Assessment of missense alterations in MLH1 and their pathogenic significance

by

Needra Sheron Perera

A thesis submitted in conformity with the requirements for the degree of Doctor of Philosophy
Department of Laboratory Medicine and Pathobiology
University of Toronto

© Copyright by Needra Sheron Perera 2010
Assessment of missense alterations in MLH1 and their pathogenic significance

Needra Sheron Perera
Doctor of Philosophy
Department of Laboratory Medicine and Pathobiology
University of Toronto
2010

Abstract

Germline mutations in mismatch repair genes predispose individuals to Lynch Syndrome, the most common colorectal cancer predisposition syndrome. MLH1 is a key mismatch repair gene that is mutated in Lynch syndrome and about a third of the genetic alterations identified in MLH1 are missense variants of unclassified clinical significance. We hypothesize that missense alterations in distinct domains of MLH1 likely affect its expression and function(s) to varying degrees. To address this we utilized several approaches to investigate the molecular basis of the pathogenicity of a panel of unclassified variants. Our results demonstrate that the MLH1 variants p.R265C and p.K618A significantly decrease the stability of the MLH1 protein, while the variant p.L749Q compromises heterodimerization of the MLH1-PMS2 complex. Given the limitations and complexity of in vitro assessment strategies, we conducted a proof-of-principle study to investigate whether missense variants in MLH1 lead to allelic imbalances at the transcriptional level. Our analysis using the PeakPicker software indicated that the missense variants c.350C>T, c.793C>T, c.1852_1853AA>GC, as well as the truncating variant c.1528C>T were associated with significantly unbalanced allelic expression. This illustrates a novel method of investigating the pathogenicity of unclassified genetic variants, which has the potential to be applied in the
diagnostic setting. Previous genetic epidemiology studies from our laboratory have demonstrated that the *MLH1* c.-93G>A promoter variant is strongly associated with the microsatellite instability phenotype in colorectal tumours. Additionally, this promoter variant was associated with an elevated risk of endometrial cancer in case-control studies. Results from our functional studies indicate that the c.-93G>A variant significantly alters the promoter activity of *MLH1*. The *MLH1* promoter is bi-directional with the *EPM2AIP1* gene located on the antisense strand. Interestingly, we observed that this variant significantly affected *EPM2AIP1* transcription as well. Furthermore, our experiments suggest that c.-93G>A variant affects transcription by altering the affinity of nuclear factors that bind this region. Combined, these findings shed light on the possible mechanisms by which missense variants affect MLH1 expression and function, which in conjunction with results from other functional assays will help increase the accuracy and efficiency of genetic testing of inherited cancers.
Acknowledgments

First and foremost I would like to thank my supervisor and mentor, Dr. Bharati Bapat for giving me the chance to train in her laboratory. Her intellectual guidance and advice were pivotal in helping me succeed in my education. I also appreciate her sincere concern and encouragement in all matters of my life.

I would like to thank all the past and present members of the Bapat laboratory, especially Vaiju Pethe, Miralem Mrkonjic, George Charames, Ken Kron, Jamie Rawson, Liyang Liu, Brian Li, Ted Chappell, Roula Raptis, Pinella Buongiorno, Susmita Esufali, Ekaterina Olkhov, Bas VanRhijn, as well as Joyce Keating for their help and advice. Over the years we have transitioned from co-workers to friends and I hope we remain so in the future.

I am grateful to my committee members, Dr. Eleftherios Diamandis and Dr. Ben Alman, for their support and guidance over the years. A special thank you to Dr. Harry Elsholtz, whose advice and encouragement has helped me through the various stages of my graduate education. I would also like to thank Dr. Hilmi Ozcelik and the members of the Ozcelik laboratory for their friendship over the years. Thank you also to all the members of the Dennis/Swallow laboratory, especially Ken Lau and Yossr Haffani. A special thank you also to my external examiner, Dr. Patricia Tonin for helping strengthen my thesis. I also appreciate the assistance of the members of the Molecular Diagnostics Laboratory and the Biospecimen Repository, especially Lily Ramyar and Soultana Tsitsikotas. I would also like to thank the genetic counselors, Kara Semotiuk and Melyssa Aronson for their help with my project. I am deeply grateful to the patients and their families for their generosity.
I would like to thank my friend Carla Rosario, who has shared with me the joys and pains of learning throughout the past eleven years. This journey has been richer because I have shared it with you.

To my family and friends, especially my parents-in-law, Anton Joseph and Manel Joackhim for their enthusiasm and interest in my education and their support over the years. A special note of thanks to my grandfather, Basil Perera whose fatal battle with cancer and faith in an 11-year old was the inspiration for this journey.

To my wonderful sister, Roshen Perera for always believing in me and reminding me that I could deal with whatever life brought my way. Thank you for your constant love and your willingness to help me in any and everything I do.

I am indebted to my beloved husband, Rohan Joackhim for his unfailing support and encouragement. I know that my quest for academic achievement has had a profound impact on your life and your sacrifices have not gone unnoticed. Thank you for everything!

I am deeply grateful to my wonderful parents, Florette and Brian Perera for giving me the confidence to reach for my dreams. No words can express my gratitude for all the sacrifices you have made for me. You have always been my greatest inspiration and strength.

Finally, my heartfelt thanks to God for making all this possible.
**Table of Contents**

ACKNOWLEDGMENTS ........................................................................................................ IV

TABLE OF CONTENTS ........................................................................................................ VI

LIST OF TABLES ..................................................................................................................... X

LIST OF FIGURES ................................................................................................................... XI

LIST OF ABBREVIATIONS ...................................................................................................... XIII

CHAPTER 1 GENERAL INTRODUCTION ............................................................................. 1

1.1 COLORECTAL CANCER ................................................................................................. 1
   1.1.1 Anatomy and physiology of the Colorectum ......................................................... 1
   1.1.2 Natural history and epidemiology of CRC ............................................................ 2
   1.1.3 Screening, staging, and prognosis ........................................................................ 3
   1.1.4 Molecular pathways in CRC ................................................................................ 10
   1.1.5 Behavioural and lifestyle risk factors ................................................................. 26
   1.1.6 Hereditary syndromes, pathology and presentation ............................................ 28

1.2 LYNCH SYNDROME ...................................................................................................... 31
   1.2.1 Overview ............................................................................................................... 31
   1.2.2 Pathology, presentation and clinical criteria ....................................................... 32
   1.2.3 Tumour spectrum ............................................................................................... 35
   1.2.4 Predictive genetic testing .................................................................................... 36

1.3 MISMATCH REPAIR ..................................................................................................... 37
   1.3.1 Overview of the mismatch repair system in prokaryotes ................................. 38
   1.3.2 Overview of the mismatch repair system in eukaryotes ..................................... 39
   1.3.3 The post-replicative error repair function ......................................................... 42
   1.3.4 The role in signaling DNA damage ................................................................. 46
   1.3.5 Other roles of mismatch repair proteins ............................................................ 50
   1.3.6 The structure and expression of the MMR protein MLH1 ............................... 52
   1.3.7 Mismatch repair and disease .......................................................................... 54

1.4 GENETIC VARIATION IN MISMATCH REPAIR ....................................................... 56
CHAPTER 2 THE ROLE OF UNCLASSIFIED VARIANTS IN INFLUENCING MLH1

EXPRESSION, STABILITY AND FUNCTION ................................. 64

2.1 SUMMARY ............................................................................ 64
2.2 INTRODUCTION ...................................................................... 64
2.3 MATERIALS AND METHODS .................................................... 69
  2.3.1 Materials and cell culture .................................................. 69
  2.3.2 Identification of MLH1 variants ........................................ 69
  2.3.3 Computational Assessment of Variants .............................. 70
  2.3.4 Site-directed mutagenesis and construction of MLH1 mutants 70
  2.3.5 RNA extraction, cDNA synthesis and sequencing ................ 71
  2.3.6 Production of recombinant proteins and western blotting ...... 71
  2.3.7 Combined coimmunoprecipitation and western blot analysis .... 72
  2.3.8 Pulse chase assays .......................................................... 72
  2.3.9 Cycloheximide experiments .............................................. 73
  2.3.10 Generation of stable inducible cell lines ............................ 73
  2.3.11 Cell survival assays ....................................................... 75
  2.3.12 Cell cycle analysis .......................................................... 76
  2.3.13 MSI analysis ................................................................. 76
2.4 RESULTS .............................................................................. 78
  2.4.1 In silico analysis of MLH1 variants ....................................... 78
  2.4.2 HCT116 cells are defective in MLH1 .................................. 82
  2.4.3 R265C and K618A decrease expression of MLH1 ................. 82
  2.4.4 Altered expression is due to decreased stability of the protein .. 83
  2.4.5 L749Q impairs ability of MLH1 to heterodimerize with PMS2 ... 88
  2.4.6 HCT116 cells express inducible wild-type and variant MLH1 .. 90
  2.4.7 Stable inducible cell lines and cell survival ......................... 95
  2.4.8 Stable inducible cell lines and the G2/M checkpoint ............... 97
2.4.9 Stable inducible expression of MLH1 & the microsatellite instability phenotype of the HCT116 cell line

2.5 DISCUSSION

CHAPTER 3 A NOVEL AND RAPID METHOD OF DETERMINING THE EFFECT OF MLH1 UNCLASSIFIED GENETIC VARIANTS ON DIFFERENTIAL ALLELIC EXPRESSION

3.1 SUMMARY

3.2 INTRODUCTION

3.3 MATERIALS & METHODS
3.3.1 Selection of Patients and Families
3.3.2 Overview of diagnostic strategy and identification of MLH1 variants
3.3.3 Microsatellite instability and immunohistochemistry
3.3.4 DNA extraction and mutation detection
3.3.5 RNA extraction and cDNA synthesis
3.3.6 PeakPicker Analysis
3.3.7 Allele specific (Asymmetric) PCR
3.3.8 In silico analysis of potential effects on splicing

3.4 RESULTS
3.4.1 The truncating variant c. 1528C>T leads to unbalanced allelic expression
3.4.2 The missense variants c. 350C>T, C. 793C>T and c.1852_1853AA>GC lead to unbalanced allelic expression
3.4.3 The missense variants c. 55A>T and c. 2246T>C do not lead to unbalanced allelic expression

3.5 DISCUSSION

CHAPTER 4 FUNCTIONAL EFFECTS OF THE MLH1-93G>A POLYMORPHISM ON MLH1 AND EPM2AIP1 PROMOTER ACTIVITY

4.1 SUMMARY

4.2 INTRODUCTION

4.3 MATERIALS AND METHODS
4.3.1 Generation of Promoter Constructs
4.3.2 Cell Culture
4.3.3 Luciferase Reporter Gene Assays ........................................................................... 146
4.3.4 Western blotting ........................................................................................................ 147
4.3.5 Electrophoretic Mobility Shift Assays (EMSA) ......................................................... 147
4.4 RESULTS ..................................................................................................................... 148
  4.4.1 Effects of the MLH1 -93G>A variant on promoter activity in colorectal cancer cell lines 148
  4.4.2 Effects of the MLH1 -93G>A variant on promoter activity in endometrial cancer cell lines 149
  4.4.3 Characterization of MLH1 and EPM2AIP1 in a panel of human cell lines .......... 149
  4.4.4 Effects of the MLH1 -93G>A variant on EPM2AIP1 promoter activity ............... 150
  4.4.5 Effect of the MLH1 -93G>A promoter variant on transcription factor binding..... 159
4.5 DISCUSSION .............................................................................................................. 164

CHAPTER 5 DISCUSSION AND FUTURE DIRECTIONS .................................................. 170

REFERENCES ............................................................................................................... 181
List of Tables

CHAPTER 1

Table 1.1 Screening tests used to detect CRC ................................................................. 6

Table 1.2 Tumour Stage Definitions According to the American Joint Committee on Cancer (AJCC) .......................................................................................................................... 8

Table 1.3 AJCC/UICC Stage Groupings ............................................................................. 9

Table 1.4 Summary of clinical criteria used to identify Lynch syndrome ....................... 34

CHAPTER 2

Table 2.1 Primers used for site-directed mutagenesis ..................................................... 77

Table 2.2 Summary of predictions generated by computational tools ............................ 77

Table 2.3 Summary of cell survival in response to temozolomide ................................. 96

Table 2.4 Summary of MSI experiments ........................................................................ 102

CHAPTER 3

Table 3.1 Primers used for PeakPicker analysis ............................................................. 120

Table 3.2 Summary of allele ratios generated by PeakPicker ........................................ 125

Table 3.3 Summary of clinicopathological characteristics .......................................... 126

Table 3.4 Predicted effects of missense variants on splicing ......................................... 127

CHAPTER 4

Table 4.1 Summary of promoter activities ..................................................................... 158
List of Figures

CHAPTER 1

Figure 1.1 The simplified progression of CRC via the CIN, MSI and CIMP pathways. .......... 24
Figure 1.2 Heterodimers formed by the MutL homologs......................................................... 40
Figure 1.3 Simplified overview of the eukaryotic MMR system. ........................................ 44
Figure 1.4 A simplified model of the MMR protein dependent DNA damage signaling network ....................................................................................................................................................... 49

CHAPTER 2

Figure 2.1 Schematic representation of the functional domains identified in MLH1. .......... 68
Figure 2.2 A Ribbon diagram of the approximate position of human MLH1 R265 on the crystal structure of E. coli MutL........................................................................................................................................... 80
Figure 2.3 Profile of time dependent expression of wild type MLH1 in HCT116 cells.......... 81
Figure 2.4 Expression of wild-type and variant MLH1 in HCT116 cells. ............................. 84
Figure 2.5 Expression profile of the MLH1 variants K618A & L749Q................................. 85
Figure 2.6 Pulse chase assays to determine half-life of wild-type and variant protein. ........ 86
Figure 2.7 Stability of MLH1 variants as assessed by cycloheximide treatments................. 87
Figure 2.8 Ability of wild-type and variant MLH1 to interact with endogenous PMS2........ 89
Figure 2.9 Screening of clones.................................................................................................. 92
Figure 2.10 Generation of the I19F MLH1 stable inducible cell line.................................. 93
Figure 2.11 Characterization of the K618A stable inducible cell line.................................. 94
Figure 2.12 Western blot to confirm levels of MLH1 in cell survival experiments.............. 96
Figure 2.13 Representative cell cycle profiles of the stable inducible and HCT116+Chr3 cell lines in response to treatment with temozolomide. 100

Figure 2.14 Expression of MLH1 and PMS2 in wild-type MLH1 cells stable inducible cells. 103

CHAPTER 3

Figure 3.1 Representative chromatograms generated by high sensitivity sequencing. 124

Figure 3.2 Pedigree of Case 5A/5B. 132

Figure 3.3 Pedigree of case 6. 133

CHAPTER 4

Figure 4.1 The schematic of the MLH1/EPM2AIP1 bi-directional promoter. 145

Figure 4.2 Effect of the MLH1-93G>A polymorphism on MLH1 transcriptional activity in the HCT116 cell line. 152

Figure 4.3 Effect of the MLH1-93G>A polymorphism on MLH1 transcriptional activity in the endometrial cancer cell lines. 154

Figure 4.4 Expression profile of endogenous MLH1 and EPM2AIP1 in a panel of human cell lines. 155

Figure 4.5 Effect of the MLH1-93G>A polymorphism on EPM2AIP1 transcriptional activity. 157

Figure 4.6 Impact of the MLH1-93 >A polymorphism on nuclear factor binding in HCT116. 160

Figure 4.7 Impact of the MLH1-93 >A polymorphism on nuclear factor binding in HEC-1-A. 161

Figure 4.8 Determination of putative factors that bind the promoter region surrounding -93G>A. 163
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-FU</td>
<td>5-Fluorouracil</td>
</tr>
<tr>
<td>A</td>
<td>Adenine</td>
</tr>
<tr>
<td>ACF</td>
<td>Aberrant Crypt Focus</td>
</tr>
<tr>
<td>AID</td>
<td>Activation-Induced Deaminase</td>
</tr>
<tr>
<td>ALT</td>
<td>Alternative Lengthening of Telomeres</td>
</tr>
<tr>
<td>APC</td>
<td>Adenomatous Polyposis Coli</td>
</tr>
<tr>
<td>ATM</td>
<td>Ataxia Telangiectasia Mutated</td>
</tr>
<tr>
<td>ATR</td>
<td>ATM and Rad3-related</td>
</tr>
<tr>
<td>BAC</td>
<td>Bacterial Artificial Chromosome</td>
</tr>
<tr>
<td>bp</td>
<td>base pair</td>
</tr>
<tr>
<td>BER</td>
<td>Base Excision Repair</td>
</tr>
<tr>
<td>BRRS</td>
<td>Bannayan-Riley-Ruvalcaba Syndrome</td>
</tr>
<tr>
<td>C</td>
<td>Cytosine</td>
</tr>
<tr>
<td>CS</td>
<td>Cowden Syndrome</td>
</tr>
<tr>
<td>CIMP</td>
<td>CpG Island Methylator Phenotype</td>
</tr>
<tr>
<td>CIN</td>
<td>Chromosomal Instability</td>
</tr>
<tr>
<td>CNV</td>
<td>Copy Number Variation</td>
</tr>
<tr>
<td>CRC</td>
<td>Colorectal Cancer</td>
</tr>
<tr>
<td>CXH</td>
<td>Cycloheximide</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>Dox</td>
<td>Doxycycline</td>
</tr>
<tr>
<td>DMSO</td>
<td>Dimethyl Sulfoxide</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic Acid</td>
</tr>
<tr>
<td>DNTP</td>
<td>deoxyribonucleoside triphosphate</td>
</tr>
<tr>
<td>EC</td>
<td>Endometrial Cancer</td>
</tr>
<tr>
<td>EMSA</td>
<td>Electrophoretic Mobility Shift Assay</td>
</tr>
<tr>
<td>EPM2AIP1</td>
<td>Laforin (EPM2A) Interacting Protein 1</td>
</tr>
<tr>
<td>ES</td>
<td>Embryonic stem</td>
</tr>
<tr>
<td>EV</td>
<td>Empty vector</td>
</tr>
<tr>
<td>Exo1</td>
<td>Exonuclease 1</td>
</tr>
<tr>
<td>FBS</td>
<td>Fetal Bovine Serum</td>
</tr>
<tr>
<td>FAP</td>
<td>Familial Adenomatous Polyposis</td>
</tr>
<tr>
<td>FOBT</td>
<td>Fecal Occult Blood Tests</td>
</tr>
<tr>
<td>G</td>
<td>Guanine</td>
</tr>
<tr>
<td>hr</td>
<td>hour</td>
</tr>
<tr>
<td>hMLH1</td>
<td>Human MutL Homolog 1</td>
</tr>
<tr>
<td>hMLH3</td>
<td>Human MutL Homolog 3</td>
</tr>
<tr>
<td>hMSH2</td>
<td>Human MutS Homolog 2</td>
</tr>
<tr>
<td>hMSH3</td>
<td>Human MutS Homolog 3</td>
</tr>
<tr>
<td>HNPCC</td>
<td>Hereditary Nonpolyposis Colorectal Cancer</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>hPMS1</td>
<td>Human Post- Meiotic Segregation 1</td>
</tr>
<tr>
<td>hPMS2</td>
<td>Human Post- Meiotic Segregation 2</td>
</tr>
<tr>
<td>IDL</td>
<td>Insertion Deletion Loop</td>
</tr>
<tr>
<td>IHC</td>
<td>Immunohistochemistry</td>
</tr>
<tr>
<td>IP</td>
<td>Immunoprecipitate</td>
</tr>
<tr>
<td>JPS</td>
<td>Juvenile Polyposis Syndrome</td>
</tr>
<tr>
<td>kDa</td>
<td>kiloDalton</td>
</tr>
<tr>
<td>LC</td>
<td>Lynch Syndrome</td>
</tr>
<tr>
<td>MAP</td>
<td>MYH- Associated Polyposis</td>
</tr>
<tr>
<td>MGMT</td>
<td>O(^6)-Methylguanine-DNA Methyl Transferase</td>
</tr>
<tr>
<td>MIM</td>
<td>Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>min</td>
<td>minute</td>
</tr>
<tr>
<td>miR</td>
<td>microRNA</td>
</tr>
<tr>
<td>MLPA</td>
<td>Multiplex Ligation-dependent Probe Amplification</td>
</tr>
<tr>
<td>MMR</td>
<td>Mismatch Repair</td>
</tr>
<tr>
<td>MNNG</td>
<td>N-Methyl-N(^\prime)-Nitro-N-Nitrosoguanidine</td>
</tr>
<tr>
<td>MNU</td>
<td>N-Methyl-N-Nitrosourea</td>
</tr>
<tr>
<td>MSI</td>
<td>Microsatellite Instability</td>
</tr>
<tr>
<td>MSI-H</td>
<td>High Frequency Microsatellite Instability</td>
</tr>
<tr>
<td>MSI-L</td>
<td>Low Frequency Microsatellite Instability</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Definition</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------</td>
</tr>
<tr>
<td>T</td>
<td>Thymine</td>
</tr>
<tr>
<td>TetR</td>
<td>Tetracycline Repressor</td>
</tr>
<tr>
<td>TNM</td>
<td>Tumour-Node-Metastasis</td>
</tr>
<tr>
<td>T-rex</td>
<td>Tetracycline-Regulated Expression system</td>
</tr>
<tr>
<td>TMZ</td>
<td>Temozolomide</td>
</tr>
<tr>
<td>V</td>
<td>Variable region of IgM antibodies</td>
</tr>
<tr>
<td>WCL</td>
<td>Whole Cell Lysate</td>
</tr>
<tr>
<td>wt</td>
<td>Wild-type</td>
</tr>
</tbody>
</table>
CHAPTER 1
General Introduction

1.1 Colorectal Cancer

1.1.1 Anatomy and physiology of the Colorectum

The colon is a component of the digestive system and in healthy individuals it functions to absorb water, sodium, and chloride, while secreting potassium and bicarbonate. The components of the colon are located either in the abdominal cavity or behind it in the retroperitoneum. Proximal and distal segments of the large intestine (defined as regions proximal and distal to the splenic flexure) have different embryological origins and biological characteristics [1]. The caecum, ascending colon and proximal two-thirds of the transverse colon derive from the midgut, while the segment comprising the splenic flexure to the upper anal canal derives from the hindgut. The distinct embryological origins of both segments are reflected in the dual blood supply and distinct innervations patterns. Furthermore, the pH of the lumen is lower in the proximal colon compared to the distal colon, a factor that can affect the growth of microbes and the metabolism of bile acids at these sites [2].

Intestinal epithelial stem cells are thought to be located at the bottom of the crypt in the colon [3]. Their differentiating progeny are thought to migrate upwards through the transit amplifying zone in the lower to middle region of the crypt, before becoming terminally differentiated and eventually shed into the lumen, a process that has been estimated to take about seven days [3]. Given this continuous loss of cells, the epithelium must replace cells at an equal rate to maintain its integrity. Thus, the intestinal epithelium has a large number of cycling cells and is one the most rapidly renewing tissues in the body [3].
1.1.2 Natural history and epidemiology of CRC

The multi-step genetic model of CRC carcinogenesis was first described in 1990 by Fearon and Vogelstein[4]. According to this model the *adenomatous polyposis coli (APC)* tumour suppressor gene is initially inactivated, followed by activating mutations in *KRAS* oncogene [4]. Subsequent mutations in TGF-β, PIK3CA, and p53 (TP53) then drive the subsequent steps in the process of malignant transformation [4-8]. Since then several refinements have been made to this original model, including the inclusion of many more genes. However, this model has helped establish several key principals that have guided our understanding and the study of CRC, as well as other solid tumours. First, multiple genetic events are required for the malignant transformation of cells; second, there exist several discrete intermediates in this process; and third, the temporal acquisition of these genetic changes is an important factor that determines its outcome [9].

The earliest identifiable lesion in the colonic mucosa is the aberrant crypt focus (ACF)[10], which has been proposed to be a putative precursor to adenocarcinoma of the colon [11]. An ACF refers to a cluster of crypts that are characterized based on several histological features such as darker staining, raised appearance and larger crypt size when compared to the adjacent normal mucosa [12, 13]. The true neoplastic potential of this lesion is still unclear, but it does appear that some of these lesions harbor mutations in *KRAS* or *APC* and can progress to adenocarcinoma [14]. While ACF has been shown to be an important predictor for CRC, there remains a need to further study these lesions and the context of their contribution to cancer [15, 16].

Worldwide about one million new CRC cases are diagnosed each year[17]. However, it is predominantly a disease that affects developed nations and is the second leading cause of cancer-related deaths in North America, affecting approximately 5% of the general population [17, 18]. It was estimated that in 2009 CRC will lead to 147,000 newly diagnosed cases as well as nearly
50,000 deaths in the United States[19]. In Canada the estimated incidence of CRC for 2010 is approximately 22,500 cases with approximately 9,100 deaths resulting from this disease [20]. CRC accounts for 13% of all cancers and is the third most common cancer in Canada [20]. The incidence and mortality of CRC within Canada varies by geographic location, with the Atlantic Provinces having the highest rates in the country and the Pacific provinces having the lowest rates. The province of Newfoundland and Labrador has approximately twice the mortality rate of CRC as the province of British Columbia[20]. In Ontario, approximately 8,100 CRC cases (4500 cases in men and 3800 cases in women) will be diagnosed in 2010[20]. The lifetime probability of developing CRC is about 6.5% and the estimated five-year relative survival rate from 2002-2004 was about 62% [21]. CRC is predominantly a disease of the elderly and more than half of all CRCs will occur in individuals over the age of 70 [20]. Furthermore, statistical data indicate only a slight difference between the incidence of CRC among males and females [20].

1.1.3 Screening, staging, and prognosis

The screening of asymptomatic persons, at an average risk of developing CRC allows for the detection of cancer at early and curable stages, consequently resulting in reduced mortality [22, 23]. Furthermore, certain screening strategies are able to detect precancerous lesions, the removal of which will result in decreased incidence of CRC [24, 25]. Currently there exist several screening strategies each with associated strengths and limitations (Table 1.1). Fecal occult blood tests (FOBT) can detect occult blood in small stool samples and is a non-invasive procedure that can be performed at home. The test detects the peroxidase activity of heme and is therefore not specific for human blood. The fecal immunohistochemical test uses antibodies against human hemoglobin, albumin, as well as other blood components that are more specific for human blood. However, studies indicate considerable variation in the performance of...
difference tests and the optimal commercial test is currently unknown[26]. Clinical trials have demonstrated that those with positive occult blood tests have a three to four times higher risk of developing cancer compared to those with negative test results[25]. Thus, colonoscopy is recommended for those with positive test results. Furthermore, randomized control studies have indicated that when administered annually or biennially the FOBT test results in the detection of cancers at an earlier stage, which over a period of 10-13 years results in a reduction of mortality from CRC of 15-33% [27-29]. Structural examinations of the colon are another type of screening method used in CRC and computed tomographic (CT) colonography is an example of this method. CT colonography generates two and three dimensional images of the colon, however it is less sensitive and specific for polyps less than 6mm in diameter[25]. Barium enema examination is another radiographic method that is able to accurately identify late-stage cancers; however, has poor sensitivity for precursor lesions[30]. Sigmoidoscopy, another method of structural examination cannot detect proximal lesions. This is important as it has been shown that 30% of patients with advanced CRC have only proximal lesions [31, 32]. Thus, colonoscopy is often the preferred method as it allows for a full colon examination with polypectomy [25]. Colonoscopy is the final assessment step in any screening program and has a high level of sensitivity. However, to date there have been no randomized, controlled trials carried out to compare colonoscopy with other screening methods. Furthermore, it is likely that some lesions are missed during this procedure, which is especially an issue for lesions larger than 10mm in diameter, as well as flat adenomas [33-35]. Thus, these issues highlight the importance of improving the quality of the colonoscopy. The recommended screening guidelines vary between nations. In Canada, the Canadian Association of Gastroenterology and the Canadian Digestive Health Foundation does not recommend screening for asymptomatic individuals under the age of 50 who do not have a family history of colon cancer [36]. For asymptomatic individuals over the
age of 50 and who do not have a family history of CRC the recommendations are: 1) FOBT test every two years, 2) a flexible sigmoidoscopy every five years, 3) flexible sigmoidoscopy combined with FOBT every five years, 4) double contrast barium enema every five years or 5) colonoscopy every ten years[36]. Individuals who have a family history of colon cancer have a higher risk of disease, thus a more stringent and focused screening regimen is recommended. The recommendations for these individuals vary depending on if they are positive for the hereditary syndromes, as well as the number, age and closeness (first-degree vs. second-degree) of relatives who have developed cancer.
Table 1.1 Screening tests used to detect CRC

<table>
<thead>
<tr>
<th>Test</th>
<th>Sensitivity</th>
<th>Advantages</th>
<th>Limitations</th>
</tr>
</thead>
</table>
| **Standard fecal occult blood test**      | Cancer: 33-50% AA: 11% | - Low initial cost  
- Can be performed at home | - Not specific for human hemoglobin  
- As specificity low, repeat annual testing recommended |
| **Sensitive fecal occult blood test (Hemoccult SENSA, Beckman Coulter)** | Cancer: 50-75% AA: 20-25% | - Low initial cost  
- Can be performed at home | - Not specific for human hemoglobin  
- As specificity low, repeat annual testing recommended |
| **Fecal Immunochemical test**             | Cancer: 60-85% AA: 20-50%[22, 23]  
Ref: [37] | - Specific for human hemoglobin  
- Low initial cost  
- Can be performed at home | - Uncertain benefit  
- Ideal number of stool samples and repeat testing regimen unknown |
| **Stool DNA**                             | Cancer: 51-80% AA: 18-40%  
Ref: [39, 40] | -Detection of specific mutations may be more accurate  
- Can be performed at home | - Costly  
- Limited to the screening of genes known to be mutated in CRC  
- Performance of new tests undetermined |
| **CT colonography**                       | Cancer: Uncertain, estimated to be >90%  
AA: estimated >90%  
Ref: [41] | - High sensitivity for detection of lesions ≥ 10 mm in diameter  
- Less invasive than endoscopy | - Detection of polyps <6 mm in diameter and flat polyps uncertain  
- Requires bowel preparation and special resources  
- Radiation exposure |
| **Sigmoidoscopy**                         | Cancer: >95% in distal colon  
AA: 70%*  
Ref: [31, 42] | - Office based and does not require sedation  
- Case-control studies show 60% reduction of mortality from cancers in distal colon | - Does not detect proximal colon cancer |
| **Colonoscopy**                           | Cancer: >95% AA: 88-98%  
Ref: [31-34, 43-45] | - Case-control studies indicate a 53-72% reduction in incidence[46] and 31% reduction of mortality[47]  
- Can detect and remove lesions during one examination | - High initial cost  
- Requires bowel preparation and special resources  
- Invasive  
- 3-5 adverse reactions per 1000 examinations |

Modified from Lieberman, DA[25]. AA: Advanced adenoma, a tubular adenoma that is 10 mm or larger in diameter/an adenoma with villous histologic features/high-grade dysplasia.* If an adenoma is detected in the distal colon the patient would undergo complete colonoscopy, which would result in detection (if any) of adenomas in both the proximal or distal colon. Ref: References that support these figures.
The stage of disease in CRC determines treatment options and is the strongest predictor for survival of CRC patients. The tumour-node-metastasis (TNM) staging system was developed by the American Joint Committee on Cancer (AJCC) and the International Union Against Cancer (UICC) to standardize staging for the uniform evaluation of data across specialties and nations [48]. In this system the designation ‘T’ refers to the local extent of the untreated primary tumour at the time of diagnosis and initial workup; ‘N’ refers to the status of the regional lymph nodes, and ‘M’ refers to distant metastatic disease (refer to Table 1.2 and Table 1.3). Pathologic classification is based on gross/microscopic examination of the resection specimen of a previously untreated primary tumour and clinical classification is based on evidence acquired through a variety of techniques that include physical examination, radiologic imaging, endoscopy, biopsy, and surgical exploration [49]. The Dukes classification system relies on the pathological classification of disease stages and is thus not as useful to clinicians as it cannot be used to preoperatively evaluate the patient [49]. Despite significant progress in our understanding over the past twenty years the 5-year survival rate for CRC remains at almost 90% for stage I cancers to 5%–15% for stage IV disease [50, 51]. Another important feature of CRCs are their histological grade. Tumours are graded into three categories based on the degree of gland-like structures and thus its resemblance to the original tissue. Accordingly, Grade 1 tumours are well-differentiated, Grade 2 tumours are moderately differentiated, Grade 3 tumours are poorly differentiated and Grade 4 tumours are undifferentiated [52]. CRC prognosis correlates well with both stage and grade [53]. Furthermore, tumour grade was also found to correlate significantly with the TNM staging system [52].
Table 1.2 Tumour Stage Definitions According to the American Joint Committee on Cancer (AJCC)

<table>
<thead>
<tr>
<th>TNM Stage</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>TX</td>
<td>Primary tumour cannot be assessed</td>
</tr>
<tr>
<td>T0</td>
<td>No evidence of primary tumour</td>
</tr>
<tr>
<td>Tis</td>
<td>Carcinoma <em>in situ</em></td>
</tr>
<tr>
<td>T1</td>
<td>Tumour invades submucosa</td>
</tr>
<tr>
<td>T2</td>
<td>Tumour invades the muscularis propria</td>
</tr>
<tr>
<td>T3</td>
<td>Tumour invades through muscularis propria</td>
</tr>
<tr>
<td>T4</td>
<td>Tumour invades other organs/peritoneum</td>
</tr>
<tr>
<td>NX</td>
<td>Regional lymph node metastasis cannot be assessed</td>
</tr>
<tr>
<td>N0</td>
<td>No regional lymph node metastasis</td>
</tr>
<tr>
<td>N1</td>
<td>Metastasis in 1-3 regional lymph nodes</td>
</tr>
<tr>
<td>N2</td>
<td>Metastasis in $\geq 4$ regional lymph nodes</td>
</tr>
<tr>
<td>MX</td>
<td>Distant metastasis cannot be assessed</td>
</tr>
<tr>
<td>M0</td>
<td>No distant metastasis</td>
</tr>
<tr>
<td>M1</td>
<td>Distant metastasis</td>
</tr>
<tr>
<td>Stage</td>
<td>T</td>
</tr>
<tr>
<td>-----------</td>
<td>----</td>
</tr>
<tr>
<td>Stage 0</td>
<td>Tis</td>
</tr>
<tr>
<td>Stage I</td>
<td>T1</td>
</tr>
<tr>
<td></td>
<td>T2</td>
</tr>
<tr>
<td>Stage IIA</td>
<td>T3</td>
</tr>
<tr>
<td>Stage IIB</td>
<td>T4</td>
</tr>
<tr>
<td>Stage IIIA</td>
<td>T1, T2</td>
</tr>
<tr>
<td>Stage IIIB</td>
<td>T3, T4</td>
</tr>
<tr>
<td>Stage IIIC</td>
<td>Any T</td>
</tr>
<tr>
<td>Stage IV</td>
<td>Any T</td>
</tr>
</tbody>
</table>

Dukes: corresponding Dukes stage, NA: Not applicable
1.1.4 Molecular pathways in CRC

CRC is a complex and heterogeneous disease with multiple genes in diverse pathways involved in its initiation, progression, invasion and metastasis. It is widely accepted that the sequential accumulation of mutations that activate oncogenes and disrupt tumour suppressor genes, combined with multiple cycles of clonal selection and evolution facilitate the process of carcinogenesis. While previous studies estimate that the disruption of about six to seven cellular processes are required for transformation[54], a recent comprehensive sequence evaluation of colon cancer genomes hints that this number may be even higher (about twelve) than previously estimated [55]. Baseline mutation rates, which are estimated to be about $10^{-9}$ [56], cannot account for the number of mutations that are required for malignant transformation. Thus, the rate-limiting step in the process of carcinogenesis is the rate at which new mutations occur, and it was proposed that cancer cells must acquire some form of genetic instability to facilitate the process of carcinogenesis [57, 58]. Genetic instability refers to a set of events capable of causing unscheduled alterations, either of a temporary or permanent nature within the genome. In CRC at least 3 major pathways that contribute to genomic instability have been described, the chromosomal instability (CIN) pathway, microsatellite (MSI) pathway, and CpG island methylator (CIMP) pathway. These pathways are not mutually exclusive; therefore a tumour may occasionally exhibit features of multiple pathways. For instance, 25% of MSI CRCs can exhibit chromosomal instabilities [59], and about 12% of CIN-positive tumours show a high level of MSI [60]. Furthermore, while CIMP accounts for the majority of MSI-positive/ CIN-negative CRCs, as much as 33% of CIMP-positive tumours can exhibit a high level of
chromosomal abnormalities [61]. Nevertheless, the biological significance of these overlapping features are not fully understood [9].

