A Survey of *hMLH1* mRNA Splicing Profiles in Human Cell Lines: Comparing Primary Cultured Fibroblasts Before and After Oxidative Stress and Transiently versus Stably Demethylated Colon Cancer Derived Cell Lines

by

Esther Tiedtke, B.A, B.Ed

A thesis submitted to the Department of Biological Sciences

in partial fulfillment of the requirements

for the degree of

Master of Science

August, 2011

Brock University

St. Catharines, Ontario

Abstract

Most human genes undergo alternative splicing, and loss of splicing fidelity is associated with disease. Epigenetic silencing of *hMLH1* via promoter cytosine methylation is causally linked to a subset of sporadic non-polyposis colon cancer and is reversible by 5-aza-2'-deoxycytidine treatment. Here I investigated changes in *hMLH1* mRNA splicing profiles in normal fibroblasts and colon cancer-derived human cell lines. I established the types and frequencies of *hMLH1* mRNA transcripts generated under baseline conditions, after hydrogen peroxide induced oxidative stress, and in acutely 5-aza-2'-deoxycytidine-treated and stably de-repressed cancer cell lines. I found that *hMLH1* is extensively spliced under all conditions including baseline (50% splice variants), the splice variant distribution changes in response to oxidative stress, and certain splice variants are sensitive to 5-aza-2'-deoxycytidine treatment. Splice variant diversity and frequency of exon 17 skipping correlates with the level of *hMLH1* promoter methylation suggesting a link between promoter methylation and mRNA splicing.

Acknowledgements

I am indebted to my family, for your unconditional love and unwavering encouragement, support, patience, and practical assistance in all its facets; for believing in me; and for modeling and instilling the tenacity to persevere in the face of adversity – this is our thesis. To the members of the Skandalis Lab and the Richards Lab, both past and present, thank you for great discussions in the lab and all things lighter outside of it. To all of you who offered assistance, insight, and encouragement during this project, there are too many of you to name; I will treasure your kindness and endeavour to pass it on. Special thanks go to Sandra Rehan, Kim Disher, Azad Eshghi, Hayam Mansour, Melanie Brown, Melissa Page, Ellen Robb, Nick Vesprini, Ai Lin Beh, and Shiri Sauday for your friendship, encouragement, clarity, and honesty; for sharing your passion for science, your expertise, and your reagents; and especially for holding on when I let go – I wouldn't have made it without you. To the past and present members of my committee, professors Jean Richardson, Jeffrey Stuart, Miriam Richards, Alan Castle, and Charles Després, your guidance has been invaluable both in terms of mentoring and scientific development. To the external examiner, Dr. Thomas Melendy, special thanks for your careful reading of the thesis and kind words. Finally, to my supervisor, Dr. Adonis Skandalis, who started it all with these fateful words: "O.K. Esther, I'm going to open the door for you." and who proceeded to teach the naïve student I was how to navigate the world of research which I have learned to love, this completed thesis is as much for you as it is for me. Adonis, you patiently held that door open to the end and you have inspired in me an unquenchable thirst for knowledge and understanding, the courage to ask for more of myself and others, and the willingness to challenge the status quo and grapple with the tough questions. Through this experience, you have forever changed my life for the better – thank you.

Table of Contents

List of Figures

Chapter 1:

Chapter 2:

Chapter 3:

Chapter 4:

List of Tables

Chapter 2:

Chapter 4:

Abbreviations

0HP	0 μM hydrogen peroxide treatment
250HP	250 μM hydrogen peroxide treatment
500HP	500 μM hydrogen peroxide treatment
3'ss	3' splice site
5'ss	5' splice site
8-oxoG	8-oxo-7,8-dihydroguanine (base lesion)
A	adenine
T	thymine
U	uracil
C	cytosine
G	guanine
AS	alternative splicing
ATM	ataxia telangiectasia mutated protein (protein kinase activated by DNA double strand breaks involved in damage sensing/response)
AzadC	5-aza-2'-deoxycytidine
BASC	BRACA1-associated genome surveillance complex
BER	base excision repair
bp	base pair
bps	base pairs
BP	branch point
BPS	branch point sequence
CBF	CELF-family Binding Factor (transcription factor)
CpG	cytosine, guanine dinucleotide
^{5me}CpG	methylated cytosine in context of CpG
CRC	colorectal cancer
CT	canonical transcript
DNA	deoxyribonucleic acid

E	exon
Δ	skipping (denotes whole or partial exon skipping)
∑	inclusion (denotes partial intron inclusion)
EST	Expressed Sequence Tag
G2/M	Gap 2 / Mitosis transition
H_2O_2	hydrogen peroxide
h	hour(s)
hMLH1	human MutL Homologue 1 gene
hMLH1	human MutL Homologue 1 protein
HNPCC	hereditary nonpolyposis colorectal cancer
hnRNP	heterogeneous nuclear ribonucleoprotein
IN	intron
MAPK	mitogen activated protein kinase
MMR	mismatch repair
mRNA	mature messenger RNA (contains exons, introns have been removed)
pre-mRNA	preliminary (unspliced) messenger RNA (contains exons and introns)
hMSH2	human MutS Homologue 2 gene
hMSH3	human MutS Homologue 3 gene
hMSH6	human MutS Homologue 6 gene
min	minute(s)
MSI	microsatellite instability
OS	oxidative stress
PCR	polymerase chain reaction
PMS1	Post Meiotic Segregated 1 gene
PMS2	Post Meiotic Segregated 2 gene
POLB	Polymerase Beta gene
POLB	Polymerase Beta protein
RNA	ribonucleic acid
RNAPII	RNA polymerase II

RRM	RNA recognition motif
s	second(s)
SF	splicing factor
SNPCC	sporadic nonpolyposis colorectal cancer
SR protein	serine-arginine rich domain protein
SV	splice variant
TF	transcription factor
TGFBR2	Transforming Growth Factor Beta Receptor Type II gene
µL	microlitre
WT	wild-type (synonymous with CT; hMLH1 protein derived from CT)

Project Overview

Background

A critical step in the fidelity of information flow in the cell occurs at the level of transcription and splicing. Splicing is a highly regulated process that involves excision of introns from pre-mRNA and ligation of exons to form mature mRNA. Loss of splicing fidelity has been linked to various diseases including colorectal cancer (CRC), a leading cause of cancer death worldwide (World Health Organization, 2011). In CRC, specific molecular phenotypes are associated with deficiency in the mismatch repair (MMR) gene human MutL Homologue 1 (*hMLH1*). Two mechanisms that contribute to *hMLH1* deficiency are transcriptional silencing by aberrant biallelic promoter cytosine methylation (Herman et al., 1998; Veigl et al., 1998) and mis-splicing of pre-mRNA (Pagenstecher et al., 2006; Tournier et al., 2008).

Although RNA splicing mechanisms and regulation lie at the crossroads of information transfer from DNA to RNA to protein, they remain incompletely understood. In particular, the factors that contribute to the regulation of alternative splicing, the process by which two or more mRNA transcripts (i.e. canonical transcript and splice variants) are generated from the same pre-mRNA transcript, are of interest as they greatly contribute to the diversity of the transcriptome and subsequently, to downstream processes including protein diversity and disease.

In order to elucidate the mechanisms and factors that contribute to alternative splicing regulation it is essential to examine the transcriptome that is actually generated. At the global level, mRNA splicing events are investigated on a genome-wide scale using exon-junction microarrays, Expressed Sequence Tags (ESTs), and *in silico* computational algorithms that screen transcripts from thousands of genes and allow regulatory networks to be discerned (e.g. coordinated tissue specific regulatory factors and sequence elements). At the opposite end of the spectrum, mini-gene constructs probe changes in splicing of individual exons within a single transcript in

response to different parameters (e.g. promoter swapping, RNAPII elongation rate, splicing factor ratios) and *in vitro* assays allow core splicing mechanisms to be studied (e.g. to determine the sequential steps involved in spliceosome assembly). All these approaches provide valuable insights into the intricacies of alternative splicing regulation however these approaches do not provide a complete mRNA splicing – profile – the types and frequencies of mRNA transcripts generated – of any one gene. A comprehensive splicing profile of an individual gene is essential where the gene is of clinical interest due to its role in disease and the impact of splice variants needs to be evaluated.

Overall Study Objective

In general, this study adds to knowledge about splicing mechanisms and regulation. Specifically, the objective is to determine how splicing regulates *hMLH1* and to evaluate the extent to which alternative splicing and mis-splicing contribute to *hMLH1* deficiency in human non-cancer and colorectal cancer derived cell lines.

Experimental Design

This study adds to existing approaches by establishing a comprehensive mRNA splicing profile of: (1) human MutL Homologue 1 (*hMLH1*), an enzyme involved in MMR that has been causally linked to CRC progression, the primary locus of interest in this study, and (2) Polymerase Beta (*POLB*), a key enzyme in base excision repair (BER), used as a control locus in this study to distinguish between global effects and gene specific (direct) treatment and/or methylation effects in CRC cell lines. Both loci are constitutively expressed under normal conditions and known to undergo alternative splicing. The method allows for full mRNA transcripts to be examined, as opposed to being limited to defined segments or single exon events, and entails the following: (1) harvesting cells grown in culture, (2) isolating total RNA, (3) using reverse transcription and PCR with oligo-dT primer to generate a cDNA library, (4) using PCR amplification with gene specific primers to isolate mRNA transcripts originating from the gene of interest, (5) transformation and cloning in *E. coli* with blue-white colony selection to isolate individual mRNA transcripts and

identify successful transformants, (6) colony PCR to produce sufficient product to visualize on an agarose gel where samples are separated and grouped based on size, and (7) restriction and sequencing to determine the identity of the transcripts.

The strength of this method lies in that it provides a comprehensive mRNA splicing profile under a certain set of conditions and allows mutually exclusive splicing patterns found among mRNA splice variants to be identified. The idea here is to (1) determine unambiguously what types of *hMLH1* transcripts are actually produced, to categorize them according to the classes of splice events they contain, and to determine their relative frequency within the sample, (2) to compare the results to those predicted by current *in silico* splicing algorithms to see which splice events are driven by known mechanisms and factors (e.g. canonical splicing, exonic splicing enhancers) that warrant further investigation in future studies, (3) to provide a baseline splicing profile before and after oxidative stress (OS) against which cancer specific splicing signatures can be compared, if present, and (4) to investigate the potential interplay between two epigenetic mechanisms, alternative splicing and DNA methylation, both of which affect *hMLH1* in the subset of colorectal cancer derived human cell lines examined in this study.

Chapter Overviews

In chapter 1, a review of the literature is conducted that explores the causal link between *hMLH1*, loss of repair function, and colon cancer progression, how oxidative stress and *hMLH1* DNA cytosine methylation (^{5me}CpG) influence *hMLH1* expression, and alternative splicing of *hMLH1* in cancer and non-cancer. Current knowledge about factors that influence splicing regulation is outlined including: core splicing mechanisms, enhancers and silencers, kinetic coupling, chromatin remodeling in stress, and exon definition with a focus on the impact of these factors on splicing and disease.

In chapter 2, a comprehensive *hMLH1* mRNA splicing profile is established in the primary untransformed fibroblast cell line, MRC5. This forms a baseline for comparison and serves as a non-cancer control. The *hMLH1* splicing profile after

250uM and 500uM hydrogen peroxide treatment (250HP and 500HP respectively), which induces oxidative stress (OS), is further characterized to determine the impact of OS on *hMLH1* and to serve as a proxy to gauge potential contributions to the splicing pattern attributable to cellular stress in subsequent comparative analyses. Also, OS is relevant to the tumor environment such that splice variants that change in response to OS might be reflected in cancer splicing profiles. This chapter focuses on identifying, describing, and categorizing the *hMLH1* mRNA transcript types generated in terms of: (1) the classes of splicing they contain (e.g. exon skipping, exon border truncations or extensions, intron inclusion), (2) splice site type (donor, acceptor, dual) and strength (strong, intermediate, weak), and (3) core splicing mechanisms (canonical, non-canonical, recursive).

In chapter 3, the potential role of enhancers and silencers found in *hMLH1* exons and flanking introns on the observed *hMLH1* splicing profiles (determined in chapter 2) is mapped and analyzed. Computationally derived high scoring cis-elements (corresponding to the RNA recognition motifs of known trans-acting factors) within *hMLH1* exons are identified, their location and distribution is compared to the splicing patterns observed in MRC5 cells (baseline, 250HP, 500HP), and the trends that emerge are discussed. The chapter concludes with a proposed model of the factors that contribute to key splicing signatures observed in MRC5 cells to be tested in future studies.

In chapter 4, the potential interplay between alternative splicing and promoter cytosine (^{5me}CpG) methylation is examined in a model system of sporadic non-polyposis colorectal cancer (SNPCC) cell lines. The *hMLH1* mRNA splicing profile is characterized in the parental (*hMLH1*$^-$) cell line after treatment with 5-aza-2'-deoxycytidine (AzadC) to induce *hMLH1* expression (transient) and in stably demethylated clonal cell lines that remain *hMLH1*$^+$ for more than 5 weeks after AzadC is withdrawn. The *hMLH1* mRNA splicing profiles in transiently and stably demethylated cell lines are compared, potential cancer specific signatures are identified (e.g. by comparison to non-cancer profiles), correlation analysis between

hMLH1 splice variants and promoter methylation level is performed, and possible treatment effects and methylation effects are discussed. In addition, the splicing profile of *POLB* (not subject to promoter methylation) is established before and after treatment as a control.

Key Findings

There are four key findings of this thesis. Firstly, *hMLH1* is extensively spliced in all cell lines and conditions examined and may be subject to regulation by both U2 and U2-independent core splicing mechanisms. Secondly, treating cells with 250HP and 500HP changes the splicing pattern. There is evidence to suggest that specific trans-acting factors including SRp55 and FOX-2 may mediate changes to the *hMLH1* splice variant distribution in response to 250HP whereas TRA2-BETA loss may down-regulate the canonical transcript (CT) in response to 500HP. Thirdly, cancer splicing profiles appear to be characterized by specific intron inclusion that is either absent or distinct from non-cancer for *hMLH1* and *POLB* respectively, suggesting that the effect is mediated by global factors as opposed to gene specific (direct) effects. Lastly, a correlation was observed between promoter methylation level and transcript diversity in the stably demethylated colon cancer cell lines (curvilinear relationship). No correlation between *hMLH1* promoter methylation status and splice variant frequency was observed, except for E17 skipping (negative correlation). However, there appears to be a drug effect of AzadC on the *hMLH1* splicing profile that is distinct from changes observed in response to HP. This suggests that cytosine methylation (or lack thereof) may influence alternative splicing but that these are global, as opposed to gene specific (direct) effects. Promising directions for future study both in terms of splicing mechanisms and potential clinical value are highlighted.

Chapter 1: Literature Review

1.1 Introduction

Mechanisms and factors that govern gene expression temporally, spatially, qualitatively, and quantitatively modulate the products and phenotypes resulting from genotype. This phenomenon is highly regulated and involves both genetic and epigenetic factors and the interplay between them. Genetics deals with the sequence of bases contained in DNA, the integrity of which is compromised by mutation. Mutations are permanent changes to DNA, some of which lead to disease when they disrupt downstream gene function such as proper protein expression and pathways in which the protein operates. For example, mutations in *hMLH1* are responsible for loss of MMR pathway function (Bronner et al., 1994; Papopoulos et al., 1994) found in ~50% of hereditary non-polyposis colorectal cancer (HNPCC) (reviewed in Lynch and de la Chapelle, 1999; Hitchens and Ward, 2009). Mechanisms that regulate gene expression either by reversible chemical modifications of nucleic acids or alternative splicing of gene transcripts fall under the purview of epigenetics. Epimutations are aberrant but reversible changes to gene regulation that are actively maintained and play an important role in the etiology of certain human cancers (Lyko and Brown, 2005). For example, the accumulation of aberrant promoter cytosine methylation (^{5me}CpG) in *hMLH1* is an epimutation that results in transcriptional silencing, loss of MMR, an increased mutation rate, and colon cancer progression in sporadic non-polyposis colorectal cancer (SNPCC) (Herman et al., 1998; Veigl et al., 1998; reviewed in Lynch and de la Chapelle; Hitchens and Ward, 2009). Treatment with the demethylating agent 5-aza-2'-deoxycytidine (AzadC) reverses silencing and restores MMR in SNPCC cell lines (Veigl et al., 1998; Hagemann et al., 2011). Genetic changes also impact epigenetic processes. For example, in one HNPCC family, a 3 base pair in-frame deletion of *hMLH1* exon 3 was expected to result in skipping of codon 71 however the deletion disrupted an exonic splicing enhancer and resulted in skipping of the entire exon (i.e. 33 codons) (McVety et al., 2006).

Similarly, 4 unrelated HNPCC patients with a G to C transversion in exon 18 of *hMLH1* that was expected to produce a glycine to histine substitution at codon 701, displayed exon 18 skipping instead (i.e. loss of 38 codons) (Pagenstecher et al., 2006). On a larger scale, Tournier et al. (2008) examined 56 unclassified allelic variants of *hMLH1* in HNPCC patients and found that a significant portion (~25%) affected mRNA splicing. These examples underscore that the genetic landscape and epigenetic mechanisms regulating gene expression influence each other. The realization that many examples of impaired hMLH1 protein function are in fact mediated by an effect on RNA splicing prompted Pagenstecher et al. (2006) to suggest that screening at the mRNA level should precede functional protein studies and triggered a renewed interest in splicing regulation of *hMLH1* and its role in hMLH1 deficiency. Although *hMLH1* is known to undergo alternative mRNA splicing (Charbonnier et al., 1995; Genuardi et al., 1998; Peasland et al., 2010), no comprehensive splicing profile based on full-length *hMLH1* mRNA transcripts has been established.

This thesis explores the impact of two specific epigenetic mechanisms; alternative mRNA splicing and promoter DNA cytosine methylation (^{5me}CpG), on the transcript profile of *hMLH1* in human non-cancer and colon cancer derived cell lines under various conditions. Loss of fidelity in either process is associated with CRC (Fackental and Godley, 2008; Kosinski et al., 2010; and Deng et al., 1999, 2001; Morak et al., 2008; Ehrlich et al., 2008 respectively); however the extent to which the processes influence each other or whether they are linked remains unknown.

1.2 Colorectal Cancer

Colorectal cancer currently remains the second leading cause of cancer death in Canada with 63% survivability after diagnosis and treatment (Canadian Cancer Society, 2011) – outcomes are improved by early diagnosis and targeted treatments. Therefore it is important to determine the factors that: (1) contribute to cancer progression in general, and (2) distinguish different types of CRC in particular.

In general, cancer is defined as uncontrolled cell growth (a tumor). The somatic mutation theory posits that cancer arises through the accumulation of mutations within a single cell; mutations that contribute to cancer progression disrupt genes involved in proliferation, differentiation, cell cycle control, damage signaling, repair, and apoptosis (reviewed in Michor et al., 2004; Baker and Kramer, 2007). This framework has led to the discovery of numerous cancer susceptibility genes including *hMLH1* (e.g. see Ghazi et al., 2010; Quan et al, 2011). Consistent with this framework, factors that contribute to cancer progression by increasing the mutation rate have been identified and include: (1) oxidative stress (increased *r*eactive *o*xygen *s*pecies, ROS), (2) loss of repair function, (3) disregulation of normal DNA methylation profiles (aberrant promoter hypermethylation and global hypomethylation of cytosine in the context of CpG dinucleotides), and (4) altered mRNA splicing.

1.2.1 Oxidative Stress and the Cellular Stress Response

Cell stress refers to the diverse range of environmental stressors that cause physical damage to cell macromolecules (i.e. proteins, lipids, DNA) resulting in cell death (reviewed in Kültz, 2005). Aerobic organisms derive energy from oxidative phosphorylation, carried out by the electron transport chain within the mitochondrial membrane, with oxygen (O_2) as the terminal electron acceptor. Oxidative phosphorylation also generates partially reduced metabolites of O_2. These include superoxide anion ($O_2^{\bullet -}$), H_2O_2, and, in the presence of transition metals, the highly reactive hydroxyl radical ($OH^{\bullet}$) that damages proteins, lipids, and DNA (reviewed in Thannickal and Fanburg, 2000). Oxidative stress (OS) refers to an imbalance between the production and removal of oxidants/ROS within a cell (reviewed in Thannickal and Fanburg, 2000). A common feature of diverse cell stresses is OS and a change in cellular redox potential referred to as oxidative burst (Kültz, 2005).

Cells monitor and respond to stress (i.e. OS, damaged proteins, lipids, and DNA) through their stress proteome. This process constitutes the cellular stress response. Minimally, the stress proteome includes products of ~40 genes with

conserved function in multicellular organisms spanning areas of redox regulation, molecular chaperones, fatty acid/lipid metabolism, energy metabolism, and DNA damage sensing and repair – including *hMSH2* and *hMLH1* (reviewed in Kültz, 2005). hMSH2 and hMLH1 form part of the BRACA1-associated genome surveillance complex (BASC), an important sensor of multiple types of DNA damage (Kültz, 2005).

The cellular stress response assesses and counteracts stress-induced damage, temporarily increases tolerance to damage, and removes individual damaged cells by apoptosis (programmed cell death) (Kültz, 2005). A long-standing view is that apoptosis can be triggered by physiological or pathological conditions/stressors and occurs in an energy dependent, regulated manner in which membrane integrity is maintained and the mitochondria play a role (reviewed in Fink and Cookson, 2005). This process is distinct from necrosis that involves groups of cells or cell layers, is pathological, and occurs in an energy independent, passive manner with loss of membrane integrity but no mitochondrial involvement (Roche Applied Science, 2011). However, recent reports suggest that necrosis, like apoptosis, can be triggered in a regulated manner that involves particular subsets of genes that may act as a default program if apoptosis fails (reviewed in Galluzi and Kroemer, 2008). In a properly functioning cell, apoptosis may be triggered in response to excessive DNA damage. In contrast, DNA base damage that is deemed repairable is removed and corrected by cellular repair pathways.

1.2.2 Repair of DNA damage

The maintenance of genomic integrity is of paramount importance to the survival of the organism. Hence, cells employ complex mechanisms, including DNA damage repair, to prevent mutations and ensure high fidelity in the transmission of genetic information (Kyng and Bohr, 2005). DNA damage that occurs under normal cell conditions is removed by one of several repair pathways, depending on the type of lesion. Oxidative stress gives rise to damage including single base lesions (e.g. 8-oxoG), the majority of which are repaired by Base Excision Repair (BER).

Spontaneous lesions alone are estimated to occur at a rate of about 10 000 lesions/cell/day (reviewed in Lindahl, 1993). In the absence of BER, the lesions are inefficiently removed (Fortini et al., 1998) which allows mutations to become fixed and compromise genetic integrity (Beard and Wilson, 2000). In addition, the mutation rate is suppressed by MMR that corrects mismatched bases and misincorporation errors that arise during replication (Blasi et al., 2006). Therefore, BER and MMR are crucial to suppressing the cellular mutation rate. Changes in POLB and hMLH1 expression were observed in CRC tumor tissue compared to normal tissue (Yu et al., 2006). POLB and hMLH1 are key components of BER and MMR respectively and each pathway is examined in turn in the sections that follow.

1.2.2.1 Base Excision Repair

The base excision repair (BER) pathway consists of multiple steps that correct aberrantly oxidized, methylated, deaminated, and spontaneously lost bases in nuclear DNA (Christensen, 2003). These bases arise spontaneously due to normal decay, as by-products of metabolism (e.g. ROS from oxidative phosphorylation and OS), or may be environmentally induced mutagens (Christensen, 2003). BER is the primary pathway responsible for removing small, non-helix distorting DNA lesions (Weaver, 2002). BER is divided into two sub-pathways defined as short patch and long patch repair that process single and 2-15 nucleotide patches of damaged DNA bases respectively (Pascucci et al., 1999; Frosina et al., 1996; Dianov et al., 1999; Fortini et al., 1998). In the predominant short patch pathway, the damaged base is first removed by a DNA glycosylase (*eg.* hOGG1 for 8-oxoG lesions) (Nilsen and Krokan, 2001). The resulting abasic or apurinic/apyrimidinic (AP) site may be processed in one of several ways as illustrated in figure 1.1 and reviewed (Christmann et al., 2003). DNA glycosylases with AP lyase activity such as hOGG1 incise the lesion containing DNA strand 3' of the lesion which is then further processed by APE endonuclease 1 (APE1). If the glycosylase does not have lyase activity, APE1 makes the incision 5' of the lesion. APE1 cleaves the phosphodiester bond 5' of the AP site generating a free 3' OH, POLB fills in the one nucleotide gap, POLB 5'deoxyribose-

phosphate lyase (5'RPase) activity removes the 5'dRP moiety, and the ligase III/XRCC1 complex reseals the DNA strand. Long patch repair occurs when the 5'dRP moiety is resistant to POLB lyase activity. While POLB may still insert the first nucleotide, subsequent nucleotides are synthesized by POL δ or ε, and the displaced strand is removed by flap endonuclease 1 (FEN1) (Prasad et al., 2000; Nilsen and Krokan, 2001). Factors required for long patch repair include PCNA, RF-C, POL δ/ε, and ligase 1 (Nilsen and Krokan, 2001).

POLB is expressed in untreated (hMLH1-) and AzadC treated (hMLH1+) SNPCC cell lines characterized in this thesis as well as in the stably demethylated (hMLH1+) clones. Thus, *POLB* was selected as a control locus in this study to provide a proxy measure of potential drug specific and/or treatment stress effects on mRNA splicing profiles of DNA repair genes in the SNPCC colon cancer cell lines examined. This allows the question of whether treatment with AzadC changes the mRNA splicing profile of genes in the SNPCC cell lines to be addressed and, depending on the splice variants observed, may point to pathways involved.

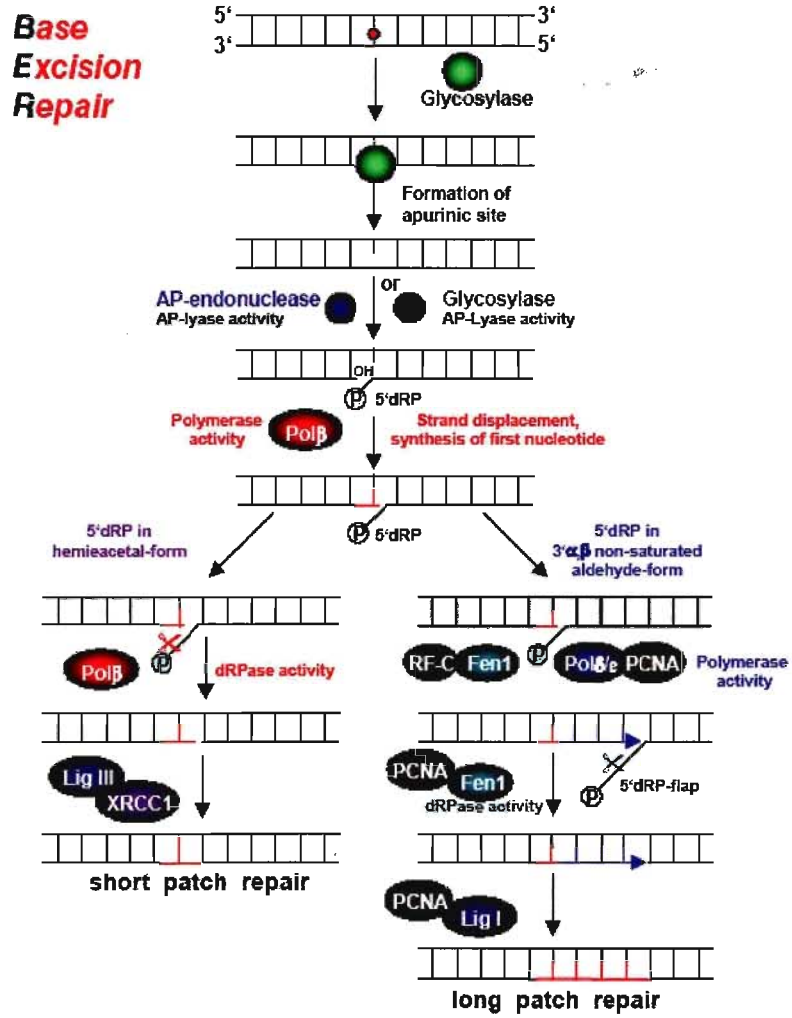


Figure 1.1: The Base Excision Repair Mechanism. Specific DNA lesions are recognized by various glycosylases that hydrolyze the *N*-glycosidic bond to remove the damaged base and from an AP site. BER then proceeds by either short patch or long patch repair as shown. Some glycosylases (*eg.* hOGG1) also have AP lyase activity and incise the DNA lesion containing strand 3' of the lesion which is subsequently trimmed by APE1 to yield a 3' OH. Alternatively, APE1 may directly incise the DNA 5' of the lesion leaving a 3'OH and a 5-deoxyribo-5'phosphate which is removed by the lyase activity of POLB after single nucleotide resynthesis by POLB, after which the DNA strand is resealed by the ligase III/XRCC1 complex. Long patch BER requires additional factors including PCNA and polymerase δ or ε for elongation (though POLB may insert the first base), flap endonuclease 1 (FEN1) to remove the displaced flap, and ligase I to reseal the nick. (Christmann *et al.*, 2003)

1.2.2.2 Mismatch Repair

In addition to their role in genome surveillance, hMSH2 and hMLH1 are key components of MMR. MMR is highly conserved and reduces the mutation rate arising from replication errors by several orders of magnitude (reviewed in Kunkel and Erie, 2005; Li, 2008). Thus, proper function of the pathway is essential for maintaining genome integrity and stability, as is evidenced by the causal role of MMR deficiency in non-polyposis CRC types (Bronner et al., 1994; Kim et al., 1994). MMR identifies and corrects (1) mismatched bases in double stranded DNA that arise through replication (e.g. due to mis-incorporation errors) or endogenous damage (e.g. damage to bases in nucleotide pool), (2) internal deletion loops, and (3) double strand breaks (homologous recombination). MMR proteins are also involved in crossovers (non-homologous end joining) during meiotic recombination (Guillon et al., 2005) however these involve a distinct subset of proteins, hMSH4-hMSH5, and the focus in the context of CRC is on repair. MMR is initiated by MutSα or MutSβ, a heterodimer of hMSH2 and hMSH6 or hMSH3, which scans, recognizes, and binds to mismatches and internal deletion loops respectively. Binding recruits another heterodimer, MutLα, consisting of hMLH1 and PMS2 that join MutS, as shown in Figure 1.2.

(Helleman et al., 2006)

Figure 1.2: Damage Recognition by MMR Repair. MutS, a heterodimer of MSH2 and either MSH6 (MutSα) or MSH3 (MutSβ), scans double stranded DNA for mismatches and internal deletion loops respectively. Upon recognition and binding, MutLα, a heterodimer of hMLH1 and PMS2, is recruited. PMS2 nicks the strand containing the mismatch 5' of the lesion stimulating excision and repair. (reviewed in Helleman et al., 2006)

PMS2 is stabilized in the presence of its binding partner hMLH1 but is degraded other wise, therefore loss of hMLH1 prevents MMR and results in an increased mutation rate. PMS2 contains latent endonuclease activity (Kadyrov et al., 2006) that nicks the DNA 5' of the lesion. This triggers excision and resynthesis of the damaged strand by Exonuclease I, RC-F, PCNA, DNA polymerase δ (mitochondrial) or ε (nuclear), DNA helicase 1, and DNA ligase 1 (reviewed in Helleman et al., 2006).

One manifestation of MMR deficiency is replication errors (expansion or deletions) at microsatellites due to polymerase slippage. Microsatellites are short repetitive DNA sequences consisting of mono- and di-nucleotide repeats of A or CA respectively. Cells with this molecular phenotype are described as displaying microsatellite instability (MSI). MSI increases the mutation rate (Eshelman et al., 1995) and that contributes to cancer progression by inactivating genes involved in cell cycle control, apoptosis, and proliferation (reviewed in Loeb and Loeb, 2000). The transforming growth factor beta receptor II gene (*TGFBR2*) is a common target (reviewed in Lynch and de la Chapelle, 1999). MMR deficiency with MSI is a hallmark of non-polyposis CRC types however the phenotype arises through both genetic and epigenetic means that require distinct tests and treatment.

