-effective. This is essential to identify such a wide range of possible variants correctly. Driver genes, which are therapeutically relevant (*KRAS*, *PIK3CA*, *BRAF*), could also be included in the panel to provide more clinically relevant information. As personalised medicine becomes more widely available in the state sector, this additional somatic molecular information will become increasingly important for therapeutic decision-making. Small panels that contain these genes relevant to CRC are commercially available for IVD. Batching samples is a major cost-saving advantage of NGS, and a large referral centre could easily receive sufficient requests to make this financially viable.

We have shown that panel-based NGS is necessary to identify the entire spectrum of MMR gene pathogenic and likely pathogenic variants in CRC in our population. Further work should focus on a cost analysis and the feasibility of providing this test at a national level.

6.5. Driver Genes and Oncogenic Signalling Pathways in pMMR CRC

Our study confirmed the findings of Alatise *et al*., (who worked on a Nigerian cohort of CRC patients) of divergent major driver gene mutation frequencies from what has been reported in the literature.[35] In particular, the *APC*, *TP53*, *KRAS* and *PIK3CA* genes were found to be less commonly mutated in pMMR CRC. We performed WES which allowed us to evaluate more potential alternative driver genes, and identified *FAT4* and *TET2* as two of the major driver genes in our South African population. These genes are emerging drivers in CRC and their mechanism of action is still being investigated.[41-47] They may provide a promising explanation for the rising incidence of left-sided CRC in sub-Saharan Africa. Further work is required to understand the functional consequences of mutations in these genes.

6.6. Strengths and Weaknesses

In this study, we have retrospectively collected a significant number of CRC resection samples with corresponding MMR protein IHC status and complete clinicopathological data. Our cohort was selected for young (<60 years of age) cases of CRC which makes it unique. We have performed several experiments, which were the first for CRC in sub-Saharan Africa. This includes the panel-based NGS of dMMR cases, WES of pMMR cases, and the dMMR/MSI pre-screening with an AI model. We have provided insights into the spectrum of mutations in MMR genes in dMMR cases. Our cohort contained a significant proportion of African and mixed ancestry cases, providing novelty for the AI and WES work.

This work had several weaknesses. Most notable was the lack of paired germline material which would have provided definitive information on the hereditary CRC cases. This data would also have made variant interpretation (particularly CNV) with the WES data more reliable. Doing molecular work with FFPE tissue has known limitations. Although we had to exclude a few samples from analysis due to poor DNA quality, those that were adequate did yield high-quality data. Another shortcoming is the lack of clinical follow-up data. This could have provided an additional prognostic element to this work.

6.7. Future Perspectives

Several directions for future work have been highlighted in the discussion above. It would be valuable to ascertain the germline status of the dMMR cohort by targeted PCR or panel sequencing. This can be performed on DNA extracted from representative FFPE tissue wax blocks of normal colonic tissue from each case. This would allow us to definitively categorise the dMMR cases as somatic or germline and further data analysis could be performed on these subgroups.

Transcription studies can be performed on the pMMR cases that were subjected to WES. Although due to the nature of the samples (FFPE), fewer cases would likely pass quality control checks, this would represent the first study of its kind from sub-Saharan Africa.

The representative tumour slides, which were scanned, have been digitally archived and would be amenable to several further studies. We would like to investigate the quantitation of TILs with artificial intelligence and compare this to what was reported by pathologists. Once the germline status of the dMMR cohort is known, further analysis could also be performed to attempt to delineate germline from somatic dMMR cases by AI TILs count. Key molecular data variables generated from the WES study could also be investigated with AI. This could include *KRAS* and *POLE* mutation status as well as identification of molecular subgroups.

The major findings from the WES study (*FAT4* and *TET2* mutations) require further investigation. This could involve immunohistochemical evaluation of tumour tissue sections to determine the functional status of their protein products.

Several other IHC-based studies could be performed on this cohort. Protein expression of the less commonly mutated MMR genes (*MLH3*, *MSH3*, and *PMS1*) could be investigated. The tumour immune cell microenvironment could also be delineated with inflammatory cell markers (CD3, CD4, CD8, CD163, etc) and key receptors of tumour-immune cell interaction (PDL1,

CTLA4, etc). As a separate study, an attempt could be made to gather available clinical follow up and outcome data. If this were to be successful, several further morphological and immunohistochemical studies could be undertaken to study novel prognostic indicators in our unique cohort.