1.1.4.1 Chromosomal instability (CIN)

CIN refers to the existence of accelerated rate of chromosomal alterations, which result in gains or losses of whole chromosomes as well as inversions, deletions, duplications and translocations of large chromosomal segments, which results in karyotypic variability from cell to cell [62]. CIN is considered to promote carcinogenesis through loss of tumour suppressors and the copy numbers gains of oncogenes [63], and CIN is observed in 65-70% of CRC. Aneuploidy refers to an abnormal karyotype and Lengauer et al have shown that aneuploidy is the result of an abnormally high rate of CIN that persists throughout the life of the tumour [62]. The APC gene located on chromosome 5q21 is commonly mutated in CRC. Furthermore, APC mutations are frequently found in dysplastic aberrant crypt foci (ACF) that show the highest potential for progressing to CRC[64]. There are two possible mechanisms by which APC mutations might contribute to the progression of CRC, the first is by over activation of the Wingless/Wnt signaling pathway and secondly by the induction of chromosome instability[64]. Wingless/Wnt signaling pathway is shown to be important for several process including organ development, cellular proliferation, morphology, motility and cell fate determination[65]. In this pathway, APC binds to a complex of proteins that includes β-catenin and glycogen synthase kinase-3β on an axin scaffold to form the β-catenin destruction complex (Figure 1.). This facilitates the phosphorylation of β-catenin by glycogen synthase kinase-3β, which leads to its degradation via the proteasome. However, in the presence of the Wnt ligand, β-catenin is no longer phosphorylated and is thus allowed to accumulate in the cytoplasm [66, 67]. β-catenin then translocates to the nucleus, and together with the members of T-cell factor(Tcf) and lymphoid enhancer factor(Lef) family of transcription factors, drive the transcription of multiple genes.
Mutations in β-catenin (CTNNB1) or APC, both of which occur in a large subset of CRCs, increase the transcriptional activity of β-catenin [68, 69]. In the second mechanism APC leads to chromosomal instability through its interactions with EB1. APC has an EB1-binding domain in its C-terminal end. EB1 associates with the growing ends of microtubules and centrosomes and has been shown to be important in the cytokinesis checkpoint in yeast [70]. APC accumulates at the kinetochore during mitosis and Apc mutant cells form mitotic spindles with an abundance of microtubules that are not able to connect efficiently with kinetochores [71]. Mouse embryonic stem cells homozygous for the mutant Min (multiple intestinal neoplasia) or Apc1638T alleles demonstrate CIN and polyploidy [71]. Although these studies implicate APC mutations in the generation of genomic instability observed in colon adenomas and CRC, the significance of these observations is still not clear[64]. Thus, it has been proposed that the inactivation of APC creates a permissive state that allows the evolving tumour to tolerate the development of aneuploidy and CIN [64, 72, 73].

The observation that cancer cells harbour an abnormal number of chromosomes was made almost a century ago [74, 75], despite this our understanding of what causes this phenomenon remains incomplete. To date several pathways and processes have been implicated in CIN including; a) pathways that ensure accurate chromosome segregation, b) abnormal centrosome number and function, c) pathways involved in telomere stability, and d) cell cycle checkpoint pathways and kinases [9, 76]. Several of these processes will be briefly discussed below.

The mitotic checkpoint is the major cell cycle control mechanism that ensures high fidelity of chromosome segregation. It is responsible for delaying anaphase until all pairs of duplicated chromatids are properly aligned on the metaphase plate. Defects in this process lead to the mis-segregation and the unequal and abnormal distribution of chromosomes into daughter cells
resulting in aneuploidy. The *mad* and *bub* genes, two highly conserved key components of the mitotic spindle checkpoint that appear to operate as checkpoint sensors and signal transducers were found to be deregulated in a subset of colon cancers[77]. MAD and BUB inhibit the anaphase-promoting complex/C (APC/C), an ubiquitin protein ligase, and also lead to cell-cycle arrest. At anaphase, APC/C normally becomes active and ubiquitinates securin, which is then degraded through the ubiquitin-proteasome pathway [78-80]. Securin functions by interacting with separins, a class of caspase-related proteases that regulate cohesins, multi-protein complexes that create physical links that maintain sister chromatids together during metaphase [81, 82]. Other than MAD and BUB, there are additional factors that play import roles in this process. For example, the securin-separin complex is regulated by an autofeedback loop that involves other proteins such as polo kinases (PLK) [83]. Mutations in two genes that control the human mitotic checkpoint, *BUB1* and *BUBR1*(MAD3), were found in a subset of CIN colon cancer cell lines [84]. Other genes involved in kinechore function are reported to be either mutated or overexpressed in colon cancer [85].

Another mechanism proposed to cause CIN is abnormal centrosome number and function. There is a strong association between abnormal centrosome number and aneuploidy in cancers [86, 87]. The centrosome plays a central role in chromosome segregation during mitosis by serving as an anchor for the reorganization of microtubules into the mitotic spindle apparatus, and it has been postulated that extra centrosomes result in the unequal distribution of chromosomes [9]. A pathogenic role has been identified for centrosome associated Aurora and Polo-like kinases. STK15/BTAK/ARK1/aurora 2, a centrosome-associated serine-threonine kinase has been showed to be amplified and overexpressed in colon cancer cell lines, and this STK15 overexpression induced aneuploidy in NIH3T3 cells [88, 89]. Polo-like kinase 1(Plk1), a serine/threonine kinase, was demonstrated to be overexpressed in 73% of primary CRCs, and
this correlated with tumour invasion, lymph node involvement and Dukes stage [90]. A recent study implicated centrosome amplification in tumourigenesis when the transplantation of extra centrosomes in the larval brain cells of flies induced the formation of metastatic tumours [91].

Telomeres refer to the segments of DNA bound by specific proteins that cap the ends of chromosomes. Thus, telomeres act as a buffer to prevent the loss of valuable genomic sequence during replication[92], as well as to prevent chromosomes fusing at the ends [76]. A RNA primer is required for the process of DNA replication. Thus, when replication proceeds from the 5′→3′ direction, it leaves a stretch of unreplicated DNA at the 5′ end. This leads to a gradual loss of telomeric repeats and the consequent shortening of telomeres after each round of replication, which is referred to as the ‘end-replication problem’ [76]. A specific enzyme, telomerase maintains telomere length. Telomerase consists of two main components; the reverse transcriptase component (hTERT), which is only expressed in cells where telomerase activity is present; and the ribonucleoprotein moiety (hTERC/hTR), which is expressed ubiquitously in all cells. In adults, telomerase activity has been observed only in immature germ cells, certain stem/progenitor cells and in a subset of somatic cells such as human fibroblasts. Upon reaching a critical short telomere length, DNA damage checkpoints trigger senescence pathways and apoptosis. Cells that survive this checkpoint activate telomerase or the Alternative Lengthening of Telomeres (ALT) pathway. It has been proposed that it is therefore important for cancer cells to regain the ability to maintain telomeres, in order to avoid senescence and extensive chromosome fusion during crisis [93, 94]. Additionally, telomere loss has been shown to inhibit tumour progression and the development of macroscopically advanced tumours [95-98]. It has been estimated that about 85-90% of human cancers have reactivated telomerase and are able to maintain telomere length[76] and increased telomerase activity has been observed in colorectal tumours[99-101]. Interestingly cancer cells that are deficient for telomerase activity are able to
maintain telomere length via ALT, possibly making use of DNA repair pathways and recombination to maintain telomere length[102]. Thus, it appears that maintaining telomere length is critical for tumourigenesis and cellular immortalization [76]. However, conflicting with this view is the observation that the telomeres of invasive human cancers are often shorter than their normal counterparts [103]. Furthermore, when the protection of telomere ends is compromised, chromosomes enter breakage-fusion-bridge cycles that can dramatically reorganize the genome [9]. Studies in telomerase deficient mice (mTERC\textsuperscript{-/-}) demonstrated that telomere shortening led to increased rates of spontaneous tumour formation, initiation of aberrant crypt foci and microadenomas in the gastrointestinal tract that[98]. Furthermore, telomeres from colon cancer cells were found to be 77%-90% shorter than those in adjacent normal tissue [100, 104-106]. Taken together this indicates that the timing at which the telomeres shortening occurs plays a crucial role in cancer development [107]. Telomere shortening promotes CIN, which in turns drives early carcinogenesis. In contrast telomerase activation and telomere stability confer immortality during the latter stages of tumourigenesis [9].

Genomic stability is maintained by checkpoints that regulate the maintenance of DNA fidelity in response to exogenous and endogenous genotoxic stress. In response to these stressors these checkpoints initiate a cascade of events that culminates in either cell-cycle arrest, which provides the cell sufficient time to repair the damage; or apoptosis/cellular senescence, if the damage cannot be repaired. These DNA damage checkpoints are traditionally considered to belong to two classes, the first group consists of checkpoints that recognize and respond to DNA damage from genotoxic agents and the second is composed of those that monitor the fidelity of replicated DNA [64, 108]. A number of DNA repair proteins have been demonstrated to play a role in human cancers, including TP53, ATM, ATR, BRCA1 and BRCA2 and hRad17 [109-112]. Of all these checkpoint proteins, TP53 has been the only one directly implicated in human colon
cancer. However, it appears that while specific TP53 mutations cause CIN, the remaining mutations likely affect other processes such as apoptosis, which has been postulated to create a permissive state that allows CIN to occur but does not directly cause CIN [113]. Furthermore, the types of rearrangements observed in primary human cancers are similar to those seen when these checkpoint proteins are inactivated in yeast[114]. This observation provides indirect evidence for a role of the replication checkpoints in the genomic instability of colon cancer.

In a subset of colorectal cancer cell lines, inactivation of the base excision repair (BER) system appears to predispose to point mutations that contribute to CRC formation by causing genomic instability at the level of the base pair [64]. BER is mainly responsible for repairing damage induced by endogenous metabolic processes such as methylation, deamination, reactive oxygen species (ROS) and hydrolysis [115]. Multiple proteins contribute to BER pathway and enable it to correct non-bulky damaged nucleotides, abasic sites as well as single-strand breaks. The process is initiated by DNA glycosylases specific for various types of damage, which recognize and cleave the N-glycosylic bond that connects the damaged base to the DNA backbone [116]. To date, 11 such DNA glycosylases have been identified in mammals [117]. Biallelic inactivation of one of the BER gene, MYH, leads to an inherited form of colorectal cancer known as MYH-associate polyposis (MAP) [118]. Tumours with biallelic MYH germline mutations demonstrate CIN but no show no differences in the frequency of mutations in genes such as TP53, SMAD4, or TGFBR2, which suggest that they have a unique molecular pathogenesis [119].

1.1.4.2 Microsatellite instability pathway (MSI)

Microsatellite instability (MSI) defined as a change of any length due to either insertion or deletion of repeating units, in a microsatellite within a tumour when compared to normal tissue
[120]. Microsatellites lengthen or shorten due to the increased propensity of DNA polymerases to slip in these regions. Defective mismatch repair activity is associated with the MSI phenotype. Instability of microsatellite repeats is seen 15% of sporadic CRCs, primarily due to hypermethylation of the promoter of the mismatch repair gene, *MLH1* [121, 122]. In addition, tumours of as many as 85% of patients with Lynch syndrome (an inherited CRC syndrome) also show MSI as a consequence of mismatch repair gene mutations [123, 124]. A number of clinicopathological features have been linked with MSI-H, such as mucinous differentiation, signet ring cell morphology, Crohn’s-like reaction, increased tumour-infiltrating lymphocytes, poor differentiation and a tendency to be located in the proximal colon [125-127]. Moreover, in CRC the MSI-H phenotype is associated with a favorable prognosis and a decreased likelihood of metastases [128].

Microsatellites are repetitive DNA sequences that occur throughout the genome. The functional impact of resulting mutations is determined by their genomic localization. While these sequences occur mostly in non-coding regions, short repetitive regions can also occur in the coding regions of tumour suppressor genes. Mutations that occur within coding regions have the potential to cause frameshift and truncating mutations, while those that occur within regulatory intronic regions may potentially alter transcript level and stability. Recently, a public database of mononucleotide repeat tracts mutated in microsatellite unstable human tumours was described (http://www.seltarbase.org/) [129]. It is expected that this will facilitate the identification of relevant genes in carcinogenesis driven by MSI. This database lists the multiple genes that display somatic mutations in MSI-H tumours including *TGFβRII, BAX, IGFIIR, BLM, CHK1, caspase-5, MSH3* and *MSH6*. The impact and frequency of mutation in several of these genes is briefly discussed below.
It has been estimated that approximately 85% of CRCs with MSI have mutated TGFβ receptor II (TGFβRII) [5, 130]. Somatic mutations that inactivate TGFβRII are mainly frameshift mutations within the 10 base-pair poly (A) repeat in the coding region (codons 125–128) [5]. Additionally, protein truncation occurs due to insertion mutations in the poly (GT)3 microsatellite region in the TGFβRII gene [5, 130]. In the colon, the ligand of TGFβRII, TGFβ1 potently inhibits the growth of normal epithelial cells. Thus, mutated TGFβRII allows tumor cells to escape these growth-suppressive effects.

Insulin like growth factor II receptor (IGFIIR) plays a critical role in cell growth, survival and differentiation; however, once mutated, these receptors play a role in tumourigenesis [131]. The IGFIIR interacts with both IGF-II and latent TGFβ1 ligands and acts as a tumor growth suppressor. The IGFIIR gene harbors several microsatellites within its coding region, one of these is a poly (G)8 repeat that is often mutated in MSI-H tumours [132].

Disruption of the APC/β-catenin pathway is a critical step in the development of CRC and other cancers. The transcription factor Tcf-4 is a Tcf-family protein, which complexes with β-catenin and plays an important role in maintenance of normal epithelial growth and development of colorectal tumours. Frameshift mutations in the poly(A)9 mononucleotide repeat in exon 17 of the Tcf-4 gene were found in MSI cells [133, 134]. This produces a short and medium isoform of the Tcf-4 protein and the properties of these isoforms are still not clear. One possibility is that mutated Tcf-4 protein may acquire increased transcriptional activity by increased binding affinity for β-catenin. These mutated Tcf-4 proteins also lose the ability to bind chromatin remodeling complexes and in doing so may facilitate the structural organization of chromatin [135]. It has been suggested that mutations in Tcf-4 may contribute only marginally to the colorectal carcinogenesis or perhaps even be passenger mutations [136].
In endometrial and colorectal cancers, mutations in \textit{PTEN} gene occur in MMR deficient cells [137]. The loss of \textit{PTEN} causes elevation of intracellular phosphatidylinositol 3-phosphate (PIP3) levels that stimulate serine/threonine kinase Akt kinase activity. The active Akt leads to tumourigenesis due to loss of proliferative and apoptotic controls of cell growth. The coding region of the \textit{PTEN} gene contains several microsatellites, among which a repeat of poly(A)$_6$ tract in exons 7 and 8, which are most susceptible to frameshift mutations[138].

Approximately 50\% of the tumours with MSI have mutations in the \textit{Bax} gene. These mutations produce frameshift mutations within a coding region of eight nucleotide stretch of guanine residues (G$_8$) [139]. Bax is a pro-apoptotic protein that forms a heterodimer with BCL2, and functions as an apoptotic activator. Bax is also thought to play a role in the release of cytochrome C from the mitochondrion. Thus, mutations in BAX contribute to tumourigenesis by decreasing apoptosis. MSI tumours also show somatic mutations in the MMR genes, \textit{MSH2}, \textit{MSH3}, which may further exacerbate the MSI phenotype [140].

However, the panel of genes mutated in diverse cancers appears to be distinct. For example a recent study found that mutations in of \textit{TGF\beta RII}, \textit{BAX}, \textit{IGFIIR}, \textit{caspase-5}, \textit{MSH3} and \textit{MSH6} are rare in urothelial bladder carcinoma and are not associated with the MSI-H phenotype [141]. Furthermore, the temporal sequence of these mutations appears to be an important feature with \textit{TGF\beta RII} and \textit{Bax} mutations reported to be later events in the adenoma to carcinoma sequence [142, 143].

The MSI status of tumours can be determined by carrying out PCR on DNA from fresh or formalin-fixed paraffin-embedded tissue. Individuals can show heritable and stable polymorphic variation within microsatellite regions. Thus, differences in number of repeats at microsatellite loci in tumour (abnormal) tissue can be only demonstrated when compared with normal tissue of
the same individual [120]. The National Cancer Institute has recommended a panel of five microsatellite markers for use in CRC, these include two mononucleotide repeats (BAT25 and BAT26) and three dinucleotide repeats (D5S346, D2S123 and D17S250)[120]. A tumour that shows instability in at least 2 of the 5 tested microsatellites is termed MSI-high (MSI-H), 1 of the 5 is MSI-low (MSI-L), and 0 of the 5 is considered microsatellite stable (MSS)[120]. If more than 5 markers are used to identify this phenotype then MSI-H is defined as instability at ≥30-40% and MSI-L as instability at < 30%-40% of the markers. This panel may not be ideal for the evaluation of MSI in endometrial carcinoma as significant differences in patterns of microsatellite abnormalities are found between endometrial and CRC in families with MLH1 or MSH2 germline mutations [144].

The independent existence of a MSI-L phenotype that is distinct from MSS has been contested in the literature. For example, one study demonstrated that the majority of CRCs show some degree of microsatellite instability when a large enough number of markers are tested, indicating that the differences between the MSI-L and MSS subgroups is merely quantitative rather than qualitative [145]. However, an emerging but significant body of work indicates that MSI-L is a distinct biological entity and is not simply a PCR artifact [146, 147]. For example one study distinguished MSI-L CRCs from the MSI-H and MSS subgroups based on gene expression profiles[148]. MSI-L tumours also show frequent instability of the trinucleotide repeat region in Ras-induced senescence 1 (RIS1) [149]. MSI-L status has also shown to be an independent adverse prognostic feature in stage III CRC in patients not treated with chemotherapy [150], particularly when associated with RIS1 mutation [149]. Despite these advances, further studies are needed to determine the biological significance of this phenotype.
MSI status may also serve as a predictive marker of the efficacy of certain chemotherapeutic agents. For example, 5-Fluorouracil (FU) is a widely-used chemotherapeutic drug that initiates apoptosis by targeting thymidine synthetase and/or by the direct incorporation of its metabolites into DNA and RNA. Despite its success at treating colon cancer several recent studies have suggested that 5-FU may not confer a survival advantage to stage II and III CRC patients whose tumours are MSI-H compared to those with MSS tumours [151]. While these findings have sparked some debate, emerging evidence indicates that the MSI status of a tumour is in fact an important predictor of 5-FU response[152].

Other than the above mentioned mechanisms that are implicated in MSI, a recent study demonstrated that the overexpression of microRNA (miR)-155 downregulated the MMR proteins, MSH2, MSH6, and MLH1, consequently induced MSI [153]. miRs are noncoding RNAs that play a role in the posttranscriptional regulation of over 30% of the human genes controlling critical biological processes. Thus, miR-155 likely acts to destabilize the expression of these MMR proteins. Furthermore, this study found an inverse correlation between the expression of miR-155 and the expression of MLH1 or MSH2 proteins in human CRC. Previous studies have noted that miR-155 is overexpressed in CRC [154, 155], and that this appears to be more frequent in MSI than MSS tumours [156].

1.1.4.3 CpG island methylator phenotype (CIMP)
In addition to the chromosomal and genetic aberrations discussed above, epigenetic modifications that affect both DNA and the associated chromatin are capable of influencing gene expression and the stability of the genome. DNA methylation or the covalent modification of the C-5 position of cytosine residues occurs primarily at the short stretches of CG dinucleotides known as CpG islands. Recent estimates suggest that there are at least 29,000 such regions in the
human genome, many of which surround the 5' ends of genes [157]. In eukaryotes methylation is thought to play a role in regulating gene expression and in silencing repeat elements in the genome [158]. Methylation also plays an important role in inactivating one copy of the X chromosome, so that equal gene dosage is maintained in the somatic cells of males and females [159]. In normal cells the pattern of expression is stably maintained following DNA replication and cell division by a maintenance enzyme, DNA methyltransferase (DNMT1). The first epigenetic mechanism implicated in carcinogenesis was DNA hypomethylation (decreased methylation) [160]. Age related decreases in DNA methylation levels occur in a tissue specific manner [161, 162], and it is likely that these changes contribute to the age-related increase in carcinogenesis [163]. CpG islands commonly occur in gene promoter regions, thus hypermethylation (increased methylation) of this region has been shown to silence gene expression [164]. Thus, given its ability to inactivate tumour suppressor genes DNA hypermethylation has been widely studied in cancer and several genes including, *MLH1*, *MGMT* and *HLTF* have been shown to be silenced by hypermethylation in CRC [64]. Furthermore, a number of genes such as *HLTF*, *SLC5A8*, *MGMT*, *MINT1*, and *MINT31* show aberrant hypermethylation in aberrant crypt foci indicating that hypermethylation is an early event in the adenoma-carcinoma sequence [165, 166].

In 1999 it was postulated that the subset of CRCs that demonstrate widespread CpG island hypermethylation arose via a distinct pathway termed CpG Island Methylator Phenotype (CIMP)[167]. However, the concept of CIMP was initially debated as it was not clear whether CIMP tumours are a unique molecular subgroup, or if it was composed of tumours that fall on the upper end of the continuum of aberrant DNA methylation; in other words if the only distinguishing characteristic was merely quantitative differences in methylation[64].
The initial report of CIMP used methylated in tumours (MINT) as markers of these tumours [167]. Since then there has also been considerable research to discover the most suitable panel of markers and at present it is not clear which markers will most accurately identify CIMP tumours. This is further complicated by the finding that CIMP tumours are not a homogenous entity, and CIMP has been classified into the subcategories CIMP-high (also referred to as CIMP-1), CIMP-low (CIMP-2) and CIMP-0 [147]. For example, the methylation markers CACA1G, CDKN2A, BRABP1, IGF2, MLH, NEUROG1, RUNX3, and SOCS1 are specific for CIMP-high but are not optimal for the identification of CIMP-low tumours[168]. In terms of biological differences, CIMP-low was shown to be associated with KRAS mutation and male sex, CIMP-high was associated with BRAF mutation and female sex, and CIMP-0 is associated with wild-type KRAS and BRAF[168]. Furthermore in a study using a large number of methylation markers, CIMP-high tumours formed a distinct group by unsupervised cluster analysis suggesting that CIMP is a distinct subtype of CRC [169].

The mechanism by which CIMP arises is currently unknown but is the subject of much research. However, it has been proposed that the synergy between KRAS and BRAF mutations in combination with specific patterns of gene methylation is necessary to facilitate the early events of tumourigenesis [147]. Additionally, it was recently proposed that all CRCs be classified based on MSI and CIMP status. However, further studies will need to be undertaken prior to implementing this classification [63, 147].
Figure 1.1 The simplified progression of CRC via the CIN, MSI and CIMP pathways.

Depicted by panel A, B and C respectively. Several key genes that are implicated in each pathway are highlighted. Reprinted from Gastroenterology, Grady, WM and Carethers, JM. Genomic and Epigenetic Instability in CRC pathogenesis, copyright 2008, with permission from Elsevier [64].
1.1.4.4 Cancer stem cell model and CRC

A cancer stem cell is a cell within the tumour that possesses the capacity to self-renew, and in doing so gives rise to the heterogeneous lineages that comprise the tumour. Cancer cells may arise therefore from tissue stem cells that have acquired mutations that render them cancerous, or it may be a more differentiated (progenitor) cell that may have ‘re-acquired’ stem cell like properties due to mutations [170]. Cancer stem cells or cancer initiating cells have been identified to date in acute myelogeneous leukemia[171], breast tumours [172], brain tumours [173, 174] and most recently in a subset of colon tumours [175]. The colon cancer initiating cells were able to induce human colon tumours that were phenotypically similar to the original patient tumours after transplantation into non-obese diabetic (NOD)/severe-combined immunodeficient (SCID) mice [175]. This cell population also self-renewed, which enabled colon cancer to be re-established in secondary and tertiary recipient mice. A recent report identified Lgr5\(^+\) cells, which are located at crypt bottoms as the putative stem cells of the small intestine and colon [176]. Furthermore, deletion of Apc in these Lgr5\(^+\) cells led to the growth of microadenomas that progressed to form adenomas in 3-5 weeks [177]. However, when Apc was deleted in the short lived transit amplifying cells, adenomas were rarely seen even 30 weeks later, which led the authors to conclude that the stem-cell specific loss of Apc results in a progressively growing neoplasia [177]. Furthermore, a recent study found that high Wnt activity functionally designates the colon cancer stem cell population[178]. Thus, based on this evidence the adenoma to carcinoma model has been refined to suggest that the first mutation in APC occurs in the gastrointestinal stem cell, which generates an APC \(^{-/-}\) mutant stem cell. These cells will then progressively colonize the entire stem cell niche, while their progenitors take over the crypts, an event that has been termed monoclonal conversion [9]. Once the second APC allele is lost, niche succession occurs again and the niche fills with APC \(^{-/-}\) cells. The subsequent
acquisition of downstream mutations in KRAS and TP53 will then occur within this cancer stem cell population.

1.1.5 Behavioural and lifestyle risk factors

Multiple studies have attempted to evaluate the relationship between CRC risk and the intake of specific dietary compounds. However, several of these studies have produced conflicting data and the precise involvement of dietary factors in decreasing or increasing the risk of CRC is currently unclear. Several epidemiologic studies have shown an association between a diet low in fruits and vegetables and increased risk of CRC [179, 180]. Furthermore, the relative risk of CRC was found to be approximately 0.5 in groups with the highest vegetable and fruit intake compared to those with the lowest [179]. Also a recent systematic review of the literature indicates that the majority of available evidence supports the association of red meat consumption with an increased risk of CRC [181, 182]. PhIP (2-amino-1-methyl-6-phenylimidazo[4,5-b]pyridine) is a bulky adduct forming agent that is commonly found in well-done meat and poultry and has been shown to cause a variety of cancers in experimental animals [183-186]. Another powerful rodent carcinogen is MNNG (N-methyl-N'-nitro-N-nitrosoguanidine), is an alkylating agent that is able to preferentially methylate the O6 position of deoxyguanosine residues in DNA [187]. Gastrointestinal cells are continually exposed to both PhIP and MNNG at varying concentrations. In a study undertaken to determine if carcinogen exposure can influence the type of instability seen in cells, it was found that cells resistant to PhIP developed a chromosomal instability or CIN phenotype. Conversely, cells resistant to MNNG exhibited MSI and associated mismatch repair defects [188]. This suggests that exposure to certain dietary carcinogens, may in fact select for cancer cells with distinct types of genetic instability, however further work is needed to confirm these findings [188].
An association between heavy alcohol consumption and increased risk of CRC has been reported in several studies [189, 190]. It has been estimated that the CRC risk was increased slightly (relative risk of 1.41) in individuals who consumed over 45g of alcohol per day, and this risk was slightly lower for those who consumed between 30 to 45g per day [191]. Interestingly, alcohol consumption was found to be associated with an increased risk of cancers of the distal colon and rectum but not with cancers of the proximal colon [192]. The specific mechanism linking ethanol and carcinogenesis is not clear. However, it is believed that the major toxic byproduct of ethanol metabolism, acetaldehyde is capable of inducing point mutations, chromosomal aberrations and alterations in protein structure [193].

Furthermore, multiple studies have examined the contribution of lifestyle factors such as smoking and physical activity to CRC incidence. Decreased regular physical activity is associated with increased risk of CRC, an effect that has been proposed to be linked to high energy intake, obesity, and the metabolic syndrome [194]. However, the precise mechanism by which physical activity exerts its protective effect is currently unclear [182]. In a recent meta-analysis smoking was associated with an increased risk of CRC incidence and mortality [195]. This study also found that this risk increased with increasing dose (increased number of pack years and cigarettes per day), but this association was only statistically significant after 30 years of smoking. Interestingly, with respect to both incidence and mortality, the association was stronger for cancer of the rectum than of the colon; suggesting that unique pathological processes at play at distinct anatomic sites.

Patients with chronic inflammatory bowel diseases involving the colon such as ulcerative colitis or Crohn’s disease are at significantly increased risk for CRC [196, 197], and special colonoscopic surveillance measures are recommended for such individuals [198]. Several
factors may influence the risk of CRC in patients with inflammatory bowel disease, including earlier onset of disease (prior to 15 years) and the severity of inflammation [182, 199].