1.2.2.3 Types of Colorectal Cancer

To facilitate accurate diagnosis and effective treatment of CRC, it is essential to identify and understand the distinct mutational pathways and markers associated with different types of CRC. Depending on how CRC arises determines whether it is classified as hereditary (also known as familial) or sporadic. Hereditary CRC arises through inherited germ line mutations and accounts for 15-20% of total CRC (Canadian Cancer Society, 2011). Molecular diagnostics have identified distinct groups of hereditary CRC. One of these groups is characterized by tumors of the right colon, diploid DNA, MSI, and characteristic gene mutations such as of *TGFBR2* (Kim et al., 1994; reviewed in Lynch and de la Chapelle, 1999). These tumors generally have a good prognosis (Yoon et al., 2011). This group includes HNPCC that is characterized by loss of MMR due to mutations or deletions affecting

one of the MMR genes (e.g. *hMSH2, hMSH3, hMSH6, hMLH1, PMS2*); the majority of cases (~90%) involve *hMSH2* (type 1) or *hMLH1* (type 2), denoted HNPCC1 and HNPCC2 respectively (Lynch and de la Chapelle, 1999).

Absence of hMLH1 protein and associated loss of MMR function has been causally linked to both hereditary and sporadic non-polyposis CRC (i.e. HNPCC2 and SNPCC) (reviewed in Lynch and de la Chapelle, 1999; Hitchens and Ward, 2009). In fact, the majority of CRC is sporadic, arising through spontaneous somatic mutations and epimutations, and accounts for about 80-85% of total CRC (Canadian Cancer Society, 2011). Among sporadic CRC, one group of cases, denoted SNPCC, exhibits a molecular phenotype that mimics HNPCC (described above) except that *hMLH1* specifically (but not *hMSH2*) is found to be epigenetically silenced at the transcriptional level by aberrant promoter DNA cytosine methylation (^{5me}CpG) without allelic *hMLH1* mutations (Veigl et al., 1998; Bettstetter et al., 2007).

Enrichment of CpG dinucleotides is a hallmark of gene promoters and the promoters of most constitutively expressed genes contain CpG dinucleotides, often referred to as CpG islands, with unmethylated cytosines (Levenson and Sweatt, 2005). Actively transcribed genes tend to contain unmethylated CpGs, or relatively lower methylation, at least at key CpG positions, as is normally the case for *hMLH1* expressing human cell lines and tissues (Deng et al., 1999, 2001). Conversely, promoter ^{5me}CpG tends to be associated with transcriptional silencing, as is the case for *hMLH1* in SNPCC. Thus, SNPCC appears to share a common carcinogenic pathway with HNPCC, as seen in figure 1.3, but accounts for a greater percentage of total CRC (10-15% versus 3-5% respectively) (Lynch and de la Chapelle, 1999; National Cancer Institute, 2011).

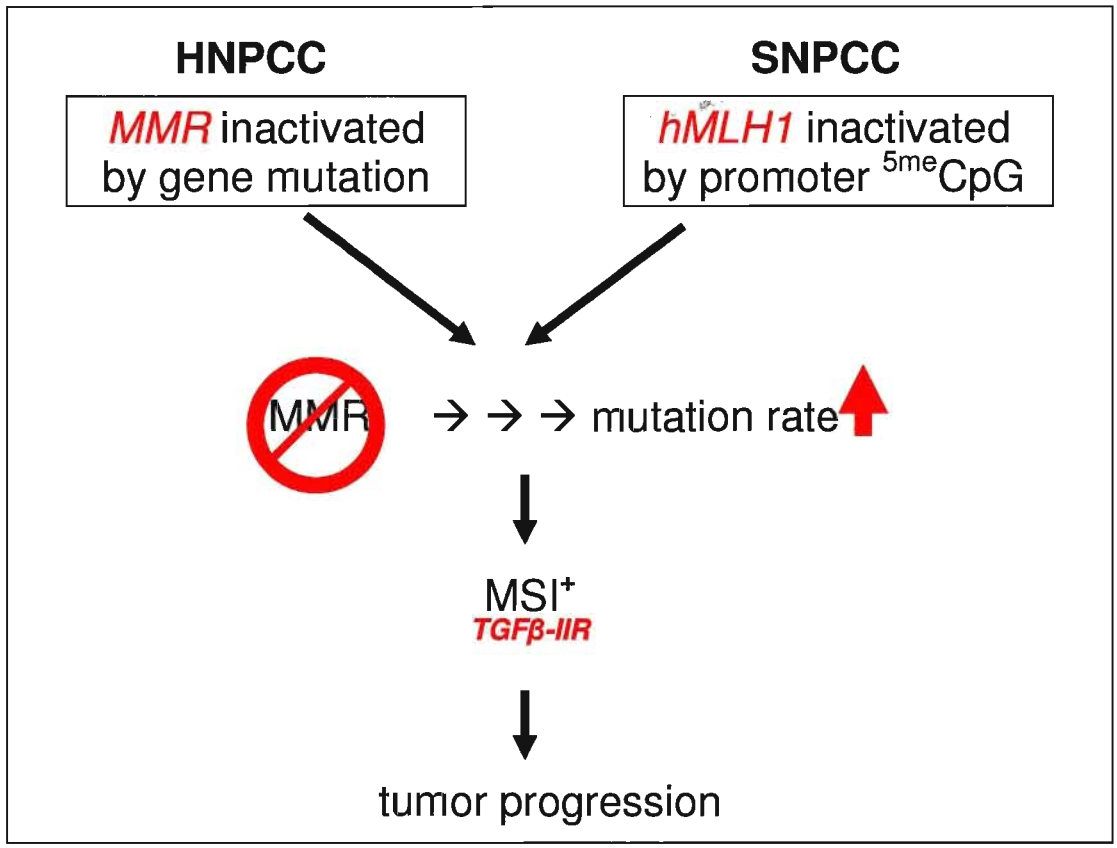


Figure 1.3: Common carcinogenic pathway in HNPCC and SNPCC.

Although HNPCC and SNPCC share a common carcinogenic pathway, SNPCC also has several distinguishing features from HNPCC. SNPCC MLH1$^-$ cell lines and tumors tend to be associated with a particular *BRAF* mutation denoted V600E that constitutively activates the mitogen-activated protein kinase (MAPK) pathway (i.e. RAS-RAF-MEK) in the absence of extra-cellular signals; thus, *BRAF* mutation further distinguishes SNPCC from HNPCC (Deng et al., 2004a; Loughrey et al., 2007). Moreover, unlike HNPCC, loss of *hMLH1* expression in SNPCC occurs epigenetically. This type of disregulation is actively maintained and reversible (e.g. by treatment with AzadC in cell culture). SNPCC requires distinct diagnostic testing, such as methylation analysis, for identification (Deng et al., 2004a). Accurate identification is of paramount importance to successful treatment outcomes as SNPCC cells (i.e. *hMLH1* deficient) are often resistant to killing by methylating

agents such as cisplatin, methyl methanesulfonate, and doxorubicin (Drummond et al., 1996; Glaab et al., 1998; Fink et al., 1998) or develop resistance to commonly used methylating agents used in chemotherapy. In these cell lines, *de*methylating agents may prove effective.

Of clinical interest, the US Food and Drug Administration recently approved AzadC based antitumor agents for the treatment of myelodysplastic cancer syndromes and clinical trials are in progress to test efficacy in recurrent CRC including certain SNPCC cases (National Cancer Institute, 2011). Although AzadC has been shown to restore *hMLH1* mRNA and protein expression in SNPCC cell lines, the extent to which AzadC treatment restores the normal splicing profile of *hMLH1* has not been determined. Therefore, the *hMLH1* mRNA splicing profile in transiently (acute AzadC treatment) and stably demethylated (remain *hMLH1*$^+$ after AzadC withdrawal) colon cancer cell lines was determined and a potential drug effect of AzadC incorporation into DNA after acute treatment is discussed (chapter 4).

1.2.2.4 Consequences of hMLH1 Loss for Cells and Tumors

Loss of *hMLH1* function has serious consequences for the cell and may facilitate tumor progression in three ways: (1) through a direct role in MMR that is essential for suppressing the cellular mutation rate by 100-1000 fold in a replication dependent manner, (2) by acting as a sensor of DNA damage and modulator of cell cycle control that triggers the G_2/M checkpoint in response to methylating agents, and (3) through involvement in a mitochondrial-mediated apoptotic response conferring sensitivity to peroxide-induced oxidative stress.

Firstly, *hMLH1* deficiency contributes to genetic instability by inactivating MMR according to the classical HNPCC associated model outlined in section 1.2.2.3 and illustrated in Figure 1.3. Inactivation of MMR increases the mutation rate (e.g. through MSI) and promotes accumulation of additional mutations that facilitate cancer progression. With respect to the consequences of deficient MMR gene function, both HNPCC and SNPCC share a common carcinogenic pathway. However, loss of MMR repair capacity may not be the only way in which *hMLH1*

inactivation contributes to cancer progression. It has been shown that as little as 10% full-length hMLH1 protein expression is sufficient to restore the repair function of MMR in isogenic *hMLH1* deficient and hMLH1 expressing kidney cells whereas a full complement is required for other functions (Cejka et al., 2003).

Secondly, *hMLH1* deficient cells have been reported to exhibit resistance to killing by methylating agents, implicating MMR in cell checkpoint activation and apoptosis. Cejka et al. (2003) show that loss of hMLH1 is linked to disregulation of cell cycle control of kidney cells through the G2/M checkpoint. The researchers developed an isogenic system using hMLH1- kidney cells in which hMLH1 expression could be induced using doxycycline, denoted hMLH1- and hMLH1+ respectively. They found that: (1) hMLH+ restores sensitivity to the DNA methylating agent N-methyl-N'-nitro-N-nitrosoguanidine (MNNG), resulting in cell death, (2) only hMLH1+ cells exhibited Cdc2 tyr-15 phosphorylation (blocks entry into mitosis and provides evidence of G2/M arrest) demonstrating a requirement for hMLH1 in checkpoint activation, and (3) low amounts of hMLH1 were sufficient to restore repair function in the presence of MNNG but activation of the G2/M checkpoint required a full complement of hMLH1. Phosphorylation at Ser-15 of p53 that leads to stabilization and activation of p53 as a transcription factor in an hMLH1 dependent manner, increased with the dose of MNNG. Moreover, Ser-15 phosphorylation peaked at 48 hours which corresponded to the time point at which the majority of cells arrested at the G2/M checkpoint (second round). Intriguingly, cell cycle activation required hMLH1 however the cell death pathway was shown to be p53 independent as p53 is present but functionally inactivated in these cells (Cejka et al., 2003). Consistent with this framework, Chen and Sadowski (2005) identified *hMLH1* and *hPMS2* as p53 target genes and mapped the p53 binding elements to the first intron of each gene. Based on their results, the researchers propose a model whereby DNA damage leads to activated p53 which in turn induces (a) cell cycle arrest through p21/WAF, and (b) expression of hMLH1 and hPMS2 that act as sensors of the extent of damage; if excessive, PMS2 stabilizes p73 which is required

for p53 dependent apoptosis. Luo et al., 2004 have shown that p53 activation in response to DNA damage is augmented by ATM mediated stabilization of MutL proteins. The ratio and localization of hMLH1-PMS1 and hMLH1-hPMS2 also appears to influence signaling, demonstrating for the first time that the hMLH1-PMS1 heterodimer (MutLβ) is involved in the signaling function of hMLH1 (Luo et al., 2004).

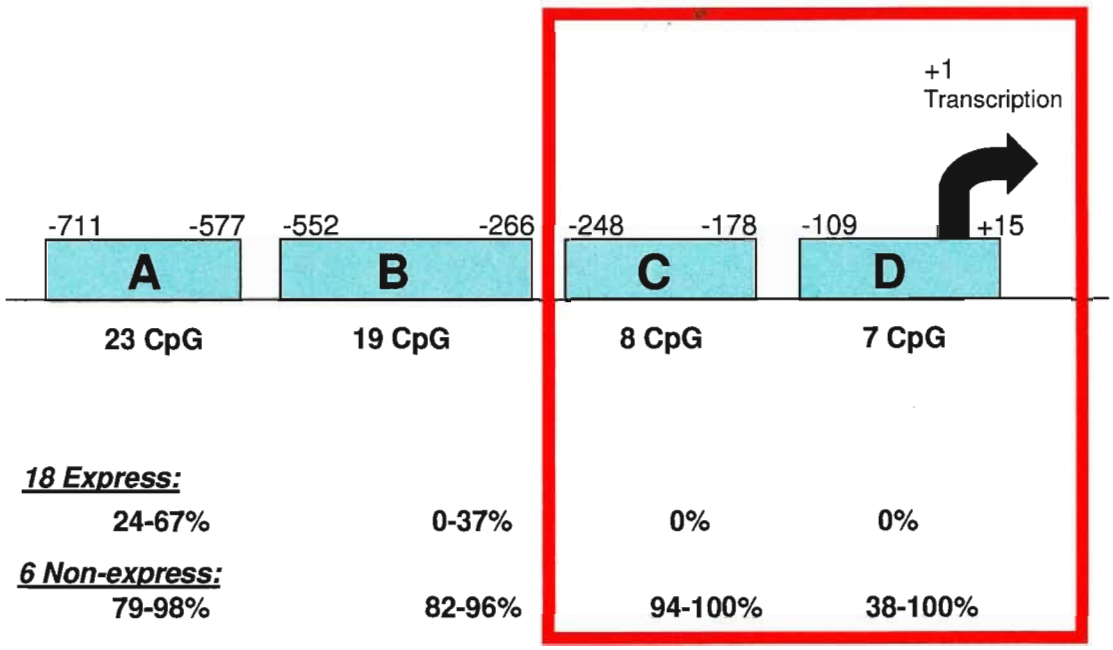
Thirdly, *hMLH1* is implicated in the response to HP-induced OS, in a manner distinct from the response to methylators, due to a requirement for hMLH1 in signaling mitochondrial-mediated apoptosis (Hardman et al., 2001). Although both MLH1+ and MLH1- cell lines show similar induction of late S-phase and G2/M cell cycle delay, MLH1+ cells show increased sensitivity to the peroxides HP and TBH due to elevated levels of cell death compared to MLH- cell lines. At this time it is not clear whether this is linked to the role of hMLH1 in removal of mitochondrial 8-oxoG (Martin et al., 2010). Taken together, the results underscore the multi-faceted roles of MMR proteins, particularly hMLH1, in genome maintenance and stability.

1.2.3 Disregulation of cytosine methylation (^{5me}CpG)

DNA methylation is an important epigenetic phenomenon that regulates a variety of functions in the genome including imprinting, X-chromosome inactivation, suppression of parasitic elements found in intronic regions, and transcriptional silencing by promoter methylation. Disregulation of normal DNA methylation profiles including aberrant promoter hypermethylation and global hypomethylation of cytosine in the context of CpG dinucleotides has been reported in cancer (Expert Reviews, 2002). In particular, aberrant promoter cytosine methylation (^{5me}CpG) of *hMLH1* has been causally linked to SNPCC progression through an increase in replication errors (MSI) resulting from impaired MMR (Herman et al., 1998), other functions of *hMLH1* notwithstanding. Silencing in SNPCC cell lines can be reversed by treatment with AzadC (Veigl et al., 1998).

How methylation of the *hMLH1* promoter interferes with transcriptional activity remains poorly understood. It is generally accepted that a certain level of

accumulated methylation leads to silencing (Ang et al., 2010) through chromatin compaction as opposed to methylation at an individual CpG position. Consistent with this framework, nucleosome occupancy at the hMLH1 promoter is associated with methylation level (Lin et al., 2007). However, azacytidine and decitabine (AzadC) have recently been shown to induce gene specific, non-random demethylation in human CRC cell lines (Hagemann et al., 2011) suggesting a specific underlying regulatory network governs CpG methylation. Consistent with this idea, Deng et al. (1999) identified a region of the *hMLH1* promoter, denoted region C (core promoter) that invariably correlates with the absence of gene expression in colorectal cancer cell lines as shown in Figure 1.4. The region includes position -178 to -248 relative to the transcription start site (+1) and contains 8 CpGs and a CCAAT element. In addition, CpG methylation at position -262, just 2bp upstream of the CCAAT element, was subsequently shown to prevent *in vitro* binding of the transcription factor CBF and inhibit transcription though it does not alone account for silencing (Deng et al., 2001). This indicates that for *hMLH1*, direct interference with particular transcription factor binding contributes to *hMLH1* deficiency and that CpGs need not be located within TF binding sites, adjacent ^{5me}CpG can inhibit binding (either directly or through conformational changes in DNA that prevent binding). It was subsequently shown that this event results in a conformational change of the DNA that restricts TF accessibility (Deng et al., 2004b). The fact that this event does not alone account for silencing underscores that various elements contribute to regulation and combinatorial control of the *hMLH1* promoter has been demonstrated (Warnick et al., 2003).



Deng et al. (1999), Cancer Research 59: 2029–2032; Deng et al. (2001), Oncogene 20: 7120-7127

Figure 1.4: CpG island methylation of the *hMLH1* proximal (to transcription start site) promoter regions C and D invariably correlates with transcriptional silencing. In 18 hMLH1 expressing cell lines, no methylation was found in promoter region C and D; 6 hMLH1 non-expressing cell lines were 38-100% methylated at the 15 CpG sites in regions C and D.

Computational scans have identified numerous potential TF binding sites in the hMLH1 promoter and binding of certain TF binding has been demonstrated experimentally (Deng et al, 2002; 2004b). Although certain TFs (e.g. CBF) contribute to an on/off or up/down regulation of transcription (i.e. of the canonical transcript) some TFs appear to influence alternative splicing. Using promoter swapping experiments in mini-genes, Cramer et al. (1997) demonstrated that changing the promoter driving the construct altered the inclusion of an exon that is subject to alternative splicing. This exon was later shown to be sensitive to changes in elongation rate, suggesting certain promoter elements can influence splicing of kinetically coupled exons (Kornblihtt et al., 2004). While promoters are not swapped in nature, they are differentially occupied by TFs in response to a variety of

parameters including tissue specificity, developmental stage, stress, and methylation. Moreover, intermediate levels of methylation have been observed in normal and matched SNPCC tissue (Ang et al., 2010) and also in cell lines demethylated by AzadC (Deng., 1999, 2001, 2004b). However, the impact, if any, on the *hMLH1* mRNA splicing profile has not been examined. Therefore, in chapter 4 of this thesis, the frequency of *hMLH1* alternative mRNA splicing events of interest (e.g. events that appear to be kinetically coupled and/or comprise domains of interest such as the repair domain) in transiently (AzadC) and stably demethylated SNPCC cell lines (for which the promoter methylation status is known) are correlated with the level of ^{5me}CpG at particular promoter regions/sites and motifs of interest are discussed.

1.2.4. Altered mRNA Splicing

Mutations that occur in elements that regulate RNA splicing can have a detrimental effect on proper protein function and downstream pathways (Pagenstecher et al., 2006). In this sense, the effect of genetic mutations is augmented by altering an epigenetic process, RNA splicing. Moreover, disregulated splicing of repair genes can lead to increased mutations in the cell as has been shown for *hMLH1*. It is estimated that at least 15% of point mutations in disease associated genes actually exert their influence on RNA splicing (reviewed in Stamm et al., 2005). An increase in mRNA splice variants is associated with cancer, suggesting that splicing regulation is impaired (Ghigna et al., 2008). Consistent with this finding, depletion of the essential splicing factor SF2/ASF has been demonstrated in specific cancers whereas over-expression is associated with transformation such that SF2/ASF is considered an oncogene. However, it is not yet clear how much of the splicing is aberrant and how much is functional or whether changes in splicing are a cause or a consequence of cancer. Since regulated alternative mRNA splicing is linked to apoptosis, an increase in apoptosis-associated isoforms is normal (not aberrant) in response to appropriate stimuli (e.g. excessive damage). Which of the observed splice variants are functional versus aberrant? Also, alternative splicing is associated with OS and OS is associated with cancer; therefore it is unclear whether

alternative splicing is a result of OS regardless of cancer status. Nevertheless, mutations that interfere with splicing regulatory elements are associated with hMLH1 deficiency in HNPCC (Pagenstecher et al., 2006). The splicing patterns either before or after AzadC treatment in SNPCC cell lines is not known. Although *hMLH1* is known to be alternatively spliced, most studies focus on quantifying the CT/WT protein but not splice variants. While it is likely that aberrant splicing of *hMLH1* mRNA occurs in normal and CRC cells, it is equally likely that some of the splice variants are functional and relate to the various roles of *hMLH1* in repair and surveillance of DNA damage in the genome. This work comprehensively characterizes those splice variants under a variety of conditions for the first time. In the sections that follow, key concepts in RNA splicing are outlined.

1.3 Splicing

The central dogma of cell and molecular biology states that in eukaryotic cells, information flows from DNA to RNA via transcription and splicing, and from RNA to protein via translation. A critical step in the fidelity of information flow in the cell occurs at the level of transcription and splicing, now understood to be coupled, and referred to as co-transcriptional splicing where splicing occurs simultaneously with transcription (Kornblihtt et al., 2004; Bentley, 2005). Splicing is carried out by the spliceosome, a dynamic multi-component complex that catalyzes intron excision and exon ligation of the nascent pre-mRNA transcript (produced by the elongating RNA Polymerase II) to form mature mRNA (Jurica and Moore, 2003) as shown in Figure 1.5.

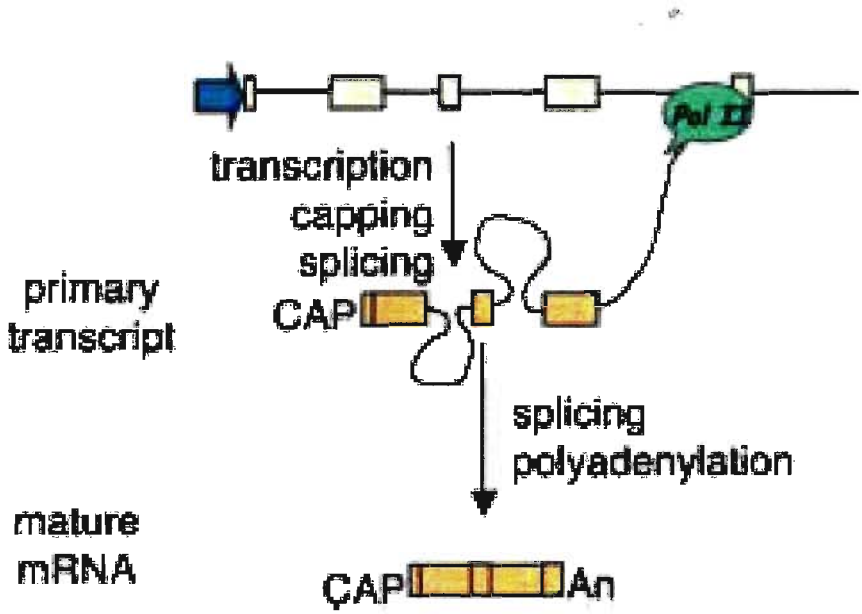


Figure 1.5: Excision of introns and ligation of exons (splicing) occurs co-transcriptionally during polymerase II elongation to form mRNA (image from Kornblihtt et al., 2004).

The spliceosome recognizes consensus sequences in the pre-RNA associated with the 5' splice site (located at the 5' intron end), the branch point, and the 3' splice site (located at the 3' intron end) with preceding polypyrimidine tract, as shown in Figure 1.6. However the sequences are degenerate and various splice site options exist that are either utilized constitutively, not utilized, utilized in error, or utilized temporally and spatially under certain conditions subject to regulated alternative splicing (e.g. in a tissue specific manner, in specific developmental contexts, in response to cellular stimuli such as stress) (Black, 2003).

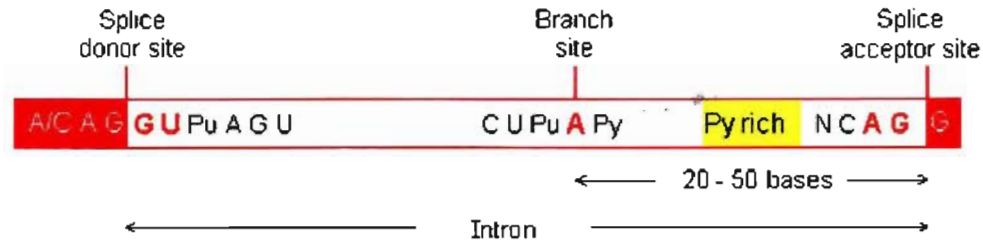


Figure 1.6: Consensus sequences recognized by the spliceosome (image from: RNA Splicing – Web Book Publications available at www.web-books.com/MoBio/Free/Ch5A4.htm).

1.4 Core Mechanisms Involved in mRNA Splicing

There are currently two core pre-mRNA splicing mechanisms that have been identified and shown to operate in humans. These include the major U2 splicing pathway and the minor U12 splicing pathway. In addition, a third, recursive splicing at non-exonic sites has been identified as a mechanism that facilitates splicing of long introns in *Drosophila melanogaster* (Burnette et al., 2005) although the phenomenon has not yet been experimentally observed in humans. Each of these mechanisms will be discussed in turn in the sections that follow.

1.4.1 The major U2 splicing pathway

The major U2 splicing pathway accounts for 95-99% of splicing in humans. Spliceosomal components, partly through complementary base pairing of small nuclear RNA (snRNA) with pre-mRNA, recognize the nearly invariant dinucleotides GU at the 5' donor end of the intron (5'ss) and AG at the 3' acceptor end of the intron (3'ss) within the context of the following degenerate consensus sequences: 5'ss MAG/GURAGU; 3'ss YAG/ preceded by a polypyrimidine tract; and the branch point adenosine at ~18-40 bps upstream of 3'ss within YNCURAY, where M is C or A, Y is a pyrimidine (C or U), R is a purine (A or G), and N is any nucleotide (A, U, C, G). Five distinct complementary snRNAs are part of protein complexes called small nuclear ribonucleoproteins (snRNPs) denoted U1, U2, U4/U6, and U5 snRNPs that form the catalytic core of the spliceosome and are aided by hundreds of additional proteins and auxiliary factors. U1 snRNP and U2 snRNP engage the 5'ss and branch point respectively through complementary base pairing. U4 and U6 pair

with each other and with U5 to form a tri-snRNP that joins the spliceosome, engages in base pairing interactions with the mRNA, and displaces U1 and U4 snRNPs. The splicing reaction occurs in two steps: (1) the branch point adenosine contacts and cleaves the 5'ss and an intron lariat is formed (lariat and 3' intron end remain attached to downstream exon), (2) 3'ss cleavage and exon ligation occur along with the release of the intron lariat (reviewed in Kramer, 1996; Black 2003; Jurica and Moore, 2003; Kalnina et al., 2005).

Most U2 introns are flanked by GU-AG, however, 1 in 20 alternative exons (i.e. may or may not be included) is a GC-AG intron (Thanaraj and Clark, 2001). Other intron types thought to be U2 spliced are observed infrequently. Although U2 splicing is responsible for excising most AU-AC introns that do occur, this intron type is also excised by the rare U12 splicing pathway which is identified by specific associated consensus sequences that are less degenerate than U2 sequences.

1.4.2 The minor U12 splicing pathway

U12 introns are estimated to comprise less than 1% of all human introns and the canonical U12 donor sites (5'ss) are GU, GC, and AU, though exceptions have been observed. U12 consensus sequences are less degenerate than U2 sequences and include the following: (1) an extended nearly invariant 5' splice site /ATATCCTT at +1 to +8 of the 5' splice junction, (2) the branch point/branch point sequence TCCTTAAC, and (3) an AC acceptor site. It is thought that for AU-AC splicing, the lack of polypyrimidine tract and a short distance between the branch site and the 3'ss in AU-AC introns commits the intron to U12 splicing. While the branch point sequence is also required, it is similar for major and minor and is thought to be able to function for either.

Interestingly, included among the examples of U12 intron containing genes is the repair gene *hMSH3* (intron 6: AT-AA) which can form part of MutS and contact *hMLH1* raising the question of whether these processes might be coordinated at the mRNA splicing level by presence or use of certain specific spliceosomal components. What determines whether a splice site will be U12 spliced or whether the balance

between U12 and U2 splicing can shift remains unknown. However, various regulatory roles have been proposed. The minor U12 splicing pathway excises a minute class of introns – the U11 snRNA pairs with 5'ss and the U12 snRNA pairs with BPS resulting in a distinct spliceosome containing U11, U12, U5 and two new snRNAs, U4atac and U6atac, along with novel proteins specific to the minor class of snRNPs but generally homologous to proteins in major class of snRNPs. Although the minor pathway processes mainly AU-AC introns, some GU-AG introns are excised by the minor pathway and other non-canonical dinucleotides have been recorded (Patel and Steitz, 2003). Mutation from G to A in the first nucleotide of the 5'ss (i.e. GU-AG to AU-AG) has been shown to inhibit the second step of U2 splicing *in vitro*, to impair gene expression, and to lead to genetic diseases such as thalassemia and Usher syndrome (reviewed in Wu and Krainer, 1999), while a simultaneous change to AU-AC seems to remove this inhibition. Thus, natural examples of AU-AG introns are expected to be processed by the minor splicing pathway. Other intron types are also observed in the U12 Data Base, albeit rarely. A change in the 5'ss from AU to CU-AC yields correct splicing in experimental models. In fact, Lin et al. (2010) have shown that the second dinucleotide of the 5'ss in U12 splice sites seem to interact differently than those of the U2 pathway (where 5' AA inhibits the second step of splicing) and tolerate greater diversity in nucleotide choice. For example, NU-AN and AN-NC (where N is any base) are tolerated and the following additional non-canonical terminal dinucleotides have recently been reported: CT-AC, GG-AG, GA-AG. Since the majority of splicing is carried out by the U2 splicing pathway, most algorithms are based on U2 (not U12) consensus sequences.

1.4.3 Recursive Splicing

Recursive splicing has been identified in *Drosophila melanogaster* where it facilitates splicing of certain long introns (usually > 10,000 bps) by providing ratcheting points (RP) at non-exonic sites that effectively subdivide long introns into smaller sections (Burnette et al., 2005). This leaves no evidence that splicing has occurred at RPs in the resulting mRNA transcript as the entire intron is still removed,

albeit in smaller sections at a time. Minimally, RPs consist of AGGU, a composite of the 3'ss and 5'ss consensus dinucleotides. Splicing at RPs regenerates 5' splice sites. The RP first functions as a 5' splice site (**AG**[GU) that splices to a downstream (3') RP (AG]**GU**) resulting in the regeneration of **AGGU** that is subsequently used as a 5'ss as shown in figure 1.7 below.

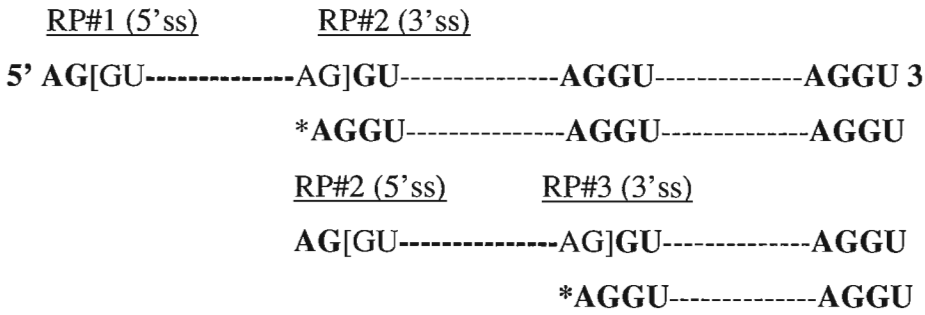


Figure 1.7: Model of Recursive Splicing in Sub-Division of Long Introns in Drosophila. Bold sections are retained as exons, light sections are removed as introns, and * denotes a regenerated 5'ss at an RP that was first used as a 3'ss and is subsequently used as a 5'ss (based on Burnette et al., 2005).

Mechanistically, recursive splicing is thought to occur via the U2 splicing pathway although this has not been explicitly verified and distinct components and factors may be involved (Burnette et al., 2005). Although recursive splicing has not yet been demonstrated in humans, it is one mechanism that could potentially contribute to the generation of splice variants. Since recursive splicing leaves no trace in the mRNA, pre-mRNA splicing intermediates of candidate genes would need to be explicitly examined in order to identify this phenomenon.