Spatial transcriptomics experiments could study the interaction of the tumour cells and the host inflammatory response. Ideally, a selection of dMMR and pMMR cases should be studied, with paired normal colon tissue acting as a control. This would be the first RNASeq study to evaluate patients of African ancestry and with early-onset CRC.

Ultimately, the goal should be to prospectively collect a cohort of CRC resection samples with paired germline material. Patients could be counselled and consented appropriately. Tissue could be collected postoperatively and stored fresh frozen to better preserve genetic material for large-scale sequencing or spatial transcriptomics studies. Follow-up data could also be more actively collected. This would yield the ideal material on which to further explore this group of tumours in our unique population.

REFERENCES (Introduction & Discussion Chapters)

1. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, Jemal A: **Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries**. *CA Cancer J Clin* 2024, **74**(3):229-263.

2. Brenner H, Heisser T, Cardoso R, Hoffmeister M: **Reduction in colorectal cancer incidence by screening endoscopy**. *Nat Rev Gastroenterol Hepatol* 2024, **21**(2):125-133.

3. Arnold M, Sierra MS, Laversanne M, Soerjomataram I, Jemal A, Bray F: **Global patterns and trends in colorectal cancer incidence and mortality**. *Gut* 2017, **66**(4):683-691.

4. Lauby-Secretan B, Scoccianti C, Loomis D, Grosse Y, Bianchini F, Straif K: **Body fatness and cancer—viewpoint of the IARC Working Group**. *NEJM* 2016, **375**(8):794-798.

5. Bouvard V, Loomis D, Guyton KZ, Grosse Y, El Ghissassi F, Benbrahim-Tallaa L, Guha N, Mattock H, Straif K: **Carcinogenicity of consumption of red and processed meat**. *Lancet Oncol* 2015, **16**(16):1599-1600.

6. Vieira A, Abar L, Chan D, Vingeliene S, Polemiti E, Stevens C, Greenwood D, Norat T: **Foods and beverages and colorectal cancer risk: a systematic review and meta-analysis of cohort studies, an update of the evidence of the WCRF-AICR Continuous Update Project**. *Ann Oncol* 2017, **28**(8):1788-1802.

7. Kyu HH, Bachman VF, Alexander LT, Mumford JE, Afshin A, Estep K, Veerman JL, Delwiche K, Iannarone ML, Moyer ML: **Physical activity and risk of breast cancer, colon cancer, diabetes, ischemic heart disease, and ischemic stroke events: systematic review and dose-response meta-analysis for the Global Burden of Disease Study 2013**. *BMJ* 2016, **354**.

8. Johnson CM, Wei C, Ensor JE, Smolenski DJ, Amos CI, Levin B, Berry DA: **Meta-analyses of colorectal cancer risk factors**. *CCC* 2013, **24**:1207-1222.

9. Collins D, Hogan AM, Winter DC: **Microbial and viral pathogens in colorectal cancer**. *The lancet oncology* 2011, **12**(5):504-512.

10. Hamid HK: **Schistosoma japonicum–associated colorectal cancer: A review**. *The American journal of tropical medicine and hygiene* 2018, **100**(3):501.

11. **WHO Classification of Tumours Editorial Board. Digestive system tumours**, vol. 1, 5th edn. Lyon (France): International Agency for Research on Cancer; 2019.

12. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis**. *cell* 1990, **61**(5):759-767.

13. Cancer Genome Atlas Network: **Comprehensive molecular characterization of human colon and rectal cancer**. *Nature* 2012, **487**(7407):330-337.

14. Poulogiannis G, Ichimura K, Hamoudi RA, Luo F, Leung SY, Yuen ST, Harrison DJ, Wyllie AH, Arends MJ: **Prognostic relevance of DNA copy number changes in colorectal cancer**. *J Pathol* 2010, **220**(3):338-347.