1.1.6 Hereditary syndromes, pathology and presentation

The most common form of inherited CRC is Lynch syndrome (MIM# 120435), also referred to as hereditary nonpolyposis colorectal cancer (HNPCC). Due to its relevance to this thesis, this syndrome will be discussed in more detail in section 1.2. Familial adenomatous polyposis (MIM# 175100) is an autosomal dominant disorder that is caused by germline mutations in the \textit{APC} gene [200]. In this syndrome multiple adenomas occur at an early age, often during preteen years and proliferate throughout the colon with malignant degeneration by the age of 40 - 50 years. Adenomas in the duodenum, which carry the risk of progressing to periampullary carcinoma are also seen in FAP [201]. Furthermore, some individuals especially those of Korean or Japanese descent also have an increase likelihood of adenomas and cancers of the stomach [202]. Most mutations detected in this gene are truncating or frameshift mutations and a smaller proportion are missense in nature. The APC tumour suppressor gene acts in a nonclassical manner in that the localization of the germline mutation determines the position as well as the type of second hit likely to occur [203, 204]. Moreover, an association was also found between the location of the mutation, the number of colonic polyps and the severity of the disease. For instance, premature stop codons that occur between codon 1250 and 1464 cause a profusion (about 5000) of polyps, while mutations that are upstream or downstream of this region are associated with a later onset of disease and lead to a lesser number of polyps (about 1000)[205]. The penetrance of this syndrome varies, with penetrance approaching 100\% in classic FAP caused by germline mutations in the \textit{APC} gene. However, penetrance is only about 10-20\% for the I1307K \textit{APC} polymorphism in Ashkenazi Jews. An
attenuated form of this disease occurs when mutations occur at the extreme 5’ end, the extreme 3’ end or in exon 9 of APC [206, 207]. These individuals had comparatively fewer polyps (<50-100), and the polyps tended to occur one to two decades later in life than classical FAP. Gardner syndrome represents one tail of the FAP phenotype. Gardner syndrome represents the association of colonic adenomatous polyposis of classic FAP with osteomas and soft tissue tumours (epidermoid cysts, fibromas, desmoid tumours). APC mutations associated with Gardner syndrome tend to occur between codons 1,403 and 1,587 [208].

Germline mutations in the BER gene, MutY human homologue (MYH) leads to an inherited form of colorectal cancer known as MYH-associated polyposis (MAP; MIM# 608456)[118]. It has been approximated that MYH is the cause of adenomatous polyposis in up to 5%–10% of individuals who have FAP. The identification of an excessive number of somatic G:C → A:T mutations in neoplasms of people with adenomatous polyposis with no detectable germline mutations in APC made way for this discovery [118, 209, 210]. This is due to the ability of reactive oxygen species to modify the C8 position of Guanine to form 7, 8-dihydro-8-oxoguanine (8-oxoG), a highly mutagenic substrate that is able base pair with adenine and cause G:C→T:A transversions [211]. Interestingly, two common missense mutations Y179C and G396D (formerly described as Y165C and G382D) account for the majority (~80%) of mutations discovered in MYH [118, 210, 212, 213]. Furthermore, compared to noncarriers monoallelic and biallelic carriers of these germline mutations have an increased risk of CRC and an increased likelihood of having first-and second-degree family members with CRC [212]. Despite the identification of germline mutations in MYH, somatic MYH mutations are not common in sporadic colorectal cancer (203, 204).
There are also several rare hamartomatous polyposis syndromes that predispose to CRC. Germline mutations in the PTEN gene are responsible for phenotypically diverse conditions which have are collectively referred to as the PTEN hamartoma tumour syndrome (PHTS). One disease that belongs to this subset is Cowden syndrome (CS, MIM#158350), a rare autosomal condition that is difficult to diagnose [214] The most important and universal indicator of CS is mucocutaneous features. Over 60% of those affected with CS are thought to develop hamartomatous polyps in the GI tract [215] and while colon cancer has been reported, its involvements in CS has not been confirmed [216]. Although originally thought to be distinct conditions, Bannayan-Riley-Ruvalcaba syndrome (BRRS, MIM#153480) is now believed to be a single entity. In the absence of definitive clinical criteria, it has been suggested that individuals with combinations of macrocephaly, lipomatosis, hemangiomas, GI polyposis, and pigmented macules of the penis be considered clinically affected[217]. Furthermore, GI, breast, thyroid and endometrial cancers are thought to be features of this syndrome[218].

Peutz-Jeghers (PJS, MIM#175200) is an autosomal dominant syndrome associated with mucocutaneous hyperpigmentation and benign intestinal polyps known as hamartomas. The characteristic pigmented macules are benign and are not thought to have malignant potential. Approximately 50% will have experienced gastrointestinal symptoms by age 20 [219]. Unlike other hamartomatous syndromes, PJS related polyps occur most frequently in the small intestine. These individuals are at a high risk of developing malignancies especially of the breast and colon [220, 221]. A majority (60-70%) of Peutz-Jeghers patients show germline mutations in the serine/threonine protein kinase STK11 (also known as LKB1) [222].

Juvenile polyposis syndrome (JPS, MIM#174900) often presents before the age of 20 and is characterized by the development of multiple juvenile polyps in the gastrointestinal tract.
Approximately 25%-50% of JPS cases are believed to be caused by *de novo* mutations[216]. Deletions in SMAD4 and BMPR1A, both components of the TGFβ pathway account for a significant proportion of JPS-causing mutations [223]. Approximately 20% of individuals with JPS are thought to have mutations in SMAD4, and a similar proportion of JPS is thought to arise due BMPR1A mutations [224]. Individuals with JPS are at risk of developing cancers in the stomach, small intestine and colon.

1.2  **Lynch Syndrome**

1.2.1  **Overview**

Germline mutations in mismatch repair genes are associated with Lynch syndrome or hereditary nonpolyposis colorectal cancer (HNPCC). Lynch syndrome, an autosomal dominant disorder that accounts for about 2% of all CRC cases, is one of the most common cancer predisposition syndromes. Instability of microsatellite repeats is seen in tumours of as many as 85% of patients with HNPCC, making it a hallmark feature of this syndrome [123, 124]. Five cardinal features have been suggested to help identify families affected by Lynch syndrome [225]: 1) an earlier age at cancer onset compared to the general population (estimated to be ~ 45 years[226] compared to ~63 years), 2) the pattern of primary cancers segregating in the family, especially CRC and endometrial cancer, 3) survival rates that differ from the norm, 4) the identification of distinguishing pathological features associated with the tumours and 5) the identification a germline mutation in the affected family members.
1.2.2 Pathology, presentation and clinical criteria

In Lynch syndrome multiple generations are affected with colorectal cancer at an earlier age with a predominance of right-sided CRC (approximately 70% are proximal to the splenic flexure). There is also an excess of synchronous CRC (defined as multiple CRCs at or within 6 months after surgical resection for CRC) and metachronous CRC (CRCs occurring more than 6 months after surgery) [226]. Additionally, there are several distinguishing pathological features that set apart Lynch syndrome tumours [127, 227-229]. Compared to sporadic CRC, Lynch syndrome tumours are more often poorly differentiated, with an excess of mucoid or signet ring features. They are also characterized by a Crohn’s-like reaction (lymphoid nodules, including germinal centers, located at the periphery of the infiltrating tumour) as well as the presence of infiltrating lymphocytes within the tumours. Furthermore, Lynch tumours undergo accelerated carcinogenesis. For example, a small adenoma may emerge as a carcinoma is 2-3 years compared to the 8-10 years this process takes in sporadic cancers [227].

In the clinical setting, diagnosis of Lynch syndrome is complicated by the lack of precise characteristic features. Clinical guidelines based on family history and age of onset of CRC, known as the Amsterdam criteria-1, were initially developed and used to identify putative Lynch families [230] (Table 1.4). These criteria were later modified to include extra-colonic cancers that were frequently seen in these Lynch kindreds (Amsterdam criteria II) [231]. The realization that Amsterdam criteria did not capture all individuals with Lynch syndrome led to the development of Bethesda criteria. These criteria help determine if individuals from families that may not meet Amsterdam criteria should undergo genetic testing for Lynch syndrome [232, 233]. Although widely used to aid in diagnosis, these criteria are limited in their ability to accurately identify patients with Lynch syndrome. Thus, the diagnosis relies primarily on the
detection of germline defects in the MMR genes of these patients [234]. It has been estimated that 40-80% of Lynch syndrome families that meet Amsterdam I criteria and 5-50% of families meet Amsterdam II criteria carry mutations in the MMR genes [235, 236]. Hitherto undiscovered genes, shared environmental influences or chance can account for families that do not show MMR mutations, as well as families that show clustering of Lynch syndrome associated cancers but do not meet these formal criteria [237]. Furthermore, the subset of families that meet Amsterdam criteria-1 but do not show MMR deficiency have been recently classified as a distinct syndrome known as familial colorectal cancer type X [238]. The risk of CRC is significantly decreased for these individuals compared to those in families with Lynch syndrome.
### Table 1.4 Summary of clinical criteria used to identify Lynch syndrome

<table>
<thead>
<tr>
<th>Criteria I [230]</th>
<th>Amsterdam Criteria I [230]</th>
</tr>
</thead>
</table>
| -To help identify Lynch syndrome families | 1. At least 3 relatives should have histologically verified CRC  
2. One should be a first-degree relative of the other 2 relatives  
3. At least 2 successive generations should be affected  
4. At least 1 relative should be diagnosed before age 50 years  
5. Familial adenomatous polyposis should be excluded |

| Amsterdam Criteria II [231] |
|-----------------------------|---------------------------------------------------------------|
| - To help identify Lynch syndrome families that may have been excluded due to the stringency of the above criteria | 1. At least 3 relatives should have cancers associated with the Lynch syndrome  
2. One should be a first-degree relative of the other 2 relatives  
3. At least 2 successive generations should be affected  
4. At least 1 relative should be diagnosed before age 50 years  
5. Familial adenomatous polyposis should be excluded  
6. Tumours should be verified whenever possible |

| Revised Bethesda Criteria [239] |
|-------------------------------|------------------------------------------------------------------|
| - To determine if tumours should be tested for the MSI Phenotype | 1. CRC diagnosed in a patient who is younger than 50 years  
2. Presence of CRCs that are synchronous or metachronous or other tumours associated with Lynch syndrome (regardless of age)  
3. CRC with MSI-H or histology diagnosed in a patient who is younger than 60 year  
4. CRC or tumour associated with Lynch syndrome diagnosed before age 50 years in at least 1 first-degree relative  
5. CRC or tumour associated with Lynch syndrome diagnosed at any age in 2 first- or second-degree relatives |
1.2.3 Tumour spectrum

In Lynch Syndrome families an individual’s lifetime risk of developing colorectal cancer (CRC) increases to about 80% [233]. In addition to cancer in the colon, Lynch syndrome is characterized by extra-colonic sites including the endometrium, stomach, urinary tract, ovaries, small bowel, hepatobiliary tract, pancreas and brain, with endometrial and ovarian cancer being the second and third most common cancers in these families [240, 241]. Individuals with Lynch syndrome may also present with sebaceous adenomas/carcinomas and keratoacan-thomas, a finding that is consonant with the Muir-Torre variant (see section 1.3.5)[226].

1.2.3.1 Endometrial cancer

The cumulative lifetime risk of developing endometrial cancer has been estimated to be about 40-60% for women in Lynch syndrome families, which equals or exceeds their risk of CRC [242]. This risk is higher in those with germline mutations in the MSH6 gene, and was estimated to be about 70% by the age of 70 [243]. The average age for the development of endometrial cancer in Lynch syndrome is around age 50 (mean age of diagnosis is over 60 years in the general population) [244]. Over 75% of women with Lynch syndrome who develop endometrial cancer present with stage 1 disease, similar to sporadic disease [244].

MMR deficiency has also been shown to give rise to sporadic endometrial cancer and about 28% of all endometrial tumours display the MSI phenotype primarily due to MLH1 promoter hypermethylation [245, 246]. Endometrial cancer is the fourth most common cancer in women in Canada, with an estimated incidence of 4500 cases in 2010 [20]. For those at average risk the lifetime probability of developing endometrial cancer is about 2.4% and the estimated five-year relative survival ratio from 2002-2004 was about 85% [21]. Although MSI-high CRCs are
generally associated with a more favourable prognosis, the clinicopathologic impact of MSI in endometrial cancers is not clear. A recent study found that the MSI-H phenotype is associated with a worse prognosis in women with early stage (I/II) disease [247]. However, more studies with larger patient cohorts are required to confirm this finding.

1.2.4 Predictive genetic testing

Advances in the understanding of the genetic basis of Lynch syndrome has allowed the clinical community to initiate and utilize predictive genetic testing programs to accurately assess disease risk for patients presenting with a family history of colorectal cancer. These programs help identify individuals with an inherited MMR mutation who are at an increased risk of developing colorectal and other Lynch syndrome associated cancers. By predicting risk prior to the progression of disease, these programs play an integral role in the effort to improve surveillance, treatment and prevention of cancer. One of the most important steps that will help with the diagnosis of a hereditary cancer syndrome, including Lynch syndrome is the compilation of a thorough family history of cancer [225, 248]. If the presence of such a syndrome is identified then molecular genetic testing will provide verification of the diagnosis [249]. The diagnosis of Lynch syndrome relies primarily on the identification of germline defects in the MMR genes [234]. The Ontario Predictive Genetic Testing Program screens individuals presenting with either a family history of colorectal cancer, and/or those who develop tumours at a young age of onset (<50). Genetic screening is initiated by performing MSI testing and immunohistochemistry (IHC) for the MMR proteins on tumour specimens when available. This is followed by germline mutation testing on lymphocyte DNA. Initially, multiplex ligation-dependent probe amplification (MLPA) is performed to detect large genomic deletions and duplications in the MMR genes. If such changes are not detected, germline DNA is screened by sequencing the *MLH1* and *MSH2*
genes, guided by MSI/IHC results of patient tumours. Once diagnosis of Lynch syndrome is established the proband’s high-risk relatives will be notified and genetic counseling and appropriate genetic testing will be performed on consenting individuals. Thereafter, surveillance measures are put in place, in attempt to reduce morbidity and mortality of the associated cancers[226]. For CRC the screening measures recommended are full colonoscopy every 1-3 years initiated between the ages of 20 and 25 for those with strong clinical evidence/ documented germline mutations [225, 250]. Furthermore, it is recommended that females in Lynch syndrome families be screened for endometrial and ovarian cancer, given the observance of accelerated carcinogenesis in such cancers. For endometrial cancer, annual transvaginal ultrasonography and endometrial aspiration for pathological assessment should begin at age 30. Screening measures for ovarian cancer should include transvaginal ovarian ultrasonography and CA 125 screening beginning at age 30. However, the screening methods available for the ovary are low in sensitivity and specificity [225]. Screening for CRC in Lynch syndrome families has been show to halve the risk of CRC and reduce the overall mortality associated with this disease by about 65% in these families [251].

1.3 **Mismatch Repair**

Mismatch repair (MMR) plays a central role in maintaining the integrity of genetic material and the basic features of this system have been highly conserved throughout evolution. The best understood function of the MMR system is the post-replicative repair of DNA polymerase errors.
1.3.1 Overview of the mismatch repair system in prokaryotes

The mismatch repair proteins were first identified in prokaryotic organisms and the *Escherichia coli* (*E.coli*) MMR system has been completely reconstituted in vitro. MutS is an ATPase that recognizes and binds (as a homodimer) to base mismatches and insertion/deletion loops [252]. The C-terminal end of MutS contains a P-loop motif for nucleoside triphosphate binding, and ATP binding/hydrolysis promotes disassociation of the MutS complex. In *E.coli* newly synthesized DNA is transiently unmethylated and this provides the signal that facilitates strand discrimination. MutH, a methylation sensitive endonuclease cleaves the unmethylated strand of a hemimethylated GATC *dam* methylation site, thereby targeting the nicked strand for exonucleolytic removal and repair. MutH can nick DNA on either side of the mismatch and depending on the side on which cleavage occurs exonucleases with either 5’→3’ polarity (RecJ or ExoVII) or 3’→5’ polarity (ExoI or ExoX) polarity excise the nicked strand [253, 254]. These nucleases are single strand specific, therefore UvrD helicase acts to unwind the duplex molecule in a directional manner beginning at the nick towards the mismatch [255]. The strand is then resynthesised by DNA Polymerase III, following which the nick is sealed by DNA ligase.

MutL couples mismatch recognition by MutS to the downstream processing steps; however, its precise role in this process has been difficult to decipher. The N-terminus of MutL is composed of ATP hydrolysis/binding domains and is highly conserved [256]. The C-terminal region is thought to contain the domain important for MutL dimerization. MutL has been shown to directly interact with MutS by affinity purification[257]. Furthermore, the MutL directly interacts with MutH and the interaction of MutL alone has been shown to be able to stimulate the endonuclease activity of MutH[258]. MutL also is able to directly interact with and activate the activity of the UvrD helicase[259]. Thus, based on these studies it has been proposed that MutL
couples the mismatch recognition/binding activity of MutS, with the helicase activity of UvrD and the mismatch specific strand cleavage activity of MutH.

1.3.2 Overview of the mismatch repair system in eukaryotes

All eukaryotes studied to date possess multiple MutS homologs (MSH proteins) and MutL homologs (MLH proteins). Unlike their prokaryotic counterparts these proteins form heterodimers with other MSH or MLH proteins. The MMR system in yeast is very well characterized and the insights gained from these studies have been helpful in elucidating the mammalian MMR system. All yeast MMR proteins, with the exception of the mitochondrial protein MSH1 have been identified in mammals. Furthermore no homologs of MutH or UvrD have been identified in eukaryotes. The central MutL homolog, MLH1 forms heterodimers with the three remaining MutL homologs PMS1 (post-meiotic segregation protein 1), PMS2 and MLH3 (Figure 1.2). The MLH1-PMS2 complex, referred to as MutLα appears to be the major player in MMR [260]. The roles of the two other MutL complexes, MLH1-PMS1 (MutLβ) and MLH1-MLH3 (MutLγ) in post-replicative error repair remains largely inconclusive, although it believed that each could act as a substitute for MutLα if the need arose [261, 262]. MutLγ has been proposed to be primarily involved in meiotic recombination [263].

The mammalian MSH2 protein forms two heterodimeric complexes. The first complex, MSH2-MSH3 (MutSβ) preferentially recognizes insertion/ deletion loops; and the second complex, MSH2-MSH6 (MutSα) recognizes both base-base mispairs and small insertion/deletion loops [264-266]. Furthermore, the stability of both MSH3 and MSH6 depends on the presence of MSH2 [267]. Like their yeast counterparts, the mammalian MSH4 and MSH5 proteins appear to be meiosis specific [268, 269].
Figure 1. 2 Heterodimers formed by the MutL homologs
1.3.2.1 Mouse models of mismatch repair

Several mouse models of mismatch repair have been developed and this has helped us gain a deeper appreciation of their roles in mammals. *Mlh1*-null (*Mlh1*−/−) mice display a wide spectrum of tumours, including lymphomas and intestinal tumours, increased MSI, and decreased survival [270]. The *Pms2*−/− mice also showed tumour susceptibility, albeit a less severe phenotype than observed in the *Mlh1*−/− mice, but did not develop gastrointestinal tumours [271]. Therefore it was proposed that there must be proteins that serve as substitutes for PMS2. *Mlh3*−/− mice developed adenomas and carcinomas in the intestinal tract and showed weak instability in mononucleotide microsatellites [272]. The slightly higher mutation rate and tumour burden in *Mlh1*−/− compared to *Pms2*−/− mice is evidence of the minor role of MutLβ in mammalian MMR [271, 273]. *Mlh3*−/− and *Pms2*−/− double-null mice showed increased tumour susceptibility and reduced survival, similar to the results observed in *Mlh1*−/− mice[272].

The phenotypes observed in *Msh2*, *Msh3* and *Msh6*-null mice can be explained by the partial redundancy between MutSα and MutSβ heterodimers [274]. Of these the *Msh2*−/− mice displayed the most severe phenotype; they developed multiple tumours in the colorectal, gastric, and endometrial regions, along with lymphoblastic lymphomas, and survived for only 2 to 5 months [275]. This is in line with MSH2s involvement in both MutS heterodimers. *Msh6*−/− mice also developed a spectrum of tumours, mainly of the gastrointestinal tract, and lymphomas but did not show any signs of MSI [276]. The *Msh3*−/− mice did not show a predisposition to CRC and developed tumours at a late age [277]. However, at 2 years the tumour incidence of *Msh3*−/− mice was comparable to *Msh3*+/+ mice. The milder phenotype associated with the *Msh3*−/− is likely due to the ability of MutSα to recognize and bind to most types of mismatches and a subset of insertion deletion loops. Mice with double homozygous deletions (*Msh3*−/−/*Msh6*−/−) had an
increased predisposition to intestinal tumours, decreased survival, and loss of MMR activity [278]. Deletions of the other MutS homologs *Msh4* and *Msh5* failed to show any predisposition to cancers or decreased survival.

1.3.3 The post-replicative error repair function

The post-replicative error repair function is capable of eliminating base-base mismatches and insertion/deletion loops that arise during DNA synthesis and have escaped the proofreading activity of the DNA polymerases. The MMR replication error repair improves the fidelity of DNA biosynthesis 100-1000 fold and reduces the error rate to one error per $10^{10}$ bases[279]. In addition to the MutS and MutL homologs this function requires the involvement of several other factors, most of which are involved in DNA replication. The first such factor, Exo1 is the only enzyme with 5'→3' directionality that have been implicated in MMR [274]. This was an intriguing finding as the strand break can be either 5’ or 3’ of the mispair, thus MMR was thought to require exonucleases with both 3’→5’ and 5’→3’ polarity. However, this disparity was resolved by the finding that the PMS2 subunit of MutLα possesses a latent endonuclease activity that can introduce single-strand DNA breaks in the discontinuous strand, thereby generating an entry point for Exo1[280]. Exo1−/− mice are prone to lymphomas but do not show gastrointestinal tumours [281]. Furthermore, the embryonic stem cells of these mice also show a weak mutator phenotype, confirming the role of Exo1 in MMR.

Another key protein implicated in MMR is proliferating cell nuclear antigen (PCNA). PCNA is loaded on to primer-template junctions by the replication factor C (RFC). Once loaded the PCNA sliding clamp functions as the processivity factor of the DNA polymerases. PCNA was also shown to directly interact with MSH3, MSH6, MLH1 and EXO1 [282, 283]. The interaction with MutSα and MutSβ via MSH6 and MSH3 indicated PCNA’s involvements in steps
preceding DNA synthesis and the presence of the MutS complexes in the DNA replication complex. Furthermore, binding of MutSα to mismatches leads to its dissociation from PCNA [284]. Therefore, it is likely that MutSα is targeted to the replication fork by binding to PCNA, then transferred from PCNA to the DNA upon encountering a mismatch [284].

The initial steps in mammalian MMR are very similar to the repair process in prokaryotes. MMR begins with the recognition of the mismatch or the insertion-deletion loop by MutSα or MutSβ (Figure 1.3). This is bound by MutLα which functions to recruit other proteins involved in this process. When this complex encounters PCNA, the binding of PCNA/RFC leads to the activation of a MutSα/PCNA/RFC/ATP-dependent endonuclease activity of MutLα [280]. This generates nicks in the discontinuous strand which function as entry points for EXO1. The rate of DNA resection by the MutSα-EXO1 complex is controlled by the replication protein A (RPA), which binds and protects the single-stranded DNA. Multiple cycles of mismatch-dependent reloading of MutSα and EXO1 and 5′→3′ degradation is thought to occur until the segment containing the mismatch is degraded [285]. It is presently believed that MutLα participates in the termination of this excision step. Following nucleolytic excision of the mismatch, repair synthesis is accomplished by DNA Polymerase β together with PCNA[286]. The final step of the mismatch repair process involves sealing of the nick by DNA ligase I. The complete MMR process was reconstituted from purified recombinant constituents by supplementing the MutSα, MutLα, RPA, EXO1, PCNA, RFC (HMGB1) system with DNA polymerase δ and DNA ligase I [285, 287]. In both these cases the reconstituted system catalyzed the repair of a heteroduplex substrate that contained a G/T mismatch and a 5′ strand break. In addition, in one of the systems replacing MutSα with MutSβ allowed the system to efficiently repair an IDL of three nucleotides [285].
Figure 1.3 Simplified overview of the eukaryotic MMR system.

Nevertheless there still some uncertainties regarding the model of eukaryotic MMR. For instance, while the transient hemimethylation of the nascent DNA strand facilitates strand discrimination in prokaryotes, it is not clear how eukaryotes ensure that DNA repair is directed towards the newly synthesized DNA strand. However, it was postulated that replication-associated strand discontinuities may serve this purpose \textit{in vivo} [288]. The molecular mechanism of mismatch recognition is also currently unclear. Based on the conservation of DNA binding motifs it appears that only MSH6 is involved in the initial mismatch recognition[289]. Thus, how MutSβ recognizes IDLs is currently unknown. Furthermore, findings indicate that mismatch binding in the presence of ATP and magnesium brings about a rapid conformational change in MutSα, which allows it to release the mismatch and move along the DNA in the form of a sliding clamp [274]. Three models of how mismatch recognition is coupled with downstream repair processes have been proposed, the molecular-switch model, active-translocation model and the DNA bending/verification model. The molecular-switch model proposes that multiple MutSα (possibly in complex with MutLα) sliding clamps diffuse bidirectionally from the mismatch [290]. When a clamp encounters a strand break tagged with other MMR factors such as PCNA it will trigger the repair process. The active-translocation model postulated that MutSα (possibly in complex with MutLα) will translocate along the DNA in a controlled manner that is dependent on ATP hydrolysis once it is released from the mismatch [291]. The DNA bending/verification model proposes that MutS complexes remain in the vicinity of the mismatch, and that communication between the mismatch and the strand-discrimination signal involves DNA bending rather than protein movement along the DNA[292]. This last model is
based on studies in E. coli and has thus far not been demonstrated in a mammalian system. Thus, current evidence favours the first two models [274].

1.3.4 The role in signaling DNA damage

MMR proteins have also been implicated in signaling DNA damage and apoptosis [293]. MMR proteins can act on DNA damage caused by oxidation, heterocyclic amines, cisplatin, UV radiation, ionizing radiation, methylating agents and alkylating agents i.e. *N*-methyl-*N*' nitro-*N*-nitrosoguanidine (MNNG), *N*-methyl-*N*-nitrosourea (MNU) and its analogues used in chemotherapy. Recognition and perhaps processing of these lesions by MMR proteins leads to activation of damage signaling pathways, which can in turn result in cell-cycle arrest, and even apoptosis. Although the role of the individual MMR proteins in post-replicative DNA repair is relatively well defined, their contributions to these other functions are not as well understood. Based on previous experiments two models have emerged to explain the role of MMR proteins in signaling DNA damage [294, 295]. In the first model, it is proposed that MMR proficient cells’ repeated futile attempts to remove mismatched bases alongside the damaged base leads to DNA strand breaks. Cellular sensors of DNA damage recognize these breaks, and this consequently leads to apoptosis. However, current experimental evidence has not been able to directly substantiate this hypothesis. The second model, which postulates that MMR proteins may be able to directly sense damage and trigger cell cycle arrest and apoptosis, without the need for further processing was put forth based on later experimental evidence.

MMR-deficient cells are around 100 times more resistant to killing by alkylating agents and around 2-4 fold more resistant to killing by cisplatin [295]. The model alkylating agent MNNG and its analogue in clinical use, temozolomide, are able to at lower doses methylate the O$^6$
position of guanine, to form O\textsuperscript{6} methylguanine (O\textsuperscript{6} MeG). In the absence of functional MMR the O\textsuperscript{6} MeG lesions persist and DNA damage accumulates but does not trigger cell death. This causes cells to be tolerant to the cytotoxic effects of methylating/alkylating agents [295].

Previous studies have illustrated that MMR proteins may be able to cause G2/M arrest [296] or cell death[297, 298] in response to alkylating agents. This G2 checkpoint, and sensitivity to these agents can be restored by the transfer of human chromosome 3 (normal location of MLH1) or chromosome 2 (MHS2), into cancer cell lines that lack functional MLH1 or MSH2 respectively [299, 300]. Treatment of cells with MNNG has been shown to phosphorylate p53 at Serine 15 and Serine 392 in a MMR protein dependent manner. These residues of p53 are preferred substrates of the ATR and ATM kinases, which play critical roles in parallel DNA damage pathways [301, 302]. Furthermore, exposure to alkylating/methylating agents have been reported to involve MutS\textalpha{} and MutL\textalpha{} dependent phosphorylation of CHK1 and CHK2, further implicating their upstream kinases ATR and ATM in these processes [303-306]. More recent evidence indicates that DNA damage results in the stabilization and accumulation of the MutL proteins, MLH1, PMS2 and PMS1 in an ATM dependent manner [307]. This study was able to establish that MLH1 is required to augment this ATM dependent phosphorylation of p53, potentially indicating the role of MMR proteins as regulators of the amplitude of the p53 response. This is in contrast to other studies that suggest that while p53 is stabilized in response to MNNG, MMR dependent apoptosis can occur independent of p53 expression [298]. In addition to the above pathways, in MMR proficient glioma and colon cancer cells, temozolomide was shown to activate the mitogen-induced p38\textalpha{} kinase and the G2 checkpoint, while no activation was observed when the glioma cells were treated with siRNA against MLH1[308]. Although the exact mechanisms by which MMR proteins are linked to the signaling pathways is unknown, collective evidence supports the notion that MMR protein likely play an important role
in signaling cell cycle arrest and apoptosis (a simplified model of MMR dependent damage signaling network is depicted in Figure 1.4).

Previous studies have suggested that the decreased levels of MMR proteins might promote tumourigenesis. For instance Msh2+/− mice are cancer prone and present with tumours that still retain the wild-type copy of Msh2 [309]. Furthermore, the embryonic stem (ES) cells from these mice did not demonstrate the MSI phenotype [310], implicating functions other than the post-replicative error repair. Subsequent studies performed using an isogenic cell line (the human embryonic cell line HEK293T) with inducible MLH1 has shown that while post-replicative error repair can occur even when as little as 10% of MLH1 is expressed, DNA damage signaling in response to MNNG requires full expression of MLH1[311].

Temozolomide is an antitumour triazene compound that under physiologic conditions is rapidly metabolized to form a methylating species that reacts with DNA to form methyl adducts. The O6 MeG lesions is thought to be the key lesion by which temozolomide exerts its toxic effects [312]. However, this adduct is subject to a single-step repair reaction that simply transfers the methyl group to a cysteine residue within the repair protein O6-methylguanine-DNA methyl transferase (MGMT), thus restoring the DNA to its original state. Thus, MGMT is a major determinant in temozolomide toxicity [312], but the MMR proteins have also been shown to be critical in this process [313, 314].
Figure 1.4 A simplified model of the MMR protein dependent DNA damage signaling network

MMR proteins recognize and bind to damaged DNA, and then recruit other DNA damage signaling kinases, including ATM, ATR and their effectors checkpoint kinases CHK2/CHK1. These kinases are able to bind to and stabilize p53. The p38 mitogen-activated protein (MAP) kinase interacts with MMR proteins and connects them with p53 and its homolog p73 another transducer of MMR dependent damage response. MMR protein, MLH1 can activate the tyrosine kinase c-Abl, which is then able to phosphorylate or stabilize p73. PMS2 is able to directly interact with p73 and stabilize it. All these pathways are able to lead to cell cycle arrest and/or apoptosis [294].
1.3.5 Other roles of mismatch repair proteins

MMR proteins have been implicated in a number of additional processes involving recombination, meiosis and somatic hypermutation. However, these processes are not as well studied as the post-replicative error repair function and many of the other proteins that cooperate with MMR proteins in these functions remain to be elucidated. Several of these processes will be briefly addressed below.

1.3.5.1 Roles of MMR in recombination

Mismatch correction in recombination intermediates

In eukaryotes, the repair of mismatches in recombination intermediates results in the phenomenon of gene conversion[315], a process by which information of one chromosome is replaced with information from the homologous chromosome. In the absence of functional MMR, mismatches in recombination intermediates are not repaired. Thus, the mismatch containing strands segregate during the first mitotic division after meiosis, a process referred to as post-meiotic segregation. In yeast with mutant *msh2*, *mlh1* or *pms1*, a decrease in gene conversion events is observed, accompanied by a concomitant increase in post-meiotic segregation[316-318]. Studies to date indicate that the same basic MMR machinery that detects replication errors is also responsible for detecting mismatches in recombination intermediates. However, it is likely that there are additional proteins that are unique to each process [253].

Regulation of recombination between nonidentical sequences

MMR has also been shown to inhibit the homologous recombination of non-identical strands. In *Saccharomyces cerevisiae* the MMR machinery can inhibit recombination between diverged sequences both in meiosis and mitosis. Furthermore, *msh2* and *mlh1* mutants appear to play
nonequivalent roles is this process, with \textit{msh2} mutants demonstrating higher rates of recombination between diverged sequences compared to \textit{mlh1} or \textit{pms1} mutants\[319, 320\]. This inhibition of crossover of diverged sequences has been proposed to be important in establishing and maintaining genetic barriers between species\[321\]. Furthermore, this activity has been proposed to promote genome stability by inhibiting interactions between diverged repetitive DNA sequences during mitosis and meiosis \[253\]. By preventing these mismatched strands from recombining, it increases the fidelity of these strand exchanges during recombination by increasing the likelihood of interaction of two homologous fragments \[274\].