1.5 Alternative Splicing

Although mRNA splicing mechanisms and regulation lie at the crossroads of information transfer from DNA to RNA to protein, they remain incompletely understood. In particular, the mechanisms that contribute to alternative splicing (AS), the process by which two or more mRNA transcripts are generated from the same pre-mRNA transcript, are of interest as they greatly contribute to the diversity of the transcriptome and subsequently to downstream processes including protein diversity (Black, 2003). Common classes of mRNA transcripts produced from nascent pre-

mRNA include the canonical transcript (CT) (also known as wild-type, WT), the transcript type that contains the exons usually observed in a given condition, and classes of splice variants that include but are not limited to exon skipping, intron retention, exon border extension, and exon border truncation (Sugnet et al., 2004) as shown in Figure 1.8. These events are not mutually exclusive (except WT) and may occur within a single unique mRNA transcript.

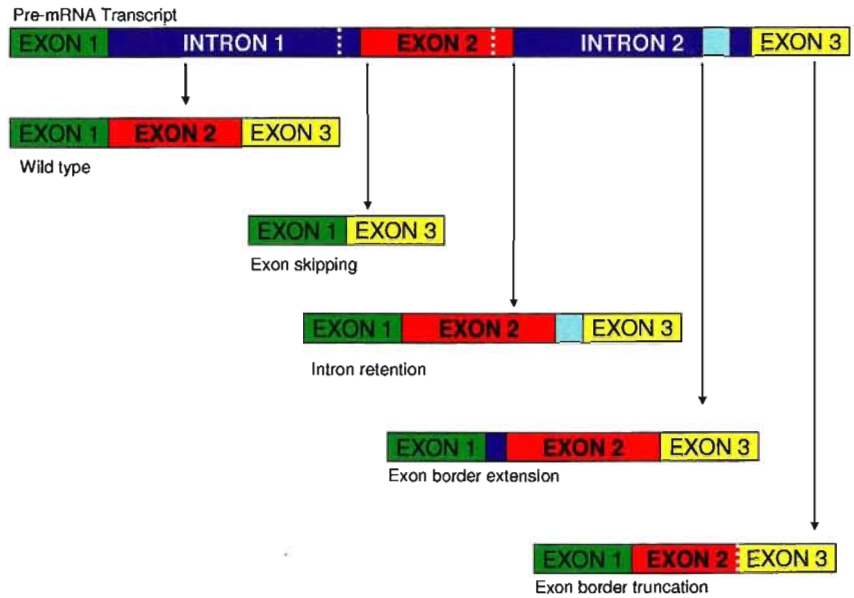


Figure 1.8: Classes of Alternative mRNA Splicing. Alternative splicing of the nascent pre-mRNA transcript yields a plethora of unique mRNA transcript types that include but are not limited to transcripts that fall into the following classes: WT, exon skipping, intron retention, exon border extension, and exon border truncation.

It is now understood that the majority of human genes undergo alternative splicing (Kim et al., 2007). Furthermore, loss of splicing fidelity has been linked to specific diseases including some cancers (Wang et al., 2003; Xu and Lee, 2003; Hui et al., 2004; Karni et al., 2007). This takes three main forms: (1) complete absence of required transcripts, (2) aberrant splicing resulting in generation of inappropriate

transcripts and/or transcript ratios, (3) mis-splicing resulting in novel non-productive transcript types, some of which may be toxic.

1.6 Splicing of *hMLH1* and *hMLH1* Alternative mRNA Transcript Types

Wild-type (WT) *hMLH1* mRNA (the canonical transcript, CT) contains a 2554 base pair coding sequence that consists of 19 exons (Genbank, 2011). *hMLH1* also undergoes alternative splicing although the functional significance (if any) of most of the splice variants remains undetermined (EMBO Alternative Splicing Data Base, 2011). Studies show that many of the *hMLH1* point mutations previously thought to be silent interfere with splicing enhancers and disrupt WT splicing of the gene (Auclair et al., 2006; Pagenstecher et al., 2006). Experimental evidence confirms that point mutations in exon 17 of *hMLH1* disrupt an enhancer element required for inclusion of exon 17 (Xu and Mattox, 2006). Exon 17 is constitutively spliced in normal tissues tested (i.e. no skipping was observed) except in colon tissue where skipping occurs in cancer and non-cancer samples (Genuardi et al., 1998). Exon 9/10 skipping has recently been shown to exert a dominant negative function on the MMR pathway (Peasland et al., 2010). Functional analysis of remaining variants remains largely unexplored despite the fact that *hMLH1* is known to undergo extensive alternative splicing and specific *hMLH1* splice variants, including exon 15 skipping, have been reported.

Early work by Charbonnier (1995) looked at the relative expression of 3 *hMLH1* transcript types: Δ9/10, Δ10/11, Δ9/10/11 in normal lymphocytes of 8 individuals and found wide ranges in relative expression (3%-84%, 21-83%, and 12-66% respectively). Genuardi et al. (1998) expanded this list to include Δ6/9, Δ9, Δ9/10, Δ9/10/11, Δ10/11, Δ12, Δ16, Δ17 and found that most were present in both normal colon and colon cancer although relative amounts were not measured, only presence or absence of the isoforms. Interestingly, Δ17 was the only transcript type found in no normal tissues (0/4) that were tested except for the normal colon where it was found in 4/5 normal colon samples tested and Δ17 was observed in all types of cancer samples tested. Could this alternative splicing event be related to the high

propensity of colon cancer found in the population relative to other types of cancer? Correct splicing of E17 has subsequently been shown to depend on enhancer activity within the exon and though regions of the exon containing the putative cis-elements relevant to skipping have been narrowed down, the specific factor(s) involved have not been confirmed (Xu and Mattox, 2006). In addition, Palmirotta et al. (1998) identified Δ15 (results in frame-shift and premature termination codon) and Δ16 (in-frame deletion) in normal lymphocytes. Recently, Peasland et al. (2010) investigated *hMLH1* splicing in a cohort of children diagnosed with acute lymphoblastic leukemia and found *hMLH1* to be highly variable under both normal and disease conditions. Alternative splicing events included exon skipping and alternative splice site use. Functional analysis of Δ9/10, one of the most abundant variants, exhibited significant dominant negative activity on MMR (in both conditions). These findings suggest that *hMLH1* is subject to complex alternative splicing regulation and raises the possibility that the relative frequencies of certain transcript types may be more relevant than whether they are present or absent in terms of their impact on disease. Based on these findings as well as the growing understanding of the importance of alternative splicing, revisiting the question of what constitutes the normal splicing pattern of *hMLH1* in colon cancer is warranted.

1.7 Factors that Influence Constitutive and Alternative mRNA Splicing

The U2 consensus sequences described previously in section 1.4.1 are necessary for U2 splicing but not sufficient as they are degenerate and various splice site options exist (Black, 2003). This raises the question of what parameters define splice site usage. In the sections that follow, key factors that influence splicing and splice site selection will be examined including splice site strength, cis-elements and trans-acting factors, promoter architecture and kinetic coupling, and chromatin structure and models of exon definition.

1.7.1 Splice Site Strength

One key determinant of splice site usage is splice site strength, a measure of how well the nucleotides at a given splice site correspond to the consensus motif for

the 5' or 3' splice site that is utilized. The consensus motifs were originally determined by Shapiro and Senapathy (1987), are built on the degree of conservation in an alignment of 1446 sequences, and assume independence between individual positions in the motif. Since the 5' consensus conforms to Watson-Crick base-pairing with the U1 snRNA (Horowitz and Krainer, 1994), measures of 5'ss strength are essentially measures of complimentarity to the 5' terminus of U1 snRNA. Roca et al. (2005) carried out a detailed analysis of 5'ss strength using mutational analyses, competition assays, and correlations with six different methods for determining splice site strength (including Shapiro and Senapathy, 1987). While the six methods generally yielded similar results, subtle differences were observed that the researchers ascribe to subclasses of 5' splice sites they denote strong (high complimentarity to U1 snRNA), intermediate (some complimentarity), and weak (only used in absence of competing strong or intermediate splice sites and probably dependent on other factors such as enhancer activity). Strong splice sites were found to be dominant when in competition with intermediate or weak sites, consistent with previous studies that found splice site strength is an important determinant of splice site selection (Roca et al., 2003). Where two strong splice sites competed, the best predictor of splicing outcome was a measure of free energy for the stability of RNA duplexes between the 5'ss sequence and the U1 snRNA terminus. Free energy in the Roca et al. (2005) study was calculated using the Turner energy rules that are based on measurements of hydrogen bonding, base stacking, mismatches, and Watson-Crick or G-U base pairs (Serra and Turner, 1995) and allow finer discrimination than the Shapiro and Senapathy method (because they take additional parameters into account). When discriminating between intermediate strength sites, the Yeo and Burge (2003) MAXENT model provided the best discrimination and, unlike the free energy model, this model takes the interdependencies between positions in the sequence motif into account. This is consistent with the idea that other sequence patterns make a dominant contribution when base-pairing is limited. In a similar vein, Nogués et al. (2002) examined 3'ss strength and demonstrated, through mutational analysis of the

polypyrimidine tract and splicing assays, that a stronger (better recognized) 3'ss results in normal splicing whereas a weaker 3'ss is more susceptible to the effects of transcriptional elongation that have been shown to increase the inclusion of an elongation sensitive (kinetically coupled) alternative exon. Taken together, these results suggest that strong splice sites exert a dominant effect but that other sequences in the proximity (or even overlapping) the consensus sequences (e.g. enhancers and silencers) are also important, especially for suboptimal exons with intermediate or weak splice sites, where they are perhaps equally or even more important than splice site strength.

In addition, dual-specificity splice sites that function alternatively as 5' and 3' splice sites have been identified (Zhang et al., 2007). When the dual splice site (AGGT) is utilized as a 3'ss, the upstream sequences are removed as an intron; when utilized as a 5'ss the downstream sequences are removed as an intron. Unlike RPs, dual splice sites can be detected in mRNA transcripts. Dual specificity splice sites are assumed to be processed by the U2 splicing machinery though this has not been verified and the factors that regulate dual splice site use have not been identified. Dual splice sites were found to have higher GC content than around constitutive splice sites, though it is not yet clear whether this is related to their regulation (Zhang et al., 2007). One factor that distinguishes dual splice sites use is splice site strength. Dual splice sites have relatively high splice site scores using both 3' and 5' splicing matrices whereas canonical splice sites tend to fall into 3' and 5' populations based on splice matrix scores. Utilization as a 5' or 3' splice site was determined in part by splice site strength (e.g. stronger 5' score versus weaker 3' score), the strength of the upstream and downstream splice sites (e.g. strong upstream 5' site favours use of downstream dual as 3' site), and other regulatory sequences (Zhang et al., 2007).

Recently, Desmet et al. (2009) developed Human Splicing Finder (HSF) as a comprehensive online bioinformatics tool to predict potential splicing signals (available at: http://umd.be/HSF/) using position-weight matrices derived from Shapiro and Senapathy; a weight is assigned to each nucleotide based on its frequency

and the relative importance of its position in the sequence motif (plus a constant for normalization) and the values are summed such that higher scores denote stronger splice sites and a score of 80 or more is considered a strong splice site. In addition to algorithms for splice site and branch point strength, algorithms to predict cis-element locations for seven different enhancers and silencers have been combined into one comprehensive tool that is freely available and easy to use allowing more of the relevant factors that modulate mRNA splicing to be examined at once. For this reason, HSF was used for the analyses in this thesis.

1.7.2 Cis Elements and Trans-Acting Factors

In conjunction with splice site consensus sequences and strength, additional cis-elements contribute to splicing regulation. These elements are comprised of specific sequences of bases, often degenerate, that are found in introns, exons, and even overlapping splice sites. Cis-elements enhance or suppress exon recognition and splice site usage when bound by their cognate trans-acting factors. Elements found in exons that enhance exon inclusion are called exonic splicing enhancers (ESEs) whereas elements found in exons that suppress exon inclusion are referred to as exonic splicing silencers (ESSs). Intronic elements that enhance or suppress inclusion of the adjacent exon also exist and are referred to as Intronic Splicing Enhancers (ISEs) and Intronic Splicing Silencers (ISSs) respectively. Enhancers and silencers are known to function in both constitutive and alternative splicing regulation and influence 5'ss and/or 3'ss selection leading to constitutive splicing, exon skipping, exon border truncations and extensions, intron inclusion, cassette exons, and mutually exclusive exons. Depending on the context, the same sequence elements may act as an enhancer to promote exon inclusion (e.g. from after the exon) or as a silencer to prevent exon inclusion (e.g. found before the exon) underscoring the challenges in predicting the action of regulatory elements and trans-acting factors even when their location is known and utilized (reviewed in Long and Caceres, 2009).

Some of the best studied trans-acting factors are the SR proteins: SRp40, SF2/ASF, SC35, and SRp55. Computational tools, such as ESE Finder have been

developed to identify putative binding sites for these factors (Cartengi et al., 2003). SR proteins derive their name from their characteristic C-terminal domain containing serine-arginine (SR) repeats that are highly phosphorylated and mediate protein-protein interactions (reviewed in Shepard and Hertel, 2009). SR proteins also contain one or two RNA recognition motifs (RRM) that contact(s) the pre-RNA regulatory motif within a particular sequence context from which putative cis-elements are computationally predicted. Numerous other SR proteins have been identified; the ones examined in this thesis are selected because their RRM has been studied and algorithms exist for predicting their locations in *hMLH1*. In addition to the four SR proteins mentioned above, recently binding sites for TRA2-BETA and 9G8 as well as the repressor hnRNP A1 have been identified and algorithms to search sequences for binding sites have been developed as part of HSF (Desmet et al., 2009). It is worth noting that additional hnRNP family members have been identified but their corresponding cis-elements and/or interacting partners remain elusive or have not yet been developed into searchable algorithms. Redundancy, antagonism, synergy, and context (e.g. proximity to 3' versus 5' splice site, location in exon versus intron) also affect splicing outcomes and add to complexity thus complicating efforts to establish definitive parameters for prediction of cis-element and trans-acting factor influence on splicing outcomes. Moreover, additional elements and trans-acting factors remain unidentified. For example several groups have identified hexamer sequences associated with silencing and enhancement computationally (Yeo et al., 2007; Wang et al., 2009), some of which correspond to known enhancers and silencers whereas others remain unidentified, highlighting the fact that splicing factors and networks remain incompletely understood.

A variety of SR proteins are known to respond to stress (e.g. H_2O_2-induced, genotoxic), apoptosis, and cancer with changes in expression, ratios, and isoforms that influence alternative splicing of their target sets of genes. These include TRA2-BETA (Takeo et al., 2009), SRp55 (Filippov et al., 2008), SC35 (Merdzhanova et al., 2008), and SF2/ASF (Li et al., 2005). Moreover, FOX-2, an RNA binding protein

that recognizes the (U)GCAUG motif, has been shown to regulate SRp55, SRp40, and a variety of other SR proteins and hnRNPs in human embryonic stem cells suggesting that it can act as a master regulator of other splicing factors (Yeo et al., 2009). In addition, expression of RNA splicing factors is controlled at the transcriptional level by promoter architecture. Chromatin remodeling by stress, mediated by the SIRT1 family of histone deacetylases, regulates key transcription factors involved in cell growth , differentiation, stress resistance, and apoptosis (reviewed in Rajendran et al, 2011; Olmos et al, 2011).

1.7.3 Promoter Architecture and Kinetic Coupling

Alberto Kornblihtt and colleagues have shown that promoter architecture influences alternative splicing through the coupling between transcription and splicing (reviewed in Kornblihtt, 2007). Originally termed 'the promoter effect', this was demonstrated using promoter swapping experiments (Cramer et al., 1997). One proposed mechanism to explain the results is recruitment of splicing factors to the transcribing RNAPII; this appears to be linked to particular features of the RNAPII carboxyl terminal domain (e.g. phosphorylation) (Kornblihtt, 2007). A second proposed mechanism to explain the results is referred to as the kinetic coupling theory of splicing and is based on a body of research that shows RNAPII elongation rate influences splice site choice of certain alternative exons that are subject to kinetic control by affecting the timing with which two competing splice sites are presented to the splicing machinery (Nogués et al., 2002; de la Mata et al., 2003; Kornblihtt et al., 2004). Kinetic coupling appears to be linked to changes in chromatin structure and other factors (e.g. pausing architecture of RNAPII) that affect transcription elongation (Kornblihtt, 2007).

1.7.4 Chromatin Structure and Exon Definition

A growing body of evidence points to chromatin structure and nucleosome organization in marking and protecting exons and facilitating splicing fidelity (Schwartz et al., 2009; Tilgner et al., 2009; Andersson et al., 2009). Nucleosomes have been found to be enriched in exons, absent from intronic regions flanking exons,

and reduced in introns in general. The CG content of exons also is higher but not condensed in the manner associated with parasitic (silenced) elements in non-incorporated intronic regions. Furthermore, the histone acetylation and methylation signature of actively transcribed genes has recently been characterized and found to correlate with exons thus marking genes that are actively transcribed (Andersson et al., 2009; Kolasinska-Zwierz, 2009)

There are currently two main models to explain how exons are defined (reviewed in Kornblihtt et al., 2009). The first model assumes that exons are defined at the DNA level by nucleosome positioning. Nucleosomes are preferentially bound to exons whereas nucleosomes are depleted in introns (Schwartz et al., 2009; Tilgner et al., 2009) and there is a distinct signature associated with actively transcribed genes at nucleosomes in the body of the gene (Kolasinka-Zwierz et al., 2009), in addition to the better characterized nucleosome signatures in promoters. It is suggested that nucleosomes slow down RNAPII for example, by providing a pausing architecture for RNAPII (Hodges et al., 2009) and facilitate co-transcriptional recruitment of splicing factors to the nascent pre-mRNA, thus improving exon definition. This model is also consistent with preferred exon length of about 147nt which corresponds to the length of bases wrapped around nucleosomes (Luger et al., 1997).

The second model relies solely on exon definition at the pre-mRNA level and posits that splicing factors are recruited to the splice sites flanking the exon in pre-mRNA that interact either directly or indirectly to support exon recognition (across the exon) and splicing (Berget et al., 1995). This model is consistent with the observations from experiments on splice site strength discussed earlier that indicate sequences in addition to the splice site consensus contribute to the decision to splice or not to splice and that these sequences may overlap part of the splice site consensus. The model also explains the selective pressure for a conserved exon length of 140-150 nucleotides in terms of the maximum distance across which protein factors can interact to define the exon. In this model, splicing factors (e.g. SR proteins) are recruited to the pre-mRNA as a consequence of ongoing transcriptional elongation by

RNAPII. Although there is strong evidence to support this model, it is likely that additional regulatory mechanisms also influence exon definition. In principal, these models are not mutually exclusive and it is likely that both contribute to RNA splicing regulation.

1.8 Current Approaches to Investigate Splicing

There are currently several approaches that are being employed to answer the big questions in alternative mRNA splicing research. Query of expressed sequence tag (EST) databases allowed the range of alternative splice variants generated under specific conditions to be identified, at least qualitatively. These data indicate the diversity and complexity that is derived from a relatively small number of human genes via alternative mRNA splicing. However, ESTs have an inherent 3' end bias and do not account for entire mRNA transcripts (e.g. due to 5' end trimming). Similarly, exon-junction microarrays allow transcript expression to be discerned on a global scale and provide valuable information about splicing networks, tissue specific expression patterns, and differences in expression in response to various conditions such as stress, drug treatment, or disease. However they generally rely on pre-determined splicing junctions and are limited by the number and location of their probes and as such do not identify novel splicing events or even all the events that occur together within a single transcript. Mini-genes and *in vitro* splicing assays have been used successfully to manipulate and study splicing of individual exons or transcripts however they cannot account for the influence of chromatin structure on splicing. A vast array of computational tools have emerged (many of which have been discussed in earlier sections) that are useful for identifying potential enhancer and silencer activity and their effect on splicing outcomes however whether these sites are actually utilized in vivo depends on a host of factors and requires verification. Cell culture and cloning approaches allow identification of full mRNA transcripts under a defined set of conditions at a given point in time and provide comprehensive mRNA expression profiles. However they are labour intensive, expensive, and are limited to single gene expression and not informative about gene

networks. As such they are appropriate where a particular gene (e.g. disease associated) is of interest, as is the case for *hMLH1*, and it is the experimental approach that is used in this study although online computational tools are also used for comparison and model building. Taken together, these approaches can provide valuable information about the regulatory mechanisms and factors that influence alternative mRNA splicing.

Chapter 2: A comparison of *hMLH1* mRNA splicing profiles in MRC5 cells under baseline conditions and H_2O_2 induced oxidative stress.

The goal of chapter two is to characterize *hMLH1* mRNA splicing profiles in MRC5 cells under basal conditions and following exposure to oxidative stress.

It is hypothesized that an important effect of cell stress is to fundamentally alter the transcriptome resulting in a decrease in the frequency of the canonical transcript (CT) and a corresponding increase in splice variants (SVs) both in terms of diversity of SVs and frequency of particular SVs.

Contributions: Performed all *hMLH1* mRNA transcript characterization (as described in Materials and Methods, 2.2.2), analysis, and chapter write-up.

2.1 Introduction

hMLH1 is involved in several cellular processes including MMR, DNA damage surveillance, G_2M checkpoint activation, and the mitochondrial-mediated cellular response to OS (reviewed in O'Brien and Brown, 2006; Mastrocola and Heinen, 2010). Functional MMR is estimated to reduce the cellular mutation rate by up to 1000 fold (O'Brien and Brown, 2006). Alterations in *hMLH1* are associated with a variety of disease processes including specific cancers. Studies show that many of the *hMLH1* point mutations previously thought to be silent actually interfere with splicing enhancers, resulting in non-canonical transcripts that cannot support proper protein synthesis (Auclair et al., 2006; Pagenstecher et al., 2006; Tournier et al., 2008). Based on this observation, it has been suggested that mRNA screening of *hMLH1* should precede functional protein studies (Pagenstecher et al., 2006).

The observation that altered RNA splicing mediates the effects of a significant number of DNA mutations underscores the importance of fidelity in RNA processing for proper cell function. This raises the possibility that non-mutagenic events (e.g. changes in splicing factor ratios in response to OS) that influence RNA splicing regulation contribute to *hMLH1* deficiency and CRC progression. Although *hMLH1* is known to undergo alternative splicing in non-cancer and cancer derived human tissues and cell lines (Genuardi et al., 1998; Peasland et al., 2010), a comprehensive splicing profile of *hMLH1* that characterizes full-length mRNA transcripts has not been established. Therefore, in this chapter, the types and frequencies of *hMLH1* mRNA transcripts generated are characterized in primary untransformed human fetal lung fibroblast cells (MRC5) propagated under standard culture conditions, to establish a baseline (non-cancer) *hMLH1* splicing profile.

It has recently been shown that *POLB* splicing specificity is sensitive to OS in 500HP treated MRC5 cells compared to baseline (Disher and Skandalis, 2007). Furthermore, H_2O_2 induced OS resulted in different types of *POLB* SVs in MRC5 compared to the lymphoblastoid cell line TK6 and its isogenic p53 null counterpart NH32, suggesting that OS affects the balance of splicing factors that govern

alternative splicing in a cell type specific manner (Disher and Skandalis, 2007). If the *hMLH1* mRNA splicing profile is sensitive to stress, this raises the possibility that OS may contribute to *hMLH1* deficiency in healthy individuals and it might explain why it leads to problems in some individuals under some conditions. This possibility was tested by looking at the response to H_2O_2 in MRC5 cells; in addition to baseline (0HP), the *hMLH1* mRNA splicing profile in characterized after 250HP and 500HP.

The goal of Chapter 2 is to characterize *hMLH1* mRNA splicing profile in MRC5 cells under basal conditions and following exposure to H_2O_2 induced OS. This is done by identifying, quantifying and categorizing the *hMLH1* mRNA transcript types generated in terms of the classes of splicing they contain, splice site type and strength, and core splicing mechanisms. It is hypothesized that an important effect of cell stress is to fundamentally alter the transcriptome resulting in a decrease in the frequency of the canonical transcript (CT) (also referred to as *w*ild-*t*ype, WT) and a corresponding increase in splice variants (SVs) both in terms of diversity of SVs and frequency of particular SVs.

2.2 Materials and Methods

2.2.1 cDNA Library Construction

In order to establish a normal (non-cancer) baseline mRNA splicing profile for *hMLH1*, a cDNA library constructed from primary untransformed human male fetal lung fibroblast MRC5 cells (9 passages), was obtained; this cDNA library was originally generated to characterize the *POLB* mRNA splicing profile in MRC5 cells under baseline (0HP) and 500HP conditions (Disher and Skandalis, 2007). Growth conditions, H_2O_2 treatment, RNA isolation, reverse transcription, and PCR with oligo-dT primer to generate a cDNA library were carried out by Kim Disher as part of the previous study according to standard lab protocols that have been described (Skandalis and Uribe, 2004; Disher, 2005; Disher and Skandalis, 2007). Briefly, MRC5 cells were grown under standard culture conditions and treated as follows: (1) cells were adjusted to $1x10^6$ cells/mL in serum-free medium, (2) cells were exposed to 0uM, 250uM, and 500uM H_2O_2 for 30 minutes, (3) H_2O_2 was removed, and (4) cells were incubated for 4h (37°C, 5% CO_2) to allow phenotypic expression. Cell survivability as measured by MTT assay (Promega) was 50% and 40% for 250HP and 500HP respectively at time of harvesting. Total RNA was then isolated from cells in each condition using the GenElute™ RNA Isolation Kit (Sigma, Canada) according to the manufacturer's instructions. Total RNA was used as a template for reverse transcription by SuperScriptIII™ (Invitrogen) with oligo-dT primer (binds to poly-A tail of mRNA transcripts), as per the manufacturer's instructions, to produce cDNA.

2.2.2 Transcript Characterization

PCR amplification of cDNA with gene specific primers (see table 2.0) was used to isolate mRNA transcripts originating from *hMLH1*. The gene specific *hMLH1* PCR product was ligated into the pGEM-T easy vector system (Promega) according to the manufacturer's instructions and incubated at 4°C overnight (12h) followed by transformation and cloning in JB110 *E. coli* cells (Invitrogen).

Specifically, 1µL of incubated ligation was added to JB110 cells following the manufacturer's instructions. Briefly, cells were electroporated and incubated (in shaker) for 30 minutes at 37°C, plated (50-80uL) on agar containing X-gal, ampicillin, and tetracycline (Fermentas) in Petri dishes and incubated at 37°C for 24 hours until discrete colonies were visible. Blue-white colony selection was used to isolate successful transformants (each representing an individual mRNA transcript) under sterile conditions. White colonies (which contain successful transformants) were selected and each was immersed in 100µL milli-Q water in a separate well of a 96-well PCR plate – 4µL were subsequently used as a template in colony PCR (25µL reactions) to produce sufficient product to visualize on an agarose gel. Reactions were as follows: 15.8µL milli-Q water, 2.5µL Thermopol™ 10X Buffer with Mg++ (*N*ew *E*ngland *B*iolabs, NEB), 1.2µL 10mM dNTPs (NEB), 0.6µL each of *hMLH1*fullFORWARD and *hMLH1*fullREVERSE primers (Sigma Aldrich) (see table 2.0), 0.3µL *Taq* polymerase (NEB), and 4µL template (containing copies of an individual cDNA clone in milli-Q water that represents a single mRNA transcript present in MRC5 cells at the time they were harvested). PCR samples (3-5µL) were separated by 1% agarose gel electrophoresis (with ethidium bromide), imaged on the Fuji phosphoimager and analyzed using Multi-Gauge 2000 software. To enhance visual discrimination between transcript types and facilitate accurate grouping of like transcript types, PCR product was digested with *Hin*dIII (NEB) (results in 3 bands for the full *hMLH1* CT) and visualized again using gel electrophoresis with 100bp DNA ladder (Invitrogen; Fermentas). Transcripts were then grouped based on size and a representative sample from each group was sequenced (Genome Quebec) to determine the identity of the variants. PCR amplification was for 40 cycles (95°C for 30 s, 60°C for 60 s, 72°C for 90 s) with an initial 2 minute denaturation period at 95°C prior to cycling and a final 10 minute extension at 72°C afterward. Oligonucleotide primers (designed in lab, ordered from Genolysis, Sigma Aldrich) for amplification and sequencing of *hMLH1* mRNA transcripts (cDNA clones) are provided in table 2.0.

Table 2.0: *hMLH1* gene specific oligonucleotide primers.

Primer Name	Location	Primer Sequence
*hMLH1*fullFORWARD[1,2]	5' UTR	5' CATCTAGACGTTTCTTGGCTCTTC 3'
*hMLH1*fullREVERSE[1]	Exon 19	5' AAGACTTTGTATAGATCAGGC 3'
*hMLH1*frontREVERSE[2]	Exon 13	5' ACAAGCTGCAGTCATTTCCTTTCG 3'
*hMLH1*backFORWARD[3]	Exon 8	5' TCAACCGTGGACAATATTCGCT 3'
*hMLH1*backREVERSE[3]	3' UTR	5' AAAGGAATACTATCAGAAGGCAAGTATA 3'

CT/WT mRNA: [1]*hMLH*1 full, 2293 base pairs (bps); [2]*hMLH1* front, 1517 bps; [3]*hMLH1* back, 1788 bps

2.2.3 Determining Consensus Motifs with Human Splicing Finder

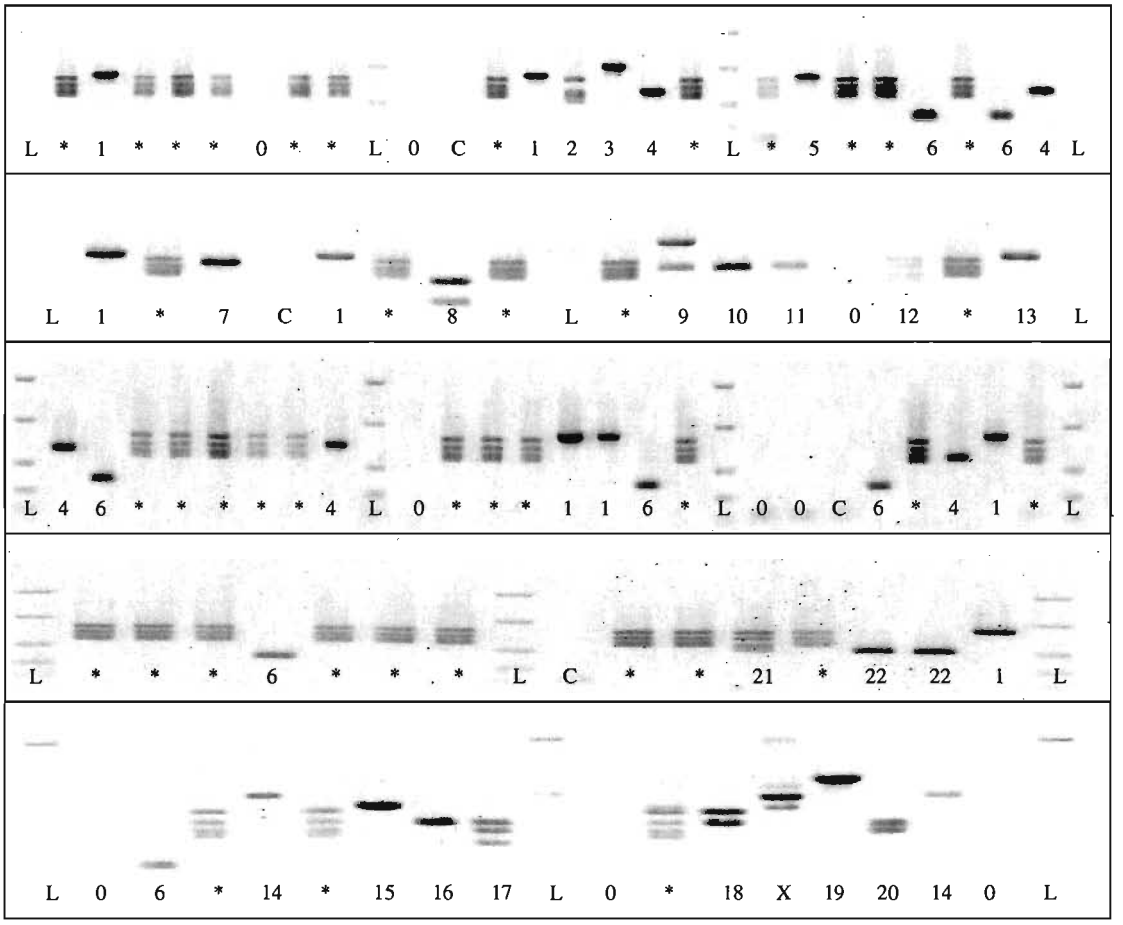
Potential donor (5') and acceptor (3') consensus splice sites for each *hMLH1* exon were generated using Human Splicing Finder (HSF). Where alternative splicing resulted in ambiguity over the actual splice site at a splice junction (based on sequence results), the bases were assigned to the upstream or downstream exon in the manner that maintains both 5' and 3' consensus splicing at HSF supported donors and acceptors respectively, where possible. HSF matrices were also used to map putative enhancer (ESE) and silencer (ESS) binding sites in and around *hMLH1* exons (see chapter 3) and for branch point (BP) analysis, as required. HSF analysis is built on U2 consensus sequence matrices and is available at: http://www.umd.be/HSF/.