15. Arends MJ: **Pathways of colorectal carcinogenesis**. *Appl Immunohistochem Mol Morphol* 2013, **21**(2):97-102.

16. Bienz M, Clevers H: **Linking colorectal cancer to Wnt signaling**. *Cell* 2000, **103**(2):311-320.

17. Müller MF, Ibrahim AE, Arends MJ: **Molecular pathological classification of colorectal cancer**. *Virchows Archiv* 2016, **469**(2):125-134.

18. Hoda SA, Hoda RS: **Robbins and cotran pathologic basis of disease**. In.: Oxford University Press US; 2020.

19. Li G-M: **Mechanisms and functions of DNA mismatch repair**. *Cell Res* 2008, **18**(1):85-98.

20. Eso Y, Shimizu T, Takeda H, Takai A, Marusawa H: **Microsatellite instability and immune checkpoint inhibitors: toward precision medicine against gastrointestinal and hepatobiliary cancers**. *J Gastroenterol* 2020, **55**(1):15-26.

21. Itatani Y, Kawada K, Sakai Y: **Transforming growth factor-β signaling pathway in colorectal cancer and its tumor microenvironment**. *Int J Mol Sci* 2019, **20**(23):5822.

22. Shevelev IV, Hübscher U: **The 3′–5′ exonucleases**. *Nat Rev Mol Cell Biol* 2002, **3**(5):364-376.

23. Seshagiri S, Stawiski EW, Durinck S, Modrusan Z, Storm EE, Conboy CB, Chaudhuri S, Guan Y, Janakiraman V, Jaiswal BS: **Recurrent R-spondin fusions in colon cancer**. *Nature* 2012, **488**(7413):660-664.

24. Rayner E, Van Gool IC, Palles C, Kearsey SE, Bosse T, Tomlinson I, Church DN: **A panoply of errors: polymerase proofreading domain mutations in cancer**. *Nat Rev Cancer* 2016, **16**(2):71-81.

25. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne CJ, Heuer ML, Larsson E: **The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data**. *Cancer Discov* 2012, **2**(5):401-404.

26. Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, Sun Y, Jacobsen A, Sinha R, Larsson E: **Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal**. *Sci Signal* 2013, **6**(269):pl1-pl1.

27. Rozek LS, Schmit SL, Greenson JK, Tomsho LP, Rennert HS, Rennert G, Gruber SB: **Tumor-infiltrating lymphocytes, Crohn's-like lymphoid reaction, and survival from colorectal cancer**. *JNCI* 2016, **108**(8):djw027.

28. Väyrynen J, Tuomisto A, Klintrup K, Mäkelä J, Karttunen T, Mäkinen M: **Detailed analysis of inflammatory cell infiltration in colorectal cancer**. *British journal of cancer* 2013, **109**(7):1839-1847.

29. Jia P, Yang X, Guo L, Liu B, Lin J, Liang H, Sun J, Zhang C, Ye K: **MSIsensor-pro: fast, accurate, and matched-normal-sample-free detection of microsatellite instability**. *Genomics, Proteomics and Bioinformatics* 2020, **18**(1):65-71.

30. Karsa L, Lignini T, Patnick J, Lambert R, Sauvaget C: **The dimensions of the CRC problem**. *Best Pract Res Clin Gastroenterol* 2010, **24**(4):381-396.

31. Snowsill T, Coelho H, Huxley N, Jones-Hughes T, Briscoe S, Frayling IM, Hyde C: **Molecular testing for Lynch syndrome in people with colorectal cancer: systematic reviews and economic evaluation**. 2017.

32. Saillard C, Dubois R, Tchita O, Loiseau N, Garcia T, Adriansen A, Carpentier S, Reyre J, Enea D, von Loga K: **Validation of MSIntuit as an AI-based pre-screening tool for MSI detection from colorectal cancer histology slides**. *Nat Commun* 2023, **14**(1):6695.

33. Echle A, Laleh NG, Quirke P, Grabsch H, Muti H, Saldanha O, Brockmoeller S, van den Brandt P, Hutchins G, Richman S: **Artificial intelligence for detection of microsatellite instability in colorectal cancer—a multicentric analysis of a pre-screening tool for clinical application**. *ESMO Open* 2022, **7**(2):100400.