1.3.5.2 Roles of MMR proteins in meiosis

Yeast MSH4 and MSH5 were the first identified examples of eukaryotic MutS homologs that have functionally diverged to carry out novel meiosis specific roles and in the process lost their ability to repair mismatches \[322\]. These proteins have been proposed to directly bind Holliday junctions and promote meiotic crossing over. Furthermore, MSH4-MSH5 proteins have been show to interact with the MLH1-MLH3 heterodimer. Consistent with its role in meiosis, MSH4 and MSH5 expression is highest in the ovaries and testis of mammals \[269, 323\]. Msh5\textsuperscript{−/−} mice are viable, but sterile \[268\]. Interestingly, mutation of MutL homologs appears to have a more severe impact on the meiotic functions of mouse models. This is supported by the fact that both male and female Mlh1\textsuperscript{−/−} mice are sterile \[324, 325\]. Examination of the testicular tissue from these mice has indicated that the spermatocytes are arrested in meiosis I. Additionally, male Pms2\textsuperscript{−/−} mice are sterile, while the corresponding females are fertile \[326\]. However, the cause behind this sex-specific difference is not clear.
1.3.5.3 Somatic hypermutation of immunoglobulins

Somatic hypermutation of the immunoglobulin (Ig) variable genes is a process that generates high-affinity antibodies in response to antigen stimulation. It is a process that allows a small amount of genetic material to be used to generate diverse antibodies, sufficient to deal with the multitude of antigens [327]. The germline encodes IgM antibodies that are composed of the variable (V), diversity (D), and joining (J) elements. However, these antibodies are not able to mount an effective response against pathogens. Once an antigen appears and stimulates B cells, they proliferate, differentiate and migrate to the germinal centers in secondary lymphoid organs and become centroblasts. In these germinal centers, centroblast B cells express large amounts of activation-induced deaminase (AID), which initiates the somatic hypermutation of the antibody V regions (corresponds to the antigen binding sites)[328, 329]. This is thought to be the first phase of this process, whereas the second step refers to the error prone repair of AID-induced mutations by BER and MMR [330]. It has been proposed that in germinal centers MMR enzymes, which normally act to maintain genomic stability, become error prone and greatly increase the number of mutations in the V region. The role of MMR in this process is debated, with some studies showing that a functional MMR system is required for full levels of hypermutation [331, 332], while other studies have demonstrated that it is not the frequency but the spectrum of hypermutation that is altered in MMR deficient mice [333].

1.3.6 The structure and expression of the MMR protein MLH1

The MutL proteins, including MLH1 are ATPases of the GHKL (gyrase/Hsp90/histidine-kinase/MutL) family [334] with the ATPase situated in the N-terminus and the MutL dimerization domain located in the C-terminal domain. MLH1 is located on chromosome 3p21 and is composed of 756 amino acids that are distributed over 19 exons. Additionally, MLH1
transcript and protein levels remain constitutively expressed throughout cell cycle progression [335].

Previous studies in mice have shown that both MLH1 and PMS2 proteins have both nuclear localization sequences (NLS) and nuclear export sequences. However, it was found that dimerization of MLH1 with PMS2 in the cytoplasm is required to unmask the NLS, which then facilitates the nuclear import of the MutLa heterodimer [336]. The same mechanism may be true in the human system, where studies have recently identified two putative NLS in MLH1 and one candidate NLS in PMS2 [337].

Moreover, MLH1 is also a component of the BRCA1-associated genome surveillance complex (BASC), which consists of proteins such as BRCA1, MSH2, MSH6, ATM, BLM and the RAD50-MRE11-NBS1 protein complex [338]. All these proteins have a role in recognizing abnormal DNA structures or damaged DNA, thus BASC may serve as a sensor for DNA damage [338].

The *MLH1* promoter region is arbitrarily divided into four regions (A- D) with respect to its translation start site; A (-711 to -577), B (-552 to -266), C (-248 to -178) and D (-109 to +15) [339]. While hypermethylation of the entire [A+B+C+D] region is associated with gene silencing and loss of MLH1 expression, this correlation is augmented when the proximal region of the *MLH1* promoter is hypermethylated [340]. Methylation of the C+D region of *MLH1* is invariably linked with gene silencing [339, 341], suggesting this to be the core promoter region of this gene. Recent evidence indicates that the promoter of *MLH1* is bi-directional with a second gene, *EPM2AIP1* located in a head-to-head orientation on the opposite strand 321 base pairs away from the *MLH1* transcription start site [342]. The ubiquitously expressed *EPM2AIP1* gene is comprised of a single large exon, 1824 nucleotides in length encoding a 607 amino acid protein
with no known homology to other proteins or structural motifs [343]. The function of EPM2AIP1 is currently not known.

1.3.7 Mismatch repair and disease

Mutations in MLH1 [344, 345] and MSH2 [346, 347] account for the more than 80% of known cases of Lynch Syndrome. Other MMR genes implicated to date include MSH6, PMS2, and possibly MLH3. MMR deficiency has also been shown to give rise to ~15% of sporadic colorectal, endometrial and gastric cancers. The majority of such sporadic CRCs associated with microsatellite instability are primarily due to epigenetic inactivation of MLH1 due to promoter hypermethylation [121, 122].

Germline MMR defects also lead to Muir-Torre syndrome (MIM #158320). Muir-Torre is characterized by tumours in the sebaceous glands or keratoacanthomas that are associated with one or more of the various visceral neoplasms, mainly colorectal, endometrial, urological, and upper gastrointestinal neoplasms. It has been proposed that Muir-Torre is a clinical variant of the Lynch Syndrome [348]. However, about 35% of tumors in individuals positive for the Muir-Torre Syndrome do not display instability of microsatellite repeats, which combined with other features of Muir-Torre has spurred the speculation that at least two distinct forms of Muir-Torre exist[349]. The first sub-type, which includes MSI positive tumors, shares genetic and pathological features with Lynch syndrome. It is characterized by early-onset colorectal carcinoma and a strong family history of cancer. The second group of MSS or MMR proficient tumors shows later ages of onset and a less pronounced family history, and is likely due to genes that are not implicated in MMR pathway. Muir-Torre syndrome is associated with germline mutations in the MSH2 gene, and to a lesser extent with mutations in MLH1 and MSH6 genes [350-353]. We recently reported a family where a complex MSH2 mutation predisposed to both
Lynch and Muir-Torre syndromes [354]. Interestingly, a subset of tumours in this family displayed neither prototypical feature of microsatellite instability or immunohistochemistry, highlighting the variable phenotypic manifestations associated with genetic alterations of the MMR genes.

Furthermore, there are also reports of individuals who are homozygous for mutations in the mismatch repair genes *MLH1, MSH2, MSH6* and *PMS2*. Such individuals develop several congenital abnormalities including haematopoietic malignancies, pediatric brain cancers, childhood leukemia, HNPCC-related cancers and multiple café-au-lait spots, a common characteristic of neurofibromatosis type 1 (MIM#162200) [355-360]. This phenotype manifests in an autosomal recessive fashion, because a mutant allele is inherited from each parent. Interestingly, the presence of early-onset gastrointestinal cancers along with the features of neurofibromatosis type 1 (NF1) was first reported in children who were positive for a homozygous missense variant in exon 18 of *MLH1* (Arg687Trp) [356]. These children had inherited each variant allele from their parents, who were heterozygous carriers of this variant.

In addition, mutations in both *APC* as well as the MMR genes, *MLH1* and *PMS2* are associated with Turcot’s syndrome (MIM#276300) [361]. There have also been reports of individuals carrying compound heterozygous PMS2 mutations who develop Turcot’s syndrome [362]. This syndrome is defined by the presence of brain tumours and multiple adenomas/colorectal cancers that occur at an early age. The central nervous system tumours in individuals with *APC* mutations are typically medulloblastoma, whereas those with MMR gene mutations are usually glioblastoma multiforme.
1.4 Genetic Variation in Mismatch Repair

1.4.1 Predictive models of germline variants in MMR genes

The ability to predict germline MMR mutations would be useful in situations where the existence of an inherited defect cannot be directly addressed in the laboratory. Furthermore, universal germline testing of all index CRC cases is not feasible given that family histories of CRC is common, and that mutations that lead to this syndrome are both rare and scattered across several genes[363]. Several groups have attempted to translate the extensive understanding of Lynch syndrome into clinically useful prediction algorithms, similar to those developed for breast[364], melanoma[365] and pancreatic syndromes[366].

The Leiden (Wijnen) model was developed to address this void in Lynch syndrome [367]. This model estimates the probability of carrying a mutation in an MMR gene based on mean age of CRC diagnoses, the presence of endometrial cancers in family members and the fulfillment of Amsterdam criteria. Despite its utility, the Leiden model does not use all available information and as well does not provide risk predictions.

Currently there are several computer-based models; MMRpro [368], prediction of mutations in \(MLH1\) and \(MSH2\) (PREMM\(_{1,2}\))[369] and MMR predict[370], which estimate the probability that an individual carries a mutation in a MMR gene. MMRpro is a model that estimates the probability that an individual carries a deleterious mutation in an MMR genes, \(MLH1\), \(MSH2\), and \(MSH6\). Data for MMRpro was obtained by meta-analysis of population-based studies. It also incorporates information from published cases along with data on the specificity and sensitivity of MSI status and germline testing. This probability is evaluated using a detailed family history of CRC and endometrial cancer, which includes information for the individual
being counseled and each of his or her first- and second-degree relatives. Furthermore, MMRpro estimates future cancer risks for unaffected individuals (mutation positive or negative), as well as untested individuals. PREMM\textsubscript{1,2} determines the likelihood that an individual carries a mutation in MLH1 and MSH2. This model used specific information about an individual’s age, sex, as well as specific details about the individual’s family history as supplied on the test submission forms at Myriad Genetics Inc., a genetic testing laboratory. MMRpredict calculates the probability than an individual carries a mutation in either MLH1 or MSH2. Data was obtained by determining the germline mutation status of 870 individuals followed by MSI and IHC testing. The presence of germline mutation was correlated with variables such as age, sex, family history, and presence of endometrial cancer to determine the weight of each predictor. Following this, information regarding MSI and IHC was incorporated into the equation.

In an independent validation study carried out in a cancer genetics clinic using a high-risk cohort, all three models performed similarly [371]. In an independent study of families with a moderately high risk of developing CRC, all three programs performed comparably, while the PREMM\textsubscript{1,2} model was found to demonstrate the best performance [363]. In fact in this study, these models performed better than Amsterdam II criteria leading the authors to propose that these models may soon replace the Amsterdam and Bethesda criteria. Unlike Amsterdam or Bethesda criteria, which are only able to identify the existence of risk, these programs are able to provide an estimation of risk. This provides the opportunity to set up probability thresholds for interventions that will assist with clinical decision making [363]. However, these cut-offs remain to be determined. It is expected that the ability to predict the presence of germline mutations will result in an increased rate of mutation discovery in the near future.
1.4.2 Spectrum of genetic alterations in the major MMR genes

Genetic alterations in the MMR genes range from large genomic deletions and duplications, truncating or missense variants to splice site alterations, which occur in distinct domains of \textit{MLH1} or \textit{MSH2}. A recent database catalogues over 1500 MMR gene variants (http://www.med.mun.ca/MMRvariants) [372]; while the International Society for Gastrointestinal Hereditary Tumors (InSiGHT, http://www.insight-group.org/) lists over 400 pathogenic mutations that have been identified in the MMR genes. MMR genes act as tumour suppressors, in that one wild-type allele is sufficient to maintain normal function. Individuals with Lynch syndrome have only one allele that is inactivated (the one bearing the germline mutation). Thus, for the process of carcinogenesis to proceed an independent somatic event must occur on the wild-type allele rendering the MMR gene inactive.

More recently, in a subset of Lynch syndrome cases constitutional epimutations of the \textit{MLH1} gene were identified[373]. These epimutations are characterized by soma-wide methylation of a single allele of the promoter and allelic transcriptional silencing. Germline \textit{MLH1} epimutations are functionally equivalent to an inactivating mutation and have been shown to produce a clinical phenotype that resembles Lynch syndrome. However the inheritance of epimutations was found to be weak, thus limiting the utility of family history in guiding screening. In addition MLH1 expression is altered by several epigenetic mechanisms. For example, the promoter region of \textit{MLH1} is hypermethylated in the majority of MSI-H sporadic cancers. Furthermore, another study found that miR-31 and miR-223 were overexpressed in colorectal tumours of patients with Lynch syndrome, indicating the potential role for miRs in regulating the expression of genes involved in susceptibility to Lynch syndrome [374].
A recent study offers another novel mechanism of Lynch syndrome susceptibility. Here heterozygous germline deletions of the last exons of TACSTD1 (encodes EPCAM); a gene directly upstream of MSH2 led to the somatic silencing of the corresponding MSH2 allele by hypermethylation. [375]. Findings from another study indicate that while deletions of EPCAM account for only a small proportion of individuals suspected of Lynch syndrome, these deletions were more frequent in individuals with genetically confirmed Lynch syndrome [376]. A recent report described somatic hypermethylation of MSH2 in panel of a Lynch syndrome tumours that were deficient for MSH2 expression [377]. In contrast MSH2 hypermethylation was not observed in normal colonic tissues or sporadic CRCs (both MSI/MSS). While a subset of these tumours had deletions in EPCAM, the majority were not associated with defects in EPCAM but instead were positive for germline mutations in MSH2 [377]. Although further studies are needed to confirm this finding, it suggests a unique mechanism for MSH2 inactivating the second, wild-type allele of MSH2.

About 16% of all MSH2 and 30% of all MLH1 alterations are missense variants of no obvious clinical significance [378]. Germline MSH2 mutations are associated with loss of MSH2 and/or MSH6 expression in the majority of cases. Consequently, a germline MSH2 missense variant in patients presenting with a family history of CRC and a MSH2 deficient CRC tumour is likely to be a pathogenic mutation. On the contrary the significance of the majority of missense variants identified in MLH1 is largely unknown. This issue is compounded by the fact that genetic alterations are dispersed across the gene, with no known hot spot regions associated with confirmed disease risk. Given that the diagnosis of Lynch syndrome relies primarily on the identification of germline defects in the MMR genes [234], detecting these unclassified genetic variants creates much ambiguity in the clinical setting, as the pathogenicity of these variants cannot be readily ascertained. This affects the accuracy of predictive genetic testing and
counseling of patients carrying these variants, as well as their family members at risk of developing Lynch syndrome.

1.4.3 Approaches to classifying unclassified variants

Several different strategies are employed to characterize the pathogenic and functional significance of MMR variants, each with specific strengths and limitations [234, 379-385]. Genetic linkage assessment is based on the co-segregation of a MMR variant with disease status in pedigrees, which is relatively straightforward to assess. However, it requires screening of multiple affected individuals in the same family, which may not always be feasible with smaller families or when clinical specimens are unavailable for all affected members in a family. MMR variants that occur concurrently with known deleterious mutations could potentially act synergistically with the deleterious mutation or have no effect on the disease phenotype. Thus, in this case a less pathogenic or benign nature is indicated, although there is no direct evidence for the lack of pathogenicity of such a variant. In addition, genetic epidemiologic studies can provide direct evidence of association with cancer risk. However, if these studies are to be informative a prohibitively large sample population is required, particularly for variants that occur with low frequency. Bioinformatics approaches that rely on different databases (e.g. SIFT or PolyPhen) are increasingly being used to predict the potential impact of a variant on MMR gene structure and/or function(s). The advantages of these approaches are that detailed family histories or large patient populations are not required. However, their predictive value is limited unless a panel of such programs are used [386]. Several groups have attempted to develop functional assays based on the sequence homology between human and yeast MLH1[387]. One assay defined variants that were unable to restore genetic stability in a mlh1 deleted yeast host strain as pathogenic. Another assay made use of the finding that overexpression of mlh1 in yeast led to defective
MMR. Based on the finding that wild-type MLH1 behaved in a dominant negative fashion and led to MMR defects, any variant that did not result in defective MMR in yeast was deemed to be pathogenic [384]. However, the utility of these assays depend upon the homology between yeast and human sequences, which is poor in the central region of the MLH1 protein. Guerrete et al., described a GST fusion assay where a radiolabelled variant of MLH1 was assessed for its ability to interact with PMS2 [388]. However this method cannot be used to evaluate variants that do not play a role in heterodimerization. Functional and biochemical studies carried out in human systems provide the most direct evidence because they are able to assess the effects of genetic variants on protein stability, expression and function(s). For this approach alterations are created by site-directed mutagenesis in plasmids containing the wild-type gene, which are then used to assay expression, MMR function or protein stability [383, 385, 389-391]. The ability of missense variants to alter the efficiency of the post-replicative error repair can be assessed using several mismatch repair assays. One prototypic approach relies on the ability of variant MLH1 protein to correct a mismatch that occurs in a DNA heteroduplex [392]. Briefly, the G•T mispair renders this heteroduplex resistant to cleavage by restriction enzymes; however if post-replicative error repair is intact then this mismatch is repaired to an A•T. This renders it sensitive to cleavage, making it possible to evaluate repair efficiency by measuring cleavage efficiency. Nonetheless, these assays can only determine a variant’s ability to alter the post-replicative error repair function. Therefore, at present it remains important to use a combination of the above approaches to accurately assess the pathogenicity associated with a variant. However, determining the most appropriate assays requires a better understanding of the regulation and function of MLH1, as well as the role of missense variants in affecting these processes. Furthermore, as our understanding of the role of MLH1 and the MMR system evolves, it is remains important to devise new assays that take this information into consideration.
1.5 Hypothesis and Objectives

I hypothesize that missense alterations in distinct domains of *MLH1* likely affect its expression and function(s) to varying degrees.

In order to characterize the possible effects of unclassified variants on MLH1 and the MMR system, the following objectives were formulated:

1. Assess the effect of unclassified variants on MLH1 expression, stability and function
2. Determine if *MLH1* genetic variants lead to unbalanced allelic expression
3. Investigate the functional effects of the MLH1 c.-93G>A variant on *MLH1/EPM2AIP1* promoter activity
CHAPTER 2

The role of unclassified variants in influencing MLH1 expression, stability and function

Sheron Perera and Bharati Bapat

Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Canada
Department of Pathology and Laboratory Medicine, Mt. Sinai Hospital, Toronto, Canada

All work in this chapter and data analysis was carried out by Sheron Perera, with the exception that Vaijayanti Pethe assisted with the generation of K618A stable inducible cell line.

Chapter 2 The role of unclassified variants in influencing MLH1 expression, stability and function

2.1 Summary
Germline mutations in the mismatch repair genes lead to the Lynch syndrome/ HNPCC, and genetic testing is offered to identify individuals at risk of developing this disease. About a third of all genetic alterations detected in the mismatch repair gene MLH1 are missense variants, the majority of which are of equivocal clinical significance. In this report, we investigate the molecular basis of the pathogenicity of four missense variants localized to distinct functional domains. Our results demonstrate that the MLH1 variants p.Arg265Cys and p.Lys618Ala significantly decrease the stability of the protein, while another missense variant p.Leu749Gln is able to compromise heterodimerization of the MLH1-PMS2 complex. Furthermore, using a stable inducible cell line model we were not able to assess the impact of these substitutions on altering MLH1 function likely due to the disruptive effects associated with MLH1 overexpression in cells.

2.2 Introduction
Lynch syndrome or hereditary nonpolyposis colorectal cancer (HNPCC), is a common cancer predisposition syndrome that is inherited in an autosomal dominant manner. Lynch syndrome is associated with germline mutations in the mismatch repair (MMR) genes. Individuals in Lynch syndrome families are at an increased risk of developing colorectal cancer (CRC), while women in these kindreds are predisposed to endometrial cancer. In the clinical setting, diagnosis of Lynch syndrome is complicated by the lack of precise characteristic features. While clinical guidelines including Amsterdam Criteria 1 and 2, and Bethesda Criteria have proven to be useful
in diagnosis, these criteria are limited in their ability to accurately identify patients with Lynch syndrome. Thus, the diagnosis of Lynch syndrome relies primarily on the detection of germline defects in the MMR genes of these patients [234].

The MMR pathway plays a central role in maintaining genomic stability by repairing both mismatches and insertion/deletion loops that arise during DNA synthesis [393]. MMR improves the fidelity of DNA biosynthesis several hundred-fold and reduces the error rate to about one error per $10^{10}$ bases [279]. Defective MMR results in the mutator phenotype, which increases the likelihood of mutations in a battery of genes that regulate growth, differentiation or apoptosis. Deficiencies in MMR lead to microsatellite instability (MSI). Tumours from the majority of Lynch syndrome patients (>85%) display MSI, making it a hallmark feature of this syndrome [123, 124].

In addition to the error repair function, MMR proteins have been implicated in signaling DNA damage and apoptosis [293]. Recognition and perhaps processing of these lesions by MMR proteins leads to activation of DNA damage signaling pathways, which can in turn result in cell-cycle arrest or apoptosis. Although the role of the individual MMR proteins in post-replicative DNA repair is relatively well defined, their contributions to signaling DNA damage is not as well understood. Emerging research indicates that unlike many of the other DNA repair pathways that mediate resistance to DNA damage, MMR is necessary to mediate the toxicity of several anti-cancer drugs [394]. Therefore, a better understanding of how the MMR proteins signal DNA damage will help us to better understand the mechanism by which MMR deficient tumours are refractory to certain classes of chemotherapeutic agents. Interestingly, previous studies with MMR mouse models have demonstrated that missense variants are capable of separating the mismatch-error repair function from the DNA-damage response function of the MMR system
This underscores the utility of missense alterations as biological tools capable of providing valuable insights into the role of MMR proteins in signaling DNA damage [380, 396].

Advances in the understanding of the genetic basis of Lynch syndrome have facilitated predictive genetic testing programs to identify individuals with an inherited MMR mutation who are at an increased risk of developing CRC and other Lynch syndrome-associated cancers. By predicting risk prior to the progression of disease, these programs play an integral role in the effort to improve surveillance, treatment and prevention of cancer. A recent database catalogs over 1500 MMR gene variants (http://www.med.mun.ca/MMRvariants [372]); while the International Society for Gastrointestinal Hereditary Tumours (InSiGHT, http://www.insight-group.org/) lists over 400 pathogenic mutations that have been identified in the MMR genes. Among these, mutations in *MLH1* and *MSH2* are responsible for over 85% of known cases of Lynch syndrome [397]. About 25% of *MLH1* alterations are missense alterations of equivocal clinical significance [378]. Detection of such an unclassified variant affects the accuracy of predictive genetic testing and counseling of patients carrying these variants, as well as their family members at risk of developing Lynch syndrome. It is likely that missense alterations of *MLH1* affect its expression and function(s) to varying degrees, thereby contributing to the variable disease phenotype observed in Lynch syndrome.

In this study we investigate the effects of four representative missense variants in *MLH1* on its expression and function, in order to clarify their pathogenic significance (Figure 2.1). We functionally characterized the *MLH1* alterations c. 55A>T leading to the substitution Ile19Phe, c.793C>T leading to Arg265Cys, both of which occur adjacent to the ATPase domains. We also studied c.1852_1853AA>GC leading to Lys618Ala, which is located in the MutL homolog binding domain. The c.55A>T variant has been reported previously however, no functional
studies have been carried out to date [398, 399]. The pathogenicity of the R265C and K618A variants remain debatable even after several analyses due to the existence of conflicting data [234, 385, 388, 400-404]. We also hypothesized that missense alterations that occur in the C-terminal region, adjacent to the MutL homolog binding domain likely affects MLH1 heterodimerization, and to investigate this we constructed an alteration c.2246T>A which leads to the missense substitution Leu749Gln and is located within the c-terminal homology domain.
Figure 2.1 Schematic representation of the functional domains identified in MLH1.

Arrows indicate the locations of the representative variants selected in this study.
2.3 Materials and Methods

2.3.1 Materials and cell culture

The human colon cancer cell line, HCT116 cell line was obtained from American Type Culture Collection (Rockville, MD) was maintained in McCoy’s 5A medium supplemented with 10% FBS. The HCT116 cell line with the additional copy of chromosome 3 was a kind gift from Drs. Minoru Koi and Richard Boland (University of Michigan Medical School, Ann Arbor, MI). The cell line was developed by transferring a single copy of human chromosome 3 from fibroblasts to HCT116 cells via microcell fusion [405]. These cells were maintained in McCoy’s medium with 10% FBS and 400μg/ml of G418 (Sigma). The vectors containing the full-length wild-type MLH1 cDNA (GenBank accession no. NM_000249.2) and PMS2 cDNA (NM_000535) was kindly provided to us by Dr. G. Plotz and Dr. R. Kolodner. Benzylguanine and temozolomide (TMZ) were purchased from Sigma Chemicals dissolved in DMSO, and stored at -80°C in single use aliquots.

2.3.2 Identification of MLH1 variants

The variants I19F, R265C and K618A were identified in patients that were screened through the Ontario Predictive Genetic Testing Program at Mt. Sinai Hospital, Toronto, Canada. Genetic screening is initiated by performing MSI testing, and immunohistochemistry (IHC) for the MMR proteins on tumour specimens of individuals presenting with either a family history of CRC, and/or those who develop tumours at a young age of onset (<50). This is followed by multiplex ligation-dependent probe amplification (MLPA), to detect large genomic deletions and duplications in the MMR genes. If such changes are not detected, germline DNA is screened by sequencing the MLH1 and MSH2 genes, guided by the MSI/IHC results of patient tumours.
Through this program we have identified many established germline mutations, known to cause aberrant expression of protein, as well as several unclassified variants.

2.3.3 Computational Assessment of Variants

We used the computational programs Polymorphism Phenotyping (PolyPhen) (http://genetics.bwh.harvard.edu/pph/), and Sorting Intolerant From Tolerant (SIFT) (http://blocks.fhcrc.org/sift/SIFT.html) to predict the pathogenicity of these variants. PolyPhen uses structural information, along with multiple sequence alignments to classify the pathogenicity of an amino acid substitution with an accuracy of about 80% [406]. Based on these criteria it predicts a substitution to be benign, possibly pathogenic or probably pathogenic respectively, in ascending order of its predicted impact. SIFT uses an algorithm that aligns closely related sequences, and calculates normalized probabilities for all possible substitutions [407]. It has been reported that SIFT is able to predict the impact of non-synonymous substitutions with an accuracy ranging from 65-92% in a range of genes associated with human disease [386]. It predicts positions with normalized probabilities of less that .05 as deleterious, while those with probabilities greater than or equal to .05 are classified as tolerated. Default settings were used for both programs.

2.3.4 Site-directed mutagenesis and construction of MLH1 mutants

We used site-directed mutagenesis to create the expression constructs according to the specifications of the manufacturer (Quick Change Site-directed Mutagenesis, Stratagene, LaJolla, CA). MLH1 cDNA coding for the variants were generated using primer pairs that contained the specific nucleotide alteration (Table 2.1). To confirm the integrity and the presence of the respective missense substitutions, all cDNA inserts were directly sequenced. We sub-
cloned these constructs into myc- His tagged expression vectors (Invitrogen, Carlsbad, CA), to use them for a wider range of experimental procedures. All mutation numbering is based on cDNA sequence, with codon 1 corresponding to the major initiation codon. Primer sequences and PCR conditions for all protocols listed above are available on request.

2.3.5 RNA extraction, cDNA synthesis and sequencing

RNA extraction was performed using TRIzol according to the manufacturer’s protocol (Invitrogen Life Technologies, Burlington, ON). Reverse transcription PCR was performed according to standard techniques as described before [408]. In brief, cDNA was generated from total RNA by use of random hexamers pd(N6) (Amersham Pharmacia, Piscataway, NJ), oligo dT (Invitrogen Life Technologies), 1 x first strand buffer, 0.5 mmol/l of deoxynucleotide triphosphate, 5 mmol/l DTT, and 200 units of SuperScript II (Invitrogen Life Technologies). cDNA was generated by incubating at 45°C for one hour and 70°C for 15 minutes. PCR was carried out to amplify the region surrounding the nonsense mutation on MLH1. Following the PCR reaction, DNA was isolated by gel electrophoresis and purified by QIAquick Gel Extraction Kit (Qiagen, Mississauga, ON, Canada). Sequencing reactions were carried out using an ABI377 sequence analyzer according to the manufacturer’s protocol (Applied Biosystems, Foster City, CA).

2.3.6 Production of recombinant proteins and western blotting

HCT116 cells were transfected using Lipofectamine 2000 reagent (Invitrogen), according to the manufacturer’s instructions with the appropriate vectors and harvested at the indicated time points. For total protein extraction, cells were washed twice with cold phosphate buffered saline (PBS). Cold radioimmunoprecipitation assay (RIPA) lysis buffer (0.5 % Nonidet P-40, 0.5%
sodium deoxycholate, 0.1% sodium dodecyl sulfate, 1x PBS) buffer containing Complete protease inhibitor tablets (Roche, Mannheim, Germany) was used to lyse the cells on ice for approximately 20 minutes. The lysates were then collected by centrifugation at 14,000 rpm for approximately 15 minutes. The expression levels for the proteins were examined by sodium dodecyl sulphate/ polyacrylamide gel electrophoresis, followed by western blot analysis with anti-MLH1 (clone 168-728; BD Biosciences) and anti-post meiotic segregation increased 2 (PMS2; MIM# 600259) (clone A16-4; BD Biosciences) antibodies. β-Actin antibody (Sigma) was used to assess equal loading in all lanes. To validate our results, we co-transfected cells with 0.15μg of pCMV β-galactosidase vector, and measured its enzymatic activity as described previously [409].

2.3.7 Combined coimmunoprecipitation and western blot analysis

For coimmunoprecipitation 1000 μg of whole cell lysate transfected with each of the constructs were incubated with 2.5 μg of anti-MLH1 antibody. The lysates were then incubated for a further 1.5 hours after the addition of 40 μl of protein A/G agarose beads (Santa Cruz Biotechnology, Santa Cruz, CA). The precipitates were then centrifuged at 1000 rpm for 2 minutes, and washed 3 times with 500μl of RIPA buffer and resolved on a 10% SDS-PAGE gel. Interactions between MLH1 and PMS2 proteins were detected by western blot analysis with anti-MLH1 and anti-PMS2 antibodies.

2.3.8 Pulse chase assays

HCT116 cells were plated in 6 cm dishes, and then transiently transfected with wild-type and variant MLH1 cDNA. 24 hours post transfection, the media was aspirated from the plates and the cells were washed twice in warm PBS. 1 ml Methionine (Met) -, and Cysteine (Cys)-free
DMEM (Sigma - supplemented with Glutamine) was added to the cells, and incubated for 30 minutes at 37°C to starve the cells of endogenous pools of Met and Cys. Thereafter 150μCi of EasyTag Express $^{35}$S (Perkin-Elmer, Waltham, MA) was added to each plate, and the cells were incubated for an additional 40 minutes at 37°C. The radioactive media was then removed and the cells were washed twice with warm PBS. For the chase, 5 ml of warm DMEM supplemented with 2mM Methionine and 2mM Cysteine was added to all plates except the ‘0’ time point. The plates were then incubated at 37°C and harvested at the specified time points from 0-30h. The harvested lysates were pre-cleared for one hour with protein A/G agarose beads. Thereafter, first total MLH1, and then total β-Actin was immunoprecipitated, and analyzed by SDS-PAGE followed by autoradiography.

2.3.9 Cycloheximide experiments

HCT116 cells were transfected with either wild-type or variant MLH1 constructs. 24 hours post-transfection, cells were treated with 25μg/ml of cycloheximide (Sigma) or vehicle (DMSO), and lysed at indicated time points as described before. The levels of MLH1 were assessed by western blotting.

2.3.10 Generation of stable inducible cell lines

A tetracycline-regulated expression system (T-rex) was used to generate HCT116 cell lines with inducible expression of wild-type or variant MLH1 according to the manufacturer’s instructions. The T-Rex system is a tetracycline regulated mammalian expression system that uses elements from the *E.coli* TN10-ecoded tetracycline operon. In the absence of tetracycline or doxycycline (a tetracycline derivative with a longer half life), the Tet repressor (TetR) binds the Tet operator.
sequences in the inducible expression vector, repressing the transcription of wt or variant MLH1. When added into the medium tetracycline (or doxycycline) binds the Tet Repressor, resulting in its release from the Tet operator sequences, which results in the induction of wt or variant MLH1 expression.

To generate an HCT116 cell line with stable expression of TetR, HCT116 cells were seeded in 6 cm plates. 24 hours later cells were transfected with 10µg of pcDNA6/TR per plate using Lipofectamine 2000 (Invitrogen Life technologies) transfection reagent according the manufacturer’s instructions. 48 hours post-transfection cells were re-seeded into 15 cm plates at low density and grown in the presence of 5µg/ml of blasticidin (Invivogen, San Diego, CA) to select for transfectants. Colonies that were established were carefully isolated, expanded and analyzed by western blotting using the TetR antibody (Invitrogen life technologies).