2.3 Results and Discussion

2.3.1 Types of *hMLH1* mRNA Transcripts in MRC5 Cells from All Conditions

The *hMLH1* mRNA splicing profile (types and frequencies of transcripts generated) was characterized in MRC5 cells after 0 μM, 250 μM, and 500 μM hydrogen peroxide (H_2O_2) treatment which induces oxidative stress (OS) heretofore referred to as baseline, 250HP, and 500HP respectively. Agarose gel visualization and analysis of *Hin*dIII digested *hMLH1* mRNA PCR product from individual clones (mRNA transcripts) supports the presence of 50 unique *hMLH1* mRNA transcript types in this study (all conditions combined). These include the canonical transcript (CT) that corresponds to wild-type (WT) hMLH1 protein and 49 splice variants (SVs). As an example, MRC5 baseline results are shown in Figure 2.0. Of the 50 mRNA transcripts observed in MRC5 from all conditions combined, 32 unique types, including the CT and 31 SVs, could be definitively identified (i.e. the entire sequence could be resolved) based on sequencing results, as shown in Figure 2.1. The 31 SVs can be divided into two categories based on mechanism as determined by the types of splice sites and exon-skipping events they contain relative to the CT. Category 1 (C1): SVs (16 types) in which *only* whole canonical exons are skipped that contain one or more individual exon-skipping events and/or skipping events of up to 4 consecutive exons provided a canonical donor (5'ss) and acceptor (3'ss) are utilized. Category 2 (C2): SVs (15 types) skipping 8 or more consecutive exons flanked by a truncated exon at the splice junction (i.e. a non-canonical splice site is utilized that is located within a canonical exon). Minimally, all of these SVs skip exons 6-13 in their entirety and at least one boundary (donor, acceptor, or both) occurs within a truncated exon outside of this range. The donor tends to occur within exon 1, 2 or 3 (found in 6/15, 3/15, and 5/15 transcript types respectively) and the acceptor tends to occur within exon 14, 15, 16, or 18 (found in 1/15, 11/15, 2/15, 1/15 SVs respectively) although exon 2, 3, and 15 appear to contain both donor and acceptor sites that may function as dual splice sites. No exon-truncations occur in exon 17 or exons 4-E13,

although these exons are skipped in their entirety. In contrast, only exon truncations of exon 1, 3, and 18 are observed but these exons are not skipped in their entirety. Exon 2, 14, 15, and 16 are skipped as whole exons in certain C1 transcript types (found in 1/16, 1/14, 5/16, and 6/16 SVs respectively) and truncated in certain C2 transcript types (for exon 15 in numerous transcript types). No intron retention or exon border extension was observed under any conditions in MRC5 cells.



Legend: * denotes CT; 1-22, SVs; 0, no amplification; X, omitted due to double bands in original PCR; C, negative control (no amplification); L, Fermentas Mid-Range Ladder/size marker (from bottom, in base pairs: 50, 200, 400, 850, 1500)

Figure 2.0: Agarose gel electrophoresis and visualization of individual *Hin*dIII digested *hMLH1* mRNA transcripts (based on gel banding pattern), cloned and PCR amplified from baseline MRC5 cells, reveals 23 unique transcript types (CT and 22 SVs); in total, 50 unique transcript types (CT and 49 SVs) are observed in MRC5 baseline, 250HP, and 500HP combined (based on gel banding pattern).

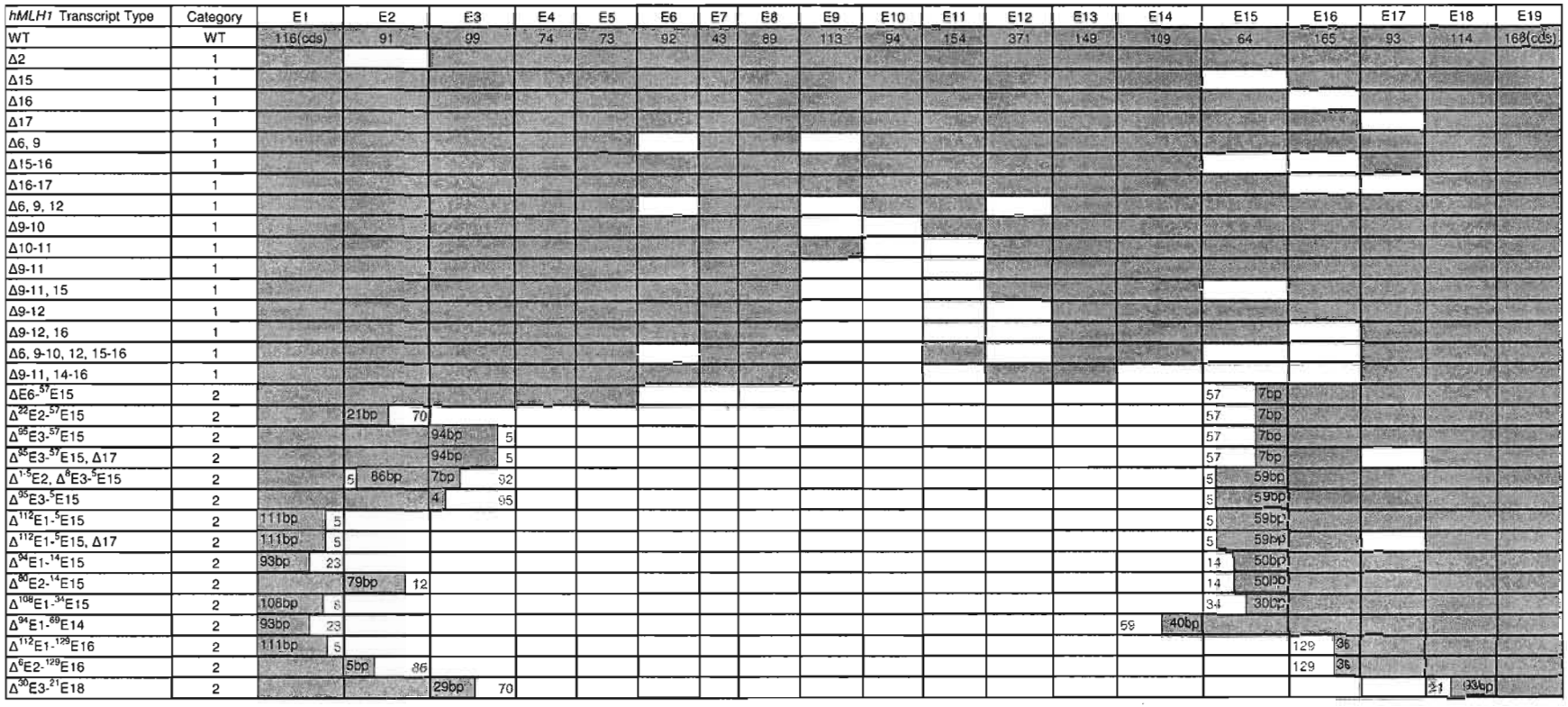


Figure 2.1: Thirty-two *hMLH1* mRNA transcript types identified in MRC5 cells with or without H_2O_2 treatment including the canonical transcript (CT). Retained exons are shaded grey and skipped exons are white. Splice variant (SV) transcript types can be divided into two groups: 16/31 that skip only canonical exons (category 1); 15/31 that skip 8+ consecutive exons with a non-canonical splice junction (category 2). 2/15 Category 2 transcript types also contain a category 1-like exon skipping event, Δ17. The number of base pairs (bp) in each canonical exon is indicated in black for CT exons; cds denotes coding sequence; Δ denotes exon (E) skipping; superscripts indicate position of skipped bp in each exon; number of bp retained for truncated exons are given in black and the number skipped in grey.

The remaining 18 SVs (of the 49 SVs observed in MRC5 overall) are denoted 'unidentified' and are observed rarely. These SVs are classified as C1 or C2 based on their agarose gel banding pattern (C1 SVs display multiple bands whereas C2 SVs display a single band after agarose gel electrophoresis) but for which full length sequencing coverage could not be obtained. That is, either sequencing failed (15-20% failure rate from sequencing facility) or only partial sequence coverage exists to support the transcript (due to the length of the *hMLH1* CT, complete sequencing coverage must be done in three segments – for some transcripts 1/3 segments was among those that failed resulting in incomplete coverage).

2.3.2 Frequency of *hMLH1* mRNA Transcripts in MRC5 Cells by Condition

The frequency with which each unique *hMLH1* transcript type occurs was determined for baseline (0HP), 250HP, and 500HP based on analysis of 81, 172, and 60 individual mRNA transcripts respectively. It was found that *hMLH1* is extensively spliced and exhibits 50% SVs under baseline conditions, 60% SVs after 250HP, and 88% SVs 500HP as shown in figure 2.2.

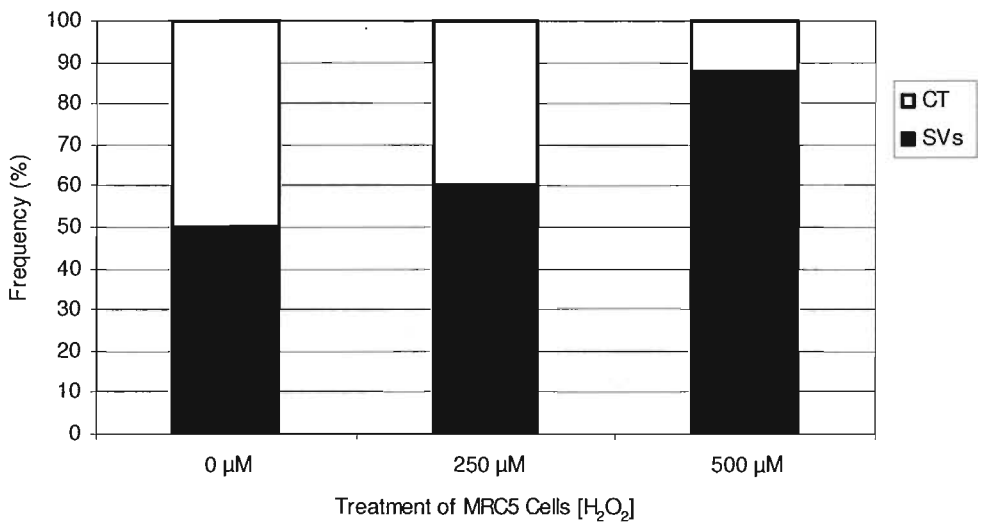


Figure 2.2: *hMLH1* mRNA canonical transcript (CT) and splice variant (SV) frequency (%) in MRC5 cells after 0uM (baseline), 250uM (250HP), and 500uM (500HP) H_2O_2 treatment.

The CT is the most frequently observed transcript type in baseline and after 250HP and the second most frequently observed transcript type after 500HP at 50%, 40%, and 12% respectively; it represents the transcript that codes for the functional protein required for MMR. All other transcript types are SVs of unknown function and total SVs equal (baseline) or exceed (250HP and 500HP) the CT in the three conditions indicating that oxidative stress changes the splicing profile of *hMLH1* in MRC5 cells. The frequencies of each of the unique transcript types (CT and 31 SVs) observed in baseline, 250HP, and 500HP are summarized in table 2.1; Δ denotes exon (E) skipping relative to the CT and superscripts indicate the position within the exon at which the splice occurs (included in the skipped section). Categories of transcript types listed include: CT, 16 unique C1 SVs, 'Other C1' SVs, 15 unique C2 SVs, and 'Other C2' SVs. The 19 transcript types (rare) that remained unidentified in this study due to incomplete sequence results were grouped as 'Other C1' and 'Other C2' based on their banding pattern on agarose gels after *Hin*dIII digestion and partial sequence results; they are included to maintain accurate overall frequencies for C1 and C2.

2.3.3 The Canonical Transcript (CT)

The canonical *hMLH1* transcript is comprised of 19 exons, each of which is followed by one of 18 intervening canonical (U2) GU-AG introns. Most introns (and exons) are flanked by strong splice sites (HSF score >80%) with the exception of the donor (5') splice site of introns 4, 6, and 18 that are characterized as moderate (HSF score of 70-80%) as summarized in Table 2.2.

Table 2.1: *hMLH1* mRNA transcript types and frequencies (%) in MRC5 cells treated with 0μM, 250μM, and 500μM H_2O_2.

			Transcript Type Frequency (%)		
Transcript Type	Frame Shift	Category	0 μM H_2O_2	250 μM H_2O_2	500 μM H_2O_2
Canonical Transcript	*No*	*CT*	*49.5*	*39.5*	*11.7*
Δ E2	Yes	1	2.5	1.7	0
Δ E6, 9	Yes	1	0	0.6	1.7
Δ E6, 9, 12	Yes	1	1.2	0.6	0
Δ E6, 9-10, 12, 15-16	Yes	1	0	0.6	0
Δ E9-10	No	1	0	1.2	1.7
Δ E10-11	Yes	1	0	0.6	0
Δ E9-11	Yes	1	0	0.6	0
Δ E9-11, 15	Yes	1	1.2	1.2	0
Δ E9-12	No	1	1.2	1.2	0
Δ E9-12, 16	No	1	0	1.2	1.7
Δ E9-11, 14-16	Yes	1	0	0.6	0
Δ E15	Yes	1	0	0.6	0
Δ E16	No	1	1.2	2.9	1.7
Δ E15-16	Yes	1	0	0.6	0
Δ E16-17	No	1	0	0.6	0
Δ E17	No	1	0	4.1	1.7
Δ Others C1 Types*	-	1	2.3	12.2	6.7
Δ E6-^{57}E15	Yes	2	1.2	0	0
Δ ^{22}E2-^{57}E15	No	2	0	2.9	0
Δ ^{95}E3-^{57}E15	Yes	2	2.5	1.2	6.7
Δ ^{95}E3-^{57}E15, 17	Yes	2	0	0	1.7
Δ ^{112}E1-^{5}E15, 17	Yes	2	0	1.2	0
Δ ^{112}E1-^{5}E15	Yes	2	12.4	15.7	41.7
Δ$^{1\text{-}5}$E2, ^{5}E3-^{5}E15	Stop, Yes	2	1.2	0	0
Δ ^{5}E3-^{5}E15	Yes	2	0	2.9	0
Δ ^{94}E1-^{14}E15	Yes	2	1.2	0	0
Δ ^{94}E1-^{69}E14	No	2	2.5	2.9	1.7
Δ ^{112}E1-^{129}E16	No	2	1.2	0	0
Δ ^{108}E1-^{34}E15	Yes	2	0	0	1.7
Δ ^{6}E2-^{129}E16	Yes	2	3.7	0	3.3
Δ ^{80}E2-^{14}E15	Yes	2	0	1.2	0
Δ ^{30}E3-^{21}E18	Yes	2	2.5	0	3.3
Δ Other C2 Types**	-	2	12.5	1.7	13.3
Total Splice Variants	-	*1&2*	*50.5*	*60.8*	*88.6*
			n = 81	*n* = 172	*n* = 60

Legend: Δ, skipping relative to canonical transcript (CT); E, exon; superscript, position in each *hMLH1* exon from start (1) to end of exon (e.g. 1-116 for E1, 1-91 for E2, 1-99 for E3, *etc.*); *n*, number of transcripts characterized; *all other category 1 (C1) transcript types (2 types & 2 transcripts, 10 types & 21 transcripts, 4 types & 4 transcripts, in baseline, 250HP, and 500HP respectively); **all other category 2 (C2) transcript types (6 types & 10 transcripts, 3 types & 3 transcripts, 5 types & 8 transcripts, in baseline, 250HP, and 500HP respectively); *n*, number of individual clones used as a template (each containing an individual *hMLH1* mRNA insert) that were successfully amplified in a 96-well PCR plate (1 plate each analyzed for baseline and 500HP, 2 plates analyzed for 250HP); Frame Shift, indicates whether the SV results in a frame shift (Yes) or maintains an open reading frame (No); Stop, generates immediate premature termination codon (PTC).

Table 2.2: *hMLH1* CT donor and acceptor splice site strength calculated using HSF U2-based matrices.

hMLH1 INTRON	U2 INTRON-TYPE	DONOR (5'SS) SCORE	ACCEPTOR (3'SS) SCORE
1	GT-AG	84.51	84.1*
2	GT-AG	94.19	80.69
3	GT-AG	96.07*	87.7
4	GT-AG	**77.1**	83.95
5	GT-AG	91.96	93.59
6	GT-AG	**77.1**	97.13
7	GT-AG	88.47	86.07
8	GT-AG	84.46	81.96
9	GT-AG	83.28	80.8
10	GT-AG	85.75	90.87
11	GT-AG	93.04	88.87
12	GT-AG	89.26	90.4
13	GT-AG	85.39	90.3
14	GT-AG	85.85	90.77
15	GT-AG	93.27*	84.57
16	GT-AG	85.94	86.53
17	GT-AG	93.22	87.88
18	GT-AG	**77.1**	87.83

Legend: Splice sites with a score greater than 80.0 are considered strong; **70.0 – 80.0, moderate**; less than 70.0, weak; splice sites denoted * occur at or near a potential dual splice site (resulting in an exon truncation) identified in this study. Transcribed *hMLH1* mRNA CT = 2524 bps & 19 exons; pre-mRNA spans 57,360 bps; 756 amino acids.

2.4 Analysis of Baseline *hMLH1* mRNA Splicing Profile in MRC5 Cells

To establish a baseline hMLH1 mRNA splicing profile in MRC5 cells, 81 mRNA transcripts generated under standard culture conditions were identified and analyzed. In all, 23 unique mRNA types were observed including the CT and 22 SVs accounting for 49.5% and 50.5% of total transcripts respectively.

2.4.1 Baseline SVs by Category

The 22 *hMLH1* mRNA SVs identified in baseline MRC5 cells can be divided into C1 and C2 as shown in Figure 2.3 and 2.4. Note that E1 and E19 must be present for a transcript to be included in this study so whole exon skipping cannot be observed for the first or last exon. Therefore, analyses of whole exon skipping exclude E1 and E19 by necessity but E1 truncations resulting from alternative 5'ss use were detected in C2 SVs and are included analyses.

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	% Skip
WT	116(cds)	91	99	74	73	92	43	89	113	94	154	371	149	109	64	185	93	114	168(cds)	49.5
Δ2																				2.5
Δ6, 9, 12																				1.2
Δ9-11, 15																				1.2
Δ9-12																				1.2
Δ16																				1.2
ΔC1 Other																				2.3

Figure 2.3: Baseline *hMLH1* C1 Splice Variants identified in MRC5 cells include transcripts with discrete exon skipping and multiple adjacent exon skipping that utilize canonical (U2) splice sites.

C1 SVs use canonical splice junctions that give rise to all GT-AG introns. Of the 7 unique C1 SVs, 5 could be identified and 2 are grouped as 'other C1'. C1 SVs account for 9% of total transcripts (7% identified, 2% unidentified) and 20% of total SVs. All individual C1 SVs are observed at low frequency as shown in figure 2.3 and include transcripts with discrete (single) exon skipping, multiple adjacent exon skipping, or both. Discrete exon skipping events observed within the transcripts include E2, E6, E9, E12, E15, and E16; multiple adjacent exon skipping is observed for E9-11 and E9-12. However, only two C1 SVs maintain an open reading frame: ΔE9-12 and ΔE16. This results in putative internally deleted proteins missing the repair domain and part of the MSH3/MSH6/PMS2 binding domain respectively.

C2 SVs represent a novel group of *hMLH1* transcripts that have not been identified in other studies (e.g. Charbonnier et al., 1995; Genuardi et al., 1998; Peasland et al., 2010).

Since these SVs exhibit multiple adjacent exon skipping of 8 or more consecutive exons, they can only be detected by examining full-length transcripts where both primers fall outside the skipped region. Of the 15 C2 SVs identified under baseline conditions, 9 were definitively identified and 6 are grouped as 'other C2' as shown in figure 2.4. C2 SVs account for 41% of total transcripts (28.5% identified, 12.5% unidentified) and 80% of total SVs. All of the C2 SVs contain reading frame shifts except for 2 SVs that maintain an open reading frame but harbor large internal deletions: ^{94}E1-^{69}E14 and ^{112}E1-^{129}E16. The 9 C2 SVs identified can be further divided on the basis of the intron-type they give rise to as follows: GU-AG – 3 types (6.2%), AU-AG – 1 type (3.7%), AA-AG – 3 types (14.8%) including ^{112}E1-^{5}E15 (12.4%) which is the most frequently observed individual SV overall and the second most frequent transcript next to the CT, AU-GC – 1 type (2.5%), and AU-CU – 1 type (1.2%).

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12	E13	E14	E15	E16	E17	E18	E19	% Skip
WT	116(cds)	91	99	74	73	92	43	89	113	94	154	371	149	109	64	165	93	114	168(cds)	49.5
Δ6-^{57}15						GU									AG					1.2
Δ^{95}3-^{57}15				GU											AG					2.5
Δ^{112}1-^{57}15		AA													AG					12.4
Δ$^{1-5}$2, ^{5}3-^{5}15			AA											AG						1.2
Δ^{94}1-^{14}15		AU												CU						1.2
Δ^{94}1-^{69}14		AU												GC						2.5
Δ^{112}1-^{129}16		AA														AG				1.2
Δ^{6}2-^{129}16		AU														AG				3.7
Δ^{30}3-^{21}18			GU														AG			2.5
ΔC2 Other																				12.5

Figure 2.4: Baseline *hMLH1* C2 Splice Variants identified in MRC5 cells.

2.4.2 Repeats at C2 Splice Junctions

All the C2 splice junctions involve exon truncation and share a common feature – there exists ambiguity in splice site assignment of 3-7 bases at the splice junction due to identical repeats at the donor and acceptor splice sites. These repeats can be assigned to either the upstream or downstream exon (i.e. they occur once in the sequenced transcript but are found at both the 5' and 3' splice site) and give rise to several possible donor and acceptor combinations. This phenomenon was not encountered previously using this method to characterize mRNA splicing profiles of other loci (e.g. *DISC1*, data not shown; *POLB* and *HPRT* in Disher and Skandalis, 2007) and it was initially thought that these transcripts represented PCR artifacts (e.g. primer extension). Sequence results revealed

bonafide *hMLH1* transcripts; these occur at a frequency unlikely to arise through random splicing error alone. The SVs, overlaps, splice junction assignment, HSF scores, and donor-acceptor dinucleotides are summarized in table 2.3 (that includes C2 SVs from all three conditions). Splice sites were assigned to preserve canonical (U2) splicing where possible however U2-independent (non-canonical) splicing characteristics are observed for many C2 SVs. These results suggest that both U2 and U2-independent splicing (e.g. U12, other) may be operating since it has been shown that a donor AA inhibits the second step of U2 but not U12 splicing and AU-AG introns are expected to be processed by the minor (U12) splicing pathway (reviewed in Wu and Krainer, 1999). Although non-canonical U12 dinucleotides are observed in U12 regulated *hMSH3* E6 (AU-AA), most of the resulting intron types shown in figure 2.4 and table 2.3 do not match the canonical U2 (GU-AG, GU-GC, AU-AC) or U12 (AU-AC, GU-AG) intron flanking dinucleotides. Although the possibility exists that some of the non-canonical C2 SVs represent artifacts resulting from reverse transcription (e.g. due to enzyme skipping over areas of secondary structure) as has been demonstrated for the short *Nox1* mRNA SV (Gieszt et al., 2004), several factors argue against this: (1) some C2 SVs are cell line/tissue specific (discussed in chapter 4), (2) C2 SVs are responsive to H_2O_2 induced OS (chapter 2), (3) there is evidence of regulation of C2 SVs by particular cis-elements/trans-acting factors (chapter 3), (4) all C2 SVs are eliminated by transient AzadC treatment (chapter 4), and (5) SuperscriptIII™ reverse transcriptase (RT) was utilized (less prone to generating artifacts compared to other RTs).

Table 2.3: Summary of *hMLH1* category 2 transcript types, repeated bases (overlap), splice site assignments, HSF splice site scores, and putative splice donors and acceptors.

Transcript ID	Overlap Region (CDS)	Overlapping Exons	# Overlapping Bases	Assignment	Donor (5')	Acceptor (3')
Δ^{22}E2 -^{57}E15 (Δc.137-1724)	137-->138 GT 1725-->1726	E2 & E15	2	/GT E2/E15	GT c.137 66.13	AG c.1724 88.1
ΔE6 -^{57}E15 (Δc.546-1724)	543-->545 CAG 1722-->1724	E6 & E15	3	CAG/ E6/E15	GT c.546 93.59^{wt}	AG c.1724 88.1
Δ^{95}E3 -^{57}E15 (Δc.302-1724)	300 --> 303 AGGT 1723 --> 1726	E3 & E15	4	AG/GT E3/E15	GT c.302 95.15	AG c.1724 88.1
Δ^{95}E3 -^{57}E15; Δ17 (Δc.302-1724; 1897-1989)	300 --> 303 AGGT 1723 --> 1726	E3 & E15	4	AG/GT E3/E15	GT c.302 95.15	AG c.1724 88.1
Δ^{30}E3 -^{21}E18 (Δc.237-2010)	233-->237 AAAGG 2007-2010	E3 & E18	5	AAAG/G E3/E18	GT c.237 70.88	AG c.2010 66.18
Δ^{6}E2 -^{129}E16 (Δc.122-1860)	120-->125 AGA TGC 1859-1864	E2 & E16	6	AG/ATGC E2/E16	AT c.122 78.21^{A}	AG c.1860 Δ9G8
Δ^{112}E1- ^{5}E15 (Δc.112-1672)	110 --> 116 AGAACTG 1671 --> 1677	E1 & E15	7	AG/AACTG E1/E15	AA c.112 65.12^{A}	AG c.1672 80.03
Δ^{112}E1-^{5}E15; 17 (Δc.112-1672; 1897-1989)	110 --> 116 AGAACTG 1671 --> 1677	E1 & E15	7	AG/AACTG E1/E15	AA c.112 65.12^{A}	AG c.1672 80.03
$\Delta^{1\text{-}5}$E2; Δ^{1}E3, Δ^{5}E3- ^{5}E15 (Δc.117-121; 208; 212-1672)	209-->212 212 --> 215 AAGA 1670 --> 1677	E3 & E15	4	207/209-211/1673 G/AAG/A E2/E3/E15	AA c.212 71.44^{A}	AG c.1672 80.03
Δ^{5}E3- ^{5}E15 (Δc.212-1672)	209 --> 213 AAGAA 1670 --> 1674	E3 & E15	6	AAG/AA E3/E15	AA c.212 71.44^{A}	AG c.1672 80.03
Δ^{112}E1- ^{129}E16 (Δc.111-1860)	108 --> 112 TGAGA 1857 --> 1861	E1 & E16	5	TGAG/A E1/E16	AA c.112 65.12^{A}	AG c.1860 Δ9G8
Δ^{94}E1- ^{69}E14 (c.Δ94-1627)	94 --> 99 ATCAAA 1628 --> 1633	E1 & E14	6	/ATCAAA E1/E14	AT	GC
Δ^{94}E1-^{14}E15 (Δc.94-1681)	92-->94 CTA 1680 --> 1682	E1 & E15	3	CT/A E1/E15	AT	CT
Δ^{80}E2- ^{14}E15 (Δc.196-1681)	196 --> 198 ACC 1682 --> 1684	E2 & E15	3	/ACC E2/E15	AC	CT
Δ^{108}E1- ^{34}E15 (Δc.108-1701)	103 --> 108 ATGATT 1697 --> 1702	E1 & E15	6	ATGAT/T E1/E15	TG	TT

Legend: Δ denotes exon (E) skipping; positions given as part of transcript ID (superscript) denote position in each *hMLH1* exon from start (1) to end of exon (e.g. 1-116 for E1, 1-91 for E2, 1-99 for E3, *etc.*); c. denotes coding sequence (CDS) from 1-2554; / indicates putative splice; for each donor and acceptor the CDS position where splice occurs is given and the HSF score (consensus value from 0-100) for donors and acceptors is indicated; ^{wt} denotes wild-type (wt); ^{A} denotes acceptor; 5*-[9] denotes overlap where the T at position 6 of the overlap motif occurs in E3 and delineates GAAGA* as the maximum bases belonging to E3; AGG denotes minimal common unit of overlap for ambiguous GT-AG transcripts; AGA denotes minimal common unit of overlap for ambiguous AT-AG and AA-AG transcripts. Note: Because dNTPs (A,T,C,G) were used for PCR, T = U found in mRNA.

2.4.3 Evidence for Non-Canonical Dual Splice Site Utilization in C2 Splice Variants

There is evidence of dual splice site use among C2 transcript types. The SV **$\Delta^{1\text{-}5}$E2**, ^{5}E3-^{5}E15 demonstrates use of the first five base pairs of E2 (**TTTAG**/ATGC) as an acceptor (3') site resulting in removal of bases 1-5^{TTTAG} as part of the upstream intron and retention of ^{6}E2-^{91}E2 as a truncated exon. The SV **Δ^{6}E2**-^{129}E16 demonstrates use of the splice site as a donor (5') site (TTTAG/**ATGC...**) that results in retention of E2 bases 1-5^{TTTAG} as a truncated exon and removal of ^{6}E2-^{129}E16 as intron 2. Use of the dual site as an acceptor is supported by a moderate strength HSF (U2-based matrix) score of 78.21 however no HSF supported donor score is observed, probably due to the non-canonical nature of the donor AT. Both the lack of supporting HSF donor score and the non-canonical nature of the dual splice site (AGAT) suggest that this dual splice site is U12 regulated (or perhaps that U2 and U12 mechanisms compete for use as a 3' versus 5'ss), a possibility that needs to be verified in future studies. If confirmed, this establishes a role for U12 mediated down-regulation of *hMLH1* at a non-canonical dual splice since $\Delta^{1\text{-}5}$E2 and Δ^{6}E2-^{129}E16 generate a stop codon and reading frame shift respectively. In this study, **$\Delta^{1\text{-}5}$E2**, ^{5}E3-^{5}E15 was observed at low frequency in MRC5 cells (1.2%) and only under baseline conditions whereas Δ^{6}E2-^{129}E16 occurs more frequently in baseline (3.7%). Intriguingly, Δ^{6}E2-^{129}E16 is observed at similar frequency in 500HP but is completely absent in 250HP suggesting it is responsive to OS (discussed later). Furthermore, studies support a role for $\Delta^{1\text{-}5}$E2 in contributing to hMLH1 deficiency under a variety of conditions including baseline (see chapter 4 of this thesis; Peasland et al., 2010). Peasland et al. (2010) identified $\Delta^{1\text{-}5}$E2 in about 50% of *hMLH1* mRNA transcripts (based on analysis of E1-6) in both leukaemic and normal bone marrow samples and report that it generates an immediate premature termination codon that is probably removed by nonsense-mediated decay (NMD). No evidence of dual splice site use was identified by Peasland et al. (2010); however the methodology used was unable to detect the SV resulting from E2 dual splice site use as a donor even if present.