34. Wagner SJ, Reisenbüchler D, West NP, Niehues JM, Zhu J, Foersch S, Veldhuizen GP, Quirke P, Grabsch HI, van den Brandt PA: **Transformer-based biomarker prediction from colorectal cancer histology: A large-scale multicentric study**. *Cancer Cell* 2023, **41**(9):1650-1661. e1654.

35. Alatise OI, Knapp GC, Sharma A, Chatila WK, Arowolo OA, Olasehinde O, Famurewa OC, Omisore AD, Komolafe AO, Olaofe OO: **Molecular and phenotypic profiling of colorectal cancer patients in West Africa reveals biological insights**. *Nat Commun* 2021, **12**(1):1-8.

36. Bucksch K, Zachariae S, Aretz S, Büttner R, Holinski-Feder E, Holzapfel S, Hüneburg R, Kloor M, von Knebel Doeberitz M, Morak M: **Cancer risks in Lynch syndrome, Lynch-like syndrome, and familial colorectal cancer type X: a prospective cohort study**. *BMC cancer* 2020, **20**:1-11.

37. Pearlman R, Frankel WL, Swanson B, Zhao W, Yilmaz A, Miller K, Bacher J, Bigley C, Nelsen L, Goodfellow PJ: **Prevalence and spectrum of germline cancer susceptibility gene mutations among patients with early-onset colorectal cancer**. *JAMA oncol* 2017, **3**(4):464-471.

38. Echle A, Laleh NG, Schrammen PL, West NP, Trautwein C, Brinker TJ, Gruber SB, Buelow RD, Boor P, Grabsch HI: **Deep learning for the detection of microsatellite instability from histology images in colorectal cancer: a systematic literature review**. *ImmunoInformatics* 2021, **3**:100008.

39. McCabe M, Perner Y, Magobo R, Magangane P, Mirza S, Penny C: **Microsatellite Instability assessment in Black South African Colorectal Cancer patients reveal an increased incidence of suspected Lynch syndrome**. *Scientific reports* 2019, **9**(1):1-10.

40. Stupart DA, Goldberg P, Algar U, Ramesar R: **Surveillance colonoscopy improves survival in a cohort of subjects with a single mismatch repair gene mutation**. *Colorectal Dis* 2009, **11**(2):126-130.

41. Wei R, Xiao Y, Song Y, Yuan H, Luo J, Xu W: **FAT4 regulates the EMT and autophagy in colorectal cancer cells in part via the PI3K-AKT signaling axis**. *J Exp Clin Cancer Res* 2019, **38**:1-14.

42. Mao W, Zhou J, Hu J, Zhao K, Fu Z, Wang J, Mao K: **A pan-cancer analysis of FAT atypical cadherin 4 (FAT4) in human tumors**. *Front Public Health* 2022, **10**:969070.

43. Cai J, Feng D, Hu L, Chen H, Yang G, Cai Q, Gao C, Wei D: **FAT4 functions as a tumour suppressor in gastric cancer by modulating Wnt/β-catenin signalling**. *Br J Cancer* 2015, **113**(12):1720-1729.

44. Li Q, Zhou X, Fang Z, Pan Z, Zhou H: **Up-regulation of FAT4 enhances the chemosensitivity of colorectal cancer cells treated by 5-FU**. *Transl Cancer Res* 2020, **9**(1):309.

45. Kriaucionis S, Heintz N: **The nuclear DNA base 5-hydroxymethylcytosine is present in Purkinje neurons and the brain**. *Science* 2009, **324**(5929):929-930.

46. Kon T, Sasaki Y, Abe Y, Onozato Y, Yagi M, Mizumoto N, Sakai T, Umehara M, Ito M, Nakamura S: **Modulation of AMPK/TET2/5-hmC axis in response to metabolic alterations as a novel pathway for obesity-related colorectal cancer development**. *Sci Rep* 2023, **13**(1):2858.

47. Rawłuszko-Wieczorek AA, Siera A, Horbacka K, Horst N, Krokowicz P, Jagodziński PP: **Clinical significance of DNA methylation mRNA levels of TET family members in colorectal cancer**. *J Cancer Res Clin Oncol* 2015, **141**:1379-1392.