To generate the doxycycline inducible 2xmyc-tagged MLH1/pcDNA4/TO construct, the 2xmyc-MLH1/pcDNA3.1 construct was digested with ApaI and EcoRV. The fragment encoding the MLH1 was subcloned into the pcDNA4/TO inducible expression vector, which was also digested with ApaI and EcoRV. The constructs for I19F, K618A were also generated in a similar manner.

To generate stable inducible clones, the HCT116-TetR cells were transfected with 2xmyc-MLH1/pcDNA4/TO as described above and transfectants were selected for in the presence of both 5µg/ml blasticidin and 100µg/ml zeocin (Invivogen). Zeocin resistant colonies were selected and analyzed by western blotting using the myc-tag antibody (9e10, Santa Cruz Biotechnology). A similar protocol was followed to generate HCT116 cell lines that stably expressed the variant MLH1 proteins. To screen resistant colonies, cells were treated with either 1 µg/ml Dox or vehicle (water). Twenty four hours later-whole cell lysates were harvested and
analyzed by western blotting using the myc-tag antibody to detect the myc-tagged wt or variant MLH1 proteins. Positive clones were maintained in the presence of 5µg/ml blasticidin and 100 µg/ml of zeocin.

Clones that show minimal levels of expression in the absence of treatment and maximal level of expression upon induction were chosen and propagated. For each of the variants, we attempted to choose the clones that show expression comparable to wild-type MLH1. Using the above process we were able to generated three cell lines expressing wild-type MLH1, I19F MLH1, K618A MLH1 and the empty vector. To characterize these cell lines, cells were induced with varying doses (0.01-1µg/ml) doxycycline to determine the lowest concentration of drug that yields robust expression of wild-type and variant MLH1. This was followed by a time course of MLH1 induction in these cell lines, where MLH1 expression was assayed 3-24 hours post-induction.

2.3.11 Cell survival assays

To assess if the missense variants in MLH1 alter cell proliferation in response to chemotherapeutic agents, HCT116 cells expressing wild-type or variant MLH1 were treated with temozolomide. Results generated from the empty vector (EV) cell line will help determine the contribution of other doxycycline responsive genes to the phenotype measured. Cells were seeded in 96-well plates at a density shown to previously to allow for exponential growth throughout the exposure period. The next day the cells were exposed to varying concentrations of temozolomide (0-500 µM) in the presence of 10-40 µM O$^6$-benzylguanine over a 5 day period. O$^6$-benzylguanine is a inhibitor of MGMT, an enzyme that efficiently removes alkylating damage. Thus, treatment was carried out to determine the independent contribution of the MMR system to DNA damage signaling. Thereafter, cell survival was measured using an MTS based
assay (CellTiter 96AQ<sub>ueous</sub> One Solution, Promega, Madison, WI) that measures the ability of metabolically active cells to cleave the tetrazolium salt MTS to the water-soluble formazan.

2.3.12 Cell cycle analysis

The profiles of cell populations in G1, S and G2 phases were assessed by flow cytometry. Briefly, cells were plated in 6 cm dishes. The next day the medium was changed to low serum (0.5%) McCoy’s medium and cultured for 24 hours to synchronize the phase of the cells. The next day, cells were treated with TMZ and O<sub>6</sub>-benzylguanine, which was added to medium that contained 10% FBS. The cells were fixed with 70% ethanol overnight at -20°C, treated with RNaseA for 30 minutes at 37°C and stained with propidium iodide (PI). The percentage of cells in each phase was determined using a BD FACSCalibur (BD Biosciences, San Jose, CA) and the resulting data analyzed using the FlowJo software.

2.3.13 MSI analysis

To minimize variability that could be introduced by a mixed cell population a single clone was selected from each stable inducible cell lines by sparsely plating cells and growing them in the presence of blasticidin and zeocin. This clone was then expanded and grown independently for 30 generations either in the presence or absence of doxycycline (the cells were not subjected to any treatment with alkylating agents). Cells were harvested at passage 0 (onset), 10, 30, and chromosomal DNA was extracted using the QiaAMP DNA mini kit (Qiagen). MSI was assessed at the mononucleotide repeat loci BAT25 and BAT26 as described before [410].
### Table 2.1 Primers used for site-directed mutagenesis

<table>
<thead>
<tr>
<th>Variant</th>
<th>Forward Primer</th>
<th>Reverse Primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>I19F</td>
<td>T7 F 5'TAATACGACTCAGTTAGGG3'</td>
<td>ML1R 5'ACCTCGAAAGCCATAGGTAG3'</td>
</tr>
<tr>
<td>R265C</td>
<td>M265F 5'AGGATAAAAACCCTAGCCTTC3'</td>
<td>ML3R 5'ATGGGTTGTGTGTTTTTGGG3'</td>
</tr>
<tr>
<td>K618A</td>
<td>ML7F 'AACTGTTCTACCAGATACTC'</td>
<td>ML7R 'CACCTCAGTGGC TAG TCG'</td>
</tr>
<tr>
<td>L749Q</td>
<td>ML8F 5'GGAGGGACTGCTATCTTC3'</td>
<td>MCSR 5'TGATCATCGGGTTAACAAGG3'</td>
</tr>
</tbody>
</table>

### Table 2.2 Summary of predictions generated by computational tools

<table>
<thead>
<tr>
<th>MLH1 Variant</th>
<th>Nucleotide Change</th>
<th>Exon</th>
<th>Functional Domain</th>
<th>SIFT&lt;sup&gt;22&lt;/sup&gt;</th>
<th>PolyPhen&lt;sup&gt;23&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>I19F</td>
<td>c.55A&gt;T</td>
<td>1</td>
<td>-</td>
<td>Intolerant</td>
<td>Possibly Damaging</td>
</tr>
<tr>
<td>R265C</td>
<td>c.793C&gt;T</td>
<td>10</td>
<td>-</td>
<td>Intolerant</td>
<td>Probably Damaging</td>
</tr>
<tr>
<td>K618A</td>
<td>c.1852_1853 AA&gt;GC</td>
<td>16</td>
<td>MutL homolog interaction</td>
<td>Tolerant</td>
<td>Possibly Damaging</td>
</tr>
<tr>
<td>L749Q</td>
<td>c.2246T&gt;A</td>
<td>19</td>
<td>C-terminal homology</td>
<td>Intolerant</td>
<td>Probably Damaging</td>
</tr>
</tbody>
</table>

Output generated by SIFT: Intolerant or Tolerant
Output generated by PolyPhen: Probably Damaging > Possibly Damaging > Benign
2.4 Results

2.4.1 In silico analysis of MLH1 variants

For this study we chose four variants located in distinct domains of \textit{MLH1}. Missense variants that lead to the I19F amino acid substitution have been reported twice in the literature [398, 399]. Variants that lead to multiple amino acids substitutions at codon 265 have been reported in the literature including the polymorphism R265H and the unclassified variant, R265C [385]; while codon 618 incurs a point mutation leading to the substitution K618T, a bona fide mutation and a two base-pair substitution leading to the variant K618A [390]. Similarly, codon 749 is shown to result in a substitution leading to L749P [411, 412].

We first employed the computational tools, SIFT and PolyPhen, to predict the pathogenicity of the missense substitutions chosen for this study (Table 2.2). SIFT has been used to predict the pathogenicity of the variants R265C and K618A in previous reports [383, 413]. However, PolyPhen is yet to be applied to the analysis of these variants. All four of the missense variants chosen were predicted to be pathogenic by either of these two programs. However, we observed that these computational databases did not always provide concordant predictions. For example, while SIFT predicts K618A as a “tolerated” alteration, PolyPhen predicts it to be possibly damaging.

To gain a better understanding of how these substitutions affect the structure and consequently the function of the protein, we mapped this variant onto the crystal structure of the MLH1 homolog in \textit{Escherichia coli}, MutL (PDB accession no.: 1B63, Ban and Yang, 1998). A high degree of sequence conservation is seen only within the N-terminal domain of \textit{E. coli} MutL (Ban and Yang, 1998; Guarne et al., 2004). The C-terminal region of \textit{E. coli} MutL shares little sequence similarity to other MutL homologs except human MutL homolog 3 (MLH3) and PMS2 (Guarne et al., 2004). In addition, these authors were unable to detect sequence conservation
between *E. coli* MutL and human MLH1 (Guarne et al., 2004). Therefore, we decided to use this approach to assess the I19F and R265C variants that occur in the N-terminal region of MLH1. The amino acid I19 corresponds to I14 in *E. coli*, and occurs on the periphery of the structure. Based on its terminal location, it likely plays a role in mediating protein-protein interactions (Figure 2.2). The amino acid R265 corresponds to R261 in *E. coli*, and it is located between an alpha helix and a beta sheet (Figure 2.2). R261 interacts with the carbonyl groups of Ser91 and Leu92; in addition it also forms an internal water-mediated salt bridge with Glu32. These interactions indicate the potential role of this residue in stabilizing the folding of the N-terminal domain. However, the level of this effect may depend on the amino acid it is substituted with, and its electrochemical properties. R261 has also been shown to destabilize the P-loop, consequently decreasing ATP binding (Ban et al., 1999).
Figure 2.2 A Ribbon diagram of the approximate position of human MLH1 R265 on the crystal structure of E. coli MutL.

R265, which is equivalent to R261 in *Escherichia coli* was mapped onto the crystal structure of the N-terminus of MutL (PDB#:1B63) using sequence homology. **B.** A more magnified view of the amino acid R265 (*E.coli* R261): Indicates its likely involvement in stabilizing the folding and the structure of the protein. **C.** R261 also forms an internal water-mediated salt bridge with Glu32.
Figure 2.3 Profile of time dependent expression of wild type MLH1 in HCT116 cells.

HCT116 cells were transiently transfected with 3 µg of pcDNA3.1+ expressing wt MLH1 and harvested at the designated time points. Total extracts were immunoblotted with either anti-MLH1 (upper panel) or anti-PMS2 (lower panel). An equal amount of the mismatch repair proficient HEK293 total cell lysate was also immunoblotted as a positive control. Results from untransfected (U) HCT116 cells indicate the absence of endogenous MLH1 expression.
2.4.2 HCT116 cells are defective in MLH1

The colorectal cancer cell line, HCT116, was chosen as a model system for this study, as it is deficient for endogenous MLH1 due to a hemizygous nonsense mutation in codon 252 [414]. The mutational status of these cells was confirmed by extracting RNA from the HCT116 cells, reverse transcribing it and sequencing the resultant cDNA. We then confirmed the absence of endogenous MLH1 at the protein level by performing western blotting assays as described below. This indicated that the mutated endogenous MLH1 is rapidly degraded and therefore not detectable by this assay, which is in line with previously published observations (Figure 2.3) [390, 405]. We also confirmed that PMS2, the cognate partner for MLH1 was present in these cells in extremely low, almost undetectable levels, as it is unstable when not heterodimerized with MLH1. We then transiently transfected in MLH1-pcDNA 3.1 vector using the Lipofectamine 2000 reagent and assayed for its expression at various time points ranging from 12 hours to 7 days. This time course experiment indicated that we were able to transiently express wild-type MLH1 from 12 hrs to 96 hrs post-transfection (Figure 2.3). However, transfection of ectopic MLH1 restored endogenous levels of PMS2. This is consistent with previous studies that indicate that PMS2 is unstable in the absence of MLH1 [260, 387, 390, 404].

2.4.3 R265C and K618A decrease expression of MLH1

Results from both tagged and non-tagged expression vectors indicate that all variant constructs are capable of producing variant MLH1 protein of the expected size (~85kDa and ~ 90kDa respectively). The R265C and K618A variants however, appear to either produce less protein or
produce protein that is less stable than the wild-type MLH1 construct (Figure 2.4). On the other hand, the expression profile for the variants I19F and L749Q were comparable to the wild-type protein (Figure 2.4 and 2.5). No significant differences in β-galactosidase transfection efficiency were seen between the wild-type and variant constructs. Representative data for K618A is depicted in Fig. 2C. (Figure 2.5)

2.4.4 Altered expression is due to decreased stability of the protein

Differences in expression could occur due to a decrease in the levels of transcription and/or translation, or due to an inherent decrease in the stability of the transcript or the protein. To address whether the stability of MLH1 is altered by the substitutions R265C and K618A, we employed pulse-chase assays to examine the rate of protein decay. Results indicate that R265C and K618A have a half-life of ~ 5.5-6.5 and ~3-4 hours respectively. These variants are thus significantly less stable than wild-type MLH1, which has a half-life about 10-10.5 hours (Figure 2.6). To verify these observations using an independent method; we treated transiently transfected cells with cycloheximide, a global inhibitor of de novo protein synthesis, and examined the stability of wild-type and variant MLH1 over time. Our experiments indicate that the half-life of R265C and K618A MLH1 is significantly decreased compared to the wild-type protein (Figure 2.7), thus corroborating our pulse chase results. To rule out technical influences the variant L749Q, which shows comparable expression to wild-type MLH1, was included in the cycloheximide assay. Our findings illustrate that the stability of the L749Q variant is comparable to that of the wild-type protein (Figure 2.7). This finding validates our pulse chase results, and demonstrates that our observations regarding decreased stability are due to the substitutions introduced at codons 265 and 618 of the MLH1 gene.
Figure 2.4 Expression of wild-type and variant MLH1 in HCT116 cells.

A. HCT116 cells were transiently transfected with wt or I19F MLH1 and harvested 24 h, 48h and 72 h post transfection. The whole cell lysates (25μg) were immunoblotted for anti-MLH1 (upper panel) and anti-β-actin (lower panel). B. Cells were transfected with 3μg (A) or 5μg of wt or R265C MLH1. U; Untransfected HCT116 cells, EV: empty vector. Results are representative of at least 3 independent experiments.
A. HCT116 cells were transiently transfected with 3µg of Myc-tagged wt, K618A or L749Q MLH1 and harvested at 24, 48 and 72 hrs. Cell lysates were probed with antibodies for anti-Myc (upper panel) and anti-β-Actin (lower panel). Results indicate decreased expression of MLH1 K618A at all time points when compared to wt, while L749Q consistently showed comparable expression to the wt vector.

B. Transfection efficiency of K618A as assessed by β-galactosidase activity: HCT116 cells were co-transfected with 0.15µg of pCMV β-galactosidase in addition to the MLH1 constructs and β-galactosidase activity was measured in cell lysates. β-galactosidase enzyme assays indicate that the K618A construct has a comparable, if not slightly higher, transfection efficiency compared to that of the wt construct (comparable transfection efficiencies were seen with the other variants). Data representative of at least 3 independent experiments and error bars represent the 95% confidence intervals.

Figure 2.5 Expression profile of the MLH1 variants K618A & L749Q.
Figure 2.6  Pulse chase assays to determine half-life of wild-type and variant protein.

HCT116 cells were transiently transfected with either wt, R265C or K618A MLH1 cDNA. The amount of radiolabelled MLH1 pulled down at each time point was analyzed by densitometry, normalized to β-Actin, and plotted against time. The line of best fit for R265C/K618A MLH1 variants are depicted. As a reference, the line of best fit that corresponds to the half-life of wt MLH1 is depicted on both graphs. Results indicate that the half-life of the variants R265C and K618A is drastically altered (half-life of ~5.5 and ~3.5 hours respectively) compared to that of the wt MLH1 protein (half-life of ~10 hours). Each experiment was performed a minimum of 3 times and representative results are shown.
Figure 2.7 Stability of MLH1 variants as assessed by cycloheximide treatments.

HCT116 cells were transfected with either wt MLH1 or R265C, K618A, L749Q variant constructs. Cells were then treated with cycloheximide (CXH) [+ or -] or vehicle [-], 24 hours post-transfection and lysed at indicated time points. MLH1 was detected by western blotting and β-actin was probed for as a loading control. R265C and K618A MLH1 shows a significantly reduced half-life (B, C), compared to wt MLH1 (A). L749Q however shows minimal effect on stability (D), compared to the wt protein. 3 independent experiments carried out and representative results are depicted.
2.4.5 L749Q impairs ability of MLH1 to heterodimerize with PMS2

Interestingly, in our transfection experiments we also observed decreased levels of endogenous PMS2 expressed when HCT116 cells were transfected with L749Q (Figure 2.8). We speculated that this may be due to the decreased ability of MLH1 to bind to PMS2, due to the L749Q missense substitution. Results from coimmunoprecipitation assays indicated that the amount of PMS2 pulled down when transfected with L749Q is strikingly low compared to the wild-type MLH1 protein (Figure 2.8). This indicates that this variant is unable to heterodimerize with PMS2 as efficiently as wild-type MLH1. R265C and K618A variants, on the other hand, did not appear to perturb the ability of MLH1 to heterodimerize with PMS2 (Figure 2.8). Taken together these results indicate that the variant L749Q perturbs the ability of MLH1 to heterodimerize with its partner PMS2, abrogates the formation of MutLα, despite having no observable effect on expression.
Figure 2.8 Ability of wild-type and variant MLH1 to interact with endogenous PMS2.

A. Levels of endogenous PMS2 expressed upon ectopic expression of MLH1 L749Q: HCT116 cells were transiently transfected with Myc-tagged wt or the L749Q MLH1 variant, and harvested 24, 48 and 72 hrs post-transfection. Cell lysates were probed with anti-Myc (upper panel) and anti-PMS2 antibodies (middle panel). β-Actin was probed for as a loading control (lower panel). Results indicate that as expected, endogenous PMS2 is stabilized in the presence of wt MLH1. However, there is a considerable decrease in the levels of endogenous PMS2 levels upon ectopic expression of the L749Q variant. B & C. Interaction of MLH1 variants with endogenous PMS2 by coimmunoprecipitation experiments: HCT116 cells were transiently transfected with 3µg of wt or variant MLH1 and harvested 48 hours post-transfection. Total MLH1 was immunoprecipitated from 1000µg of whole cell lysate (WCL) using the MLH1 antibody. Levels of endogenous PMS2 and ectopic MLH1 pulled down in each sample were probed for by western blotting. The upper and lower panels show PMS2 and MLH1 profiles respectively, on ectopic expression of the respective MLH1 constructs. In the right panel, equal amounts of WCL are loaded, to establish levels of protein. Expression of L749Q reveals decreased PMS2 pulled down as an indicator of its disrupted ability to heterodimerize with PMS2, albeit high levels of MLH1 expression (B). In contrast, the variants R265C and K618A did not show disrupted ability heterodimerize with MLH1. As expected R265C and K618A show decreased expression of MLH1 in the WCL samples when 25µg cell lysate are loaded. However, there appeared to be no observable differences in the total level of MLH1 immunoprecipitated from 1000µg of cell lysates when transfected with wt, R265C or K618A, likely due to the detection level of the antibody being saturated (C). Data representative of at least 3 independent experiments.
2.4.6 HCT116 cells express inducible wild-type and variant MLH1

A potential limitation of carrying out transient transfection experiments is that we cannot control for the effects of variable transfection efficiency. Furthermore, transient transfection experiments cannot be used to carry out treatments and measure phenotypic differences over longer periods of time. Thus, we opted to generate stable inducible cell lines to assess the effects of these variants on the functionality of MLH1. Inducible cell lines were chosen to minimize the variability that could be introduced by the cells adapting to the presence of MLH1, as would be the case if an uninducible stable cell line was generated. The major benefit of this model system therefore would be that it would allow us to regulate the expression of MLH1. We generated four different cell lines, EV, wtMLH1, I19F MLH1 and K618A MLH1. The empty vector cell line was generated to help determine the contribution of other doxycycline responsive genes to the phenotype measured. The I19F variant was chosen as the molecular mechanism by which this substitution impacted MLH1 and MMR function was still unclear. In our previous assays it neither affected MLH1 expression nor heterodimerization. Interestingly, when mapped to the crystal structure of MLH1 it was located on the exterior surface of MLH1, and its location hinted that it might be involved in interacting with other proteins. The K618A variant was chosen as it showed altered protein stability, and as previous studies had indicated that DNA damage signaling in response to an alkylating agent required full expression of MLH1 [311]. Clones that show minimal levels of expression in the absence of treatment and maximal level of expression upon induction were chosen and propagated. However, it was challenging to generate such clones, especially those that successfully showed differential expression of wtMLH1. We able to generate two such clones for the wtMLH1 cell line (out of about 90 clones screened), two for K618A (out of about 30 clones) and three I19F (out of about 30 clones) (Figure 2.9, 2.10). Of
these we chose the clone that showed maximal induction of wtMLH1 (clone 25); and variants that showed comparable MLH1 expression to that of the wild-type clone (I19F-clone 20, K618A-clone 18). To further characterize these cell lines we induced cells with varying doses (0.01-1μg/ml) of doxycycline to determine the lowest concentration of drug that yields robust expression of wild-type and variant MLH1(Figure 2.11). All cell lines showed equivalent induction at all concentrations. Thus, the lowest concentration tested, 0.01μg/ml was used in subsequent experiments unless otherwise noted. We also carried out a time course of MLH1 induction in these cell lines, which showed a gradual increase in expression over time as expected (Figure 2.11)
Resistant clones of the Tet-repressor expressing cells or wild-type MLH1 expressing cells were screened for by western blotting. Clones that showed the maximal induction of protein and the least leakiness (assessed by minimal/negligent expression in the absence of Dox) were expanded for subsequent experiments. [+] induced with Dox (doxycycline), [-] treated with vehicle, water. Antibodies against the Myc-tag of MLH1 was used in these experiments.
Figure 2.10 Generation of the I19F MLH1 stable inducible cell line.

A. Stable clones generated for the I19F variant. [+] induced with Dox (doxycycline), [-] treated with vehicle, water. B. Ribbon diagram of the approximate position of human MLH1 I19 on the crystal structure of E. coli MutL: I19, which is equivalent to I14 in *Escherichia coli* was mapped onto the crystal structure of the N-terminus of MutL (PDB#: 1B63) using sequence homology.
Figure 2.11 Characterization of the K618A stable inducible cell line.

**A.** The doxycycline concentration course for K618 and wild-type MLH1. Indicates that all concentrations tested efficiently induced the expression of MLH1. **B.** Time course experiment demonstrating induction of K618A MLH1 cell line. Similar profiles were seen with all stable inducible cell lines with maximal induction evident as 24 hours post induction. U: Uninduced.
2.4.7 Stable inducible cell lines and cell survival

When cell survival was assessed in the wtMLH1 stable inducible cell line in response to a range of doses of temozolomide we observed no difference between induced and uninduced cells. Similarly when the ratio of induced to uninduced cells expressing wild-type MLH1 was compared to this ratio in the EV cell line no statistically significant differences were seen (Table 2.3). To determine if hampered expression was responsible for this lack of effect, we carried out a western blot to confirm MLH1 expression in a subset of these conditions. Cells for western blotting were plated and treated at the same time as the cell survival experiments and results demonstrated unimpeded expression of MLH1 (Figure 2.12). Furthermore, these experiments confirmed that the addition of temozolomide and O6-benzylguanine did not interfere with the induction of MLH1 by doxycycline.
### Table 2.3 Summary of cell survival in response to temozolomide

<table>
<thead>
<tr>
<th>Cell line</th>
<th>TMZ dose</th>
<th>Relative Cell # (I/U)</th>
<th>Std Dev</th>
<th>95% CI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>EV</td>
<td>0</td>
<td>0.95</td>
<td>0.40</td>
<td>0.50 - 1.39</td>
<td></td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>1.21</td>
<td>0.33</td>
<td>0.84 - 1.59</td>
<td></td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>1.20</td>
<td>0.23</td>
<td>0.94 - 1.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.72</td>
<td>1.13</td>
<td>0.44 - 3.01</td>
<td></td>
</tr>
<tr>
<td>wt MLH1</td>
<td>0</td>
<td>0.88</td>
<td>0.29</td>
<td>0.54 - 1.21</td>
<td>0.635</td>
</tr>
<tr>
<td></td>
<td>125</td>
<td>1.65</td>
<td>0.46</td>
<td>1.13 - 2.17</td>
<td>0.273</td>
</tr>
<tr>
<td></td>
<td>250</td>
<td>2.37</td>
<td>1.15</td>
<td>1.06 - 3.67</td>
<td>0.186</td>
</tr>
<tr>
<td></td>
<td>500</td>
<td>1.43</td>
<td>0.32</td>
<td>1.07 - 1.79</td>
<td>0.755</td>
</tr>
</tbody>
</table>

Data representative of 3 independent experiments. TMZ: Temozolomide, I: induced, U: Uninduced, CI: confidence interval

**Figure 2.12** Western blot to confirm levels of MLH1 in cell survival experiments.

TMZ: Temozolomide. BG: O\(^6\)-Benzylguanine, Dox, Doxycycline.
2.4.8 Stable inducible cell lines and the G2/M checkpoint

Previous studies have illustrated that MLH1 is important for the G2/M checkpoint arrest. In these experiments we used the HCT116+Chr3 cell line as our positive control. This cell line contains an extra copy of the entire chromosome 3 and as a consequence includes an extra copy of MLH1. Thus, these cells express MLH1 and MMR assays have demonstrated that this cell line was able to restore MMR proficiency when compared to MMR deficient HCT116 cell line [405].

Previous studies have demonstrated that this arrest is evident only when the phases of these cells are synchronized prior to treatment with an alkylating agent [306]. Thus, in accordance with these studies the cells were synchronized with low serum media (0.5% FBS) and then treated with 100 or 500 µM temozolomide. At both these doses, 3 and 4 days after treatments the HCT116+Chr3 cell line arrested in G2/M phase, while the parental HCT116 cell line did not do so. However, the stable inducible cell lines expressing wt or variant MLH1 did not arrest 3 or 4 days after treatment (representative results depicted in Figure 2.13). Similar results were observed in experiments carried out with varying concentrations of TMZ (100-500 µm).
A. EV - Untreated

B. EV - TMZ Treated

C. HCT116+Chr3 - Untreated

D. HCT116+Chr 3 - TMZ Treated
E. WT - Untreated

![Graph showing cell count vs. FL2-H: FL2-PI Height for WT - Untreated.]

F. WT - TMZ Treated

![Graph showing cell count vs. FL2-H: FL2-PI Height for WT - TMZ Treated.]

G. I19F - Untreated

![Graph showing cell count vs. FL2-H: FL2-PI Height for I19F - Untreated.]

H. I19F - TMZ Treated

![Graph showing cell count vs. FL2-H: FL2-PI Height for I19F - TMZ Treated.]

Figure 2.13 Representative cell cycle profiles of the stable inducible and HCT116+Chr3 cell lines in response to treatment with temozolomide.

Cells were treated with 100 µM temozolomide, untreated cells were treated with an equal concentration of vehicle, DMSO and the profile of the cells were assed 3 days later. The first peak shows the relative proportion of cells in G1 phase expressing 2n DNA content. The second peak shows the cells in G2 and M phases of the cell cycle expressing 4n DNA content. The red arrow highlights the sample that displays a G2/M arrest profile. Representative results are depicted here.
2.4.9 Stable inducible expression of MLH1 & the microsatellite instability phenotype of the HCT116 cell line

Given that the role of MLH1 in response to DNA damage is not as well defined, we wanted to determine if stable inducible expression of MLH1 corrected the MSI-H phenotype of the HCT116 cell line. The MSI phenotype is a surrogate marker of MMR activity and previous studies have used similar approaches to determine the post-replicative error repair activity of MMR proteins in cells [311]. However, when we assessed the stability of the BAT25 and BAT26 markers, no change in instability was observed in the induced cells compared to uninduced cells both at passage 10 and passage 30 (Table 2.4). Induction of MLH1 should restore MMR activity, which should in turn restore stability at the BAT25/26 markers. Given that the outcome of these experiments hinged on the continued expression of MLH1, we harvested protein from a subset of cells at passage 20 and passage 30. Western blotting indicated that the protein was continually expressed (Figure 2.14) and was therefore not responsible for this lack of effect. However, the level of protein that was expressed by the cells appear to be lower at passage 30 (at the conclusion of experiment) compared to the levels at passage 20.
### Table 2.4 Summary of MSI experiments

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Passage 0 Uninduced</th>
<th>Passage 0 Induced (Dox)</th>
<th>Passage 10 Uninduced</th>
<th>Passage 10 Induced (Dox)</th>
<th>Passage 30 Uninduced</th>
<th>Passage 30 Induced (Dox)</th>
</tr>
</thead>
<tbody>
<tr>
<td>wt MLH1</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
</tr>
<tr>
<td>I19F MLH1</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
</tr>
<tr>
<td>K618A MLH1</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
<td>MSI-H</td>
<td>No change</td>
</tr>
</tbody>
</table>

Grown independently for 30 generations either in the presence (induced) or absence (uninduced) of Doxycycline (Dox). Cells were harvested at passage 0 (onset of experiment), 10, 30.
Figure 2.14 Expression of MLH1 and PMS2 in wild-type MLH1 cells stable inducible cells.

Cell lysates were harvested after the EV and WT stable inducible cell lines were grown for 20 or 30 passages in the presence (I) or absence (U) of doxycycline. The upper panel shows MLH1 expression, the middle panel shows PMS2 expression and the lower panel shows the β-Actin expression levels in these cells. HCT116+Chr 3 was included as a control to assess relative expression levels.
2.5 Discussion

The overall aim of this study was to employ a series of biochemical assays to elucidate if missense variants in *MLH1* are able to significantly alter protein expression/stability, heterodimer formation and/or its function in signaling DNA damage. Our expression assays illustrated that the variants R265C and K618A altered levels of MLH1 expression, leading us to postulate that this might be due to decreased protein stability. Although, it has been previously suggested that missense variants may decrease the stability of the MLH1 protein, to our knowledge, no attempts have been made to date to assess the stability of the variant proteins [391]. As several mechanisms may account for differences in expression, we chose to investigate the stability of these variant proteins using both pulse-chase assays and cycloheximide experiments. Both assays indicate that compared to wild-type MLH1, both R265C and K618A missense substitutions decrease the half-life of the protein. It is important to note that we cannot expect to observe identical half-lives for the respective proteins, given the technical differences between the two protocols. The pathway responsible for the degradation of MLH1 is currently unknown; therefore it is difficult to explain the mechanism responsible for this observed decrease in stability. However, we speculate that missense substitutions may affect the tertiary structure of the protein, which may lead to defective folding and the consequent degradation of MLH1. The location of the residue on *E. coli* MutL and its putative interactions hint that such might be the case for R265C.

In addition to the above strategies, we also used SIFT and PolyPhen to determine if these programs can be used to predict the pathogenic nature of these variants. All three variants were predicted to be deleterious, indicating that in the clinical setting these computational programs are likely to be helpful in classifying the risk associated with novel variants. A recent study by
Chan et al., demonstrated that the predictive value of these programs can be further increased by using a combination of computational methods [386]. However, the two programs provided discordant predictions with regard to K618A. Thus, in instances such as this where discordant predictions are made, it remains important that missense alterations are assessed using a combination of in-vitro/in-vivo strategies until improved models of prediction are available.

Several attempts have been made to assess the pathogenicity and/or biological significance of the R265C and K618A variants. In this case R265C was identified in a proband that showed deficient MLH1 expression by IHC, while the K618A variant was identified in a proband that showed intact expression by IHC (details for IHC data included in Chapter 3). However, R265C and K618A have been identified in multiple families both in our clinic as well as others, and the IHC findings from these families are not always identical. Quantitative MMR assays carried out with R265C in *Saccharomyces cerevisiae* indicate that this variant retained its ability to repair defects, albeit at lower proficiency than wild-type and ‘bona fide’ mutants of MLH1[401]. These observations lead the authors of this study to refer to this class of variants as ‘efficiency polymorphisms’. Interestingly, a study published less than a year later evaluated R265C using an *in vitro* MMR assay and found it to be competent in post-replicative repair [385]. A recent study employed conversion analysis technology to assess this variant. Conversion analysis is able to expose masked mutations by converting the human chromosome complement to a haploid state by fusing it with recipient E2 mouse cell line, after which these hybrid cell lines are used to assay RNA expression. This study demonstrated that R265C leads to the skipping of exons 9 and 10, as well as a decrease in the levels of mRNA [234]. However, evidence for the skipping of *MLH1* exons 9 and 10 alone does not necessarily delineate the status of this variant as a disease causing mutation; as previous studies have reported the existence of low levels of alternatively spliced transcripts of *MLH1* that lack these exons [415]. Therefore, it is likely that the
detrimental effects of exon skipping depend on the level of the transcript that is deleted. It is also possible that this and other variants contribute to HNPCC susceptibility by several mechanisms including defective splicing and altered protein stability. Taken together these studies imply the potentially pathogenic nature of this variant, and the associated risk for individuals carrying the R265C variant. Our results provide further insight into the molecular basis of the pathogenicity of R265C, by demonstrating that in addition to causing splicing defects, this variant may also decrease the efficiency of MMR by destabilizing MLH1.