2.4.4 Potential Canonical Dual Splice Site Utilization in C2 SVs

In addition to the dual splice site in E2, several potential canonical dual splice sites emerge in baseline MRC5 cells. For these splice sites, transcript evidence to support use as both an acceptor and a donor is lacking (preventing identification as bonafide dual splices). However, use of these splice sites results in AS (exon truncations with large internal deletions) and splice site strength predicts use as a donor or acceptor consistent with observations by Zhang et al. (2007) who first described dual splice sites and their characteristics. This study does not exhaustively characterize all mRNA transcripts generated in MRC5 and additional (rare) SVs are likely to exist that escaped detection. Therefore, the possibility exists that the alternative splice sites utilized in C2 SVs generate additional transcripts that would support classification as dual splice sites but that remained undetected in this study. Alternatively, these splice sites might serve as initiating or terminal ratcheting points to divide long introns (a phenomenon not yet identified in humans) or they may simply function as alternative splice sites (e.g. regulated by particular splicing factors). At the very least, the regulation of these alternative 5' and 3' splice sites found at 95-E3, 57-E15, and 30-E3 is of interest as they contribute to *hMLH1* deficiency. This assertion is based on the finding that the three C2 SVs, ^{95}E3-^{57}E15, E6-^{57}E15, and ^{30}E3-^{21}E18 that contain the alternative splice sites, conform to canonical U2 splicing rules (i.e. give rise to a GU-AG intron) but generate a reading frame shift suggesting they down-regulate *hMLH1*. The first C2 splice sites of interest, 95-E3 and 57-E15 occur together as a donor and acceptor pair in Δ^{95}E3-^{57}E15. 95-E3 has an HSF supported AG acceptor = 79.1 and GT donor = 95.15 supporting preferential use of this site as a donor resulting in skipping of the last 5bps of E3. The 95-E3 donor = 95.15 precedes the WT donor = 96.07 at the end of E3 by 5bps. The WT splice site is only slightly stronger suggesting that factors other than competition contribute to the regulation of this splice site (e.g. enhancers/silencers, specific splicing factors). Interestingly, there is also a potential dual site at the end of E3 (AG acceptor = 75.63; WT donor = 96.07) suggesting multi-faceted regulation of E3. Δ^{95}E3-^{57}E15 provides transcript evidence for use of 57-E15 as an acceptor (AG acceptor = 88.1; GT

donor = 80.48). Subsequent use of the constitutive WT donor at the end of E15 (WT donor = 93.27) results in retention of the last 7bps of E15 (GTTATCG). The third site of interest occurs in **30E3**-21E18. 30-E3 is used as a donor (AG acceptor = 75.67; GT donor = 70.88) even though the acceptor score strength is somewhat higher (both moderate strength) emphasizing that splice site strength is not the only factor contributing to regulation at these sites; 21E18 is not a dual splice site but is supported by a weak AG acceptor score of 66.18. Although these C2 splice sites appear to conform to U2 splicing rules, the possibility exists that a U12 mechanism excises them. There is a U12 5' consensus sequence located in intron 3 within 40bp of the end of exon 3 (WT) and a U12 branch point consensus sequence within 30bps of the start of E15 (and a U2 branch point within exon 15; a U2 branch point consensus is thought to be able to function for both U2 and U12) although the significance, if any, is currently unknown and may be entirely coincidental. However, U12 branch point sequences are found in introns prior to E14, E15, and E16 that act as acceptors in truncated SVs but not throughout *POLB* for which no similar pattern of truncations with large internal deletions and non-canonical splice junctions is observed.

As a single splicing event, Δ95E3-57E15 gives rise to a GT-AG intron of 41,276 bps where the average intron size for U2 and U12 introns is about 3600bps and 4400bps respectively. This raises the possibility that recursive splicing is also going on as the resulting RPs would show no evidence in the mRNA transcript while effectively sub-dividing the large intron into smaller sections. In fact, many potential ratcheting points (AGGT) exist within hMLH1 introns and exons however, two sites stand out in that they combine both the 3' and 5' consensus: (G)**AGGT**GAG at the end of E3 and (A)**AGGT**GAG found within intron 13, the longest intron (11,253bps). Dual splice sites and recursive splicing are distinct but not mutually exclusive such that some dual splice sites may also serve as RPs and RPs could be used between dual splice sites. The regulation of both types of splice sites, both in terms of core mechanism and regulatory cis-elements and proteins (or even in terms of purpose and when utilized) has not yet been determined. Data here at least warrant consideration of the possibility that U12

splicing and/or other mechanisms contribute to hMLH1 regulation at dual splice and/or alternative splice sites. The other alternative is that transcription and splicing have become uncoupled and these SVs are spliced post-transcriptionally by a U2 (or U2-independent) mechanism.

2.5 Analysis of 250HP *hMLH1* mRNA Splicing Profile in MRC5 Cells

This section examines the characteristics of the *hMLH1* mRNA splicing profile in response to 250uM H_2O_2 induced OS in MRC5 cells. The frequency of splicing in 250HP was found to be 40% CT and 60% SVs, a 10% decrease in CT splicing and a corresponding increase in SV frequency compared to baseline (a trend that does not reach statistical significance). However, the distribution of SVs between C1 and C2 changes with 250HP. Twenty-six C1 transcript types are observed (16 identified, 10 unidentified) that account for 30% of total splicing and 50% of SVs. Ten C2 transcript types are observed (7 identified, 3 unidentified) that account for 30% of total splicing and 50% of SVs. Intriguingly, the types and frequencies of identified C2 variants do not change that much. Instead, the high diversity of unidentified C2 transcript types in baseline (about 12%) disappears with 250HP and is replaced by high diversity of unidentified C1 transcript types (about 12%). Moreover, diversity of identified C1 SVs increases with 250HP and all of the C1 SVs that were identified in baseline are also observed in 250HP. In addition, 11 identified novel transcript types are observed in 250HP that were not seen in normoxia all of which occur at low frequency (less than 2%) with one exception: E17 skipping, which generates an internally deleted protein, occurs at about 4% with 250HP and constitutes the highest frequency C1 transcript type. Although all C1 SVs are generally observed at low frequency, two additional C1 SVs stand out in 250HP: (1) E9/10 skipping, postulated to exert dominant negative function on MMR (Peasland et al., 2010) emerges with 250HP at a frequency of about 1%, and (2) E16 skipping increases from about 1% in baseline to about 3% in 250HP. The frequency of the main C2 SV, Δ^{112}E1-^{5}E15 increases slightly (by about 4.5%). As mentioned previously, the overall frequency of C2 SVs decreases but is driven largely by disappearance of the 6 unidentified transcript types that account for about 12% of C2 splicing in baseline. The

new breakdown of identified C2 SVs (28% total) with 250HP is as follows: GU-AG – 2 types (4% total), AU-AG (none), AA-AG – 3 types (19.8%) including the main SV Δ^{112}E1-^{5}E15 at 16.9%, AU-GC – 1 type (2.9%), AU-CU (none) – Δ^{94}E1-^{14}E15 disappears, and AC-CT – a new C2 SV Δ^{80}E2-^{14}E15 (1.2%). The frequency of the main C2 SV, Δ^{112}E1-^{5}E15 , increases slightly (up by about 4.5%). The Δ^{112}E1-^{5}E15 splice event also occurs twice with ΔE17 in same transcript in 250HP suggesting two distinct mechanisms (not mutually exclusive) or factors contribute to the generation of the two splice events since both can occur individually or together.

Intriguingly, C2 SVs involving E2 display a shift from splicing in front of the exon to splicing in latter parts of exon. Two C2 SVs, 6E2-129E16 & 30E3-21E18 are exclusively absent in 250HP. In contrast, two C2 SVs, 22E2-57E15 & 80E2-14E15 are found exclusively in 250HP. Taken together, the results suggest that there is an increase in the diversity of canonical (U2) SVs in 250HP. This suggests that some of the effects of OS on the *hMLH1* splicing profile may be mediated by changes in factors that regulate U2 splicing at the RNA level such as RBPs (e.g. SR proteins) that act at enhancers and silencers to direct splice site choice (the subject of chapter 3). In contrast, the frequency of proposed (identified) U12 generated C2 SVs remains constant (with the exception of splicing involving E2) suggesting that this mechanism is regulated independently of the 250HP induced splicing changes.

2.6 Analysis of 500HP *hMLH1* mRNA Splicing Profile in MRC5 Cells

This section examines the characteristics of the *hMLH1* mRNA splicing profile in response to 500uM H_2O_2 induced OS in MRC5 cells. The frequency of splicing in 500HP was found to be 12% CT and 88% SVs, a significant decrease in CT relative to baseline and 250HP (38% and 28% respectively) and a corresponding increase in total SVs. The distribution of SVs between C1 and C2 changes with 500HP relative to 250HP, proportionally resembling that seen in baseline but with significantly greater SVs overall driven mainly by 2 specific C2 SVs: Δ^{112}E1-^{5}E15 and Δ^{95}E-^{57}E15. Nine C1 transcript types are observed (5 identified, 4 unidentified) that account for 15% of total transcripts and 20% of SVs. Ten C2 transcript types are observed (7 identified, 3

unidentified) that account for 73% of total transcripts and 80% of SVs. Skipping of ^{112}E1-^{5}E15 more than doubles to about 42%. Assuming authentication, this SV may down-regulate *hMLH1* expression in response to OS to prevent cell death. It has been demonstrated that hMLH1 expressing cell lines have increased sensitivity to peroxides relative to non-expressing cell lines resulting in cell death (p53 independent) that is thought to be mediated by a requirement for hMLH1 in the transduction of a mitochondrial-mediated apoptotic signal (Hardman et al., 2001). Consistent with this idea, loss of hMLH1 and POLG has been shown to result in an increase in mitochondrial (but not nuclear) 8-oxo-G accumulation that is proposed to result in mitochondrial DNA damage (single strand breaks) and trigger apoptosis (Martin et al., 2010). However, POLG (but not POLB) is upregulated in hMLH1 deficient colon tumor tissue compared to matched normal controls (Martin et al., 2010) leading to the intriguing possibility that hMLH1 downregulation (so long as sufficient hMLH1 is present to facilitate repair) can stimulate increased POLG expression and enhance associated repair of mitochondrial 8-oxo-G, thus preventing cell death. This suggests that the ratio of hMLH1 and POLG in mitochondria may be a determinant of mitochondrial repair pathway outcomes, at least with respect to 8-oxo-G and mitochondrial mediated apoptosis. MMR is proficient at low WT hMLH1 concentrations but activation of the G_2/M checkpoint requires a full complement of hMLH1 and is accompanied by p53 phosphorylation (Cejka et al., 2003) suggesting an additional hMLH1 signaling function in response to nuclear DNA damage (e.g. by genotoxic agents such as methylators and alkylators) perhaps involving DNA double strand breaks.

Additional SVs of interest in 500HP include Δ^{95}E3-^{57}E3 that increases from about 2% in baseline and 250HP to about 8% in 500HP. The increase with 500HP raises the possibility that this SV may be triggered by cross-talk with genotoxic stress pathway (e.g. p53 dependent) resulting from increased oxidative DNA damage (e.g. 8-oxo-G). E2 splicing reverts to that seen in baseline. Two C2 SVs, Δ^{6}E2-^{129}E16 & Δ^{30}E3-^{21}E18 exclusively absent in 250HP reappear. Two C2 SVs, Δ^{22}E2-^{57}E15 & Δ^{80}E2-^{14}E15 exclusively present in 250HP disappear. This suggests certain E2 SVs (Δ^{6}E2-^{129}E16 &

Δ^{30}E3-^{21}E18) are specifically down-regulated in response to 250HP (but not baseline and 500HP) and may reflect specific changes unique to the OS response pathway (or a particular level of OS). Possible candidates include RBPs/SR proteins known to change in response to OS (e.g. TRA2-BETA, SRp55, FOX-2) that will be considered further in chapter 3. One possibility is that 250HP triggers an apoptotic response to OS, perhaps accompanied by changes in splicing factors such as SRp55 that have been shown to be associated with apoptotic pathways resulting in the increased C1 SV diversity and change in the distribution of SVs between C1 and C2 with only a small change (decrease) in overall CT frequency. In contrast, the significant decrease in CT frequency and return of the SV distribution to baseline proportions may reflect necrotic death (e.g. necroptosis pathways) triggered by excessive damage in 500HP.

2.7 Comparative Analysis of hMLH1 mRNA Splicing Profiles in MRC5 Conditions

Transcript types identified in all three conditions include WT, ΔE16, Δ^{95}E3-^{57}E15, Δ^{112}E1-^{5}E15, and Δ^{94}E1-^{69}E14. This group includes the two transcript types that occur most frequently in all three conditions: WT, which shows a trend of decreasing frequency with increasing OS (shown in figure 2.2) and Δ^{112}E1-^{5}E15 that shows a trend of increasing frequency with increasing OS (shown in figure 2.5), however the changes are small after 250HP and large after 500HP. Trends for both variants reach statistical significance after 500HP (compared to 250HP) using Fisher's (two-tailed) exact test (WT compared to SVs, p=0.0001; Δ^{112}E1-^{5}E15 compared to all other SVs, p=0.0116). Note that after 500HP, the magnitude of the decrease in WT (-26%) can be accounted for in terms of the magnitude of the increase in Δ^{112}E1-^{5}E15 (+26%) although other changes in splicing profile also occur. For example, Δ^{95}E3-^{57}E15 which occurs infrequently in baseline and 250HP, is the third most frequently observed transcript type after 500HP although the increase in frequency from 250HP to 500HP (+5.5%) does not reach statistical significance.

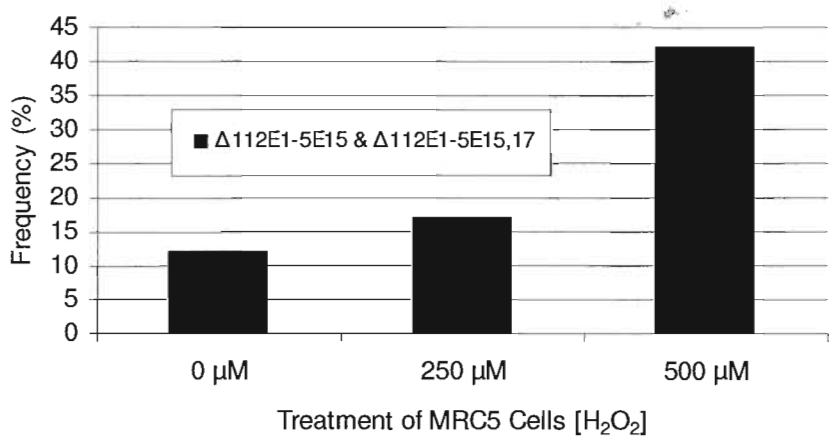


Figure 2.5: Frequency (%) of Δ^{112}E1-^{5}E15 and Δ^{112}E1-^{5}E15, 17 in MRC5 cells after 0uM, 250uM, and 500uM H_2O_2 treatment (baseline, 250HP, and 500HP respectively). The frequency of Δ^{112}E1-^{5}E15 containing transcript types increases by half its level (~5%) in baseline after 250HP and more than doubles (up ~ 25%) after 500HP.

2.7.1 *hMLH1* Splice Variants Present in Baseline

SVs seen only in baseline include ΔE6-^{57}E15, $\Delta^{1\text{-}5}$E2, ^{5}E3-^{5}E15, $\Delta\ ^{94}$E1-^{14}E15, and $\Delta\ ^{112}$E1-^{129}E16. All of these transcripts are rare (i.e. occur only once) and contain truncated exons although all the donor and acceptor splice sites are also utilized in different combinations in other transcript types and treatments. This suggests that none of the SVs are baseline specific, rather that OS changes the diversity, frequency, distribution of SVs as will be shown in subsequent sections.

SVs seen in baseline and 250HP (but not 500HP) include ΔE2; ΔE6, 9, 12; ΔE9-11, 15; ΔE9-12. The two C1 SVs containing only discrete exon skipping are observed at lower frequency after 250HP than the two C1 SVs with multiple consecutive exon skipping, however none of the four transcripts occur after 500HP. Note that due to the smaller sample size after 500HP relative to baseline and 250HP these results should not be taken to indicate that the transcript types do not occur after 500HP, only that they were not detected in this study. What is of interest is that there appears to be a trend within C1 transcript types that reflects a shift towards decreased discrete exon skipping and

increased multiple adjacent exon skipping with 250HP (but possibly not 500HP) relative to baseline that warrants further examination.

2.7.2 *hMLH1* Splice Variants Seen Only After Oxidative Stress but Not in Baseline

SVs seen after 250HP and 500HP (but not in baseline) include ΔE6, 9; ΔE9-10; ΔE9-12, 16; and ΔE17. All of these transcripts occur at low frequency and use canonical splice junctions for skipping events revealing a subtle influence of OS on the splicing profile of *hMLH1* in MRC5 cells in terms of C1 SVs. Note that E17 skipping occurs in 2 transcript types (described earlier) that each contains a separate splice event flanked by an exon truncation; in one of the transcripts exons 16-17 are skipped together. Similar to ΔE17 alone, these three transcript types occur after 250HP or 500HP but not in baseline. Taken together, total E17 skipping from all transcript types was not observed in baseline but was observed at low frequency after 250HP and 500HP (5.8% and 3.3% respectively). These reports are consistent with the literature to the extent that E17 skipping was not observed in any of the normal tissues or cell lines examined with the exception of colon tissue (Genuardi et al., 1998). The fact that OS induces E17 skipping is of interest because it points to a possible connection between oxidative stress and a loss of (in this case a reduction in) *hMLH1* repair capacity, both of which are associated with cancer progression. E17 skipping results in an internally deleted protein although the function, if any, of this protein has not been established. However, it has been demonstrated that the hMLH1 binding partner, PMS2 is not transcriptionally down-regulated in a coordinated fashion with hMLH1 rather than it requires hMLH1 to stabilize it, such that PMS2 is down-regulated by the absence of its binding partner, hMLH1. Therefore, one possibility is that E17 skipping provides a mechanism for fine tuning the repair system and results in the down-regulation of PMS2 without disrupting production of other repair components. Could the propensity for E17 skipping under normal conditions be a contributing factor to the high incidence of colon cancer in the population? Since the Genuardi study did not quantify E17 skipping but only reported the presence or absence of different splice isoforms, addressing this question requires

quantifying E17 skipping in a colonic system, which is the subject of chapter 4, so this possibility will be discussed further in that context.

SVs seen only after 500HP include Δ^{95}E3-^{57}E15& Δ17 and Δ^{108}E1-^{34}E15 both of which are rare (i.e. occur only once) although the transcript Δ^{95}E3-^{57}E15& Δ17 is related to another transcript found in all three conditions in that it contains the same splice unit, Δ^{95}E3-^{57}E15, but also skips exon 17 in its entirely and as such is not unique to 500HP. Conversely, the appearance of Δ^{108}E1-^{34}E15 which is unique to 500HP may point to factors (e.g. RBPs) that are specifically up-regulated or down-regulated in 500HP and as such warrants further investigation. Analysis of *cis*-elements at splice sites and potential RBPs that may contribute to the observed splicing pattern of this and other SVs is the subject of chapter 3.

2.7.3 Distinct Splicing Signature Emerges With 250HP

SVs seen only after 250HP include ΔE6, 9-10, 12, 15-16; ΔE10-11; ΔE9-11; ΔE9-11, 14-16; ΔE15; ΔE15-16; ΔE16-17; Δ^{112}E1-^{5}E15, 17; Δ^{22}E2-^{57}E15; Δ^{5}E3-^{5}E15; and Δ^{80}E2-^{14}E15. Of these, two-thirds (7/11 types) use canonical splice junctions for skipping events (C1) and occur rarely (i.e. once) whereas one-third (4/11 types) contain exon truncations (C2) and occur infrequently but more than once (i.e. 2-5 times). Note that the transcript Δ^{112}E1-^{5}E15, 17 is related to the most frequently observed splice variant (found in all conditions) in that it contains the same splice unit, Δ^{112}E1-^{5}E15, but also skips exon 17 in its entirety. These observations, taken together with splicing events described previously, suggest that there are distinctions between transcripts delineated as C1 versus C2 but that some factors (e.g. splicing pathway, SR protein concentration) may influence the generation of certain transcript types within both C1 and C2 such that distinct (and/or overlapping) mechanisms are operating that are not necessarily mutually exclusive.

SVs seen in baseline and 500HP (but not 250HP) include Δ^{6}E2-^{129}E16 and Δ^{30}E3-^{21}E18. Both of these transcripts contain exon truncations at both splice sites and are observed more than once (i.e. 2-3 times) in baseline and 500HP but not after 250HP. SVs seen in other conditions but not 250HP are of interest because of the greater sample size

investigated for 250HP. That is, SVs that are seen in both baseline (normal sample size) and 500HP (smallest sample size) but not in 250HP (largest sample size) might be important for discerning what distinguishes changes in splicing profile that occur after 250HP in particular. These variants reflect the distinct change in splicing signature that is observed after 250HP.

Taken together with SVs that occur only after 250HP and transcripts that do not occur in 250HP but are observed in both baseline and 500HP, a distinct splicing signature emerges in 250HP. Diversity of C1 SVs increases (sample size doubles while diversity measured as the number of unique transcript types identified triples) and diversity of C2 SVs decreases (by one-third) as shown in Table 2.4.

Table 2.4: Diversity (# of *hMLH1* transcript types) in MRC5 by category and condition.

	MRC5 0uM	MRC5 250uM	MRC5 500uM
CT	**1**	**1**	**1**
C1 (Identified)	5	16	5
C1 (Other)	2	10	4
C1 TOTAL	**7**	**26***	**9**
C2 (Identified)	9	7	7
C2 (Other)	6	3	5
C2 TOTAL	**15**	**10**	**12**
OVERALL TOTAL SVs	*22*	*36*	*21*

* Baseline to 250HP: C1 Sample size doubles; diversity of transcript types triple (data not adjusted for sample size)

Similarly, in terms of frequency, C1 increases by two-thirds and C2 decreases by one-quarter in 250HP compared to baseline. After 500HP (compared to 250HP), the frequency of C1 decreases by one-half and C2 increases by one-and-one-half (150%) as shown in Table 2.5.

Table 2.5: Frequency (%) of *hMLH1* in MRC5 by category and condition.

	MRC5 0uM	MRC5 250uM	MRC5 500uM
CT	**49.5**	**39.5**	**11.7**
C1 (Identified)	7.4	18.6	8.4
C1 (Other)	2.3	12.2	6.7
C1 TOTAL	**9.7**	**30.8**	**15**
C2 (Identified)	28.4	27.9	60
C2 (Other)	12.5	1.7	13.3
C2 TOTAL	**40.9**	**29.6**	**73.3**
OVERALL TOTAL	*100.1**	*99.9**	*100*

*total differs from 100% due to rounding error

2.7.4 Distribution of C1 and C2 Splice Variants Changes in Response to Different Levels of Oxidative Stress

As a group, C1 and C2 SVs respond differently to OS, as shown in Table 2.5, even though overall, SVs increase as oxidative stress increases (and the CT decreases). In the absence of H_2O_2 treatment, total SVs equal CTs in terms of frequency (~50% each) and splice variant composition is comprised of 10% C1 and 41% C2 such that C2 transcripts outnumber C1 transcripts 4 to 1. After 250 uM H_2O_2 treatment, total SV frequency exceeds WT by 20% (60% and 40%) and the splice variant frequency is equal for C1 (30%) and C2 (30%) or 1 to 1. After 500 uM H_2O_2 treatment, overall splice variants account for most splicing at 88% compared to only 12% WT (i.e. SVs outnumber WT by 6:1) and the splice variant composition is comprised of 18% C1 and 70% C2 such that C2 transcripts once again outnumber C1 transcripts by 4 to 1 (as in 0uM H_2O_2).

Exceptions include Δ^{112}E5-^{5}E15 that is observed at high frequency (for a single SV) in baseline (12%) and increases slightly (to 17%) after 250HP and then further increases (dramatically, to 42%) after 500HP such that it rivals the CT (nearly the equivalent of CT in baseline) and suggests a shift in mechanism or severe disturbance in the normal mechanism has occurred. Interestingly, that shift does not appear to be accompanied by an increase in mis-splicing so much as a down-regulation of the *hMLH1* CT. Could this be a default for stressful conditions? Could p53 regulation be involved? Intron 1 of *hMLH1* contains a p53 response element (Chen and Sadowski, 2005) that appears to resemble an activating as opposed to repressing response element (Wang et al., 2009) however the effect on splicing and/or expression of the CT (up or down regulation) has not been verified. It has been shown that p53 modulates splicing of *POLB* in response to H_2O_2 induced OS (Disher and Skandalis, 2007). It is possible that *hMLH1* is similarly regulated. Although no candidate SVs stand out (that cannot be better explained by other factors) in this study, the effects may be indirect or influence expression level of the CT instead of SVs.

Another skipping event, ΔE17, either alone or in SVs with other skipping events, is not seen in baseline but occurs at low frequency after 250HP and 500HP (6% and 3% respectively). ΔE17 contributes to the overall pattern of distribution changes in C1 and C2. The complete absence of E17 skipping in any capacity in normoxia suggests that OS induces E17 skipping although not in a dose dependent way. It is possible that global changes to splicing factors that promote E17 splicing fidelity (e.g. by binding to the enhancer in E17 to promote inclusion) are affected by OS and are mediating the effect. If the results were due to direct damage at the DNA level, a dose dependent response would be expected (i.e. increased OS, increased DNA lesions, increased E17 skipping) but a similar level (slight decline) is observed instead pointing to a global change (e.g. SR protein shift).

2.7.5 *hMLH1* Exon-Internal Donors Preferentially Pair with Acceptors in E15

Exon truncations in E1, E2, or E3 appear to act as splice donors that are joined to exon truncations in E14, E15, E16, or E18 that appear to act as splice acceptors; 15/32 transcript types exhibited this splicing signature (as shown in table 2.6). Partial E15 was by far the most utilized acceptor with 11/15 transcript types terminating in E15, although 4 different positions (splice units) within E15 are utilized as follows: 57-E15 (4 types), 5-E15 (4 types), 14-E15 (1 type), and 34-E15 (1 type). All initiating truncated exons preferentially pair with E15 such that partial E1, E2, and E3 join most often to partial E15 as seen in 4/6, 2/3, and 4/5 transcript types respectively; the 11th transcript type, ΔE6-^{57}E15, also terminates in E15 but initiates with a canonical splice that results in ΔE6.

Table 2.6: Number SV types containing exon truncations by donors (E1, E2, E3, E6) and acceptors (E14, E15, E16, E18) that give rise to C2 splice junctions.

5'ss / 3'ss	**E14**	**E15**	**E16**	**E18**	***TOTAL***
E1	1	4	1	0	***6***
E2	0	2	1	0	***3***
E3	0	4	0	1	***5***
E6	0	1	0	0	***1***
TOTAL	**1**	**11**	**2**	**1**	***15***

2.7.6 E1 and E15 are Hotspots of Exon-Internal *hMLH1* Splicing

In terms of frequency (shown in table 2.7), E1 and E15 are splicing hotspots for exon truncations that warrant further investigation. Overall E2 splicing, whole exon and truncations combined, occurs infrequently in all three conditions and does not appear to be significantly impacted by oxidative stress leading to speculation that E2 may be regulated differently from E1 and E3. E3 truncations, while less frequent than E1 truncations, do nevertheless respond similarly to treatment in that their frequency remains relatively steady in baseline and 250HP but increases dramatically (more than double) after 500HP. Similarly, some specific splice sites, including 112-E1, 95-E3, 5-E15, and 57-E15, are utilized more frequently than others overall, as shown in table 2.8. Conversely, splice sites within E2 are differentially utilized in response to 250HP: 6-E2 use decreases (dual splice site not utilized either as an acceptor or donor), whereas 22-E2 and 80-E2 increase (appear exclusively in 250HP).

Table 2.7: Frequency (%) of SVs with exon truncations in E1, E2, E3, E14, E15, E16, and E18 found in MRC5 cells by condition.

	MRC5 0μM H_2O_2	MRC5 250μM H_2O_2	MRC5 500μM H_2O_2
E1	17.2	19.8	45.0
E2	4.9	4.7	3.3
E3	6.2	4.1	11.7
E14	2.5	2.9	2.7
E15	18.5	25.0	51.7
E16	4.9	0	3.3
E18	2.5	0	3.3

Table 2.8: Frequency (%) of donors and acceptors in truncated exons by position found in MRC5 cells by condition.

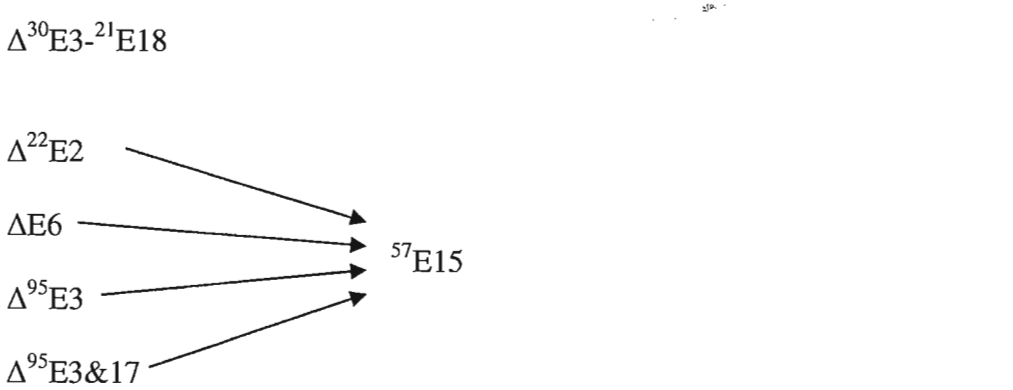
Initiating Partial Exon (5' Splice Unit)	MRC5 0μM H_2O_2	MRC5 250μM H_2O_2	MRC5 500μM H_2O_2
112-E1	13.4	16.9	41.7
108-E1	0	0	1.7
94-E1	3.7	2.9	1.7
6-E2*	3.7	0	3.3
1-5E2** with ^{5}E3-^{5}E15	1.2	0	0
22-E2	0	2.9	0
80-E2	0	1.2	0
95-E3	2.5	1.2	8.3
5-E3	1.2	2.9	0
30-E3	2.5	0	3.3
Terminating Partial Exon (3' Splice Unit)			
5-E15	13.6	19.8	41.7
57-E15	3.7	4.1	8.3
14-E15	1.2	1.2	0
34-E15	0	0	1.7
129-E16	4.9	0	3.3
69-E14	2.5	2.9	1.7
21-E18	2.5	0	3.3

* and ** share the same splice site; *skips only the first 5bps of E2; **retains only the first 5bps of E2 and skips the rest (this splice event occurs in a transcript that also contains ^{5}E3-^{5}E15 skipping)

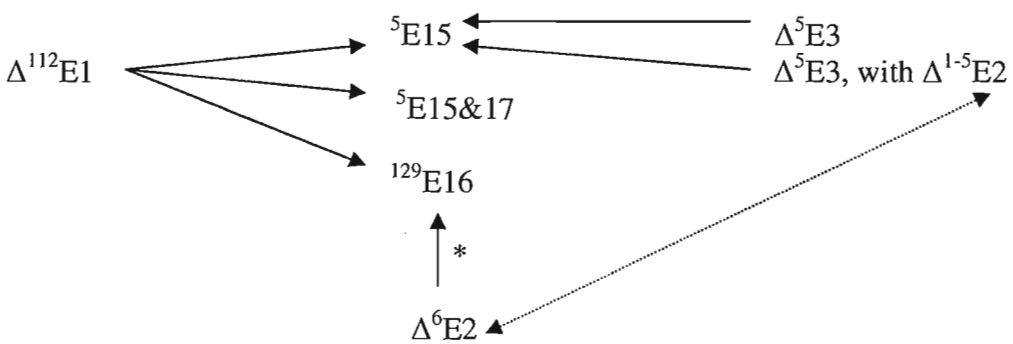
2.7.7 Interrelationships Among *hMLH1* C2 5' and 3' Splice Site Pairs

Certain C2 donor and acceptor splice sites (e.g. 112-E1, 95-E3; 5-E15, 57-E15 respectively) tend to be paired and occur at high frequency within one transcript type (e.g. Δ^{112}E1-^{5}E15, Δ^{95}E3-^{57}E15) but also occur in different combinations (e.g. 5-E3, 129-E16) in other transcript types (e.g. Δ^{5}E3-^{5}E15, Δ^{112}E1-^{169}E16), albeit at lower frequency. Some donor and acceptor sites are only observed together within a single unique transcript type (e.g. 30-E3 and 21-E18 in Δ^{30}E3-^{21}E18; 108-E1 and 34-E15 in Δ^{108}E1-^{34}E15) and occur at low frequency in this study. The resulting map of inter-relationships between the 5' and 3' splice sites that form splice junctions involving truncated exons are diagramed in Figure 2.4. Of the splice site pairs identified, 5 utilize standard U2 splicing (i.e. give rise to a GT-AG intron) and have HSF supported donor and acceptor splice site scores such that their splice junctions are considered to be unambiguously resolved (see table 2.7) . One pair gives rise to an AT-AG intron and is supported by HSF donor and acceptor scores. This pair is inter-connected to 5 pairs with AA-AG splicing however it should be noted that while there is no HSF supported donor score there is an HSF supported acceptor score at the donor site (the relevance of which is currently unknown). Also, AA at the donor site has been shown to impede the second step of U2 splicing raising the possibility that the minor U12 splicing pathway is responsible for the splicing events or that another U1 independent mechanism is modulating the splicing outcome. The remaining three inter-connected splice unit pairs and one unrelated splice unit pair give rise to a non-standard splice where neither the donor nor acceptor are HSF supported. Although irregular splice sites are observed for U12 splicing, reliable algorithms for de novo prediction of U12 splicing remain to be determined. Nevertheless, the *hMLH1* pre-mRNA sequence was scanned for U12 consensi and both 5'ss and branch point sequences are found throughout *hMLH1* but not *POLB*. Moreover, U12 branch point sequences are found prior to E14, E15, and E16. These exons form part of the region that contacts *hMSH3* – a MMR protein that contains a U12 regulated exon (E6) with a non-canonical (for U12) acceptor site (AT-AA). Taken together, the results point to both U2 and U2-independent mediated splicing of *hMLH1*.

GT-AG Splice Donor (5') and Acceptor (3') Junctions:



AT-AG* and AA-AG Splice Donor (5') and Acceptor (3') Junctions:



Irregular Splice Donor (5') and Acceptor (3') Junctions:

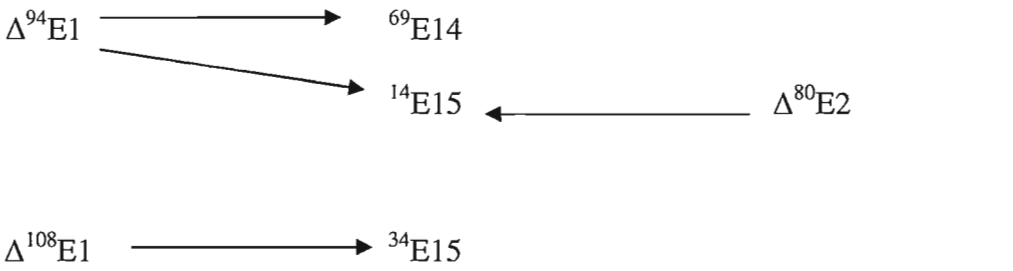


Figure 2.6: Map of truncated donor and acceptor splice sites that form splice junctions (black arrows) and their interrelationships. Two-sided arrow with dotted line denotes utilization of the ^{6}E2 splice site as either an acceptor (as in $\Delta^{1\text{-}5}$E2) or a donor (as Δ^{6}E2-129E16) supporting the existence of a dual splice within E2.

2.8 Overall Discussion

In order to establish a baseline *hMLH1* mRNA splicing profile, the types and frequencies of 81 *hMLH1* mRNA transcripts generated were characterized in MRC5 cells grown under standard culture conditions. Baseline *hMLH1* splicing resulted in 50% CT and 50% SVs. This is comparable to results reported for *POLB* (44% CT and 56% SVs) in the same cell line (Disher and Skandalis, 2007) and shows that *hMLH1* and *POLB* both exhibit a similar frequency of constitutive splicing (i.e. CT) and undergo extensive AS (i.e. generate a high frequency of SVs) under baseline conditions. In contrast, *HPRT* exhibits high constitutive splicing and infrequent AS (99% CT and 1% SVs) in baseline MRC5 cells (Disher and Skandalis, 2007) indicating that genes differ significantly in their baseline level of constitutive splicing and AS. Furthermore, *hMLH1* and *POLB* respond to 500HP with a change in the frequency (increase) and distribution of SVs; both are constitutively expressed 'housekeeping genes' involved in the removal of oxidized base damage (in MMR and BER respectively). On this basis, *POLB* was selected as an appropriate control locus for use in subsequent analyses (chapter 4). Although various disease processes including specific cancers are associated with RNA splicing disregulation (reviewed in Ghigna et al., 2008), the results obtained and compared here underscore the importance of interpreting cancer associated splicing patterns of individual genes within the context of comparable non-cancer (baseline) profiles. For example, an increase in *HPRT2* AS indicates altered splicing whereas either an increase or a decrease in AS indicates altered splicing for *POLB* and *hMLH1* where AS is the norm. Disregulation of RNA splicing is governed by the signaling pathways to which particular genes are linked (or in which they participate/are effectors). Splicing of distinct sets of genes will be altered depending on the pathway(s) affected in specific cancers. This is driven by TFs/promoter architecture, kinetic coupling, and RNA splicing factors/RBPs that influence splicing outcomes in response to inputs from cell signals. The significance of the observed *hMLH1* splicing profile lies in that cell lines with near perfect *hMLH1* splicing (i.e. a high frequency of CT/constitutive splicing and low SVs) are not 'normal' in the sense that they differ from baseline. Therefore, observing a high

level of canonical *hMH1* splicing (e.g. in response to drug treatment in cancer cell lines, chapter 4) may actually represent an aberrant or stressed state and/or an altered splicing response (that may or may not be pathogenic). The caveat here is that the baseline *hMLH1* CT frequency and SV distribution must be confirmed in various cell lines, tissues, and under various conditions (e.g. physiological oxygen concentrations) in addition to the non-cancer baseline profile established in MRC5 cells in this study.

In the present study, splicing at repeats in C2 SVs resulted in large internal deletions flanked by exon truncations in effect mimicking a situation in which part of the upstream and downstream exon is recognized as part of a very large intron. These findings are reminiscent of recent reports that point to a role for functionally interacting pairs of cis-regulatory elements (usually found in introns) in facilitating exon definition (Friedman et al., 2008). Moreover, Ke and Chasin (2010) show that intronic motif pairs in pre-mRNA, many of which resemble the binding sites of known hnRNP proteins, cooperate across exons to promote splicing raising the possibility that hnRNP proteins (best known for their role as splicing silencers though examples of context depending splicing enhancement have also been reported) may actually facilitate exon definition for example from particular repeats (binding sites) within introns (possibly in a U1 independent manner). Work by Martinez-Conteras et al. (2006) shows hnRNP A1 binding at intron ends can bring the two ends together to promote intron definition. As such, C2 *hMLH1* SVs and in particular the repeats identified warrant further investigation. Consideration should be given to whether the observed C2 splicing events are hnRNP regulated; if so, by which hnRNPs and in what manner. Other possibilities include that: (1) U12-splicing mediates certain *hMLH1* splicing events, (2) recursive splicing occurs under some conditions, (3) dual splice site use governs certain *hMLH1* splice events, (4) transcription and splicing have become uncoupled (i.e. splicing occurs post-transcriptionally instead of co-transcriptionally in C2 SVs), and/or (5) novel factors selectively inhibit specific steps of spliceosome assembly in ways not previously characterized to modulate *hMLH1* C2 SVs.

Exon skipping events correspond to the 3 domains of *hMLH1*. Since *hMLH1* exons 6-13 contain the MutS homologues (*MSH2*, *MSH3*, and *MSH6*) binding domains, skipping of this region as seen in C2 SVs will result in down-regulation of *hMLH1* WT mRNA in general driven by a specific underlying reduction in competence to bind MutS homologues originating at the mRNA level. That is, the effect in terms of reduction of hMLH1 WT protein is the same but the driving factors originating at the mRNA level can differ. Similarly, *hMLH1* exon-internal splicing within exons 1, 2, or 3 falls within the ATPase domain and splicing within exons 14-19 corresponds to the PMS2/MLH3/PMS1 interaction domains. It is tempting to speculate that the manner in which *hMLH1* is down-regulated at the mRNA level corresponds to specific factors that affect coordinated sets of genes in a common regulatory network, though verification of this idea is beyond the scope of this project. As in C2 SVs, skipping of the central portion of *hMLH1* that encompasses E9-12 is also observed in C1, however C1 is distinguished from C2 in this study by a maximum limit of 4 adjacent exon skipping events (i.e. E9-12) while C2 is defined by a minimum of 8 adjacent exon skipping events (i.e. E6-13). This falls within the region associated with MutS and points to a possible link to *hMLH1* repair pathway function. Notably, the Δ9/10 isoform that has recently been shown to exert a dominant negative effect on the MMR pathway (Peasland et al., 2010) and has been identified in normal and cancer cell lines and tissues in other studies (Charbonnier, 1995; Genuardi et al., 1998; Peasland et al., 2010) was not observed in baseline MRC5 cells and occurred at low frequency (about 1%) after 250HP and 500HP. This suggests that it is not a significant contributor to *hMLH1* deficiency in MRC5 cells either in baseline or in response to OS. This transcript type is classified as a C1 variant in this study and while E9/10 skipping occurs in combination with other exon skipping events in alternative C1 SVs, the ΔE9/10 isoform that is associated with dominant negative activity occurs infrequently, highlighting the importance of employing studies that allow full transcript types and their relative frequencies to be examined, at least for transcripts thought to exert particular effects. It is likely that known SR proteins are contributing to the

observed splicing patterns in MRC5 cells under baseline conditions and in response to OS and this is investigated in chapter 3.

2.9 Summary

In chapter 2, the baseline (non-cancer) splicing profile of *hMLH1* in MRC5 cells was characterized using full length mRNA transcripts and the response to 250HP and 500HP was determined. *hMLH1* was found to be extensively spliced both before and after oxidative stress and the splicing pattern appears to be driven by both U2 and U2-independent mechanisms that warrant further investigation. The first step would be to determine which, if any, of the C2 splice variants are artifacts. As hypothesized, the diversity and frequency of hMLH1 SVs was influenced by OS. The *hMLH1* transcriptome changed in response to different levels of H_2O_2 (250HP and 500HP). The frequency of the CT decreased in response to H_2O_2 accompanied by a corresponding increase in SVs. The distribution of SVs between C1 and C2 changed specifically with 250HP and was driven by an increase in C1 SV diversity. This is the first study to identify and quantify the extent to which the C2 splice variant signature contributes to alternative *hMLH1* mRNA transcript generation in MRC5 cells and to document the changes in splice variant distribution between C1 and C2 variants with changes in H_2O_2 induced oxidative stress.

In addition to the CT, particular C1 and C2 SVs of interest were identified. For C1, these include: E9/10, E9-12, E16, and E17 skipping. For C2 these include: (1) Δ^{112}E1-^{57}E15 and Δ^{95}E3-^{57}E15 that exhibit a similar frequency in baseline and 250HP but double in response to 500HP; and (2) E2 truncations that change in response to 250HP: 1-5E2 and 6E2-129E16 dual splice site use disappears; and 80E2-14E15 and 22E2-57E15 splice site use appears.

In this chapter, the *hMLH1* mRNA splicing profiles were identified, quantified and categorized in terms of the classes of splicing they contain, splice type and strength, and core splicing mechanisms. In addition, regulatory cis-elements found in pre-mRNA that function as enhancers and silencers of splicing are known to contribute to AS outcomes; these are considered in the next chapter.

Chapter 3: Computationally identified enhancers in *hMLH1* exons correspond to H_2O_2 induced mRNA splicing patterns in MRC5 Cells.

The goal of Chapter 3 is to examine whether computationally derived *cis*-elements (i.e. enhancers and silencers) identified in hMLH1 exons by Human Splicing Finder are associated with the observed mRNA splicing pattern in a consistent manner in MRC5 cells under baseline and oxidative stress conditions.

It is hypothesized that stress responsive exons will contain higher strength cis-elements that correspond to specific factors involved in modulating the cellular response to OS.

Contributions: Conceptualized chapter (goal, hypothesis, methods); performed all analyses and chapter write-up.

3.1 Introduction

The mRNA splicing profiles characterized in chapter two reveal highly variable and stress-sensitive splicing of *hMLH1* in untransformed human MRC5 cells however the factors that modulate the *hMLH1* splicing pattern under basal and stressful conditions remain unknown. Computationally, the best predictors of splicing outcomes are splice site strength (examined in chapter 2) and *cis*-elements referred to as enhancers and silencers found in pre-mRNA that promote or suppress exon inclusion respectively (Culler et al., 2010; Wang et al., 2009; Zhang et al., 2007; Roca et al., 2003; 2005). In this context, *cis*-elements are specific sequences of RNA bases that correspond to the RNA recognition motif (RRM) of a particular RNA binding protein (RBP). Enhancer and silencer motifs found within exons in pre-mRNA are known to regulate constitutive and alternative splicing when bound by their cognate trans-acting factors or RBPs. Among RBPs, the serine-arginine (SR) proteins SRp40, SRp55, SC35, SF2/ASF, TRA2-BETA, 9G8 and the heterogeneous nuclear ribonucleoprotein A1 (hnRNP A1) are among the best studied splicing factors and algorithms have been designed to predict the location of their RRMs within exons and introns (Cartengi et al., 2003; Desmet et al., 2009). These algorithms have recently been combined into a single tool, Human Splicing Finder (HSF) (Desmet et al., 2009); HSF computationally derived *cis*-elements found in *hMLH1* exons are analyzed in this work. SR proteins typically act as enhancers from within exons whereas hnRNPs tend to act as silencers. However, antagonistic effects of SR proteins have been reported for SF2/ASF and SRp55 versus SC35 and SRp40 (Chandradas et al., 2009). Moreover, splicing outcomes depend on the ratios and regulation (e.g. phosphorylation state, cellular localization) of splicing factors such as SF2/ASF and hnRNPA1, and on the context of *cis*-elements within the exon (e.g. location near 3' versus 5' splice site, element strength, overlapping elements) such that combinatorial effects are observed (reviewed in Ghigna et al., 2008; Long and Caceres, 2009). What is the contribution of exonic enhancer and silencer utilization by known SR proteins and hnRNP A1 to *hMLH1* mRNA splicing profiles? Based on

comparison of the actual *hMLH1* splicing profiles observed in this thesis (chapter two) and the location of potential factor RRMs within *hMLH1* exons obtained using HSF algorithms (Desmet et al., 2009) (shown in results and discussion section of this chapter), a model is proposed to explain key splicing signatures and methods for experimental verification in future studies are discussed. It is hypothesized that *hMLH1* exon skipping frequency will correspond to: (1) the strength or relative strength of cis-elements utilized under a given condition, and (2) the level of cognate trans-acting factors present that regulate the exon. Therefore, similar high scoring *cis*-elements that occur in exons that share a similar skipping profile in response to OS (compared to baseline) point to the involvement of particular trans-acting factors *in vivo* that can be experimentally verified in future studies.

3.2 Method

Firstly, algorithms included in HSF http://www.umd.be/HSF/ (Desmet et al., 2009; Cartengi et al., 2003) were used to identify potential RRMs of trans-acting factors in pre-mRNA that occur within *hMLH1* exons for SRp40, SF2/ASF, SC35, SRp55, TRA2-BETA, 9G8, and hnRNP A1. In addition, RRMs were examined in the 1000 base pairs of flanking intron sequence upstream and downstream of each exon to analyze the possibility that factors are acting from flanking introns as opposed to from exons (data not shown). However, the focus in this chapter is on potential motifs acting from within exons and intronic motifs are described and discussed only where they occur within 100bps of exon boundaries and appear to relate directly to a specific splicing pattern.

Secondly, high scoring motifs were identified for each exon from HSF 'percent (%) variation' scores, a measure of element strength. A threshold is set within the algorithm for each *cis*-element that delineates the cutoff for hits expected by chance. Since the length of *cis*-elements varies between the factors, affecting the raw scores, HSF includes the % variation score, calculated as the difference between the raw score and threshold divided by the difference between 100 and the threshold

expressed as a percent. This allows the strength of the *cis*-elements to be compared. A score of at least 40% or more above threshold was considered high scoring based on arbitrary selection that appeared to distinguish a manageable number of high scoring motifs from lower scoring motifs for preliminary analysis. Note that exon 1 is included in the analysis since the forward primer is located within the 5' untranslated region (i.e. transcripts with exon 1 truncations are included but it is assumed that whole exon 1 skipping does not occur) whereas exon 19 is excluded from analyses since the reverse primer is located within exon 19 (as opposed to the 3' untranslated region) such that transcripts must contain the final exon (exon 19) in order to be included in this study. High scoring motifs were mapped to the *hMLH1* exons in which they occur (see Results and Discussion Figure 3.1). Thirdly, high scoring *cis*-elements corresponding to particular and/or frequently observed splice variants, exon skipping events, and splicing signatures (identified in chapter two and summarized in Figure 3.2) were identified by the following means: (1) the high scoring motif sequence (5-7 bases) was recorded by exon for each RRM/corresponding trans-acting factor, (2) exon skipping events that tend to occur together within transcripts or in response to a particular level of OS were selected and RRMs in common (present or absent) were identified, (3) the core common motif sequence of RRMs present in all such events was identified, (4) the hMLH1 pre-mRNA sequence was searched for additional core motifs within exons not identified by HSF and these were added to the analysis on the assumption that all such core motifs have the potential to be active, and (5) a proposed model of factors that contribute to *hMLH1* splicing outcomes was compiled.

3.3 Results and Discussion

3.3.1 High Scoring RNA Recognition Motifs Identified in *hMLH1* Exons

Computationally derived high scoring cis-elements corresponding to the RRM of known trans-acting factors (or splicing factors) were identified in *hMLH1* exons using HSF. In general, SR proteins found within exons act as enhancers to promote exon inclusion and production of the canonical transcript (CT). Therefore, down-regulation of a splicing factor that normally enhances splicing will result in skipping of the exons it regulates. Since redundancy exists among factors, the effect may be apparent, or it may be masked by other factors that are present, depending on the context of the *cis*-elements within the exon (e.g. 5' versus 3' location, location relative to other cis-elements). It is assumed that higher scoring motifs will be preferentially utilized in the presence of their cognate trans-acting factors (other factors being equal). The presence or absence of computationally derived high scoring potential factor binding sites for SRp40, SF2/ASF, SC35, SRp55, TRA2-BETA, 9G8, and hnRNP A1 found within *hMLH1* mRNA varies by exon and several trends are observed; the results are summarized in Figure 3.1.

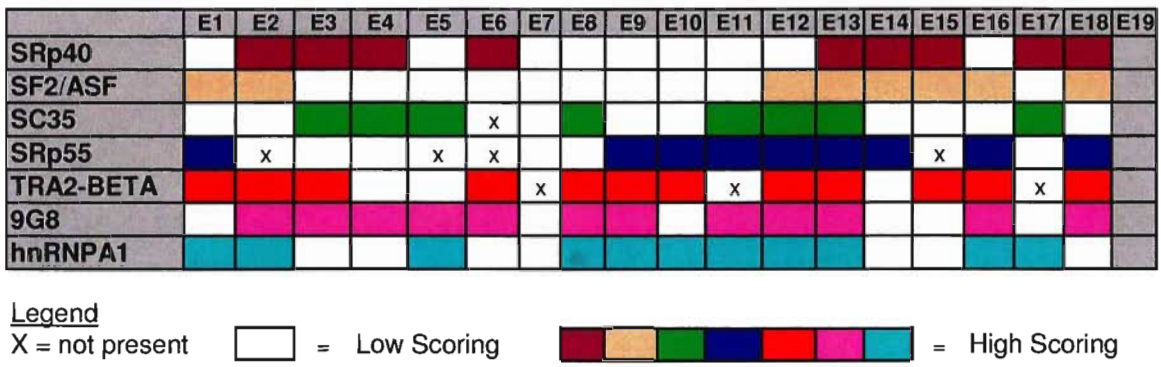


Figure 3.1. Computationally derived high scoring factor-binding sites for SRp40, SF2/ASF, SC35, SRp55, TRA2-BETA, 9G8, and hnRNP A1 found in *hMLH1* exons.

Firstly, three or more high scoring motifs were identified in all exons except E7. No high scoring motifs are found within E7, the smallest *hMLH1* exon, although all motifs except TRA2-BETA (completely absent) are found with low scores. E7

also has three high scoring hnRNP A1 motifs containing 'AAGG' within the 100 base intron flanking region (upstream, 1 motif; and downstream, 2 motifs). Although the relevance, if any, of this observation is unknown it is mentioned here because it has been postulated that small exons may be subject to intronic regulation in a way that larger exons are not due to their small size which places constraints on the enhancers that can be found within the exon without impairing coding sequence requirements. Therefore, hnRNP A1 may be acting from within introns 6 and 8 to bring the ends of E7 together to promote splicing. In contrast, all factors display high scoring motifs in E12 and E13. For E12, this result may be attributable to the large size of E12 (371 bps) relative to other *hMLH1* exons but this same rationale cannot account for the heavy regulation by high scoring motifs in E13 (149 bps). The presence of high scoring motifs for all factors in E13 may account for the fact that E13 is not internally spliced in this study and tends to be skipped only as part of C2 multiple adjacent units (not individually, or as part of a C1 multiple adjacent unit), presumably because numerous high affinity enhancer and/or silencer sites exist resulting in redundancy such that any of several other factors can compensate for loss of another factor thus promoting whole exon inclusion or skipping. Also, E13 (149 bps) is the only *hMLH1* exon of optimal size for wrapping around a nucleosome (147 bps of DNA) and nucleosomes have been shown to be positioned at internal exons of actively transcribed genes (Andersson et al., 2009); therefore it is possible that exon definition at both the DNA and RNA level contributes to the high fidelity of E13 splicing (inclusion) in C1.

Secondly, SRp40, SF2/ASF, 9G8, and hnRNP A1 motifs are found in all *hMLH1* exons although the presence or absence and distribution of high scoring versus low scoring motifs varies for each factor. Focusing on the distribution of high scoring motifs, the constitutive splicing factor SRp40 is found throughout *hMLH1* in E2-4, E6, E13-15, and E17-18 whereas the distribution of high scoring SF2/ASF is overrepresented in the latter half of *hMLH1* being found in E1-2, E12-16, and E18. Several features stand out: (1) SRp40 is high scoring in exons 6 and 13 that comprise

minimum boundary associated with large internal deletions of whole exons in C2 and in E17 that is only skipped as a whole exon; in contrast, SF2/ASF is only found as low scoring throughout the bulk central portion of hMLH1 mRNA spanning E3-11 but is high scoring in all truncated acceptors identified (i.e. E2, E14, E15, E16, E18), (2) SF2/ASF has a high scoring motif in E16 but not E17 in direct contrast to SRp40 which has a high scoring motif in E17 but not in E16 such that differences in splicing pattern that are directly attributable to changes in SRp40 and SF2/ASF levels should be distinguishable on the basis of their impact on E16 and E17 skipping, and (3) both SRp40 and SF2/ASF have a high scoring motif in E15 that overlaps the 3' splice site of the 57-E15 dual splice site utilized in certain C2 splicing signatures either of which are expected, when bound, to prevent utilization of the site.

Thirdly, SC35, SRp55, and TRA2-BETA motifs are completely absent from certain exons: SC35 is completely absent from E6; SRp55 is completely absent from E2, E5-6, and E15; and TRA2-BETA is completely absent from E7, E11, and E17. This allows a certain level of discrimination between the factors as the high frequency skipping observed for E15 cannot be attributed to SRp55 binding and changes in TRA2-BETA levels cannot directly impact E17 skipping. Do high scoring factor sites also allow discrimination between SVs classified as C1 and C2 or between sub-classes of C2? Comparison of *hMLH1* exon skipping in MRC5 cells by category and type (whole exon versus exon truncation) shown in figure 3.2 with the location of high scoring motifs shown in Figures 3.1 and 3.3 suggests that use of specific high scoring factor binding sites distinguishes certain C1 and C2 SVs as well as certain sub-classes of C2 SVs.

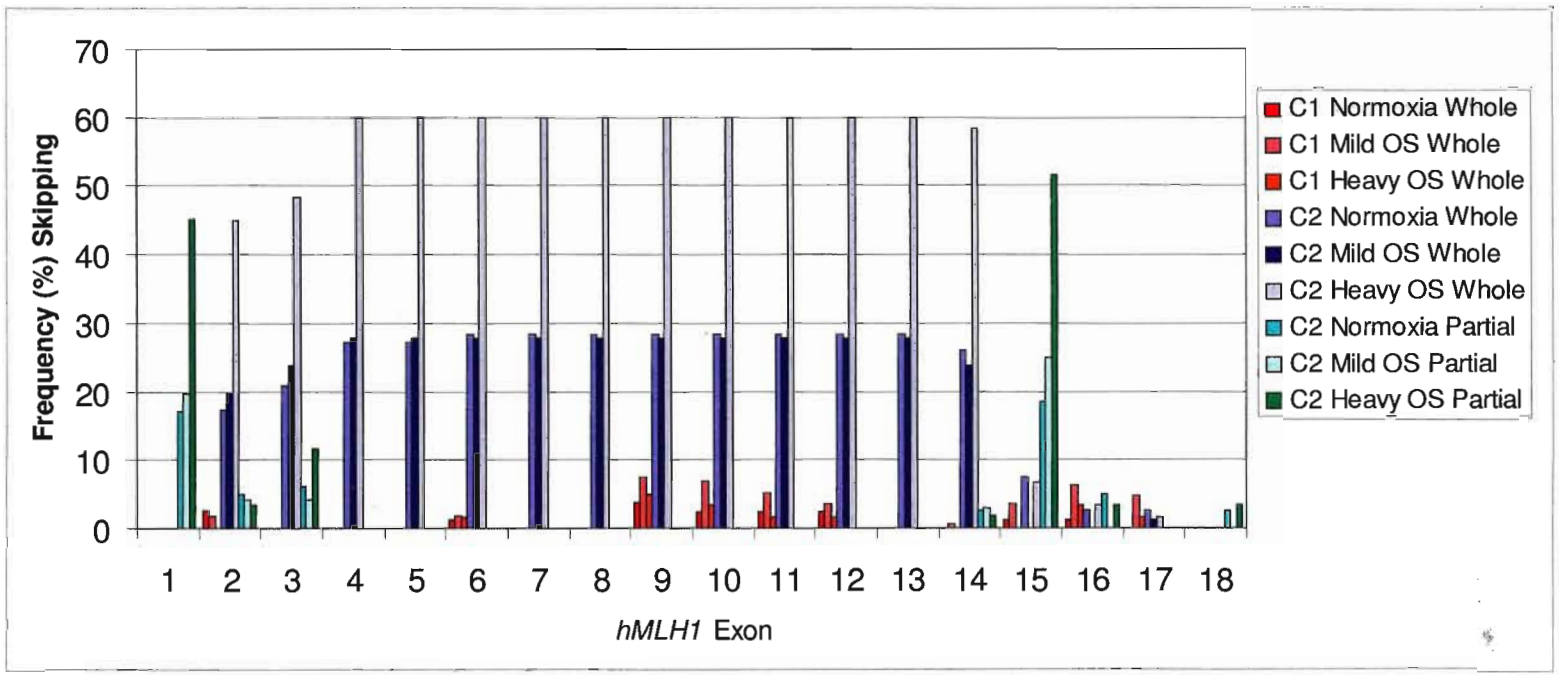


Figure 3.2: *hMLH1* exon skipping frequency by category, treatment, and whole versus partial skipping in MRC5 cells.

3.3.2. High Scoring SRp55 motifs containing TGCA occur in *hMLH1* exons 9-12 that are frequently skipped in C1 SVs and define the C1 splicing signature.

Examination of high scoring SRp55 motifs reveals distinct motif compositions that correspond to C1 exon skipping as shown in Figure 3.3; X indicates the absence of SRp55 motifs and L the presence of low scoring SRp55 motifs.

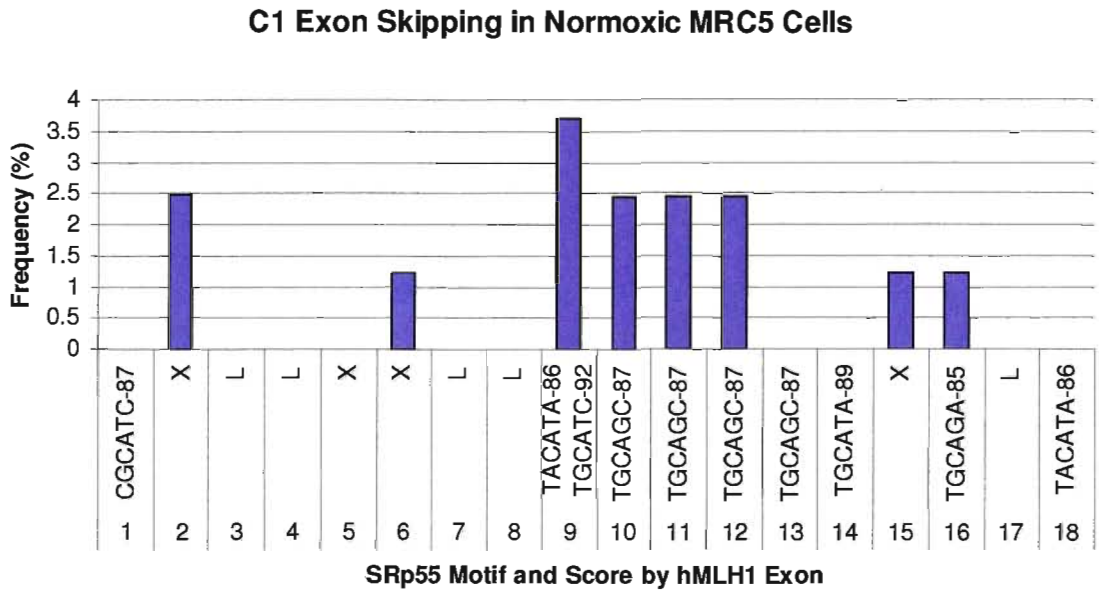


Figure 3.3: Frequency of *hMLH1* C1 Whole Exon Skipping in Baseline MRC5 Cells reveals elevated splicing of E9, E10, E11, and E12 with high scoring HSF SRp55 motifs containing TGCA. High-scoring HSF SRp55 motifs contain TGCA and are found in *hMLH1* exons 1, 9-14, 16, and 18; numerical value indicates HSF % above threshold score, X=absent, L=low scoring and non-TGCA-containing.

When the high scoring motifs that occur in each exon were recorded and examined, high scoring SRp55 motifs containing TGCA showed a distinct trend in that they occurred in E9-12. Interestingly, these 4 exons comprise the maximum adjacent exon skipping observed for the C1 splicing signature. No other factor showed this level of correspondence suggesting that SRp55 may be the factor driving the C1 splicing signature and lending support to the idea that C1 and C2 splicing events are in fact distinct splicing categories reflecting specific responses to cellular signals controlled by splicing factors. Consistent with this idea, E9 that contains the highest scoring TGCA

containing motif also exhibited the highest frequency of skipping. Why then, is no skipping observed for E13 which also contains the TGCA motif? Closer inspection of the context of the SRp55 motif within each exon and in relation to other high scoring motifs showed that whereas in E9-11 the SRp55 motif is highest scoring without other overlapping motifs, in E13, which contains multiple redundant high scoring motifs for all seven splicing factors, the high scoring SRp55 motif is overlapped by two other high scoring motifs, SRp40 and SC35. Since the SRp40 motif has an equivalently high score compared to SRp55, it is likely that SRp40 which is normally present, prevents binding of SRp55 at this site or that the other multiple redundant factors throughout the exon overcome the ability of SRp55 to disrupt the splicing machinery to the extent that exon skipping occurs. This results in E13 inclusion, even in the presence of SRp55 and sets the upper boundary of the multiple adjacent splice unit, E9-12, that defines the C1 splicing signature. The lower boundary is established by the absence of a high scoring SRp55 motif containing TGCA in E8 (associated with exon skipping) such that E8 is retained in C1 splicing. Thus, the defining feature of C1 splicing, maximum multiple adjacent exon skipping of E9-12, can be explained in terms of context dependent high scoring SRp55 motifs containing TGCA where the presence of bound SRp55 results in exon skipping of E9, 10, 11, and 12.