The K618A substitution replaces a charged amino acid with a neutral one. This is important as K618 occurs alongside four charged amino acids that are well conserved in mammals, as well as in Drosophila melanogaster [402]. Fearnhead et al., have proposed a link between the K618A variant and an enhanced risk of multiple adenomas [402]. Additionally, in a recent study carried out in the ovarian cancer cell line A2780, K618A showed normal levels of expression, but was unable to reverse the mutator phenotype [400]. This variant was also unable to reverse tolerance to methylating agents, a characteristic feature indicative of defective MMR. However, a study carried out in the human embryonic kidney cell line 293T transfected with K618A demonstrated that neither the expression of MLH1 nor the efficiency of repair activity as assessed by an in vitro MMR assay was altered [383]. Furthermore, when Guerrette et al. examined the ability of variants expressed in vitro to interact with PMS2 in a GST fusion protein interaction assay, they found that K618A reduced the binding efficiency to PMS2 by > 85% [388]. In contrast to this, we observed no effect of K618A on the ability of MLH1 to bind to PMS2 using coimmunoprecipitation assays. Our observations are supported by the finding that in mlh1<sup>+/−</sup>; pms2<sup>−/−</sup> murine embryonic fibroblasts, K618A did not display compromised ability to stabilize PMS2 [404]. Moreover, results from yeast-two hybrid assays demonstrated that K618A was only partially defective in its ability to interact with the partner proteins of MLH1 [403]. Although we
cannot exclude the role of this variant in compromising binding and heterodimerization, it is plausible that the differences in experimental conditions between the GST pull down assays and coimmunoprecipitation assays may contribute to this discrepancy. Thus, based on these functional assays alone it is difficult to determine if K618A is truly a pathogenic mutation or a rare functional polymorphism. Our observation that K618A decreases the stability of the protein, therefore offers an additional piece of evidence regarding the role of this variant in disease susceptibility.

The discrepancies observed between these various analyses could be due in part to the diverse experimental protocols, the specificity of cell types, and the sensitivity of the methods. We chose to carry out our assay in a human colorectal cancer cell line to minimize such effects of cell type or tissue specificity, and to enhance the relevance of our findings. In addition, another issue which arises in the interpretation of findings from mismatch repair assays is the difficulty in defining the threshold below which repair activity is abnormal or deficient. This leads to a situation where researchers have to set up arbitrary cut-off values for normal MMR activity, which influences the verdict of the role of these missense variants in compromising repair. It can be postulated that irrespective of the functionality of the protein, decreased stability likely leads to a situation where lower levels of protein are present in the cell. Although the levels of protein essential for normal cellular function and the threshold below which cancer susceptibility occurs has not been established; it is likely that decreased levels compromise to some degree function(s) carried out by these protein complexes. For MLH1, possible functions affected in this manner are not limited to the repair of mismatches, but may also extend to other functions such as DNA damage signaling and recombination. It is also possible that varying threshold levels of MLH1 exist for the different functions of the MMR system, even in the same cells or tissues. The observation that less MLH1 expression is required for the error-repair activity compared to the
DNA damage response, might explain how these variants may still increase an individual’s susceptibility to HNPCC, without necessarily decreasing MMR activity. Thus, by establishing that both R265C and K618A decrease the stability of the MLH1 significantly, our results now demonstrate an independent mechanism by which the pathogenicity of these variants can be established.

It is prerequisite that MLH1 heterodimerize with PMS2, to form the MutLα complex, prior to carrying out any of its function described to date. Therefore, any variant that perturbs this interaction could potentially compromise the efficiency of diverse functions carried out by this complex. We observed that the variant L749Q disrupts binding to PMS2, despite proficient MLH1 expression. This variant is located just outside the PMS2 interaction domain, but within the C-terminal homology (CTH) domain of MLH1, which is exclusive to eukaryotic MutL homologs [345, 387]. A study by Mohd et al., showed that a truncating mutation at the same codon (MLH1 L749X) was capable of interacting with, but incapable of stabilizing PMS2 [404]. The results of this study lead the authors to conclude that the CTH domain plays a role in stabilizing the MutLα complex, leading to the consequent stabilization of PMS2. Hence, it is possible that a similar mechanism may underlie our observations regarding the L749Q variant.

Previous studies using a cell line with inducible MLH1 has shown that while post-replicative error repair can occur even when as little as 10% of MLH1 is expressed, DNA damage signaling in response to MNNG requires full expression of MLH1[311]. This indicates the possibility that MLH1 variants that show differences in stability may have differential levels of resistance/tolerance to chemotherapeutic agents, which can be exploited to study the role of MLH1 in these processes. We generated stable inducible cell lines of wild-type and variant MLH1 in order to minimize the confounding effects that could occur due to variable, transient
expression. Cells deficient in MMR are tolerant to alkylating drugs such as temozolomide and treating MMR proficient cells with increasing concentrations of temozolomide was shown to lead to decreased cell survival [314, 416]. Despite continued expression of MLH1, we did not see altered cell survival and/or G2/M cell cycle arrest in the stable inducible cell lines expressing wild-type or variant MLH1. The positive control HCT116 + Chr3 has been shown to display MLH1 expression and correct the MSI-H phenotype of these cells [405]. In contrast to our stable inducible cells, treatment with temozolomide induced G2/M arrest in the HCT116+Chr3 cells, which is in agreement with previously published studies. In our MSI experiments we did not observe a change in the level of stability at the BAT25 or BAT26 loci, despite sustained MLH1 expression. However, this result should be interpreted cautiously, as given the nature of these experiments (the requirement to serially passage cells in the presence of doxycycline) we did not have an appropriate positive control. Thus, it appears that although the stable inducible cell lines generated express MLH1, the functionality of this protein is limited. Previous attempts to generate HCT116 with stable expression of MLH1 led to a growth inhibitory effect and the inability to restore MMR proficiency despite the detectable expression of the MLH1 protein (unpublished results, personal communication Dr. Bert Vogelstein). Furthermore, in a previously published study using a stable inducible cell line in the HEK293T background, expression of MLH1 resulted in G2/M arrest [311]. However, it appears that even in this cell line the sensitivity to damaging agents was not evident in clones that showed overexpression of MLH1 compared to PMS2; and only in cases where PMS2 expression was comparable to MMR proficient cell lines (in this case the parental cell line 293T) did they observe MNNG sensitivity (personal communication Dr. P. Cejka). Based on these findings the most likely reason for this lack of response is that the overexpression of MLH1 negatively affects the cell and possibly disrupts the function of the MMR system. A similar phenotype has been described in
Saccharomyces cerevisiae, where the overexpression of MLH1 was shown to lead to the mutator phenotype and defective MMR [417]. A possible mechanism for this is that the overexpressed MLH1 is not able to heterodimerize with PMS2 efficiently and the remaining excess partner-less MLH1 disrupts MMR activity in this in vitro system. The low level of PMS2 stabilized by MLH1 in these cell lines is evident in Figure 2.14. A possible explanation for MLH1’s inability to heterodimerize with PMS2 could be that the Myc-tag interferes with correct protein folding. Therefore, these studies indicate that the level of MLH1 is carefully regulated in cells. Thus, a biologically meaningful increase or decrease in MLH1 expression impairs the functions carried out by the MMR system.

In summary, we observed that the two naturally occurring variants R265C and K618A decreased the stability of MLH1. In addition a third variant generated by us, L749Q, affected MLH1’s ability to heterodimerize with PMS2. Although no single test or assay can definitively indicate the pathogenicity of a variant, combined data from a series of assays with regard to a specific variant or to comparable variants that occur in close proximity will likely help elucidate the pathogenicity associated with unclassified variants identified in HNPCC patients.

Acknowledgements

We would like to thank Dr. Yosr Haffani for her assistance with mapping the variants on the crystal structures. We would also like to thank Dr. Ken Lau, Annie Bang, Kelly Fenwick and Ted Chappell for their technical assistance.
CHAPTER 3

A novel method of determining the effect of MLH1 unclassified genetic variants on differential allelic expression

Sheron Perera\textsuperscript{1,2,3}, Brian Li\textsuperscript{2}, Soultana Tsitsikotas\textsuperscript{3}, Aaron Pollett\textsuperscript{1,3}, Kara Semotiuk\textsuperscript{4}, and Bharati Bapat\textsuperscript{1,2,3}

\textsuperscript{1} Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON, Canada; \textsuperscript{2} Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, ON, Canada; \textsuperscript{3} Department of Pathology and Laboratory Medicine, Mt. Sinai Hospital, Toronto, ON, Canada; \textsuperscript{4} Dr. Zane Cohen Digestive Diseases Clinical Research Centre, Mt. Sinai Hospital, Toronto, ON, Canada

The work in this chapter was primarily contributed by Sheron Perera. BL, an undergraduate 4th year project student worked with SP and helped with DNA and RNA extractions. ST, a technologist extracted RNA from a subset of patient lymphocyte samples. AP, a pathologist confirmed IHC staining patterns and KS, a genetic counselor provided access to family history and clinical information of the probands.

Chapter 3 A novel and rapid method of determining the effect of MLH1 unclassified genetic variants on differential allelic expression

3.1 Summary

Germline mutations in mismatch repair genes predispose patients to Lynch Syndrome and the majority of these mutations are seen in two key genes, *MLH1* and *MSH2*. In particular, about a third of the missense variants identified in *MLH1* are of unknown clinical significance. Using the PeakPicker program we have conducted a proof-of-principle study to investigate whether missense variants in *MLH1* leads to allelic imbalances. Lymphocyte RNA extracted from patients harboring known *MLH1* variants was used to quantify the ratio of variant to wild-type transcript, while lymphocyte DNA was used to establish baseline allelic expression levels. Our analysis indicates that the missense variants c.350C>T, c.793C>T, c.1852_1853AA>GC, as well as the truncating variant c.1528C>T led to significantly unbalanced allelic expression. However, the variants c.55A>T and c.2246T>C did not demonstrate an allelic imbalance. This illustrates a novel and efficient method of investigating the pathogenicity of unclassified genetic variants discovered in mismatch repair genes, as well as genes implicated in other inherited diseases. In addition, the PeakPicker methodology has the potential to be applied in the diagnostic setting, which in conjunction with results from other assays will help increase the accuracy and efficiency of genetic testing of colorectal cancer, as well as other inherited diseases.
3.2 Introduction

Lynch syndrome or hereditary nonpolyposis colorectal cancer (HNPCC; MIM 120435), which accounts for about 2% of all colorectal cancer cases, is one of the most common cancer predisposition syndromes. Those affected have an 80% lifetime risk of developing colorectal cancer[233]. In addition, malignancies of the endometrium, skin, bladder, ovaries, kidney, and small intestine are also associated with this syndrome.

Lynch syndrome is an autosomal dominant condition that is characterized by germline mutations in mismatch repair (MMR) genes. The MMR system is a post-replicative DNA repair system that recognizes and repairs base-base mismatches and insertion/deletion loops that occur during DNA replication and escape the proofreading activity of the DNA polymerases. MMR proteins have also been implicated in signaling DNA damage and apoptosis. Germline mutations of the MMR genes MLH1 (MIM 120436), MSH2 (MIM 609309), PMS2 (MIM 600259), PMS1 (MIM 600258), and MSH6 (MIM 600678) have hitherto been implicated in Lynch syndrome. However, about 90% of all known mutations are found in the two MMR genes MLH1 and MSH2[397], which appear to play essential and non-redundant functions of the MMR system.

Tumours with defective MMR function are characterized by expansion/contraction of the microsatellite regions, leading to microsatellite instability (MSI), a hallmark feature seen in about 85% of Lynch syndrome tumours[123]. Individuals diagnosed at a younger than average age and/or with familial clustering of colorectal or other Lynch syndrome associated cancers are generally referred to predictive genetic testing programs. Here, MSI testing is carried out on tumour DNA, which acts as a surrogate marker of defective MMR function. Typically, immunohistochemistry (IHC) is also performed to assess if MMR protein expression is lost.
Although MSI testing is a robust assay of mismatch repair activity, it is not a universal indicator of pathogenicity as it is possible for mutations to affect other functions of the MMR pathway [418]. Similarly, the loss of protein expression by IHC does not definitively link the unclassified variant to the loss of protein expression. Additionally, some tumours show equivocal staining patterns with intermediate levels of expression, which cannot be clearly distinguished by IHC. Furthermore, methylation of the MLH1 leads to the MSI phenotype and the loss of MLH1 expression in about 15% of all sporadic colorectal cancers [121, 122]. Thus, to confirm the presence of Lynch syndrome, the MMR genes of these individuals are screened for germline mutations. Similar to the other MMR genes, diverse sequence variants, including nonsense, missense, splice site, and frameshift mutations, have been reported throughout the coding region of the MLH1 gene. Many of these mutations are catalogued in the LOVD and MMR Variant Databases (http://chromium.liacs.nl/LOVD2/colon_cancer/home.php and http://www.med.mun.ca/MMRvariants, last accessed December 16, 2009) [372]. About 24% of mutations detected in MLH1 are missense substitutions [372]. The majority of these missense variants are of unknown clinical significance. Given that the diagnosis of Lynch syndrome relies primarily on the identification of germline defects in the MMR genes [234], detecting these unclassified genetic variants creates much ambiguity in the clinical setting as the pathogenicity of these variants cannot be readily ascertained.

Upon detection of missense variants several strategies are employed to characterize their pathogenic significance. These include linkage analysis, association studies, computational prediction tools and functional assays, each with specific strengths and limitations [234, 382-385, 389, 390]. Linkage analyses is often the most straightforward way to determine the relationship between the disease phenotype and the segregation of the variant, but is often not feasible due to smaller family sizes or unavailability of samples from relevant family members.
Computational programs, while helpful in classifying novel variants in the clinical setting, are limited in their predictive value unless a panel of such programs are used[386]. However, at present it is important to use a combination of in silico and experimental strategies until improved models of prediction become available. We and other groups have attempted to devise in vitro assays to characterize the pathogenic and functional significance of MLH1 variants. A common approach is to create these alterations in vitro by site-directed mutagenesis, after which these constructs are used to assay expression, MMR function or protein stability [383, 385, 389-391]. However, this approach requires a considerable investment of time and effort and cannot be routinely performed in a clinical diagnostic service.

Determining unbalanced allelic expression offers an alternative method of investigating the pathogenic nature of a variant. Several studies in other genes have demonstrated that single nucleotide substitutions are capable of altering the levels of mRNA transcripts [419-421]. Furthermore it has been proposed that such cis-acting functional variants can affect transcription, messenger RNA (mRNA) processing and mRNA stability [420]. We have conducted a proof-of-principle study to investigate whether missense variants in MLH1 lead to an allelic imbalance by measuring levels of transcript present in lymphocytes of individuals carrying the variants c.55A>T (p.I19F), c. 350C>T (p.T117M), c.793C>T (p.R265C), c.1528C>T (p.Q510X), c.1852_1853AA>GC (p.K618A), and c.2246T>C (p.L749P), using the PeakPicker method.

3.3 Materials & Methods

3.3.1 Selection of Patients and Families

These probands were referred to the Molecular Genetics Laboratory as part of the Provincial Cancer Genetics Program for the assessment of the possible diagnosis of Lynch syndrome. Predictive genetic testing was offered to clinically affected and at risk subjects, with pre- and
post-test genetic counseling as described previously[422]. In addition to complete follow-up information, clinical and histopathological data were collected retrospectively on all affected patients. Informed consent was obtained from all subjects and all studies were performed according to guidelines of the Ethics Committee of the University of Toronto.

3.3.2 Overview of diagnostic strategy and identification of MLH1 variants
The variants chosen for this study were identified in patients that were screened through the Ontario Predictive Genetic Testing Program at Mt. Sinai Hospital, Toronto, Canada based on a series of predetermined referral criteria. This program screens individuals presenting with either a family history of colorectal cancer, and/or those who develop tumours at a young age of onset (<50). Genetic screening is initiated by performing MSI testing and IHC for the MMR proteins on tumour specimens when available. This is followed by germline mutation testing on lymphocyte DNA. Initially, multiplex ligation-dependent probe amplification (MLPA) is performed to detect large genomic deletions and duplications in the MMR genes. If such changes are not detected, germline DNA is screened by sequencing the MLH1 and MSH2 genes, guided by MSI/IHC results of patient tumours. Through this program we have identified many established and novel germline mutations, as well as several unclassified variants.

3.3.3 Microsatellite instability and immunohistochemistry
MSI testing and IHC analysis for MMR proteins were performed as described previously.[410] Matched normal and tumour DNA were assessed using a panel of five microsatellite markers as recommended by the National Cancer Institute[232]. These include –mononucleotide BAT25 and BAT 26 and dinucleotide D2S123, D5S346 and D17S250 microsatellite markers. Each case was designated as either microsatellite unstable (MSI-H; ≥30% markers unstable), MSI-low (MSI-L; < 30% markers unstable), or microsatellite stable (MSS; no unstable markers). IHC
analysis of the respective MMR proteins was performed on formalin-fixed, paraffin-embedded tissues.

3.3.4 DNA extraction and mutation detection

Blood samples were obtained from patients and lymphocytes were isolated using NH$_4$Cl–Tris. DNA was extracted from lymphocytes using the QIAGEN or saturated salt-out method as described previously [408]. DNA from patients, whose tumours showed deficiency for a MMR protein, was subjected to exon by exon sequencing of genomic DNA to screen for alterations in $MLH1$ and $MSH2$ on an ABI 377 DNA Sequencer (Applied Biosystems, Foster City, CA). Sequence information of the coding region was derived from RefSeq NM_000249.2 ($MLH1$), NM_000251.1 ($MSH2$). PCR conditions and primer sequences are available upon request.

3.3.5 RNA extraction and cDNA synthesis

RNA extraction was performed using TRIZOL according to the manufacturer’s protocol (Invitrogen Life Technologies, Burlington, ON). The RNA then treated with DNaseI for 30 minutes using the RNase-Free DNase set (Qiagen), which allows for digestion of DNase prior to purification using the RNeasy MinElute kit (Qiagen). Reverse transcription PCR was performed according to standard techniques as described before [408] and cDNA was generated in triplicate from each RNA sample to minimize variation that may be introduced during the process of reverse transcription.

3.3.6 PeakPicker Analysis

3.3.6.1 Primer design and PCR

Primers were designed to amplify the exonic regions flanking the variant of interest so as to allow the same pair of primers to be used for the amplification of both the genomic (gDNA) and complementary DNA (cDNA) to ensure unbiased amplification (Table 3.1). PCR products were
then visualized on an 1.7% ethidium bromide/agarose to verify the presence of product. PCR products were treated with 5 units of Shrimp Alkaline Phosphotase and 2 units of Exonuclease I per 15µl of PCR product for 37°C for 1 hour to remove unincorporated dNTPs and excess amplification primers. The enzymes were then deactivated by incubating at 75°C for 15 min.

3.3.6.2 Sequencing and PeakPicker Analysis

High sensitivity sequencing was carried out as described before [423], using 0.5 µL of Big Dye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) with 2 µL of PCR product, 1.75 µL of 5x sequencing buffer and 10pmol of sequencing primer in a 10 µL reaction using the manufacturer’s protocol. The products were separated on an Applied Biosystems 3130XL DNA analyzer. Sequencing was carried out with both the forward and reverse primers whenever possible, unless the variant was located too close to the start of the primer.

The PeakPicker software was developed by Ge et al., for quantitative allele ratio analysis [423, 424]. This software is publicly available at http://genomequebec.mcgill.ca/publications/pastinen/ (last accessed April 1, 2010). It relies on multiple alignments of sequence traces generated by high sensitivity sequencing. The reference sequence (gDNA) is used to select the peaks that represent the single nucleotide polymorphism (SNP) or variant, as well as control peaks. The program then identifies SNP and control peaks in all sequences, and analyzes the peaks in parallel. The allelic ratio of the gDNA samples are set to 1 as no allele specific expression is expected. However, by concurrently analyzing the gDNA sample in triplicate, we were able to increase the accuracy of this normalization step. A file with SNP allele ratios normalized to gDNA peak heights is generated as the output of this program. Allelic ratios were calculated by dividing the peak height of the variant allele by that of wild-type allele. Three independent PCR and sequencing reactions were carried out from each gDNA and RNA (one reaction from each
independent cDNA) sample and a two-tailed Student’s *t*-test was used to assess statistically significant differences in allelic expression.

3.3.7 Allele specific (Asymmetric) PCR

Primers were designed to amplify the exonic regions surrounding the variant and the c. 655A>G polymorphism. Differential amplification of alleles was achieved by designing two unique reverse primers that contained either of polymorphic bases at its 3’ end. Given the stringent requirement for complementarity at this location, this allowed us to preferentially amplify each of the alleles. PCR was then carried out as described above using a common forward primer, except that twice as much reverse primer was used as the forward primer to ensure successful amplification of the product. PCR products were then purified with Shrimp Alkaline Phosphotase and Exonuclease and sequenced using the Applied Biosystems 3130XL DNA analyzer.

3.3.8 *In silico* analysis of potential effects on splicing

We used the computational programs NNSPLICE (http://www.fruitfly.org/seq_tools/splice.html, last accessed April 13, 2010) and GeneScan (http://genes.mit.edu/GENSCAN.html, last accessed April 13, 2010) to determine the predicted effect of the splice site mutation. We also used the program ESEfinder (http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home, last accessed April 19, 2010), RescueESE (genes.mit.edu/burgelab/rescue-ese/, last accessed April 19, 2010) and PESX (http://cubweb.biology.columbia.edu/pesx/, last accessed April 19, 2010) to assess if these variants created or abolished exonic splice enhancer (ESE) or exonic splice silencer (ESS) sites.[425, 426] Default settings were used for all programs.
<table>
<thead>
<tr>
<th>Variant/Exon</th>
<th>Primers (5’-3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I19F/1</td>
<td>F: GGTGAACATCTAGACGTTTCT&lt;br&gt;R: TTCTCAATCATCTCTTTTGATAGCA</td>
</tr>
<tr>
<td>T117M/4</td>
<td>F: ATATAT CTT TGG CCA GCA TAA GCC&lt;br&gt;R: CTG TAT GCA CAC TTT CCA TC</td>
</tr>
<tr>
<td>I219V/8</td>
<td>F: CAA GGA GAG ACA GTA GCT GA&lt;br&gt;R: CGACTAACAGCATTTCCAA</td>
</tr>
<tr>
<td>Q510X/13</td>
<td>F: TCTGATGTGGAAATGATGGA&lt;br&gt;R: CCTCATGTCCCTGCTCATTA&lt;br&gt;CATGTCCCTGCTCATTAATTTCTTCCT</td>
</tr>
<tr>
<td>K618A/16</td>
<td>F: CAGCACCGCTCTTTGACC&lt;br&gt;R: TTTCCAAAGAGAAATAGTCTGCAA</td>
</tr>
<tr>
<td>L749P/19</td>
<td>F: CCAAACTCCTGGAGTGGGAC&lt;br&gt;R: TCCACAGTGCAATAAATACCA</td>
</tr>
</tbody>
</table>

Primers surrounding the R265C variant could not be designed as it was located to the intron-exon boundary, thus primers surround I219V was used. F: Forward primer, R: Reverse primer
3.4 Results

3.4.1 The truncating variant c. 1528C>T leads to unbalanced allelic expression

In this study we determined the allelic expression of *MLH1* in lymphocytes obtained from individuals carrying missense variants c.55A>T, c.350C>T, c.793C>T, c.1852_1853AA>GC, c.2246T>C and the truncating variant c.1528C>T in *MLH1*. Genomic DNA (gDNA) was used as a reference to establish baseline levels of expression, and cDNA generated from three independent reverse transcription reactions was used to measure mRNA transcript levels. Identical primer pairs were used to PCR amplify the region surrounding the variant in both gDNA and cDNA samples, followed by high sensitivity sequencing of the amplicons (Figure 3.1). The sequence information was then analyzed using the PeakPicker program. This program quantifies peak heights of the two polymorphic bases and uses this information to compute their relative abundance in the sample. As a component of this step this program compares the heights of adjacent peaks to mitigate any effects the efficiency of the sequencing reaction might have on overall peak height. The ratio of two variant bases is set to 1 in the gDNA samples and the heights of the peaks are normalized to the values derived from gDNA. In accordance with previously published studies, we set ≤ 0.5 as a threshold value below which allelic ratios were considered abnormal [427, 428]. This ratio corresponds to a situation where twice as much wild-type transcript is present compared to the transcript carrying the variant base.

Case 4 was diagnosed with CRC at the age of 42. Exonic sequencing indicated that this individual carries the nonsense mutation c.1528C→T in exon 13 of *MLH1*. This mutation led to an allelic ratio of 0.38 (p = 0.0023). Results generated from PeakPicker analysis, as well as clinicopathological variables are summarized in Table 3.2 and Table 3.3 respectively.
3.4.2 The missense variants c.350C>T, C.793C>T and c.1852_1853AA>GC lead to unbalanced allelic expression

Case 2 presented for at-risk testing and is a carrier of the c.350C>T variant, which occurs in exon 4. Her father (the proband) was diagnosed with MSI-H, MLH1 deficient adenocarcinoma of the colon and was found to carry c.350C>T. However, his lymphocyte DNA and RNA was not available for analysis. This variant led to an allelic ratio of 0.19 \((P = 0.015)\) (Table 3.2). Case 2 was also heterozygous for the c.655A>G (p.I219V) polymorphism. We therefore reanalyzed this case using a primer set that flanked this polymorphism to rule out the possibility that the observed allele ratio was primer specific. When PeakPicker analysis was carried out using primers that flanked this polymorphism in exon 8 of MLH1, the allelic ratio was found to be 0.07 \((P = 0.0004)\), which is in agreement with the results obtained using primers that flank the variant in exon 4.

The variant c.793C>T, which occurs in exon 10 was detected in an individual who developed CRC at age 55 (case 3). Given the location of this substitution in the third nucleotide at the beginning of the exon, it was not possible to use identical primers to amplify the region surrounding this variant from both gDNA and cDNA. However, case 3 was heterozygous for the c.655A>G polymorphism, which enabled us to use primers flanking this region to assay its allelic ratio. Using allele, specific PCR we were able to show that c.655A occurs on the wild-type allele (c.793C). PeakPicker analysis demonstrated that this variant was associated with an allelic ratio 0.5 \((P = 0.025)\).

The MLH1 variant c.1852_1853AA>GC was the only genetic alteration discovered in a proband that developed small bowel cancer at the age of 34 years (Case 5A). This variant represents a two nucleotide substitution in exon 16 of MLH1. PeakPicker analysis revealed that the proband’s allelic ratio was reduced to 0.05 \((P = 0.0009)\) (Figure 3.1). This proband was also
heterozygous for a common MLH1 polymorphism c.655A>G (where c.655A occurs on the wild-type allele). Using primers that flank this polymorphic region the MLH1 allelic ratio was determined to be 0.13 ($P = 0.0004$), confirming our previous finding. This proband’s father (case 5B), who presented for at-risk testing, is also a carrier of c.1852_1853AA>GC variant. By PeakPicker analysis case 5B’s MLH1 allelic ratio was demonstrated to be 0.5 ($P = 0.0135$). In keeping with the results from case 5A, this variant appears to be associated with an allelic imbalance. However, as this individual was not heterozygous for the c.655A>G polymorphism we were unable to repeat the analysis as with case 5A. The predicted effects of these alterations on splicing are summarized in table 3.4.

3.4.3 The missense variants c. 55A>T and c. 2246T>C do not lead to unbalanced allelic expression

By exonic sequencing we identified the missense variant c.55A>T in exon 1 of an individual who developed endometrial cancer at the age of 47 (case 1). Results from the PeakPicker analysis demonstrated that the allelic expression of the variant allele was 0.86 ($p = 0.564$) (Table 3.2), which was not found to be statistically significant when assessed using a two-tailed Student’s $t$-test.

The c.2246T>C variant was discovered in a proband who developed colorectal cancer at the age of 24 (case 6). This variant occurs in exon 19 of MLH1 and led to an allelic ratio of 1.02 ($p = 0.8253$, Figure 3.2).
Figure 3.1 Representative chromatograms generated by high sensitivity sequencing.

The left panels (A, C) indicate genomic DNA (gDNA) and the right panels (B, D) indicate complementary DNA (cDNA). Case 5A demonstrates decreased levels of the variant (GC) allele relative to the wild-type (AA) allele, while case 6 shows no change in the levels of the variant (C) allele relative to the wild-type (T) allele. Arrows highlight the heterozygous peaks. Both chromatograms shown were generated using the respective forward primer. (Blue: C, Green: A, Red: T, Black: G)
<table>
<thead>
<tr>
<th>Case</th>
<th>Sample</th>
<th>Av. variant/wt peak height&lt;sup&gt;A&lt;/sup&gt;</th>
<th>95% CI</th>
<th>Allele Ratio</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>gDNA</td>
<td>1</td>
<td>(0.96-1.05)</td>
<td>0.85</td>
<td>0.5640</td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.86</td>
<td>(0.47-1.25)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>gDNA</td>
<td>1.02</td>
<td>(0.82-1.23)</td>
<td>0.19</td>
<td><strong>0.0155</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.2</td>
<td>(0.17-0.22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2#</td>
<td>gDNA</td>
<td>1</td>
<td>NC</td>
<td>0.07</td>
<td><strong>0.0004</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.07</td>
<td>(0.04-0.11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>gDNA</td>
<td>1</td>
<td>(0.97-1.05)</td>
<td>0.5</td>
<td><strong>0.0256</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.5</td>
<td>(0.32-0.56)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>gDNA</td>
<td>1</td>
<td>(0.97-1.03)</td>
<td>0.38</td>
<td><strong>0.0023</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.38</td>
<td>(0.31-0.46)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5A</td>
<td>gDNA</td>
<td>1</td>
<td>(0.94-1.07)</td>
<td>0.05</td>
<td><strong>0.0009</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.05</td>
<td>(0.03-0.06)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5A#</td>
<td>gDNA</td>
<td>1</td>
<td>(0.97-1.03)</td>
<td>0.13</td>
<td><strong>0.0004</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.13</td>
<td>(0.12-0.13)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5B</td>
<td>gDNA</td>
<td>1</td>
<td>(0.93-1.08)</td>
<td>0.5</td>
<td><strong>0.0135</strong></td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>0.5</td>
<td>(0.39-0.61)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>gDNA</td>
<td>1</td>
<td>(0.98-1.02)</td>
<td>1.02</td>
<td>0.8250</td>
</tr>
<tr>
<td></td>
<td>cDNA</td>
<td>1.02</td>
<td>(0.9-1.13)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>A</sup> Average height of variant peak/ height of wild-type peak

Allele ratios = Average height of variant peak/ height of wild-type peak in cDNA

Average height of variant peak/ height of wild-type peak in gDNA

P-values are calculated using a two-sided Student’s t-test. P-values ≤ 0.05 are shown in bold.

All PCR reactions were carried out with primers that flank the region surrounding the variant, unless indicated by a # sign, which denotes results obtained using primers that flank the polymorphic region on Exon 8. CI: Confidence interval, WT: wild-type, NC: Not computable as replicates generated identical values.
Table 3.3 Summary of clinicopathological characteristics

<table>
<thead>
<tr>
<th>Case</th>
<th>Alteration</th>
<th>Predicted effect on protein</th>
<th>Clinical diagnosis</th>
<th>MLH1 IHC Status</th>
<th>Tumour MSI status</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>c.55A&gt;T</td>
<td>Ile19Phe</td>
<td>Endometrial cancer at 47</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>2</td>
<td>c.350C&gt;T</td>
<td>Thr117Met</td>
<td></td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>3</td>
<td>c.793C&gt;T</td>
<td>Arg265Cys</td>
<td>CRC at 55</td>
<td>Deficient</td>
<td>MSI-H</td>
</tr>
<tr>
<td>4</td>
<td>c.1528C&gt;T</td>
<td>Gln510Ter</td>
<td>CRC at 42</td>
<td>Deficient</td>
<td>MSI-H</td>
</tr>
<tr>
<td>5A</td>
<td>c.1852_1853AA&gt;GC</td>
<td>Lys618Ala</td>
<td>Small bowel cancer at 34</td>
<td>Intact</td>
<td>MSI-H</td>
</tr>
<tr>
<td>5B</td>
<td>c.1852_1853AA&gt;GC</td>
<td>Lys618Ala</td>
<td>At-risk testing</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>6</td>
<td>c.2246T&gt;C</td>
<td>Leu749Pro</td>
<td>CRC at 24</td>
<td>Intact</td>
<td>MSI-H</td>
</tr>
</tbody>
</table>

*Case 2’s father (the proband) carried the same variant and presented with a MLH1 deficient, MSI-H tumour but his DNA/RNA was not available for analysis; 5A and 5B are members of the same kindred. IHC: immunohistochemistry; Intact: intact expression of MLH1 by IHC; Deficient: MLH1 expression absent by IHC; MSI-H: Microsatellite instability-high; ND: Not determined (No access to tumour sample); CRC: colorectal cancer; N/A: Not applicable.*
Table 3.4 Predicted effects of missense variants on splicing

<table>
<thead>
<tr>
<th>MLH1 Variant</th>
<th>AI</th>
<th>Effect on Splicing</th>
<th>Rescue ESE&lt;sup&gt;a&lt;/sup&gt;</th>
<th>PESX&lt;sup&gt;b&lt;/sup&gt;</th>
<th>ESEfinder&lt;sup&gt;c&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>c.55A&gt;T</td>
<td>No</td>
<td>NC</td>
<td>NC</td>
<td>-</td>
<td>3.07/-</td>
</tr>
<tr>
<td>c.350C&gt;T*</td>
<td>Yes</td>
<td>4/2</td>
<td>ESE</td>
<td>4.60/3.06</td>
<td></td>
</tr>
<tr>
<td>c.793C&gt;T*</td>
<td>Yes</td>
<td>NC</td>
<td>NC</td>
<td>-</td>
<td>3.85/4.21</td>
</tr>
<tr>
<td>c.1852_1853</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AA&gt;GC*</td>
<td>Yes</td>
<td>4/1</td>
<td>ESE</td>
<td>2.38/-/2.86</td>
<td></td>
</tr>
<tr>
<td>c.2256T&gt;C</td>
<td>No</td>
<td>NC</td>
<td>NC</td>
<td>3.04/3.8</td>
<td></td>
</tr>
</tbody>
</table>

* These in silico predictions have been published in previous reports that assessed the impact of unclassified variants on splicing [429, 430]

<sup>a</sup>The number of ESE motifs predicted by Rescue ESE for the wild-type sequence/variant sequence
<sup>b</sup>The number of ESE/ESS motifs predicted by PESX for the wild-type sequence/variant sequence
<sup>c</sup>The score for ESE motifs recognized by 4 SR proteins as predicted by ESEFinder for wild-type/variant
<sup>d</sup>This variant has been reported to lead to the skipping of exon 10 [430]

AI: Allelic imbalance, defined as having an allelic ratio ≤0.5 by PeakPicker analysis; NC: No change predicted;
3.5 Discussion

The discovery of a predisposing mutation in Lynch syndrome families and the concomitant increase in surveillance has been shown to significantly reduce mortality [251, 431]. Therefore it is of paramount importance to classify the pathogenicity associated with unclassified variants to determine their role in cancer predisposition. PeakPicker analysis provides a relatively rapid and robust method of differentiating potentially pathogenic variants that lead to unbalanced allelic expression. Furthermore, no specialized tools or complex technical skills are required to perform this analysis, thus this technique has the potential to be implemented in a diagnostic setting for the routine assessment of unclassified genetic variants.

The c.55A→T variant has been reported in two previous independent studies; however, no functional studies have been carried out to date [398, 399]. Although PeakPicker analysis demonstrated that this variant generated an allelic ratio that was less than 1, this value was not found to be statistically significant. In this study considered an allelic ratio ≤ 0.5, combined with statistical significance to be indicative of altered allelic expression/stability. A previous study documented that the normal range of allelic variation in MLH1 is between 0.81 and 1.23[432], which is in line with our observation that this variant likely does not lead to an allelic imbalance.

It is possible that some genetic variants may only mildly affect mRNA levels and thus lead to intermediate allelic ratios (1 - 0.5). The accurate interpretation of these variants in the clinical setting will require that the arbitrary cut off established in this study be considered in the light of other available evidence (i.e. MSI/IHC information of the proband’s tumour). In this case we did not have supplementary data from MSI and IHC. However, we observed expression levels comparable to that of wild-type MLH1 protein by in vitro expression assays (Figure 2.4). Thus, combined these observations suggest that altered mRNA expression/stability is unlikely to be the
mechanism by which this variant mediates its pathogenicity but highlights the utility of PeakPicker analysis in cases where incomplete clinical information is present.

Based on phylogenetic, structural and functional evidence, as well as segregation data the c.350 C>T variant is usually considered to be a deleterious mutation [430]. This variant has been shown to affect the efficiency of MMR function by several studies [413, 433]. However by conversion analysis technology this variant (detected in a proband from a separate family) did not show decreased allelic expression [234]. Conversion analysis separates alleles by converting the human chromosome complement to a haploid state by fusing it with recipient E2 mouse cell line. The resulting hybrid cell lines contain only a subset of human chromosomes and are used to assay cDNA sequence changes. While the benefit of this approach is that it allows the maternal and paternal alleles to be analyzed independently, the hybrid nature of the cell may affect the regulation and expression of mRNA. Thus, this discrepancy may have arisen due to the differences between these two methodologies, another possibility is that other unique genetic influences may have led to the distinct allelic profiles observed in the two individuals.

The c.793C>T variant have been reported to lead to reduced MMR efficiency in several assays [401, 413, 433, 434]. Furthermore, in experiments in carries out in our laboratory we were able to show that this variant decreased the stability of the MLH1 protein [389]. This variant has also been associated with abnormal splicing, with two independent studies demonstrating that it induced partial skipping of exon 10 (Table 3.4) [429, 430]. Given its proximity to the intron-exon border this highlights the role of cis-acting elements in flanking sequences in regulating splicing and defining exons. Interestingly, this variant also decreased the expression of the variant allele when measured using conversion analysis [234]. Given the distinctions of the two
assays, this finding provides evidence of the sensitivity and accuracy of the PeakPicker methodology.

Previous functional studies carried out with the c.1852_1853AA→GC variant have produced conflicting pieces of evidence regarding its pathogenicity [383], making it difficult to ascertain if c.1852_1853AA→GC is a pathogenic or benign variant. In a previous study using pulse chase assays and cycloheximide treatments we demonstrated that the c.1852_1853AA→GC variant drastically decreases the stability of the MLH1 protein (Figure 2.6). However, these assays are time consuming and therefore not practical to be applied to the clinical setting. Furthermore, results from in vitro findings do not always mirror the effect of the variant in vivo. The results of the PeakPicker analysis demonstrate that this substitution significantly decreases the expression/stability of the variant allele in the individual’s lymphocyte RNA. The IHC findings indicate that this individual’s tumour was positive for MLH1 expression. This may be due to the fact that the variant leads to intermediate levels of expression, which cannot be accurately determined by a qualitative method such as IHC. Another possibility is that the mutation that affected the second allele of MLH1 may have compromised the functionality of the protein without altering its expression (for example by affecting the ability of MLH1 to bind to PMS2 or to translocate to the nucleus). The differential effect we see in these two family members could be attributed to unique genetic changes or environmental exposures that might have influenced the levels of MLH1 transcribed or the stability of this transcript. The pedigree of this family is depicted in Figure 3.2. However, it is not possible to rule out if the unbalanced allelic expression profile we observed is the result of a hitherto undiscovered mutation that lies in regulatory regions or deep within intronic regions that is in linkage with the variant allele.
The c.2246T→C variant was discovered in a proband that developed colorectal cancer at the age of 24. The pedigree of this proband is depicted in Figure 3.3. This variant lies within the c-terminal homology domain that is thought to be important in stabilizing interaction between MLH1 and PMS2[404] and while this variant has been previously reported, no functional studies have been done to date [411, 412, 435]. Previous studies carried out in our laboratory have demonstrated that another missense variant affecting the same nucleotide, MLH1 c.2246 T→A did not affect protein expression or stability but showed decreased efficiency of heterodimerizing with PMS2 [389]. Thus, it is likely that this variant affects heterodimerization of MLH1 rather than its expression or stability.
Figure 3.2 Pedigree of Case 5A/5B.

The proband in this pedigree was found to carry the unclassified variant K618A. Solid shading denotes cancer at the time these pedigrees were drawn. The proband is indicated with an arrow. V+, indicates individuals positive for variant.
Figure 3.3 Pedigree of case 6.

The proband in this pedigree was found to carry the unclassified variant L749P. Solid shading denotes cancer at the time these pedigrees were drawn. The proband is indicated with an arrow.
Truncating mutations are the most common type of mutation identified in both the *MLH1* and the *MSH2* genes [436]. Truncating mutations lead to transcripts with premature termination codons that are generally degraded by nonsense mediated RNA decay, which functions to eliminate truncated proteins that may have deleterious effects [437]. However, depending on the location of the termination codon, truncating mutations may or may not lead to decreased mRNA. The finding that the truncating mutation, c.1528 C→T leads to diminished allelic expression/stability, confirms PeakPicker’s sensitivity and its versatility in recognizing the effects of diverse types of genetic alterations.

Unclassified variants in *MLH1* have been shown to alter protein expression, stability and turnover [383, 389]. However, not much attention has been paid to the role of unclassified variants in leading to allelic imbalances despite its potential to be a noteworthy cause of MMR deficiency. Allelic imbalances could arise due to genetic alterations that affect the efficiency of transcription, alter transcript expression or affect mRNA stability and this study provides initial evidence that a subset of unclassified missense variants in *MLH1* can and do affect these processes. Similarly, a previous study showed that a missense variant in *RB1* was able to significantly impair the allelic balance in a pedigree with retinoblastoma [421]. While the extent of the effect may depend on the missense alteration in question, it offers another important clue regarding the pathogenicity of ambiguous variants. Missense variants may alter transcript expression or stability by leading to incorrect processing or folding of the transcripts.

Another mechanism by which variants may potentially contribute to altered mRNA expression is through abnormal splicing, and several missense mutations in *MLH1* that result in splicing abnormalities have been shown to be associated with RNA defects [438]. Disruption in the fidelity of splicing can create unstable mRNAs by activating the NMD system. While none of
the variants used in this study occur in the invariant splice donor and acceptor sites, several other loosely defined cis-acting sequences have been shown to be important for accurate exon recognition [429]. Thus, the disruption of these sequences is a likely mechanism by which unclassified variants may lead to disease susceptibility. An example of this is the c.793C>T variant, which has been shown to lead to exon skipping in several studies. However, the splicing prediction tool, NNSPLICE and the gene prediction program, GeneScan did not predict that these variants would lead to splicing abnormalities. In addition, previous studies have shown that pathogenic missense mutations in the MLH1 and MSH2 genes are located in exonic splice enhancer (ESE) sites significantly more frequently than expected [439]. ESEs are short (6-8 base pair) degenerate exonic splicing regulatory sequences that serve as the binding sites for serine/arginine-rich (S/R) proteins. Therefore, alteration of ESEs offers another mechanism by which these variants may potentially lead to splicing defects. The three in silico approaches we employed provided dissimilar predictions (Table 3.4). Our analysis indicates that several of the variants are predicted to lead to the creation, loss and/or the strengthening of putative ESEs. However, given the redundancy and the complexity of the regulation of splicing, further studies are required to confirm these predictions as previous studies have indicated that these tools have limited specificity and sensitivity [429, 430]. Nevertheless, this analysis offers some insight into the mechanisms that may have led to altered allelic ratios.

A limitation of the PeakPicker strategy is that it is difficult to assay variants that occur close to exon boundaries or those that occur in relatively small (<80 base pair) exons. Furthermore, heterozygosity of the variant allele is required to carry out this analysis. However, these issues can be circumvented if the individual carrying the unclassified variant is heterozygous for a known, common polymorphism such as c.655A>G. This polymorphism is suitable for PeakPicker analysis as it has the highest level of heterozygosity among exonic MLH1
polymorphisms [432], and occurs within exon 9 of MLH1 away from intron/exon boundaries. In this situation, primers flanking the polymorphic region can be used to confirm the effect of the uncharacterized variant. This is more cost efficient than designing unique primers surrounding each exon, and can be adopted when a patient population shows a high frequency of a common SNP. Alternatively in the case of short exonic regions, it may also be possible to use primers with linker sequences so that useful sequence information is not lost in the process of sequencing. A short linker sequence was added on to the primers flanking c.350 C>T, which enabled us to assay this variant, which occurs in a relatively short exon. Importantly, PeakPicker analysis cannot be used as a stand-alone test, as there may be several variants that do not lead to allelic imbalances but may possibly affect other functions. Thus, this analysis should be used in conjunction with a panel of assays that will help determine the putative effect of genetic variants on expression and/or function. Additionally, this should be combined with other relevant clinical and family history information indicating tumour MMR deficiency status to obtain a more accurate understanding of a variant’s pathogenicity.

Another strength of the PeakPicker analysis is that it will aid in discovering mutations in individuals that are classified as mutation-negative using conventional diagnostic techniques. A major problem that arises in the detection of mutations is the possible masking of alterations in one allele by the normal sequence present on the other allele. Techniques such as conversion technology and methods based on single nucleotide primer extension are able to analyze the maternal and paternal alleles separately and have been applied to the detection of MMR alterations [234, 427, 432]. However, conversion technology is cumbersome and cannot be readily implemented in the diagnostic setting. Using an allelic expression assay based on single nucleotide primer extension, Renkonen and colleagues were able to identify the MLH1 R100X nonsense mutation in a mutation-negative family [427]. PeakPicker analysis, which is very
similar in principal to single nucleotide primer extension assays, offers a very efficient method of
detecting hitherto unidentified genetic changes by measuring imbalances in allelic expression.
Importantly, PeakPicker analysis and the benefits it offers can be extended to the study of other
genes implicated in inherited cancers, as well as other diseases. BRCA1 and BRCA2 are examples
of two such genes that can benefit from this type of analysis as the Breast cancer Information
Core (BIC) database lists about 1500 variants, mostly missense substitutions of unknown clinical
significance [440].

In summary, we demonstrate that the publicly available software tool, PeakPicker was
able to identify MLH1 variants that led to unbalanced expression of mRNA in a time and cost
efficient manner, making it amenable to implementation in the diagnostic setting. Information on
decreased transcript levels, in combination with other assays will help increase the confidence
with which a variant is deemed to be pathogenic, especially if limited clinical data is present.
This technique essentially generates a snapshot of the gene expression profile present in the
individual’s cells; therefore unlike in vitro assessment strategies this assay takes into
consideration the unique environmental or genetic influences that may be specific to the
individual. The PeakPicker strategy will also open avenues to further analyze mutation-negative
cases. Moreover the strategy described here can also be utilized to help classify genetic
alterations identified in several hereditary diseases, allowing for the improved accuracy and
efficiency of predictive genetic testing programs.

Acknowledgements

We thank the patients and their families for participating in this study. We are grateful to Dr.
Tomi Pastinen (McGill University, Montreal, QC, Canada) and Dr. Bing Ge for providing us
with the PeakPicker software and for helping us set up the program. We also gratefully
acknowledge the assistance provided by the staff of the Molecular Pathology Laboratory, the Biospecimen Repository and the Familial GI Cancer Registry of the Mt. Sinai Hospital for their help with this study. In addition, we would like to thank Siwei Chen and the staff of the Analytical Genetics Technology Centre (University Health Network) for assistance with high sensitivity sequencing.
CHAPTER 4

Functional Effects of the MLH1-93G>A Polymorphism on MLH1 and EPM2AIP1 Promoter Activity

Sheron Perera, Miralem Mrkonjic, James B. Rawson and Bharati Bapat.

Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, ON
Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, ON
Department of Pathology and Laboratory Medicine, Mount Sinai Hospital, Toronto, ON

The work presented in this chapter was primarily contributed by Sheron Perera. MM, a former Ph.D. student, and JBR, a summer student, assisted in the creation of the promoter constructs.

A modified version of the data presented in this chapter has been submitted for publication.
Chapter 4 Functional Effects of the MLH1-93G>A Polymorphism on MLH1 and EPM2AIP1 Promoter Activity

4.1 Summary

Defective mismatch repair (MMR) leads to the microsatellite instability (MSI) phenotype of colorectal cancer (CRC). We have previously shown that the MLH1 c.-93G>A promoter polymorphism is strongly associated with MSI tumours, suggesting a modifier role for this polymorphism in CRC. The MLH1 promoter is bi-directional with the EPM2AIP1 gene located on the antisense strand. To evaluate the functional impact of this polymorphism we transfected a panel of CRC and endometrial cancer cell lines with luciferase promoter constructs containing either the -93G or -93A allele. We transfected a core MLH1 promoter construct that spans nucleotides -301 to +99, and a shorter MLH1 promoter construct that spans nucleotides -113 to +99. Furthermore, we used a construct in reverse orientation to assess the effect of this polymorphism on EPM2AIP1. The luciferase activity was compared using a two-sided Student’s t-test. Using electrophoretic mobility shift assays (EMSA), we evaluated whether differential protein binding was responsible for differences in promoter activity. With the core promoter constructs, we observed higher activity with the -93G allele compared to the -93A allele in HCT116 colorectal cancer cells. With the shorter promoter constructs, we consistently observed higher level of activity with the -93G allele in all cell lines examined. This polymorphism also affected the transcription of the EPM2AIP1 gene. EMSA suggests that this polymorphism alters the affinity of nuclear factors that bind this region. This indicates that the -93G>A polymorphism modifies the efficiency of MLH1/EPM2AIP1 transcription.
4.2 Introduction

With over a million cases diagnosed each year, colorectal cancer (CRC) is a significant cause of morbidity and mortality worldwide. CRCs can be generally subdivided into two clinically relevant groups depending on their mismatch repair (MMR) status. The DNA mismatch repair system guards the integrity of the genome and contributes to the overall fidelity of DNA replication by targeting mispaired bases and insertion-deletion loops that occur through replication errors, during homologous recombination, and as a result of DNA damage [441]. Defective MMR is also implicated in Lynch syndrome, a cancer predisposition syndrome that accounts for 2-3% of CRCs. MMR deficiency has also been shown to give rise to sporadic endometrial cancer with 28% of all endometrial tumours displaying loss of MMR [245]. Moreover, endometrial cancer is also the most common extra-colonic malignancy observed in Lynch syndrome [442].

Cancers deficient in MMR exhibit genome-wide instability at microsatellite sequences compared to matched normal tissue DNA. When 30% or more of microsatellite markers tested show instability, the tumour is defined as high-frequency microsatellite instability or MSI-H. About 15% of sporadic CRCs and >85% of Lynch syndrome tumours display MSI-H phenotype [227, 443]. Furthermore, the majority of MSI-H CRCs occur due to the epigenetic silencing of the MLH1 promoter [444].

Recent evidence has shown the MLH1 promoter to be bi-directional with a second gene, EPM2AIP1, located in a head-to-head orientation on the opposite strand 321 base pairs away from the MLH1 transcription start site [342]. Methylation of this promoter region results in transcriptional silencing of both MLH1 and EPM2AIP1 [342]. Furthermore, MSI-H tumours show a significant reduction in MLH1/EPM2AIP1 expression compared with normal colonic
mucosa and microsatellite stable (MSS) tumours indicating that EPM2AIP1 expression is turned off in the majority of sporadic tumours with MLH1 hypermethylation [445]. Although these findings potentially implicate EPM2AIP1 in CRC, its function is unknown. The MLH1 gene promoter has been empirically subdivided into four regions based on CpG dinucleotide clusters and the relationship between their methylation status and MLH1 gene expression. These regions are termed: A (-711 to -577 base pairs from the MLH1 translation start site), B (-552 to -266), C (-248 to -178), and D (-109 to +15) [341]. The core promoter region consists of regions C and D, since methylation of this region is consistently correlated with loss of MLH1 expression in CRC tumours [339].

A previous genetic epidemiology study carried out in our laboratory aimed at assessing common single nucleotide polymorphisms identified an association between the MLH1-93G>A promoter variant and MSI-H CRCs in two separate Canadian populations from the provinces of Ontario and Newfoundland [446]. Since then these findings have been independently replicated by other studies as this polymorphism was shown to be associated with increased risk of hyperplastic polyps and adenomas in smokers [447], as well as MSI-H CRCs [448]. The MLH1-93G>A variant was also shown to be associated with CRCs that display loss of MLH1 protein expression [449]. Furthermore, the MLH1-93G>A variant was shown to increase the risk of developing endometrial cancer [450]. The correlation between the -93G>A SNP and cancer has not been limited to CRC or endometrial cancer, with previous studies reporting an association with increased risk of ovarian [451] and squamous cell lung cancer [452]. Taken together, these studies provide evidence that the -93G>A variant is likely to mediate functional effects promoting carcinogenesis. However, there has been no systematic investigation of the functional contribution of this variant to date, and the limited data that exists are conflicting [453, 454]. Furthermore, no study has evaluated cell-type specific influences of this variant, which are
important given differences in transcription factor expression and regulation in diverse host cell
types [455].

We carried out a comprehensive investigation of the role of the -93G>A variant in
altering MLH1 transcription in a subset of cell lines representative of colorectal and endometrial
cancers. In addition, we extended our analysis to determine if this variant affects transcription of
the EPM2AIP1 gene, in order to elucidate a role in gene in carcinogenesis.

4.3 Materials and Methods

4.3.1 Generation of Promoter Constructs

pGL3-Basic plasmids containing the C+D region (-301 to +99) of the MLH1 promoter
(ENST00000231790) were gifts from Dr. Hiromichi Hemmi (Toho University School of
Medicine, Japan). These plasmids, which contain either the G or A nucleotide at the -93 location,
have been described previously [454].

To generate plasmids of the MLH1 promoter D-region with the G or A allele, DNA
fragments corresponding to the MLH1 promoter D region (-113 to +99) were PCR amplified
from the appropriate C+D promoter constructs. The PCR reaction contained 10X Taq Buffer
HiFi, 2 mM MgCl₂, 0.2 mM dNTPs, 1 mM primers, 14 ng template DNA, and 1 unit high-
fidelity Taq DNA polymerase (Fermentas Life Sciences, Burlington, ON, Canada). The PCR
conditions were as follows: initial denaturation (95°C for 10 minutes), denaturation( 95°C for 30
seconds), annealing (57°C for 45 sec), extension (72°C for 1 min), and final extension (72°C for
8 min). The denaturation-annealing-extension steps were repeated 35 times. The 5’ to 3’
sequences of the forward and reverse primers, respectively, are:
To generate plasmids with the EPM2AIP1 promoter G or A allele, DNA fragments corresponding to the -513 to +122 region of the MLH1/EPM2AIP1 promoter were amplified from lymphocyte DNA of CRC patients homozygous (for G or A) at the -93 locus. The PCR conditions were identical to that described for the MLH1 promoter D-region except that 40 ng of template DNA and 2.5 mM MgCl₂ were used. The 5’ to 3’ sequences of the forward and reverse primers, respectively, are: CCTCGTCGACTTCCATCTTGCTTCTTTT and CCGTACCAGTTCTCAATCATCTCTTTGAT. A second, nested PCR was performed using the PCR fragment generated above as templates. The PCR conditions were identical to the above reaction except that 20 ng of template DNA was used. The 5’ to 3’ sequences of the forward and reverse primers, respectively, are: AAAAAAAGAGCTTCCACAACGCGTCCGTCGTAAGCTA (XhoI site underlined) and AAAAAAAGAGCTTCCACAACGCGTCCGTCGTAAGCTA (HindIII site underlined). Following all PCR reactions, DNA was isolated by gel electrophoresis and purified by QIAquick Gel Extraction Kit (Qiagen, Mississauga, ON, Canada). MLH1 D-region promoter constructs, EPM2AIP1 promoter constructs, and pGL3-Basic vector were digested with XhoI and HindIII restriction endonucleases (Fermentas). The restriction digestion reaction used 10X Buffer R with BSA, 12.5 units of each restriction enzyme, and a 3:1 insert to vector ratio. The reaction mix was incubated at 37°C for 18 hrs. A second round of restriction enzymes was added at 8 hrs to ensure maximum digestion. Following the digestion, 5μl of shrimp alkaline phosphatase (SAP) was added to prevent religation of the plasmid and the mixes were incubated at 37°C for 1 hr. Following enzyme inactivation at 80°C for 20 min, DNA was purified using QIAQuick PCR Purification Kit (Qiagen). The integrity of all constructs was verified by DNA sequencing.
Figure 4.1 The schematic of the MLH1/EPM2AIP1 bi-directional promoter.

The MLH1 promoter is empirically divided into four regions A-D as shown [341]. The nucleotides demarcating the start and end of each region are indicated (numbering is with reference to the MLH1 translational start site). The core promoter region, which comprises the C and D regions, is shared between the MLH1 and EPM2AIP1 genes. The translational start sites of MLH1 and EPM2AIP1 are shown with arrows. The promoter constructs used in luciferase reporter assays are indicated below. The C+D construct extends from -301 to +99, the D construct from -113 to +99 and the EPM2AIP1 construct from -513 to+122. The asterisk represents the -93G>A polymorphism located in D-region. The figure is not drawn to scale.
4.3.2 Cell Culture

Tissue culture reagents and fetal bovine serum (FBS) were purchased from Invitrogen Life Technologies (Burlington, ON, Canada). HCT116 and HEC-1-A cells were cultured in McCoy’s 5A media. SK-UT-1B was maintained in Eagle’s minimum essential medium (ATCC, Manassas, VA). All cell culture media was supplemented with 10% FBS. All cell lines were incubated at 37°C in a humidified atmosphere of 5% CO₂.

4.3.3 Luciferase Reporter Gene Assays

All transfection experiments were carried out using Lipofectamine 2000 reagent (Invitrogen Life Technologies) according to manufacturer’s instructions. Cells were plated in 24-well plates at a density of 100,000 cells per well 24 hours prior to transfection. The promoter-less pGL3-Basic (empty vector) was used as a negative control. Luciferase constructs (500 ng) containing either pGL3-Basic, MLH1 C+D promoter constructs (-93G or A), MLH1 D promoter constructs (-93G or A), or EPM2AIP1 promoter constructs (-93G or A), were transfected into cells to determine luciferase reporter activity [454, 456]. Cells were additionally co-transfected with an internal control (5 ng of pRL-SV40, Promega, Madison, WI) for monitoring transfection efficiency and normalization. Luciferase and Renilla activities were measured 24 hr after transfection using the Dual Luciferase Reporter Assay System (Promega), which allows the firefly and Renilla luciferase activities to be quantified in the same sample. Luminescence was quantified using a Berthold 96-well microplate luminometer (Berthold, Wildbad, Germany). All reporter assays were done at least in triplicate, with at least three independent experiments performed. Statistically significant differences in promoter activity were assessed using a two sided Student’s t-test. Error bars represent standard deviation of the mean (SD).
4.3.4 Western blotting

Cold radioimmunoprecipitation assay (RIPA) lysis buffer (0.5 % Nonidet P-40, 0.5% sodium deoxycholate, 0.1% sodium dodecyl sulfate, 1x PBS) buffer containing Complete protease inhibitor tablets (Roche, Mannheim, Germany) was used to lyse the cells on ice for approximately 20 minutes. The lysates were then collected by centrifugation at 14,000 rpm for approximately 15 minutes. The expression levels for the proteins were examined by sodium dodecyl sulphate/ polyacrylamide gel electrophoresis, followed by western blot analysis with anti-MLH1 (clone 168-728; BD Biosciences) and anti-EPM2AIP1 (developed by Julie Turnbull and Dr. Berge Minassian, University of Toronto)

4.3.5 Electrophoretic Mobility Shift Assays (EMSA)

The following double stranded probes (variant shown in bold within brackets) that were labeled with biotin on the 5’ end were purchased from Integrated DNA Technologies (Coralville, IA): -93G sense, 5’-TAAGCTACAGCT[G]AAGGAAGAAGTG-3’; -93G antisense, 5’CACGTTTCTTCTT[C]AGCTGTAGCTTA-3’; -93A sense, TAAGCTACAGCT[A]AAGGAAGAAGTG-3’; -93A antisense, 5’-CACGTTTCTTCTT[T]AGCTGTAGCTTA-3’. The probes generated for the competition reaction were identical, except they did not contain a 5’ biotin group.

Nuclear extracts of the cell lines were prepared using the NE-PER kit (Pierce, Rockford, IL) according to the manufacturer’s instructions. EMSA was carried out using the LightShift Chemiluminescent EMSA Kit (Thermo-Fisher, Rockford, IL) according to the instructions of the manufacturer. We incubated 15 µg of nuclear proteins from the HCT116 cell line and 8 µg of nuclear protein from the HEC-1-A cell line with 60 fmol biotin-labeled oligonucleotide for 25 min at room temperature in binding buffer (10 mM Tris, 50 Mm KCl, 1 mM dithiothreitol, pH
7.5), 50 ng/µl poly (dI-dC), 0.05% NP40, and 3.125% glycerol. Binding was competed by 200, 400, 800 and 2000-fold excess unlabeled oligonucleotides. Binding complexes were resolved by electrophoresis using 5% TBE Criterion gels (Bio-Rad, Hercules, CA), transferred to Biodyne B pre-cut modified nylon membranes (Pierce), UV cross-linked, and visualized using the Chemiluminescent Nucleic Acid Detection system (Pierce). The EMSAs were performed at least at least three times independently.

Supershift experiments were conducted using specific antibodies (rabbit anti-DEC2 and anti-Max Santa Cruz Biotechnology, Santa Cruz, CA). Reactions were identical to gel shift reaction conditions except 2µg or 4µg of antibody along with the binding reaction mixtures was incubated for 60 minutes on ice before the addition of labeled probe. Once the probe was added the reaction mixtures were incubated for 25min at room temperature prior to electrophoresis as described above.

4.4 Results

4.4.1 Effects of the MLH1 -93G>A variant on promoter activity in colorectal cancer cell lines

We examined the effects of the MLH1-93G>A promoter polymorphism on MLH1 promoter activity using two different constructs, one that spans a core promoter C+D region (-301 to +99) and another that spans only the shorter D region of the promoter (-113 to +99; Figure 4.1). In the HCT116 cell line, the -93G allele shows ~1.28 fold higher transcriptional activity of MLH1 compared to the -93A allele when transfected with the core promoter construct (P = 0.0009; Figure 4.2). When the D construct was transfected, the activity of the -93G allele was ~5 fold higher than the -93A allele (P = 0.0025). Promoter activities of all constructs are summarized in Table 4.1.
4.4.2 Effects of the \textit{MLH1} -93G>A variant on promoter activity in endometrial cancer cell lines

We assessed the effect of this polymorphism on the transcriptional activity of \textit{MLH1} in the human endometrial cancer cell lines, HEC-1-A and SK-UT-1B (Figure 4.3). For both cell lines, there was no significant difference in luciferase reporter activity between the two alleles when the core promoter construct containing the C+D region was used (Figure 4.3). However, with the D region promoter construct, the -93G allele showed a 4.3 fold ($P = 0.0005$) and a 9.5 fold ($P = 0.001$) higher activity compared to the -93A allele in HEC-1A and SK-UT-1B respectively (Table 4.1).

4.4.3 Characterization of MLH1 and EPM2AIP1 in a panel of human cell lines

\textit{EPM2AIP1} shares the same promoter with \textit{MLH1}. However, neither the function of this gene nor its expression status in colon and endometrial cancer is known. As a preliminary analysis, we determined the expression of EPM2AIP1 in a panel of human cell lines derived from cancers of different sites, as well a non-tumourigenic cell line. Our results suggest that EPM2AIP1 protein is expressed in a range of cell lines as a doublet (Figure 4.4). Furthermore, this expression correlates with the mutation/methylation status of MLH1 in these cell lines. HCT116 cells have a hemizygous nonsense mutation [414], thus its deficient for endogenous MLH1, but expression of EPM2AIP1 is not affected. The SW620 cell line is neither methylated nor mutated (Unpublished observations, Dr. Richard Hamelin, INSERM, France); therefore both MLH1 and EPM2AIP1 are expressed. The shared \textit{MLH1/EPM2AIP1} promoter is hypermethylated in the embryonic kidney cell line, HEK293T [385] making it deficient for both of the proteins. The prostate, endometrial and breast cancer cell lines (PC3, SK-UT-1B and MCF-7 respectively) appear to show expression of both MLH1 and EPM2AIP1.
4.4.4 Effects of the \textit{MLH1} -93G>A variant on \textit{EPM2AIP1} promoter activity

Because of the bi-directional nature of the \textit{MLH1/EPM2AIP1} promoter [342], we evaluated the effects of the -93G>A polymorphism using the reverse promoter constructs (Figure 4.1). In the HCT116 ($P = 0.007$) colorectal cancer cell line the -93G allele showed a statistically significant decrease in promoter activity compared to the -93A allele (Table 4.1, Figure 4.5).

In the endometrial cancer cell line HEC-1-A the -93G allele displayed lower activity than the -93A allele ($P = 0.004$), similar to the colorectal cancer cell lines. However, in the SK-UT-1B cell line this trend was reversed, with the -93G allele showing statistically significantly higher activity ($P = 0.024$, Figure 4.5).
Figure 4.2 Effect of the MLH1-93G>A polymorphism on MLH1 transcriptional activity in the HCT116 cell line.

Effects of the -93G>A polymorphism were evaluated using luciferase constructs encompassing the C+D region (A, C) or the D region (B, D) containing either the G or A allele at the -93 location. Results shown for the -93 G or -93A allele were derived from firefly luciferase activity normalized to Renilla, and expressed relative to the EV plasmid. Each sample was assayed at least in triplicate, with three independent experiments performed. Statistically significant differences in promoter activity were assessed using the Student’s t-test and are indicated by an asterisk. In bar graphs error bars represent standard deviation from the mean, and in dot plots error bars represent 95% confidence interval. EV: empty vector (pGL3-Basic) construct, C+D: construct containing the C and D region of MLH1 promoter, D: construct containing the D region of MLH1 promoter.
Figure 4.3 Effect of the MLH1-93G>A polymorphism on MLH1 transcriptional activity in the endometrial cancer cell lines.