Recently, several studies have documented changes in specific splicing factor ratios in response to H_2O_2 to induced OS. Among these, SRp55, FOX-2, and TRA2-BETA (Tsukamoto et al., 2001; Takeo et al., 2009) are all inducible however the timing, mechanism, and dose dependence of the response varies between the factors. If SRp55 is responsible for the observed C1 skipping of E9, 10, 11, and 12, then treating MRC5 cells with different levels of H_2O_2 to induce OS might be expected to influence skipping of SRp55 regulated exons. In fact, E9, 10, 11, and 12 skipping increases in response to 250HP (about double) as shown in figure 3.4. However, the response differs after 500HP; a slight increase in E9 and E10 skipping is observed relative to normoxia whereas a slight decrease in E11 and E12 skipping is observed. Since E9/10 skipping has been shown to exert a dominant negative effect on MMR (Peasland et al., 2010), the

changes observed may reflect a shift away from SRp55 driven splicing in response to 250HP and towards a p53 regulated apoptotic response to greater levels of genotoxic damage with 500HP (or perhaps a shift from apoptosis to necrosis). Filippov et al. (2007, 2008) have shown that SRp55 can be transiently up-regulated in response to sub-lethal doses of genotoxic agents (in the absence of p53) and participates in the DNA damage response by changing the isoform ratios of target genes including the *Fas* (p53 inducible) cell death receptor to the soluble form that staves off activation of the cell death signal transduction cascade. Similarly, the results shown in figure 4.4 show a trend towards increased hMLH1 E9-12 skipping (about double) with 250HP consistent with a transient role (within five hours of treatment) for SRp55 up-regulation in response to moderate genotoxic damage that attenuates in the presence of 500HP (reminiscent of p53 pathway activation in response to greater DNA damage). In the Filippov studies, the response was only observed in the absence of p53 (in p53 null cell lines) and the level of genotoxic stress induced was higher. If the SRp55 driven response can also occur in the presence of p53 at sub-lethal doses, such as 250HP, that fail to trigger the p53 regulated response, then the findings in this study are consistent with the literature and lend support to the idea that SRp55 up-regulation is contributing to the observed C1 splicing pattern in response to OS.

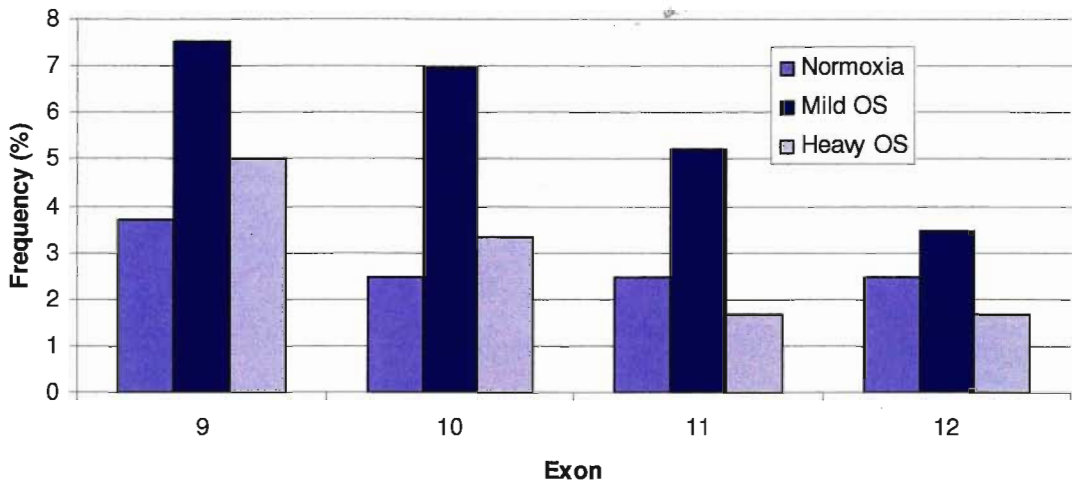


Figure 3.4: The frequency of *hMLH1* E9, 10, 11, and 12 that contain high scoring SRp55 motifs containing TGCA show a trend towards increased skipping in response to mild but not heavy oxidative stress (OS) relative to normoxic MRC5 cells.

If SRp55 is driving the C1 splicing pattern, are exons 14, 16, and 18 that also contain high scoring SRp55 motifs similarly regulated? In fact, the evidence suggests that they are, with the strongest support found for exon 16 that contains the extended TGCAG motif also found in exons 10, 11, and 12. First, in the presence of 250HP, there is an increase in transcript diversity that specifically reflects various combinations of SRp55 bearing exons (found at low frequency) including skipping of E9-10, E10-11, E9-11, and E9-12, E9-12&16. Moreover, E16 skipping (as the only exon skipped in the transcript) doubles in the presence of 250HP relative to normoxia and 500HP. These transcript types constitute a unique subset of C1 transcript types that appear to be derived from SRp55 regulated splicing and this notion is supported by the observation that the increase in transcript diversity in the presence of 250HP cannot be accounted for only in terms of the increase in sample size. That is, sample size doubled and transcript diversity tripled (resulting in some of the aforementioned transcript types). Also, most of the transcript types described above follow the trend seen in figure 3.4 in terms of increased frequency in response to 250HP relative to normoxia and 500HP as seen for E9-12 that

are proposed to be SRp55 regulated. Furthermore, when the overall occurrence of E14 and E16 skipping is plotted as a frequency based on the sum of all C1 transcript types containing an E14 and/or E16 skipping event as shown in figure 4.5, the same trend of increased exon skipping with 250HP compared to normoxia or 500HP is again observed in MRC5 cells.

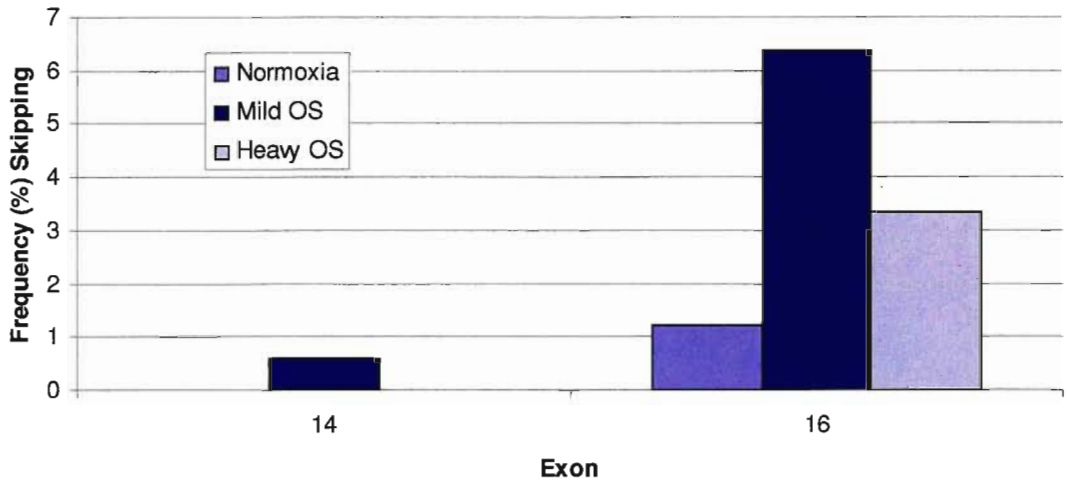


Figure 3.5: Frequency (%) of C1 *hMLH1* E14 and E16 skipping that have a high scoring SRp55 binding site containing TGCA in the motif show a trend towards an increase in frequency with 250HP compared to normoxia and 500HP in MRC5 cells.

It is unclear why E14 is skipped infrequently compared to the other TGCA containing SRp55 motif containing exons. It is possible that the location of the SRp55 motif at the front of E14 compared to the second half of the exon as in all other cases influences the splicing outcome. SRp55 has been shown to be enriched in exons near 5'splice sites and may exert a corresponding context dependent effect (Wang et al., 2005; Xiao et al., 2007). Alternatively, it is possible that motifs ending in CATA behave differently, as in the high scoring motif found in E18, TACATA, which does not contain TGCA, for which no skipping was observed in any of the three conditions in MRC5 cells. It is also possible that some other factor preferentially binds the E14 site and prevents usage of the site by SRp55 (a possible candidate would be FOX-2, to be discussed later). Finally, it is possible that in fact E14 is SRp55 regulated but that the transcript types

remain unidentified in this study and are among those 'other C1 types' for which incomplete sequence data prevented definitive identification of transcript types.

Despite the high-scoring SRp55 motif in E18, TACATA, no exon skipping was observed. This suggests that the observed pattern depends specifically on TGCA containing motifs. Based on this observation, the *hMLH1* pre-mRNA sequence was searched for additional TGCA containing motifs in exons or in flanking intronic regions that were not identified as part of a computationally derived SRp55 motif. This analysis revealed several additional motifs of interest, underscoring both the value and limitations of employing computationally derived motifs as the basis for model building. Notably, E2 contains a TGCA motif at position 6-9 of the exon (TTTAG/A**TGCA**A...). In the presence of 250HP, Δ^{6}E2-^{129}E16 completely disappears but it is present in normoxia and 500HP, arguing for SRp55 driven splicing for this C2 SV as well. Consistent with this idea, the E16 high scoring motif occurs just downstream of the 129-E16 acceptor (....TGAG/ATCGT**TGCA**G). A similar pattern is observed for Δ^{30}E3-^{21}E18, although only the 30-E3 dual splice site contains a nearby downstream TGCA (E18 does not contain a TGCA motif and the high-scoring TACATA is further downstream, near the 5'ss). This raises the possibility the TGCA (and SRp55, if shown to bind as predicted) are involved in dual splice site regulation and if correct, suggests that dual splice site use may be linked to the cellular response to 250HP (p53 independent), an intriguing idea that warrants further investigation. The analysis also revealed a TGCAG motif immediately prior to E13 and E14. Moreover, E13 has 5 TGCA motifs within and around it in flanking region, one before and one after, three within, providing another possible reason why E13 is skipped as a whole exon, but not truncated.

Conversely, there is another possible candidate for binding to the TGCA element and that is FOX-2. It is also possible that both factors compete for binding and/or utilize the same sites. Such a scenario could explain the observed splicing of E17 in response to 250HP. There is a conserved FOX-2 binding element in intron 17 within 40 bps of the end of E17 (CA**TGCAT**G) and FOX factors have been shown to be enriched near alternative exons and regulate their inclusion (Ponthier et al., 2006; Yeo et al., 2007b).

FOX-2 elements have been shown to promote inclusion when located downstream of the exon (versus exclusion when found upstream of the intron or in the exon), to mediate epithelial cell-specific fibroblast growth factor receptor 2 exon choice (reviewed in Kuroyanagi, 2007; Zhou et al., 2007; Zhou and Lou, 2008) and to play a role in developmental contexts (Yeo et al., 2009) suggesting the likelihood of its presence in MRC5 cells is high and providing a possible explanation for the high fidelity of E17 inclusion in normoxia. FOX-2 may also be responsive to differing levels of OS and/or genotoxic stress and it has been shown to regulate other splicing factors including SRp55 in certain contexts (Yeo et al., 2009). If the SRp55/TGCA motif binding factor is also upregulated, as appears to be the case in 250HP, this factor could compete with the FOX-2 site (and vice versa) resulting in the appearance of E17 skipping. However, there is also a TGCA motif within 10bps of the start of E17 (in intron 16) that is not within the preferred FOX-2 context suggesting binding of SRp55 at this site in 250HP may attenuate the effect of FOX-2 binding downstream of E17 (that normally promotes inclusion) resulting in the observed skipping. Since SRp55 effect is expected in 250HP (but not, or to a lesser extent) in 500HP, E17 skipping would still be expected to occur with 500HP but to a lesser extent, which is what is observed. In addition, such interactions may explain the splicing of E9, and E14 in which the TGCA motif in TGCAT that resembles the FOX-2 element but all these possibilities would have to be verified. Whatever the case, SRp55 and FOX-2 appear to be likely candidates involved in the regulation of *hMLH1* splicing, and as such they warrant further investigation.

3.3.3. Investigating SRp55 and FOX-2 involvement in *hMLH1* alternative mRNA splicing.

In order to investigate this idea, the first step would be to determine whether in fact SRp55 and FOX-2 trans-acting factors are present in the MRC5 cell line and whether their expression responds to HP-induced OS in a dose dependent manner. In order to ensure the results are meaningful, active versus inactive phosphorylation states (if applicable) must also be determined and measured. Similarly, potential changes in compartmentalization must be considered and measured if applicable. Next, a mini-gene

construct encompassing E17 with flanking introns containing the regulatory elements could be tested for skipping in response to different levels of each factor under a variety of conditions.

Alternatively, if SRp55 and FOX-2 are shown to be present in a cell line, characterizing the *hMLH1* splicing pattern in response to over-expression and knock-down under basal and OS conditions could also provide evidence depending on how profile shifts. RNA expression and hMLH1 protein expression should also be quantified alongside this characterization to gauge the extent to which a full complement of WT hMLH1 is expressed in the cell since this has been shown to be important for the various functions of hMLH1 (Cejka *et al.*, 2003; reviewed in O'Brien and Brown, 2006; Mastrocola and Heinen, 2010). Inclusion of *POLB* exon alpha could be monitored as an additional control in response to 250HP. This alternative exon ends in the TGCAG motif and is predicted to be up-regulated by SRp55 under mild but not 500HP, based on results in the *hMLH1* profiles and on *POLB* splicing patterns discussed in chapter four of this thesis.

3.3.4. Evidence for TRA2-BETA involvement in C2 splicing

Do any of the known SR proteins correspond to the C2 splicing signature? Of all the factors, the most likely candidate, based on the distribution of high scoring motifs appears to be TRA2-BETA both for the exons in which it is found to be high scoring and due to the complete absence from E17, the only factor examined to be completely absent from E17. E17 skipping is interesting in that it is not truncated in any C2 SVs, suggesting that whatever factor is absent from E17, may be linked to exon truncations such as those found in E1, E2, E3, E14, E15, E16, and E18. Also, TRA2-BETA is high scoring in E1, 2, 3 that are truncated, but not in E4 or 5. It is also high scoring in E6 & E13 that form the boundaries of the minimum multiple adjacent splicing unit observed as part of the C2 splicing signature and is present in most exons (except E7 and E11) which make it a good candidate for a splicing factor that, when absent, results in multiple adjacent exon skipping. It is also present at low score in E14 and at high score in E15, 16, and 18 that contain exon truncation in C2 SVs. Notably, the acceptor in E15 for the

main C2 SV, ^{112}E1-^{5}E15, occurs within a TRA2-BETA site with 100% homology to the computationally identified HSF AAGAA (as shown in Figure 3.6), such that loss of TRA2-BETA would be consistent with E15 being a splicing hotspot in terms of C2 SVs, both in terms of diversity of acceptors and in terms of the main SV that is occurs at high frequency. In addition, TRA2-BETA has recently been shown to occur in two conformations, one of which is part of a stem loop structure (Tsuda et al., 2010) and it is possible that the repeat at 112-E1 and 5-E15 may form such a structure, hinting at why this particular site occurs with such high affinity, if for example, certain conditions lead to the abolition (or promotion) of one kind of binding but not another. Moreover, TRA2-BETA has been shown to be upregulated in response to OS in rat mucosa (Takeo et al., 2009) in a manner consistent with the results of this study in human cells. Unfortunately, the rules governing TRA2-BETA regulation have proved elusive in previous studies and a variety of factors are expected to mediate the effects of TRA2-BETA on splicing. For example, phosphorylation state, alternative isoforms of TRA2-BETA (Stoilov et al., 2004), ratios of other SFs including hnRNP G (Nasim et al., 2003) and SRp55 and FOX-2 (Zhou et al., 2007) have all been shown to mediate the impact of TRA2-BETA on splicing often with opposite effects, depending on context including location of cis-elements and tissue specificity (Venables et al., 2005). However, there is strong circumstantial evidence to suggest it is involved. Pursuing this line of investigations will require careful planning and methodology as the interactions between factors may prove difficult to tease apart.

3.3.5. FOX-2 provides a link between the C1 and C2 splicing signatures.

Despite the disparate characteristics of the C1 and C2 splicing signatures, it is possible that they are regulated by a common factor. The Fox-1/Fox-2 family of evolutionarily conserved RNA-binding proteins recognizes the (U)GCAUG element in RNA (reviewed in Kuroyanagi, 2009) that resembles a composite of the SRp55/TGCA and Fox-2/TGCATG motifs thought to be active in this work. Intriguingly, global analysis of Fox-2 targets has identified splicing regulators including SRp40 and SRp55 in human embryonic stems cells (Yeo et al., 2009). Thus, it is possible that changes in Fox-

2 alter splicing of SRp55, SRp40, and other factors that regulate *hMLH1* splicing. Consistent with this idea, high scoring SRp40 motifs are also aligned with distinguishing exons of the C2 splicing signature (e.g. E6 and E13). Therefore, it is possible that changes in SRp40 (and interactions with other factors) are driving the C2 splicing signature. Moreover, Fox-1 and Fox-2 have been implicated in U2-dependent, U1-independent splicing (e.g. through inhibition of the E-complex formation) and alternative 3' end formation of specific genes respectively (Fukumura et al., 2009; Kuroyanagi, 2009 and references therein; Yeo et al., 2009). Theoretically, U2-dependent, U1-independent splicing is what would be expected if recursive splicing were mediating the subdivision of long introns such as those observed in this work; it may be worth investigating whether this phenomenon occurs in a Fox-2 regulated manner. Based on the variety of mechanisms and factors that modulate *hMLH1*, this gene provides a valuable model system for studying alternative splicing regulation.

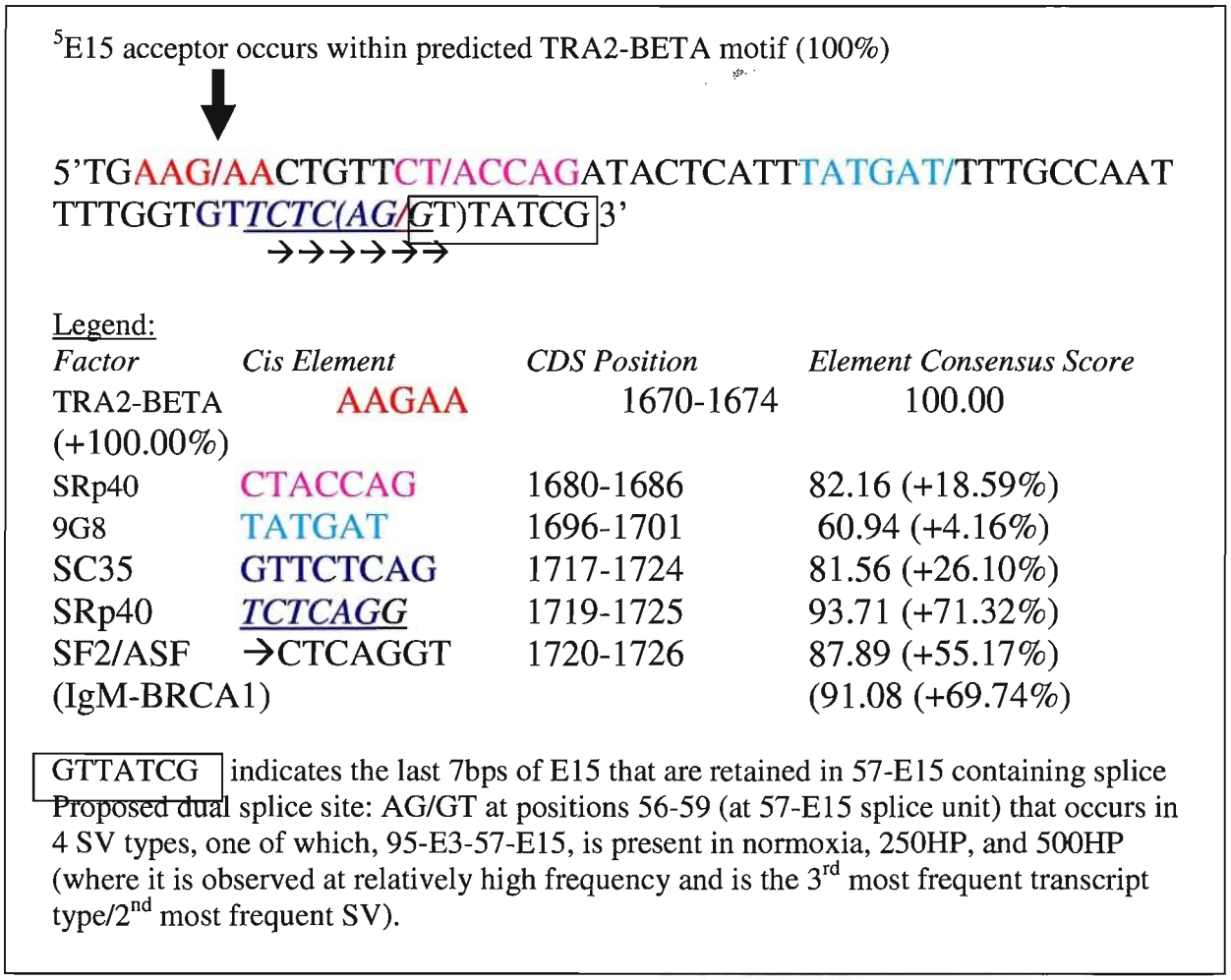


Figure 3.6: Regulatory Motifs Identified by HSF in *hMLH1* Exon 15. Exon 15 positions 1-64 correspond to CDS positions 1668-1731. The 5E15 acceptor is the most frequently utilized splice site MRC5 cells under normoxia, 250HP, and 500HP and occurs within a TRA2-BETA motif with the highest homology (100%) to the RRM of any of the computationally identified factor binding sites found in E15.

Chapter 4. Differences in *hMLH1* mRNA splicing profiles in transiently and stably demethylated sporadic non-polyposis colon cancer cell lines and relationship to promoter ^{5me}CpG methylation level.

The goal of Chapter 4 is determine whether different levels of *hMLH1* promoter methylation influence the splicing profile of *hMLH1*.

It is hypothesized that intermediate levels of *hMLH1* promoter methylation will result in a decrease in frequency of the CT and a corresponding increase in the diversity and frequency of splice variants.

Contributions: Generated cDNA library from total RNA (provided by collaborators) for all colorectal cancer cell lines and characterized the mRNA splicing profiles of *hMLH1* and *POLB* as described in methods; collated raw methylation data (provided by collaborators) into tables; performed all analyses and chapter write-up.

4.1 Introduction

In chapter 4, the potential interplay between alternative splicing and promoter cytosine (^{5me}CpG) methylation in colon cancer cell lines is examined. Loss of fidelity in either process is associated with cancer. Do alternative splicing and promoter methylation also influence each other? Although promoter ^{5me}CpG hypermethylation is associated with transcriptional silencing, both tumors and normal matched tissues display mosaic patterns of *hMLH1* promoter ^{5me}CpG and intermediate levels exist in *hMLH1* expressing cells. Whether differential levels of *hMLH1* promoter ^{5me}CpG influence *hMLH1* alternative splicing patterns has not been studied. In order to investigate this question, a system is required in which both mechanisms operate. Such a situation exists in the subset of sporadic non-polyposis colon cancer wherein *hMLH1* is epigenetically silenced at the transcriptional level by aberrant promoter cytosine methylation. In Vaco432 cells, acute treatment with the demethylating agent 5-aza-2'deoxycytidine (AzadC) transiently restores *hMLH1* mRNA expression and wild-type (WT) protein expression and allows the *hMLH1* mRNA splicing profile to be determined and potential cancer specific signatures to be identified (e.g. by comparison to non-cancer profiles). Since *hMLH1* is transcriptionally silenced prior to AzadC treatment, neither methylation or potential drug effects on mRNA splicing can be identified by a standard before and after treatment comparison. As a proxy, stably demethylated clonal cell lines of Vaco432 were used. The clonal cell lines were established by seeding and propagating AzadC treated (*hMLH1* expressing) Vaco432 isolates (single cells) in standard culture for 4 weeks by which time most of these clonal cell lines no longer expressed *hMLH1* (due to re-methylation of the promoter and associated transcriptional silencing) except in rare cell lines that remained stably demethylated and expressed *hMLH1* (for >1yr in continuous culture). Total RNA was obtained from 3 such stably demethylated cell lines: V20 (biallelically derepressed), V5 (paternal allele derepressed), and V25 (maternal allele derepressed) – the expressing (derepressed) *hMLH1* alleles are distinguished by a coding region polymorphism at codon 219 that allows their origin (paternal, maternal, or both) to be identified by sequencing. Therefore, this model system also allows differences, if any,

between alternative splicing patterns attributable to the maternal and paternal alleles to be identified.

Since silencing of hMLH1 has been causally linked to tumor progression in these cell lines, hMLH1-/- cells are associated with the cancer phenotype whereas hMLH1+/+ cells are not. Therefore, comparing the hMLH1 splicing profile in the stably and transiently demethylated VACO cell lines should allow cancer associated changes in hMLH1 mRNA splicing to be identified (taken together with the comparison of splicing in non-cancer and transiently demethylated cancer cell lines).

The methylation status of the core *hMLH1* promoter that is responsible for transcriptional silencing is known for RKO (AzadC), Vaco432 (AzadC), V20, V5, and V25 and the cell lines differ in their level of promoter methylation (raw methylation data and total RNA from corresponding cell lines supplied by collaborators, Dr. David Sedwick and associates, in Cleveland, Ohio, USA). This allows the question of whether or not there is a potential link between the two epigenetic mechanisms to be addressed by determining whether a correlation exists between the alternative splicing pattern generated and the level of promoter methylation of *hMLH1* in the cell lines examined.

To monitor for a possible drug effect of AzadC on *hMLH1* splicing, an additional transiently demethylated cell line, RKO was examined. Thus, significant changes in WT frequency and splice variants present (or absent) in all three stably demethylated clones that are absent (or present) in both Vaco432 and RKO indicate the possibility of a drug effect on splicing. Understanding the influence of AzadC on cellular function including splicing outcomes is of clinical interest as AzadC based treatments are used in certain aggressive cancer syndromes. Differences in *hMLH1* mRNA splicing profile between Vaco432 and RKO also serve as a cell line control to distinguish between features unique to a particular cell line versus those that may apply to *hMLH1* epigenetically silenced colon cancer cell lines in general. As an additional control, the *hMLH1* WT splicing frequency was determined in the *hMLH1* +/+ SW480 colon cancer cell line before and after AzadC treatment.

To distinguish between potential gene specific effects and global effects of AzadC treatment and/or different levels of promoter methylation on *hMLH1* mRNA splicing, the splicing profile of *POLB* was established as a control. *POLB* is constitutively expressed before and after AzadC treatment in RKO and Vaco432, and in the clonal cell lines V20, V5, V25, and V17 (a fourth clonal cell line of Vaco432 that is *hMLH1-/-* and included as a control). In addition, the splicing profile of *POLB* was established in the colon cancer derived cell line SW480 before and after AzadC treatment. SW480 is not subject to transcriptional silencing of *hMLH1* by epigenetic means and serves as a colon cancer cell line control that allows methylation specific mechanisms (direct effects) to be distinguished from global mechanisms (indirect effects).

4.2 Materials and Methods

4.2.1 Cell Lines

Total RNA from the human colon cancer derived cell lines, RKO (untreated), RKO (AzadC), Vaco432 (untreated), Vaco432 (AzadC), V20, V5, V25, V17, SW480 (untreated), and SW480 (AzadC) was a generous gift from Dr. David Sedwick at the University of Cleveland in Cleveland, Ohio, USA. In untreated RKO and Vaco432, *hMLH1* is transcriptionally silent due to biallelic promoter hypermethylation. The cell lines were cultured as previously described (Veigl et al., 1998). Cells were treated with 0.45 uM (0.1 ug/mL) 5-aza-2'-deoxycytidine on days 2 and 4 (with washout on days 3 and 5) and harvested on day 6, resulting in transient *hMLH1* mRNA re-expression, as evidenced by illumino-fluorescence imaging that detects the presence (or absence) of *hMLH1* expression on a per cell basis. A coding region polymorphism at codon 219 allows identification of the maternal and paternal alleles in the Vaco cell lines. From the parental cell line, single cell clones were seeded and described previously. Although most clones reverted to the transcriptionally repressed state of *hMLH1* within three weeks, some clones remained stably demethylated after 5 weeks. From the stably demethylated clones, 3 cell lines were established as follows: (1) V20 – biallelically

derepressed, (2) V5 – paternal allele (GTC) derepressed, (3) V25 – maternal allele (ATC) derepressed. In addition a fourth cell line, V17 was established as a control in which *hMLH1* mRNA expression was biallelically re-repressed (silenced) via promoter hypermethylation. *hMLH1* in expressed both before and after AzadC treatment in SW480 cells.

4.2.2 *hMLH1* and *POLB* mRNA Splicing Profiles

Methods used to generate mRNA splicing profiles have been described in chapter 2, methods (see also Disher and Skandalis, 2007; Skandalis and Uribe, 2004). Total RNA from the cancer derived cell lines was provided by Dr. David Sedwick and associates (Cleveland, Ohio). As before, reverse transcription and PCR with oligo-dT primer was used to generate a cDNA library and PCR amplification with gene specific primers (as given in chapter 2, table 2.0 for *hMLH1* and in table 4.0 that follows for *POLB*) was used to isolate mRNA transcripts originating from the gene of interest.

Table 4.0: *POLB* gene specific oligonucleotide primers.

Primer Name	Location	Primer Sequence
*POLB*uniFORWARD	Exon 1	5' CCGCAGGAGACCCTCAACGG 3'
*POLB*uniREVERSE	Exon 14	5' CCACTGGATGTAATCAAAAATGTC 3'

Wild-type *POLB* mRNA: 967 bps

4.2.3 Promoter Methylation Profile

Raw methylation data for the core/proximal (to transcription start site) *hMLH1* promoter region (C and D) shown to be important for transcriptional silencing (Deng et al., 2001; see also Chapter One, Figure 1.4), and corresponding to the cancer cell lines from which total RNA was obtained, were provided by collaborators, Dr. David Sedwick and colleagues at the University of Cleveland, Ohio, USA. The methodology used to map the methylation sites in the *hMLH1* promoter was as follows: (1) Isolate, denature and treat DNA with sodium bisulfite to convert unmethylated cytosine to uracil, (2) PCR amplify promoter region and clone into the pCR72.1-TOPO vector so that each vector contains the *hMLH1* promoter region from a single allele, and (3) Sequence 10-15 clonal isolates to discern the pattern of CpG methylation (15 sites). Biallelic versus monoallelic

expression of hMLH1 VACO432 subclones was determined using a coding region polymorphism at codon 219. The raw data were provided in Excel format for each of the 15 CpG positions, in each of the isolates, by cell line.

4.3 Results and Discussion

4.3.1. Level of *hMLH1* Promoter ^{5me}CpG

To investigate whether the level of core *hMLH1* promoter ^{5me}CpG influences the alternative mRNA splicing of *hMLH1*, a model system of sporadic non-polyposis colon cancer cell lines in which both factors operate was utilized. The pattern of *hMLH1* promoter methylation is partially known, based on analysis of raw methylation data (from collaborators) indicating the presence or absence of methylation at each of the 15 CpG sites comprising region C (core) and D (proximal) of the *hMLH1* promoter, and the cell lines differ in their level of promoter methylation. The data were collated and the average percent (%) methylation was determined for each position based on 5-18 clonal isolates as shown in Table 4.1; several trends emerge.

Table 4.1: Collated methylation data showing the average percent (%) methylation at the fifteen CpG sites by position in the core (C) and proximal (D) *hMLH1* promoter (relative to the transcription start site) by cell line and condition.