Luciferase constructs containing either the G or A allele at the -93 encompassing the C+D region or the D region were transfected into HEC-1-A (A, B, E, F) and SK-UT-1B (C, D, G, H) cell lines. Results shown for the -93G or -93A allele were derived from firefly luciferase activity normalized to Renilla, and expressed relative to the EV plasmid. Each sample was assayed at least in triplicate, with three independent experiments performed. Statistically significant differences in promoter activity are indicated by an asterisk. In bar graphs error bars represent standard deviation from the mean and in the dot plots error bars represent the 95% confidence interval.
Figure 4.4 Expression profile of endogenous MLH1 and EPM2AIP1 in a panel of human cell lines.

Total extracts were immunoblotted with either anti-MLH1 (upper panel) or anti-EPM2AIP1 (lower panel) (Note that the blot was cut in half and then immunoblotted for the two proteins). HCT116 cells have a hemizygous nonsense mutation[414], thus its deficient for endogenous MLH1, but expression of EPM2AIP1 is not affected. The SW620 cell line is neither methylated nor mutated (Unpublished observations Dr. Richard Hamelin); therefore both MLH1 and EPM2AIP1 are expressed. The shared \textit{MLH1/EPM2AIP1} promoter is hypermethylated in the HEK293T cell line [385], making it deficient for both of the proteins. The prostate, endometrial and breast cancer cell lines (PC3, SK-UT-1B and MCF-7 respectively) appear to show expression of both MLH1 and EPM2AIP1.
Figure 4.5 Effect of the MLH1-93G>A polymorphism on EPM2AIP1 transcriptional activity.

Constructs that span the shared bidirectional promoter, which contain either the A or G nucleotide at the -93 position (relative to *MLH1*) were transfected into HCT116 (A, D), HEC-1-A (B, E), and SK-UT-1B (C, F). Results shown for the -93 G or A allele were derived from firefly luciferase activity normalized to Renilla, and expressed relative to the EV plasmid. Each sample was assayed at least in triplicate, with three independent experiments performed. Statistically significant differences in promoter activity are indicated by an asterisk. In bar graphs error bars represent standard deviation from the mean and in dot plots error bars represent the 95% confidence interval.
### Table 4.1 Summary of promoter activities

#### MLH1 promoter activities

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Construct</th>
<th>Relative Luciferase Activity</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-93G ± SD</td>
<td>95% CI</td>
</tr>
<tr>
<td>HCT116</td>
<td>C+D</td>
<td>90.38±2.99</td>
<td>87-93.76</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>69.38±4.32</td>
<td>64.49-74.27</td>
</tr>
<tr>
<td>HEC-1-A</td>
<td>C+D</td>
<td>138.08±10.11</td>
<td>126.64-149.52</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>15.85±0.34</td>
<td>15.47-16.23</td>
</tr>
<tr>
<td>SK-UT-1B</td>
<td>C+D</td>
<td>100.45±11.21</td>
<td>87.76-113.14</td>
</tr>
<tr>
<td></td>
<td>D</td>
<td>83.52±2.45</td>
<td>80.75-86.29</td>
</tr>
</tbody>
</table>

#### EPM2AIP1 promoter activities

<table>
<thead>
<tr>
<th>Cell Line</th>
<th>Relative Luciferase Activity</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-93G ± SD</td>
<td>95% CI</td>
</tr>
<tr>
<td>HCT116</td>
<td>139.73±7.39</td>
<td>131.37-148.09</td>
</tr>
<tr>
<td>HEC-1-A</td>
<td>26.94±1.14</td>
<td>25.65-28.23</td>
</tr>
<tr>
<td>SK-UT-1B</td>
<td>255.15±11.83</td>
<td>241.76-268.54</td>
</tr>
</tbody>
</table>

n=3 independent experiments with each samples assayed at least in triplicate. P –values are calculated using a two-sided Student’s t-test. SD: standard deviation, CI: confidence interval.
4.4.5 Effect of the *MLH1* -93G>A promoter variant on transcription factor binding

We carried out EMSA to assess whether the differences observed in the luciferase reporter activity between the -93G and the -93A allele were due to differential binding of nuclear factors. These experiments revealed that multiple factors bound the probe. We used increasing amounts of unlabelled probe as a competitor to verify the specificity of binding interactions. In reactions carried out with the nuclear extract of the HCT116 cell line, multiple factors bound to the labeled probe (Figure 4.6). In general, the specificity appeared to be very high with 800-2000x unlabelled probe required to compete out the binding reaction. In addition, these factors appear to have different affinities for the binding sequence and consequently seem to exhibit competitive binding. Addition of the unlabelled competitor probe diminished certain interactions and strengthened others. When the binding patterns with the -93G probe were compared to that of the -93A probe, we observed that one of the factors (arrowhead, Figure 4.6) had higher affinity for the -93G probe. Furthermore, the highest concentration of competitor was unable to compete out binding with this factor, while this was not so with the -93A allele.

To investigate the effect of this polymorphism on a different cell type, EMSA was also performed with nuclear extract from the HEC-1-A cell line (Figure 4.7). Here too, we observed the binding of multiple factors to the probe, as well as competition between the different factors that bound the region. One nuclear factor (arrowhead, Figure 4.7) appeared to have higher affinity for the -93A allele, as it remained bound even in the presence of 2000x unlabelled probe, while with the -93G probe 800x unlabelled probe was sufficient to compete out this reaction. We also observed that the -93A unlabelled probe was unable to compete out the reaction with the -93G labeled probe (and vice versa), which provided further evidence for the specificity of these interactions.
Figure 4.6 Impact of the MLH1 -93 >A polymorphism on nuclear factor binding in HCT116.

At least 3 independent EMSAs were carried out to determine the effects on binding and a representative experiment is depicted. Unlabelled competitor was added in decreasing concentrations (2000x, 800x and 400x respectively) to verify the specificity and the affinity of the interactions. (+) reactions carried out in the presence of nuclear lysate; (-) reactions carried out without nuclear extract or without competitor. The arrowhead highlights factor(s) that appear to show differential binding between the -93A and -93G alleles.
Figure 4.7 Impact of the MLH1 -93 >A polymorphism on nuclear factor binding in HEC-1-A.

At least 3 independent EMSAs were carried out to determine the effects on binding and a representative experiment in depicted. Unlabelled competitor was added in decreasing concentrations (2000x, 800x and 400x respectively) to verify the specificity and the affinity of the interactions. (+) reactions carried out in the presence of nuclear lysate; (-) reactions carried out without nuclear extract or without competitor. The arrowhead highlights factor(s) that show differential binding between the -93A and -93G alleles.
<table>
<thead>
<tr>
<th></th>
<th>-93G</th>
<th>-93A</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unlabelled competitor</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Antibody</td>
<td>- 3 4</td>
<td>- 3 4</td>
</tr>
<tr>
<td><strong>B</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unlabelled competitor</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Antibody</td>
<td>- 1 2</td>
<td>- 1 2</td>
</tr>
<tr>
<td><strong>C</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Extract</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Unlabelled competitor</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Antibody</td>
<td>- 1 2 3 4</td>
<td>- 1 2 3 4</td>
</tr>
</tbody>
</table>
Figure 8  Determination of putative factors that bind the promoter region surrounding -93G>A.

Max and DEC2 antibodies were added to the binding region to determine if these transcription factors bound the region surrounding the variant in nuclear lysates from the HEC-1-A cell line (A, B) and the HCT116 cell line (C). Unlabelled competitor was added to verify the specificity and the affinity of the interactions. (+) reactions carried out in the presence of nuclear lysate/unlabelled competitor; (-) reactions carried out without nuclear extract or without competitor. (1): 2µg of Max antibody, (2): 4µg of Max antibody, (3): 2µg of DEC2 antibody, (4): 4µg of DEC2 antibody.
Max is a member of the basic helix-loop-helix leucine zipper (bHLHZ) family of transcription factors and forms homodimers and heterodimer with other family members. DEC2 is a basic-helix-loop-helix (bHLH) transcription factor and both proteins have been implicated in the regulation of MLH1. In an attempt to shed light on the identity of the factors that bind this region we carried out supershift assays with 2μg or 4μg of anti-DEC2 or anti-Max antibody in HEC-1-A and HCT116 nuclear lysate (Figure 4.8). In these assays we did not observe a supershift or an interference of the specific bands indicating that these factors are not involved in binding this region in these cell types.

4.5 Discussion

In this study, using constructs of two different lengths spanning the MLH1/EPM2AIP1 promoter region, we were able to show that the -93 G>A promoter polymorphism affects the transcription of MLH1 in cell lines derived from cancers of the colorectum and endometrium. Similar effects were also seen in cell lines derived from non-tumourigenic tissues as well (unpublished data included in manuscript that is currently under review, experiments performed by Miralem Mrkonjic, a former PhD student in our laboratory). In the CRC cell line HCT116; we saw that the -93G allele was significantly more active than the -93A allele when the core promoter (C+D) construct was examined. We consistently observed increased promoter activity for the -93G allele with the D constructs in the three cell lines examined. Similar results were observed in several other CRC cell lines as well (unpublished data, Miralem Mrkonjic). Thus, the effect of the MLH1 -93G>A polymorphism on promoter activity was much more pronounced in the shorter promoter constructs. This was likely due to its abbreviated length, as the longer
core promoter construct likely harbors multiple transcription factor recognition sites that might eclipse the subtle effect of the -93G>A polymorphism. As expected, the overall activities of the D-region promoter constructs were considerably lower compared to the core promoter constructs. This is likely due to differential expression and regulation of transcription factors. A universal dramatic effect on the transcription of MLH1 at the level of the core promoter would likely have a larger impact on the associated cancer risk. If the -93G>A were to have such a detrimental effect, then it would likely be categorized as a disease causing mutation rather than a polymorphism. This is supported by the fact that about 23% of the general, unaffected, population are carriers of this variant. Therefore, our finding that this polymorphism has a subtle but significant and consistent effect at the level of the minimum D promoter region is more in line with the epidemiological data.

In addition to its role in MLH1 transcription, analysis of the reverse promoter constructs indicate that the -93G>A promoter polymorphism also plays a role in transcription of the EPM2AIP1 gene. Our colorectal and endometrial cancer cell lines demonstrated differences in the level of EPM2AIP1 transcribed by the -93G and -93A alleles. However, in two of the cell lines, HCT116 and HEC-1-A the -93A allele demonstrated higher activity. However, in SK-UT-1B, there was a significant decrease in the activity of the -93A allele. This discrepancy could have arisen due to the cell specific expression of factors involved in transcription. These results indicate that while the overall effect of this variant on the transcription of EPM2AIP1 may differ in individuals and/or cell types, the presence of this polymorphism alters the level of EPM2AIP1 transcribed. Interestingly, the -93G>A polymorphism is located in a region that was previously shown to be crucial for EPM2AIP1 transcription and promoter methylation that silences the MLH1 gene also results in loss of EPM2AIP1 gene expression [342]. Nonetheless, the function(s) of the EPM2AIP1 is currently unknown [343]. A genome-wide breast cancer
association study identified *EPM2AIP1* gene as a high-priority candidate that modifies breast cancer risk in BRCA2 mutation carriers [457]. EPM2AIP1 was found to interact with Laforin (EPM2A), which was implicated in the Wnt signaling pathway as well as rapid onset tumourigenesis in mice [458]. Nevertheless, it is unclear whether EPM2AIP1 plays a role in Laforin-mediated tumourigenesis. Our results offer evidence that genetic alterations in the promoter region of *MLH1*, hitherto considered only for their potential effect on *MLH1* transcription can potentially regulate the transcription of *EPM2AIP1* as well.

Given the observed effect of the -93G>A variant on both the *MLH1* and *EPM2AIP1* genes in CRC and endometrial cancer cell lines, it is important to assess how this variant might influence transcription. One possible mechanism is by altering transcription factor binding sites. Our EMSA results indicate that the region surrounding the -93G>A variant was bound by multiple nuclear factors and that the -93G>A variant altered the affinity and binding pattern of these factors in both CRC, endometrial and non-tumourigenic cell lines. EMSA allows for the assessment of the abundance, relative affinity and binding specificity of DNA-binding proteins[459]. These results corroborate the findings of our promoter activity assays. However, our experimental design cannot determine the identity of the nuclear factors. Several studies have analyzed the promoter region of *MLH1* for potential transcription factor binding sites. One study identified a protein-binding site at the -93G>A location, which suggests that the region surrounding the variant is important for protein binding. However, no known proteins were predicted to bind this region according to the TFSEARCH and DNASIS software [456]. While knowing the identity of the factors will help shed further light on the mechanism by which this promoter polymorphism mediates its effects on gene transcription, identifying these factor(s) will not necessarily help interpret the risk associated with this variant. Given the different binding patterns observed between the colorectal, endometrial and normal colonic cell lines, it is possible
that considerable variation exists in the nuclear factors that bind this region between different host cell lines. The results from our EMSA provide further evidence that the -93G>A variant alters transcription of the two genes, *MLH1* and *EPM2AIP1*, that are regulated by a common promoter.

Two proteins implicated in binding the *MLH1* promoter are DEC2 and Max. The transcription factor Max has been shown to interact with MSH2 both *in vitro* and *in vivo* [460]. Recent studies implicated Max in repressing *MLH1* and *MSH2* under hypoxic conditions. This repression was shown to occur via a dynamic shift in occupancy from activating c-Myc/Max to repressive Mad1/Max and Mnt/Max complexes at the proximal promoters of both the *MLH1* and *MSH2* genes [461]. These factors recognize E-Box motifs in the DNA. The hypoxia-inducible transcription repressor, DEC2 has been shown to regulate the transcription of *MLH1* by binding to E-box-like motif(s) in the *MLH1* promoter region [462]. While little is known about the microenvironmental factors that may lead to genetic instability in the tumour microenvironment, emerging evidence suggests that hypoxia plays a key role in this process. It was shown that hypoxia downregulates the expression of the *MLH1* gene at the transcriptional level [463]. An E-Box motif (AACGTG), is located 8 nucleotides downstream from the -93G>A polymorphism. However, results from our supershift assays did not provide evidence of DEC2 or Max as possible factors that bound the region around the -93 variant of the *MLH1* promoter. However, it is possible that these interactions occur only under hypoxic conditions or that these interactions are specific to certain cell types. Our analysis does not exclude the possibility of these interactions as it could occur outside of the 25 base pair region spanning our probe.

The main limitation of our study is our inability to conclude whether the proteins that bind the region surrounding the -93G>A variant in the EMSA experiments are activators or
repressors. Neither are we able to specifically delineate if these factors regulate *MLH1*, *EPM2AIP1*, or both genes. In addition, given our observation that multiple factors bind the labeled probe, it is unclear whether a subset of the observed proteins represents multi-protein complexes of coactivators or corepressors, or distinct transcription factors.

A previous study found the -93G>A variant to be associated with *MLH1* methylation in endometrial and colorectal tumours [464, 465], which raises the possibility that this polymorphism affects binding of methylation machinery and therefore results in gene silencing. Thus, in addition to modifying the binding of transcription factors, this polymorphism appears to be able to alter promoter methylation in a context dependent manner. Furthermore, it is likely that these effects are not necessarily independent of each other and that they act in conjunction in certain instances. For example, differential methylation of the *MLH1* promoter has been observed in CRC [339, 466]; however, differences in methylation between the C and D regions have not been studied [340]. Hypothetically, if a large part of the *MLH1* promoter C region is methylated, then the role of the D region in mediating transcription might be enhanced, leading to a greater effect of the -93G>A variant. Biologically, this effect could occur in individuals whose *MLH1* promoter is partially methylated leaving the C-region inactivated, while the promoter D-region remains unaffected.

This study has several strengths. This is the first study to examine the effects of the *MLH1*-93G>A promoter polymorphism on promoter activity in *MLH1* promoter constructs of two different lengths. In addition, we demonstrate the functional effects of this polymorphism in a panel of cell lines representative of colorectal and endometrial cancer. This is also the first study to evaluate the effects of this variant on transcription of EPM2AIP1, demonstrating that a genetic alteration simultaneously affects both the regulation of *MLH1* and its antisense gene. We
demonstrate that $MLH1$ -93G>A is a functional polymorphism in which the A allele results in decreased $MLH1$ transcription. This builds on our previous finding that the -93A allele is associated MSI-H CRCs, and in combination with previously published genetic epidemiology studies, supports its role in modifying the risk of cancer development. Further functional characterization of risk alleles identified in population-based studies will enhance our understanding of cancer initiation and progression and may possibly serve as potential tools for screening and prognosis.

Acknowledgements

We would like to thank Julie Turnbull and Dr. Berge Minassian (Hospital for Sick Children, Toronto) for sharing with us the anti-EPM2AIP1 antibody that they generated.
CHAPTER 5

Discussion and Future directions

Germline mutations in the MMR genes lead to Lynch syndrome and genetic testing is offered to identify individuals at risk of developing this disease. About a third of all genetic alterations detected in the mismatch repair gene MLH1 are missense variants, the majority of which are variants of unclassified clinical significance. In this thesis, our overall aim was to investigate the mechanism(s) by which a panel of missense variants localized to distinct domains impacted the regulation, expression and function(s) of MLH1.

5.1 Stability and dimerization of MLH1

The work presented in this thesis demonstrates for the first time that two missense substitutions, p.R265C and p.K618A alter the stability of the MLH1 protein. Although altered stability has been proposed as mechanism by which unclassified variants affect MLH1 expression, this had not been definitively demonstrated. However, the pathway responsible for the degradation of MLH1 is currently unknown and as such it is difficult to explain the mechanism responsible for this observed decrease in stability. Based on our findings, we speculate that missense substitutions may affect the tertiary structure of the protein, which may lead to defective folding and the consequent degradation of MLH1. It can be postulated that irrespective of the functionality of the protein, decreased stability likely leads to a situation where lower levels of protein are present in the cell. Although the levels of protein essential for normal cellular function and the threshold below which cancer susceptibility occurs has not been established; it is likely that decreased expression compromises to some degree function(s) carried out by these
protein complexes. It is also possible that varying threshold levels of MLH1 exist for the different functions of the MMR system, even in the same cells or tissues. This was underscored in a previous study that indicated that less MLH1 expression is required for the post-replicative error repair activity of MLH1 compared to the DNA damage response function[311]. This illustrates how these variants may potentially impact one function, while leaving another intact. If this situation holds true in Lynch syndrome patients, it will explain how genetic variants not associated with the MSI-H phenotype in tumours (a surrogate marker of defective post-replicative error repair) could potentially lead to disease susceptibility.

Our findings also offer some insight into the stability of wild-type MLH1. However, the mechanism responsible for MLH1 is degradation and turnover is currently unknown. Given that the sustained expression of MLH1 is required to stabilize the expression of the other MutL homologs [467], further studies regarding the mechanism responsible for MLH1 degradation is warranted. These studies will likely shed light on the mechanisms by which MLH1 is regulated, a topic on which currently very little is known.

Another unclassified variant, p.L749Q impacted the ability of MLH1 to heterodimerize with PMS2. All MMR functions described to date require that the MutL homologs be localized to the nucleus of the cell. Furthermore, it has been shown that the dimerization of MLH1 and PMS2 is necessary for the nuclear localization of MutLα [336]. Additionally, a recent study found that the nuclear localization of MLH3 is dependent on MLH1, and is competitive with PMS2 [468, 469]. This raises the possibility that missense variants in MLH1 that alter its ability to bind to specific MutL homologs, alter the relative abundance of MutL complexes in the cells and in doing so alter the functional capabilities of this system.
5.2 The role of MLH1 in DNA damage signaling

Using a stable inducible cell line model, we were not able to ascertain if MLH1 variants alter MLH1’s ability to signal DNA damage. However, in spite of these technical setbacks this remains an important question to be addressed. Numerous studies have linked MLH1 and the MMR proteins as important players in signaling DNA damage caused by a range of agents. However, the precise mechanism by which this occurs remains to be resolved. Gross genetic and or biochemical approaches such as knocking down the entire MLH1 protein might interfere with more than function of the MMR system, leading to potentially confounding findings. The ability of missense variants in MSH2 to uncouple the functions of the MMR system [381], underscores the utility of missense alterations in evaluating the diverse functions of the MMR proteins. The lack of functionality associated with MLH1 overexpression in our cell line system demonstrates that increased expression, like decreased expression can potentially be harmful to the cell. Given our experiences, in future studies it will be prudent to generate cell lines carrying bacterial artificial chromosomes (BAC) of wild-type and variant MLH1. BACs are useful as it maintains not only the coding sequence but also the regulatory sequences upstream of it, including various promoter sequences that may potentially govern a gene's expression level. Given that the level of MLH1 appears to be maintained at a delicate balance, a system of this nature will minimize the potentially negative effects of artificially altered expression. If this approach identifies variants that are associated with cell cycle arrest or apoptosis, then microarray analysis can be performed on these cell lines to determine the panel of genes that are up- or down-regulated in these cells. The benefit of this experimental system would be the ability to compare two, otherwise identical cell lines that express a protein that differ by only one amino acid, which should at least theoretically yield a smaller subset of candidate genes than comparing two independent cell
lines. It is expected that this will help identify novel regulators of the MMR mediated DNA damage signaling pathway. Moreover, if this approach shows interesting results, then mouse models of these variants can be generated and their phenotypes’ can be compared to Mlh1−/− and Mlh1+/− mice. The studies that demonstrated the uncoupling of the DNA damage signaling function from that of the post-replicative error repair function was carried out in this manner using mice that carry the G674A missense mutation [381]. In addition, ES cells from these mice can be used to both investigate the presence of genetic instability, as well as perform experiments evaluating the functional consequences of these variants under controlled but physiologically relevant conditions.

Moreover, the majority of the efforts to date that aim to characterize unclassified variants in MLH1 have focused on the post-replicative error repair function. While this is undoubtedly a principle function of the MMR proteins, in the face of emerging evidence, it is unlikely to be the only mechanism by which variants mediate their pathogenicity. However, at present there are no widely-applicable and clinically relevant ways of assessing the impact on these functions. A better understanding of the role of MLH1 and the MMR system in responding to DNA damage can be utilized to develop relevant functional assays. A recent study described a highthroughput system for generating BAC transgenes and stably transfecting them into cell lines [470]. If such a method can be adopted, then this will allow for multiple variants to be screened simultaneously. Furthermore, a better understanding of the role of MLH1 in signaling DNA damage will help develop anti-cancer treatments that exploit the MMR expression status of a tumour.
5.4 Allelic imbalances in MLH1 and its implications for predictive genetic testing programs

The work presented in this thesis also demonstrates that a subset of missense variants in *MLH1* lead to allelic imbalances. Allelic imbalances could arise due to genetic alterations that affect the efficiency of transcription, alter transcript expression and/or affect mRNA stability and our study provides initial evidence that a subset of unclassified missense variants in *MLH1* affect these processes. Another mechanism by which variants may potentially contribute to altered mRNA expression is through abnormal splicing, and several missense mutations in *MLH1* that result in splicing abnormalities have been reported. Further studies focused on determining the mechanism by which the MLH1 transcript is regulated will help understand if particular motifs are associated with transcript regulation and stability. Identification of such motifs will allow for the more streamlined classification of variants based on their location.

It is possible that the MLH1 expression profile in lymphocytes differs from colon cells. Colon tissue, if available, would be the ideal source for RNA based analysis. However, unlike lymphocytes that are readily available from both symptomatic and asymptomatic individuals, colon tissue is often not easily obtained. Thus, lymphocytes are a useful tool especially for predictive genetic testing. Furthermore, in current diagnostic practice RT-PCR analysis of RNA extracted from lymphocytes is used to determine the putative effect of splicing alterations underscoring its applicability for interpreting the functional implications of genetic alterations. However, given differences in tissue specific expression, these results should be interpreted with caution until data comparing the expression profile of MLH1 in colon cells and lymphocytes are available.
In addition, the PeakPicker methodology that is described in this thesis has the potential to be applied in the diagnostic setting for the routine assessment of unclassified genetic variants, both in the MMR system as well as in other genes. The PeakPicker strategy can also be used to further analyze cases where no genetic lesion has been identified through conventional methodologies. Even among our panel, two of the variants assessed did not demonstrate allelic imbalances. Thus, it is likely that a significant proportion of unclassified variants will not demonstrate allelic imbalances, highlighting the need to use PeakPicker in conjunction with a panel of functional/biochemical assays. Many of the current functional assessment methods, including the DNA stability and binding assays described above require molecular cloning, bacterial and cell culture limiting their applicability in the diagnostic setting. Recently a quantifiable cell-free MMR activity that can be used for the diagnostic assessment of MLH1 variants was described [434]. However, this assay focuses on the post-replicative MMR function. Usage of such novel assays, in conjunction with PeakPicker and standard diagnostic methods such as MSI analysis and IHC will allow unclassified variants to be characterized more accurately. This also raises another consideration that is important for this field of study, in that it is unlikely that a single method will provide a clear cut picture of the pathogenicity associated with an unclassified genetic variant. Thus, an ideal situation would be to compile a systematic panel of assays, which can be either extended or compacted based on the clinical and familial information that is available.

5.5 Functional mechanisms underlying -93G>A variant and its implications for carcinogenesis

The work presented in this thesis also demonstrates that the MLH1 c.-93G>A promoter polymorphism alters the transcriptional activity of MLH1. A previous genetic epidemiology
study carried out in our laboratory has shown that the MLH1 -93G>A promoter variant is strongly associated with MSI tumours [446], a finding that has since been replicated in several independent populations. Furthermore, -93G>A has been shown to be associated with an increased risk of endometrial cancer [450]. With the core promoter constructs, we observed higher activity with the -93G allele compared to the -93A allele in HCT116 colorectal cancer cells. With the shorter promoter constructs, we consistently observed higher level of activity with the -93G allele in all cell lines examined. Our findings suggest that this polymorphism has a subtle but significant and consistent effect at the level of the minimum D promoter region, a finding that is in line with current epidemiological data. A drastic and universal effect on the transcription of MLH1 at the level of the core promoter would likely have a larger impact on the associated cancer risk, which is not consonant with the finding that about 23% of the general, unaffected, population are carriers of this variant [446]. Less than 5% of all colorectal cancers are associated with dominant inheritance, while some type of genetic predisposition accounts for 10-25% of all colorectal cancers. Most familial risk is probably attributable in some degree to inherited low-penetrance alleles that act in combination with environmental risk factors, such as dietary carcinogens. This is interesting given that the -93 G>A variant has been linked to lifestyle factors such as smoking [471]. Thus, the finding that the -93G>A variant can alter transcript levels of MLH1 has important implications for understanding mechanisms at play in moderate- to low- risk individuals. A previous study found this variant to be associated with MLH1 methylation in endometrial and colorectal tumours [464, 465]. Thus, it is possible that this variant also acts by altering methylation levels. Our findings also strengthen the genetic epidemiological findings regarding the association of this variant with endometrial cancer. In view of this, it may be possible to use MLH1 -93G>A in conjunction with a panel of other relevant polymorphisms, as prognostic indicators of endometrial cancer risk.
Our studies also indicate that nuclear proteins bind the region of the *MLH1* promoter surrounding the -93G>A variant. In addition, we observe that the -93G>A variant alters the affinity of these protein interactions. However, these effects are subtle, as would be expected for a common genetic variant and these findings were replicated in a recent study that looked at the effect of -93G>A on protein binding in the choriocarcinoma cell line, JEG3 [472]. Despite several attempts, the identity of the putative factors that bind this region remains to be elucidated. Our supershift experiments indicate that DEC2 and Max are unlikely to play a significant role in this process, at least under non-hypoxic conditions in the cell lines examined. A previous study was able to demonstrate that the *MSH6* promoter is regulated by the Sp transcription factors [473]. A similar approach can be utilized to uncover the identity of the factors that bind the *MLH1* promoter region. The identities of these factors will provide clues regarding the mechanisms regulating MLH1 expression. Moreover, given the different binding patterns we observed between the colorectal and endometrial cell line, it is possible that considerable variation exists in the nuclear proteins that bind this region. Accordingly, this should be taken into consideration when designing experiments that aim to address this question.

This is also the first study to demonstrate that a genetic alteration simultaneously affects both the regulation of *MLH1* and the adjacent gene, *EPMAIP1*. The cellular role(s) of EPM2AIP1 is currently not known; although a recent genome-wide association study identified *EPM2AIP1* as a high-priority candidate that modifies breast cancer risk in *BRCA2* mutation carriers [457]. EPM2AIP1 was also found to interact with Laforin, a protein implicated in the Wnt signaling pathway [458]. A recent study reported that deletions in *TACSTD1* (the gene directly upstream of *MSH2*) led to the somatic silencing of the corresponding *MSH2* allele by hypermethylation [375]. Similarly, it is possible that genetic alterations in *EPM2AIP1* can potentially affect MLH1 expression and activity. Alternatively, it is possible that EPM2AIP1 plays a role in
carcinogenesis. This is particularly interesting given that MSI-H tumours show a significant reduction in MLH1/EPM2AIP1 expression compared with normal colonic mucosa and microsatellite stable (MSS) tumours [445]. Although methylation of the shared promoter region is likely to be responsible for this loss of expression in the majority of cases, it is important to address if EPM2AIP1 independently contributes to neoplasia [342]. Initial studies to address this should attempt to assess EPM2AIP1 expression by IHC in a subset of MSI-H tumours that show intact expression of the other known MMR genes. Furthermore, EPM2AIP1’s role in Lynch syndrome can be evaluated by sequencing the promoter and coding region of EPM2AIP1 in individuals with no other known MMR mutations. A knock out model of Epm2aip1 has been developed by one of our collaborators, Dr. Berge Minassian and his group at the University of Toronto. It would be of interest to observe if these mice develop tumours, and if so characterize their phenotype. It is also plausible that EPM2AIP1 does not act independently but cooperates with MLH1 in tumourigenesis and future studies should address this possibility as well.

Another important theme that arises in this thesis is the classification of genetic alterations. All the variants assessed in this study are missense substitutions, and with the exception of Q510X they do not lead to the premature truncation of the protein. However, these genetic variants span the spectrum in terms of their association with disease risk. For instance, some are the only genetic alteration identified in kindreds that meet the stringent criteria for Lynch syndrome, others occur in individuals with suspected but unconfirmed Lynch syndrome, and yet others occur in individuals unaffected by cancer. They also vary in frequency in that some are novel, others while rare have been reported in families from different parts of the world, and yet others occur in a large proportion of the population. With regard to these differences, T117M is classified as mutation, I19F, R265C, K618A, L749P as unclassified variants and -93G>A as a single nucleotide polymorphism (SNP). SNPs have traditionally been defined as phenotypically
neutral sequence alterations that are present in the general population with a frequency of >1\% [474]. While lack of involvement in disease is likely true for the large majority of SNPs, various studies support the notion that a subset of SNPs can potentially contribute to disease pathogenesis. In fact characterizing unclassified variants in the context of predictive genetic testing is challenging as it difficult to interpret if they are disease causing mutations responsible for the phenotype observed or rare, benign polymorphic variants. Our studies demonstrate that at least with regard to cancer, genetic alterations cannot always be neatly compartmentalized into defined categories as their consequences form an overlapping continuum. Even with regard to a single gene such as \textit{MLH1}, it is likely that the distribution of genetic alterations on this continuum depend on the complex interactions that occur between other genes and environmental factors.

\textit{MLH1} is an integral component of the MMR system and MMR proteins play a central role in maintain the stability of the genome. However, \textit{MLH1} deficiency and for that matter MMR deficiency leads to tumours only in specific tissues. This could occur if the threshold of \textit{MLH1} required for the diverse functions of the MMR proteins vary between tissues. It could also result due to the ability of other DNA repair systems to substitute for defective MMR in specific tissues and the overlap between the functional capabilities of the multiple DNA repair pathways. Another possibility is that a specific type of error or damage may be more common in certain tissues, which may in turn influence that the type of repair critical for the integrity of that tissue. Thus, a combination of these factors may influence the tumour spectrum that arises due to \textit{MLH1} deficiency.

In conclusion, the findings presented in this thesis have important implications for the assessment and study of rare variants identified in suspected Lynch syndrome cases. Our
findings also shed light on the functional mechanisms underlying a common susceptibility variant. These studies also provide several insights into the regulation and expression of MLH1, which have implications for furthering our understanding of the MMR system.

The discovery of a predisposing mutation in Lynch syndrome families and the concomitant increase in surveillance has been shown to significantly reduce mortality [251, 431]. Therefore it is of paramount importance to classify the pathogenicity associated with unclassified variants to determine their role in cancer predisposition. Overall, an accurate assessment of the pathogenic nature of unclassified variants in the MMR genes will greatly increase the accuracy of molecular diagnostic testing and facilitate more efficient surveillance and prevention of associated cancers.
References