CELL LINE	#	+15	-6	-21	-54	-62	-86	-109	-178	-193	-204	-223	-229	-231	-241	-248
V432-UN	10	90	90	100	100	100	80	90	100	100	100	100	100	100	100	100
V432-AzadC	18	39	39	39	39	22	28	39	39	50	39	39	50	39	39	39
V20	10	10	10	10	10	10	10	20	10	10	10	0	0	0	0	0
V5 (GTC)	16	88	88	75	94	94	94	81	63	56	63	63	63	63	63	63
V25 (ATC)	10	80	10	20	20	20	20	90	10	90	100	100	100	100	100	100
V17	24	96	92	54	63	79	67	92	83	96	100	100	100	100	100	92
RKO-UN	5	100	100	100	60	100	100	100	100	100	100	100	100	100	100	100
RKO-AzadC	15	67	40	73	53	47	93	87	80	87	87	87	93	93	73	80
					D							C				

Legend: Cell lines highlighted in yellow (V432-UN, RKO-UN, and V17) are hMLH1- (untreated and stably re-methylated respectively), the remainder are hMLH1+. AzadC denotes transient demethylation whereas Vaco432-Azad clones (V20, V5, and V25) are stably de-methylated. The number of individual clonal isolates analyzed for each condition is given in red (#).

The overall level of *hMLH1* promoter methylation (C&D) from highest to lowest is: RKO-AzadC (76%), V5 (74%), V25 (64%), Vaco432-AzadC (39%), V20 (7%). However, rank order results differ between the two promoter regions as follows: Region C: V25 (88%), RKO-AzadC (85%), V5 (62%), Vaco432-AzadC (42%), V20 (4%); and Region D: V5 (88%), RKO-AzadC (66%), V25 (37%), Vaco432-AzadC (35%), V20 (11%). The two transiently demethylated cell lines differ in their level of methylation overall and by region in a consistent fashion: RKO-AzadC retains a high level of methylation (greater than 50%) whereas Vaco432-AzadC exhibits low methylation (less than 50%). Both biallelically de-repressed Vaco432-AzadC (transient) and V20 (stable) consistently exhibit low promoter methylation. Among stably demethylated clones, both V5 and V25 have higher promoter methylation compared to V20 overall but differ by region. Whereas V5 exhibits high methylation in both regions C (62%) and D (88%), V25 exhibits high methylation in region C (88%) and low methylation in region D (37%).

4.3.2 *hMLH1* mRNA Splicing Profiles in Colon Cancer Cell Lines

The *hMLH1* mRNA splicing profiles were characterized in the colon cancer derived cell lines RKO (AzadC), Vaco432 (AzadC), V20, V5, and V25 based on a total analysis of 293 mRNA transcripts (73, 39, 83, 42, and 56 respectively). It was found that *hMLH1* is extensively spliced and exhibits 25% SVs in RKO, 54% SVs in Vaco432, 46% SVs in V20, 50% SVs in V5, and 57% SVs in V25. The CT is the most frequent transcript type in all cell lines (found at 75%, 46%, 54%, 50%, and 43% respectively) as shown in Figure 4.1. Agarose gel analysis supports the presence of 41 unique transcript types in this study of which twenty-four (CT and 23 SVs) were definitively identified based on sequence results as shown in Table 4.2.

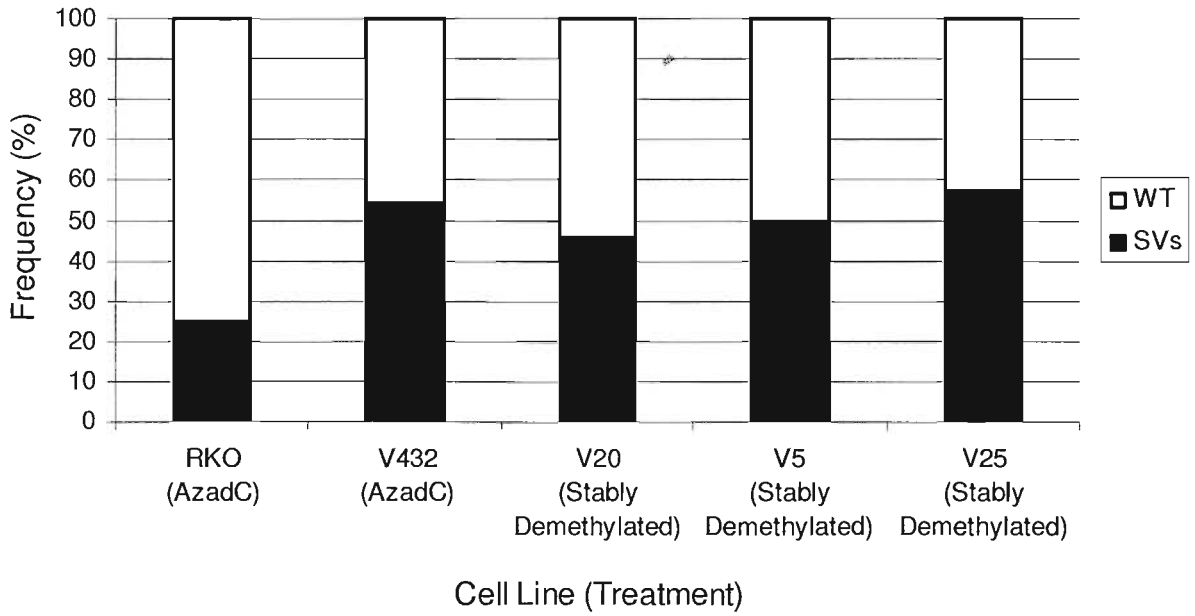


Figure 4.1: Frequency (%) of *hMLH1* WT transcripts and total SVs in the transiently demethylated (acutely AzadC treated) colon cancer derived cell lines RKO and V432, and in stably demethylated clones of Vaco432: V20 (biallelically derepressed), V5 (maternally derepressed), and V25 (paternally derepressed).

Table 4.2: *hMLH1* mRNA splice variant frequencies (%) in RKO cells after acute treatment with 5-aza-2'-deoxycytidine treatment, Vaco432 cells after acute treatment with 5-aza-2'-deoxycytidine, and in cells derived from three stably demethylated clones of Vaco432: V20 (biallelically derepressed), V5 (paternal allele derepressed), V25 (maternal allele derepressed).

		Splice variant frequency (%)				
Splice variant type	Category	RKO	V432	V20	V5	V25
Canonical Transcript	*CT*	*75.3*	*46.15*	*54.22*	*50.00*	*42.86*
Δ exon 2	1	0	2.56	2.41	2.38	1.79
Δ exon 6, 9	1	0	0	0	0	1.79
Δ exon 6, 9, 12	1	0	0	1.20	0	1.79
Δ exon 6, 9-10, 12, 15-16	1	0	0	0	0	0
Δ exon 9-10	1	4.1	7.69	1.20	0	0
Δ exon 10-11	1	0	0	1.20	0	0
Δ exon 9-11	1	0	0	0	2.38	0
Δ exon 9-11, 15	1	1.4	2.56	0	2.38	0
Δ exon 9-12	1	0	2.56	0	0	5.36
Δ exon 9-12, 16	1	0	0	0	4.76	0
Δ exon 9-11, 14-16	1	0	0	0	0	0
Δ exon 15	1	0	0	0	0	0
Δ exon 16	1	6.8	7.69	4.82	4.76	8.93
Δ exon 15-16	1	0	0	0	0	0
Δ exon 16-17	1	0	0	0	0	0
Δ exon 17	1	8.2	15.38	9.64	4.76	3.57
Δ Others C1 Types*	1	1.4	15.38	3.61	7.14	10.71
Δ E6-^{57}E15	2	0	0	0	0	0
Δ ^{22}E2-^{57}E15	2	0	0	0	0	0
Δ ^{95}E3-^{57}E15	2	0	0	7.23	11.90	12.50
Δ ^{95}E3-^{57}E15, 17	2	0	0	7.23	2.38	0
Δ ^{112}E1-^{5}E15, 17	2	0	0	0	0	0
Δ ^{112}E1-^{5}E15	2	0	0	0	2.38	0
Δ$^{1\text{-}5}$E2, ^{5}E3-^{5}E15	2	0	0	0	0	0
Δ ^{5}E3-^{5}E15	2	0	0	0	0	0
Δ ^{94}E1-^{14}E15	2	0	0	0	0	0
Δ ^{94}E1-^{69}E14	2	0	0	0	0	0
Δ ^{112}E1-^{129}E16	2	0	0	0	0	1.79
Δ ^{108}E1-^{34}E15	2	0	0	0	0	0
Δ ^{6}E2-^{129}E16	2	0	0	0	0	0
Δ ^{80}E2-^{14}E15	2	0	0	0	0	0
Δ ^{30}E3-^{21}E18	2	0	0	0	0	0
Δ Other C2 Types**	2	0	0	0	2.38	8.93
Δ ^{19}E3-^{57}E15†	2	0	0	2.41	0	0
$\sum$ ^{124}IN1(1167...1290 IN1)†	3	1.4	0	0	2.38	0
Δ E16, $\sum$ $^{-1,-2,-3(IN16,CAG)}$E17†	3	0	0	1.20	0	0
$\sum$ $^{1\text{-}3}$IN6, ΔE7-^{57}E15, Δ16†	3	0	0	1.20	0	0
Δ $^{1\text{-}5(TTTAG)}$E2, $\sum$ $^{-1...-17(IN8)}$E9, ΔE10-11†	3,4	0	0	1.20	0	0
Δ $^{1\text{-}5(TTTAG)}$E2, Δ $^{95\text{-}99(GTGAG)}$E3, ΔE16†	4	0	0	1.20	0	0
Δ $^{95\text{-}99(GTGAG)}$E3, Δ E16†	4	1.4	0	0	0	0
		n=73	*n*=39	*n*=83	*n*=42	*n*=56

Legend: *overall % frequency of all other C1 types (14 unique types total, sum of all cell lines); **overall % frequency of all other C2 types (3 unique types total, sum of all cell lines); †found only in colon cancer derived cell line(s) but not in MRC5 (either with or without OS); 3 or C3 denotes Category 3, partial intron retention (intron retention); 4 or C4 denotes Category 4, partial exon skip (exon truncation) without C2 event.

The distribution of *hMLH1* SVs summarized in Table 4.2 can be divided into 4 categories (C1, C2, C3, and C4) based on the classes of splice events they contain as shown in figure 4.2. C1 and C2 are as described previously, C3 denotes SVs containing intron retention events (e.g. exon border extension), and C4 denotes SVs with exon truncations without a C2 splice unit.

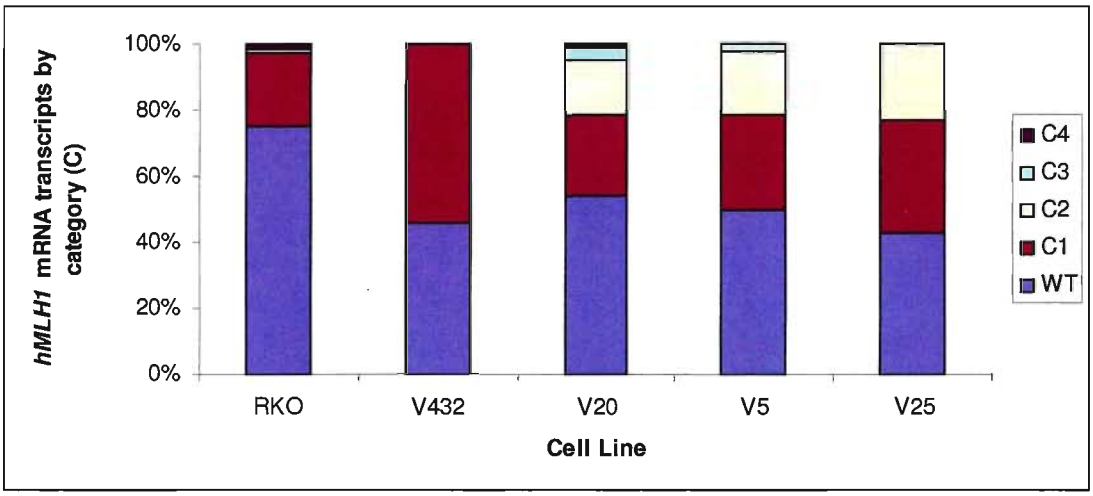


Figure 4.2: Distribution of *hMLH1* WT and SVs by category in the acutely AzadC treated cell lines RKO and Vaco432 and the stably demethylated clones of Vaco432: V20, V5, and V25.

Three unique features in SV distribution are observed in the colon cancer derived cell lines compared to MRC5 (chapter 2). Firstly, two new classes of splice events, C3, intron retention (specifically, exon border extension of canonical *hMLH1* exons) and C4, exon truncations (without C2) are observed in RKO and V20, albeit at low frequency. With identification of the V432 C1 SVs that presently remain unidentified (about 15%) and an increase in sample size of V5 and V25, it is possible that C3 and C4 SVs would also be found in the remaining cell lines (V432, V5, and V25). In contrast, no intron retention was observed in MRC5 under any conditions and $^{95}E3^{GTGAG}$ (C4) was not skipped independently but occurred at the start of a C2 SV. Intron retention is interesting because cancer specific intron retention events, if present, have diagnostic potential in terms of testing (since they are the addition of something not normally present, as opposed to the absence).

Secondly, there is a complete absence of C2 transcript types in the transiently, but not stably demethylated cell lines. In the latter (VACO clones) the main C2 SV Δ^{95}E3-^{57}E15 occurs with and without E17 skipping a frequency of 12-15% but few other C2 transcript types are generated. This differs from the MRC5 where greater diversity of C2 transcript types is observed; the main MRC5 C2 SV, Δ^{112}E1-5E15 is only observed once (V5) in all 293 colon cancer transcripts screened. The response to AzadC treatment also differs from the H_2O_2 induced response in MRC5. Whereas AzadC treatment results in complete disappearance of C2 SVs in RKO and VACO432; MRC5-250HP results in redistribution of C1 and C2 SVs while the main C2 SV remains constant. Furthermore, MRC5-500HP induces doubling of both the main SV and Δ^{95}E3-^{57}E15; both preferentially occur without E17 skipping like in V5 and V25 (but not V20), although the latter SV occurs at higher frequency in the stably demethylated clones. In contrast, there is a higher frequency of Δ9/10 skipping in the transiently (RKO, 4%; VACO, 8%) but not stably demethylated (V20, 1%) cell lines compared to MRC5. Taken together, these results point to a potential drug effect of AzadC on the *hMLH1* splicing profile that manifests in loss of the C2 splicing signature and an increase in ΔE9/10 associated with dominant negative MMR function (Peasland et al., 2010). However, it cannot be ruled out that these differences are associated with the underlying cancer profile, the expression of which is induced by transient demethylation. In this scenario, restoration of *hMLH1* function in stably demethylated clones restores the splicing profile over time and represents the normalized condition (i.e. reappearance of C2 and disappearance of Δ9/10 with stable demethylation as opposed to disappearance of C2 and appearance of Δ9/10 with AzadC drug treatment). The absence of these SVs in MRC5 baseline and their inducibility in response to 500HP argue against this line of reasoning. Comparison of *hMLH1* splicing profiles in additional non-cancer colon cell lines and matched normal and colon cancer tissues should help clarify this issue.

Thirdly, the only two distinct C1 transcript types that were identified in all colon cancer cell lines are ΔE16 and ΔE17 and these are observed at considerably higher frequency than in MRC5. Transcripts defined as only ΔE16 or ΔE17 are the most

frequently observed transcript types in RKO and Vaco432 (but not in V20, V5, and V25) and the most frequent C1 transcript types in all cell lines. In addition, a variety of other transcript types contain exon skipping events accompanied by either ΔE16 or ΔE17 both but not both. This may indicate cell/tissue specific differences in splicing factors/ratios.

Fourthly, the SV frequency of RKO is significantly lower than that found in any VACO cell lines, irrespective of transient or stable demethylation (also lower than MRC5). It is possible that this low diversity of transcript types is linked to the high level of promoter methylation. In contrast, the distribution of C1, C2, C3, and C4 as a percentage of total SVs (as opposed to of total transcript types including SVs and WT) is most similar for the transiently demethylated cell lines, RKO (AzadC) and Vaco432 (AzadC), and for the three stably demethylated VACO cell lines; this feature is driven mainly by the difference in C2 splicing signature.

4.3.3 Summary of *hMLH1* SVs with Distinct Expression Profiles in Colon Cancer

hMLH1 SVs with distinct expression patterns in the colon cancer cell lines (i.e. observed at higher frequency compared to MRC5) are summarized in table 4.3. The SVs contain particular splice events of interest; some were identified in MRC5 ($E2^{TTTAG}$, $E3^{GTGAG}$) but occur in a different class or category of SV in colon cancer cells (i.e. C4: with or without C1 event; with C3 intron retention).

Table 4.3: *hMLH1* SVs with Distinct Expression Patterns in Colon Cancer Cell Lines

Splice Variant Identity	**Frequency (%)**				
	RKO	**V432**	**V20**	**V5**	**V25**
Δ exon 9/10	4.11	7.69	1.20	0	0
Δ exon 9-12, 16	0	0	0	4.76	0
Δ exon 16	6.85	7.69	4.82	4.76	8.93
Δ E16, ∑ -1,-2,-3(IN16,CAG)E17†	0	0	1.2	0	0
∑ 1-3IN6, ΔE7-57E15, Δ16†	0	0	1.2	0	0
Δ 1-5(TTTAG)E2, Δ 95-99(GTGAG)E3, ΔE16†	0	0	1.2	0	0
Δ 95-99(GTGAG)E3, Δ E16†	1.37	0	0	0	0
TOTAL ΔE16 Skipping All Sources	**8.22**	**7.69**	**8.42**	**9.52**	**8.93**
Δ exon 17	8.22	15.38	9.64	4.76	3.57
Δ 95E3-57E15, 17	0	0	7.23	2.38	0
TOTAL Δ17 Skipping All Sources	**8.22**	**15.38**	**16.87**	**7.14**	**3.57**
Δ 95E3-57E15	0	0	7.23	11.90	12.50
Δ 95E3-57E15, 17	0	0	7.23	2.38	0
TOTAL Δ 95E3-57E15 Skipping All Sources	**0**	**0**	**14.46**	**14.28**	**12.50**

Legend: Δ denotes exon (E) skipping, ∑ denotes intron (IN) inclusion, ∑1-3IN6 denotes inclusion of bps 1-3 of intron 6 whereas ∑ -1, -2, -3 E17 denotes inclusion of the last 3bps of intron 16.

Several points of interest arise when comparing the MRC5 splicing profile to that of the colon cancer cell lines. Firstly, the dominant negative isoform, ΔE9/10 that occurs rarely in MRC5 is also observed rarely (V20) in the stably demethylated clones. In contrast, ΔE9/10 occurs more frequently in the acutely AzadC treated cell lines RKO and Vaco432 (about 4% and 8% respectively) suggesting that these cell lines are experiencing some sort of disregulation or regulated program that differs from that which maintains normal cell function (e.g. cell death program activation). It is tempting to speculate that this is a cancer associated phenomenon but this idea would require verification and careful quantification since the isoform also occurs under normal circumstances and has been identified and investigated in previous studies.

Secondly, ΔE16 occurs at a similar level in all the colon cancer cell lines but is observed much more frequently than in MRC5 suggesting possible tissue specific and/or cancer specific regulation. Skipping of E17 also occurs at higher frequency in all colon cancer cell lines relative to MRC5 however frequency of expression differs between the cell lines in a way that corresponds to promoter methylation levels; Vaco432 and V20 with low methylation have the highest frequency of E17 skipping whereas the other 3 cell lines with high promoter methylation exhibit a lower frequency (approximately 50% less) of E17 skipping, irrespective of transient versus stable demethylation. It is unclear what the significance of these results is at present, however it may relate to disruption of E17 enhancer activity (regardless of whether it is linked to promoter methylation or not). Alternatively, it is possible that level of promoter methylation is indicative of other methylation patterns (e.g. hypomethylation in the body of the gene, methylation in certain splicing factors or cis-elements) or of some other factor in the cell that is related to E17 splicing outcomes.

Thirdly, unlike in MRC5, intron retention (exon border extension) is seen in the colon cancer cell lines involving intron 6 (3bps adjacent to E6) and intron 16 (3bps preceding E17). Moreover, the splice units TTTAG and GTGAG that are observed as part C2 SVs in MRC5 also occur in non-C2 transcript types in the colon cancer cell lines. Of particular interest, GTGAG, the last 5 base pairs of E3 are observed as a discrete unit

in the colon cancer cell lines, whereas this unit only occurs as part of an internally deleted splice event in MRC5. Also, here these splice events occur together with E16 skipping, pointing to a link in regulatory mechanism. In contrast, the Δ^{95}E3-^{57}E15 SV is the only C2 SV observed at high frequency (only in stably demethylated clones) that is also observed in MRC5. In the stably demethylated clones 95E3-57E15 occurs together with E17 skipping, but not with E16 skipping suggesting 2 distinct, coordinated mechanisms or factors may be operating. The distinction between transcripts containing E16 and E17 skipping suggest reduced activity of SF2/ASF (E16) and SRp40 (E17) based on analysis of computationally derived cis-elements discussed in chapter three.

4.3.4. Promoter Methylation Level of *hMLH1* and mRNA Transcript Diversity

One way to generate the diversity of splice variants identified in the colon cancer cell lines summarized in Table 4.2 is through differential promoter occupation (combinatorial control) of transcriptional activators and repressors that also influence co-transcriptional alternative splicing (e.g. by impacting elongation rate either through kinetic coupling and/or factor recruitment that in turn result in differential inclusion of exons regulated in this manner). Promoter methylation may interfere with binding of transcriptional activators and/or repressors in two ways: (1) by direct interference, and (2) by inducing a more closed chromatin structure that prevents factor binding. If differential promoter occupancy contributes the extensive mRNA splicing identified for *hMLH1* in a particular manner, then the level of promoter methylation should correlate with: (1) SV diversity, and (2) frequency of inducible (e.g. kinetically coupled) SVs or splicing events. To investigate the first possibility, *hMLH1* transcript diversity – the number of unique transcript types (CT and SVs) – identified in each colon cancer cell line were tabulated as shown in Table 4.4. Since the sample sizes differ between the cell lines, the projected number of transcript types expected in a sample of 100 was calculated as follows: # of type observed in cell line divided by the number of transcripts analyzed in the cell line times 100. This provides an estimate of the number of transcript types that would be observed if 100 transcripts were characterized based on the rate of transcript identification observed in each cell line. Although this is only an approximation – the

results are skewed by sample size (i.e. projected transcript diversity is over-estimated to a greater extent in smaller sample sizes) – the approach captures the trends observed in the data and allows them to be compared.

Table 4.4: *hMLH1* transcript diversity (CT and SVs) in colon cancer cell lines.

Cell Line - treatment	RKO-AzadC	V432-AzadC	V20	V5	V25	TOTAL
Number of Unique Transcript Types	8	13	17	15	14	41
Total Transcripts Analyzed	73	39	83	42	56	293
Projected # of Types per 100 Transcripts	11	33	20	36	25	14

Legend: Projected # of types/100 (yellow); actual number of types/293 transcripts x 100 (orange); V, VACO. Note: Total includes RKO; VACO only combined: 40/220=18 projected types/ 100 transcripts.

Two features stand out: (1) RKO exhibits lower transcript diversity than the VACO cell lines, irrespective of transient or stable demethylation, and (2) among VACO cell lines, VACO432-AzadC and V5 exhibit higher projected diversity compared to V20 and V5 that exhibit lower projected diversity. To examine whether the trends in projected diversity correspond to the overall promoter methylation level within the cell lines, correlations were performed between (1) transiently demethylated cell lines, and (2) stably demethylated cell lines (data not shown). In transiently demethylated cell lines, there is a negative relationship between promoter methylation (overall and by region) and transcript diversity such that low methylation is associated with higher diversity (VACO432-AzadC) and high methylation is associated with low diversity (RKO-AzadC). Conversely, for the stably demethylated clones, there is a positive relationship such that promoter methylation (overall and by region) and diversity increase (or decrease) together. However, when all cell lines are plotted together, regions C&D behave differently. For region C, a curvilinear relationship appears to exist such that both low and high levels of methylation are associated with lower diversity whereas intermediate levels of methylation are associated with high diversity as shown in figure 4.3. If verified, this is consistent with both activators and repressors controlling region C.

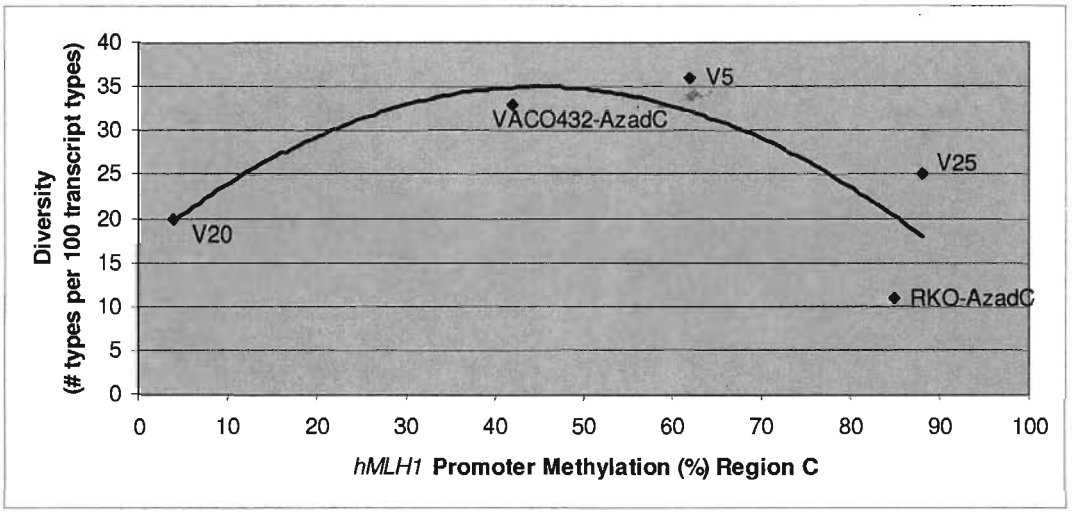


Figure 4.3: There appears to be a curvilinear relationship between *hMLH1* promoter methylation of region C and splice variant diversity in the colon cancer cell lines.

For region D, the overall trends that were discussed hold true but are more pronounced. On this basis, it is tempting to speculate that region D may be related to the regulation of C2 SVs, since the transiently (no C2) and stably (C2 present) demethylated cell lines exhibit opposite trends with respect methylation level of region D and diversity. This idea is consistent with the disappearance of specific E2 SVs with 250HP only in MRC5 that were all spliced close to a TGAG motif found in region D.

The *hMLH1* promoter has been shown to be regulated by a variety of factors and contains a variety of putative TF binding sites. Therefore, regions C and D may contain numerous regulatory elements relating to particular factors. Although region C is most important for expression it contains additional regulatory elements. Moreover, region D contains a putative nf2 site containing TGAG, reminiscent of motifs observed in certain SVs including at the end of exon 17 and the exon 3 (GTGAG) that are frequently skipped. Moreover, exon 3, exon 15, and exon 17 motifs contain CpG dinucleotides adjacent to frequently spliced regions suggesting methylation may participate in their regulation. Therefore, the frequency of these particular events was examined in relation to promoter methylation level of *hMLH1* region C and D. Exon 17 skipping correlates negatively with promoter methylation level overall and by region with the best correlation for region C (consistent with involvement in C1 inducible splicing) shown in

Figure 4.4. This suggests that methylation plays a role in E17 inclusion however a correlation does not imply a causal relationship. It is possible that promoter methylation correlates with other factors involved in E17 regulation. Still, E17 is an inducible exon and correlates well with promoter methylation. Therefore this SV warrants further investigation to determine whether promoter methylation is involved in alternative splicing of E17.

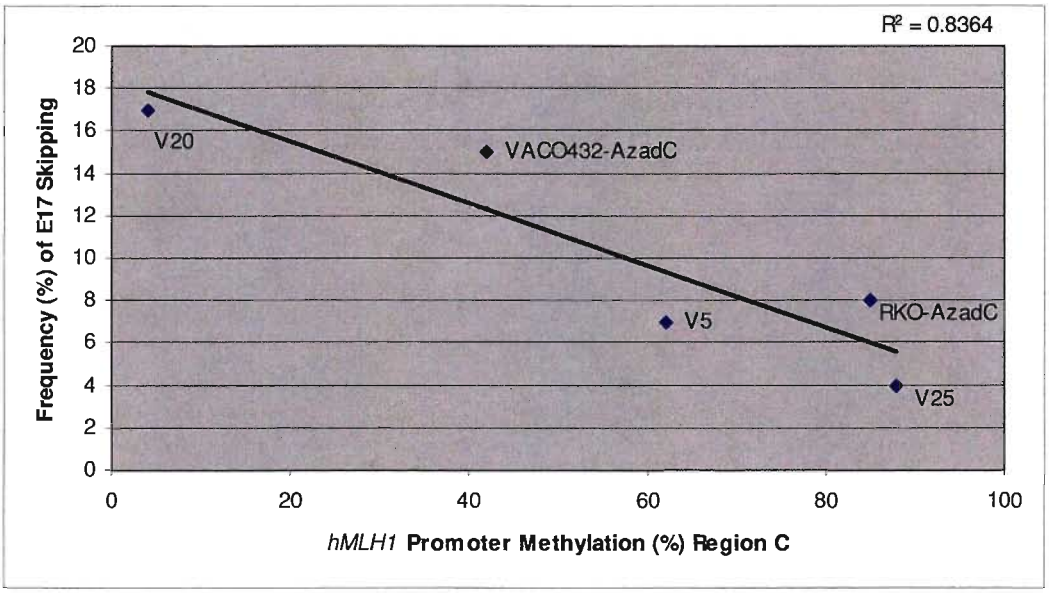


Figure 4.4: There appears to be a negative relationship between promoter methylation level and exon 17 skipping for region C.

Taken together, the results suggest that examining the role of promoter methylation in relation to *hMLH1* transcript diversity and frequency of specific splicing events is warranted. Future studies should also consider potential methylation effects in the body of the gene. In addition, RNA structure is known to contribute to splicing regulation though this area remains largely unexplored as it is methodologically difficult to study.

4.3.5. Sequence Motifs Related to *hMLH1* Splicing in Colon Cancer

In order to shed light on the mechanisms that may be contributing to the distinct *hMLH1* SV expression patterns observed in the colon cancer cell lines, the *hMLH1* sequence was searched for GTGAG motifs and the corresponding pairing motif, CTCAC – motifs of interest found in Δ^{95}E3-^{57}E15. Note that these sequences are essentially the 5'ss and branch point consensus sequences for U2 splicing but also correspond to hnRNP RRMs raising the tantalizing possibility of U1 independent splicing regulation. Search results reveal that GTGAG and the pairing motif CTCAC occur frequently throughout *hMLH1* but that specific repeats of this motif and pairing motif (shown in table 4.5) are located in introns adjacent to exons with distinct expression patterns in the colon cancer cell lines. While the significance of this finding is unknown at this point, it suggests that RNA structure may be mediating the resultant splicing outcomes and significant additional pairing was found around the motifs in some cases.

Table 4.5: GTGAG and CTCAC pairing motifs in *hMLH1* pre-mRNA sequence.

MOTIF 5' --> 3'	# FOUND	LOCATION	PAIRING MOTIF 5' --> 3'	# FOUND	LOCATION
GTGAG	78		CTCAC	75	
GTGAGCTGAGA	6	INTRONS 2,2,4,6,10,12	TCTCAGCTCAC	6	INTRONS 1,3,5,6,6,13
ATGAGCTGAGA	1	INTRON 15			

In addition, the repeat AGGT that occurs at the donor and acceptor of the ΔGTGAG-GTTATCG (^{95}E3-^{57}E15) transcript is of interest as it suggests this regulation maybe involved in dual splice site use and regulation. Moreover, cytosine methylation at 'CG' sites within exons may also be contributing to regulation for example by interfering (or promoting when absent) certain factor binding, probably involving hnRNPs; these possibilities await further study.

4.3.6. Sequence Motifs Related to *hMLH1* Splicing Found in *POLB*

The *POLB* pre-mRNA sequence was searched for the pairing motifs (shown in table 4.6) that were identified in *hMLH1*. The location of the motifs shows remarkable correspondence to adjacent exons that are spliced in the colon cancer cell lines, including intron retention events that are not observed in MRC5 and thus distinct to the colon cancer cell lines (possibly tissue specific or cancer related).

Table 4.6: GTGAG and CTCAC pairing motifs in *POLB* pre-mRNA sequence.

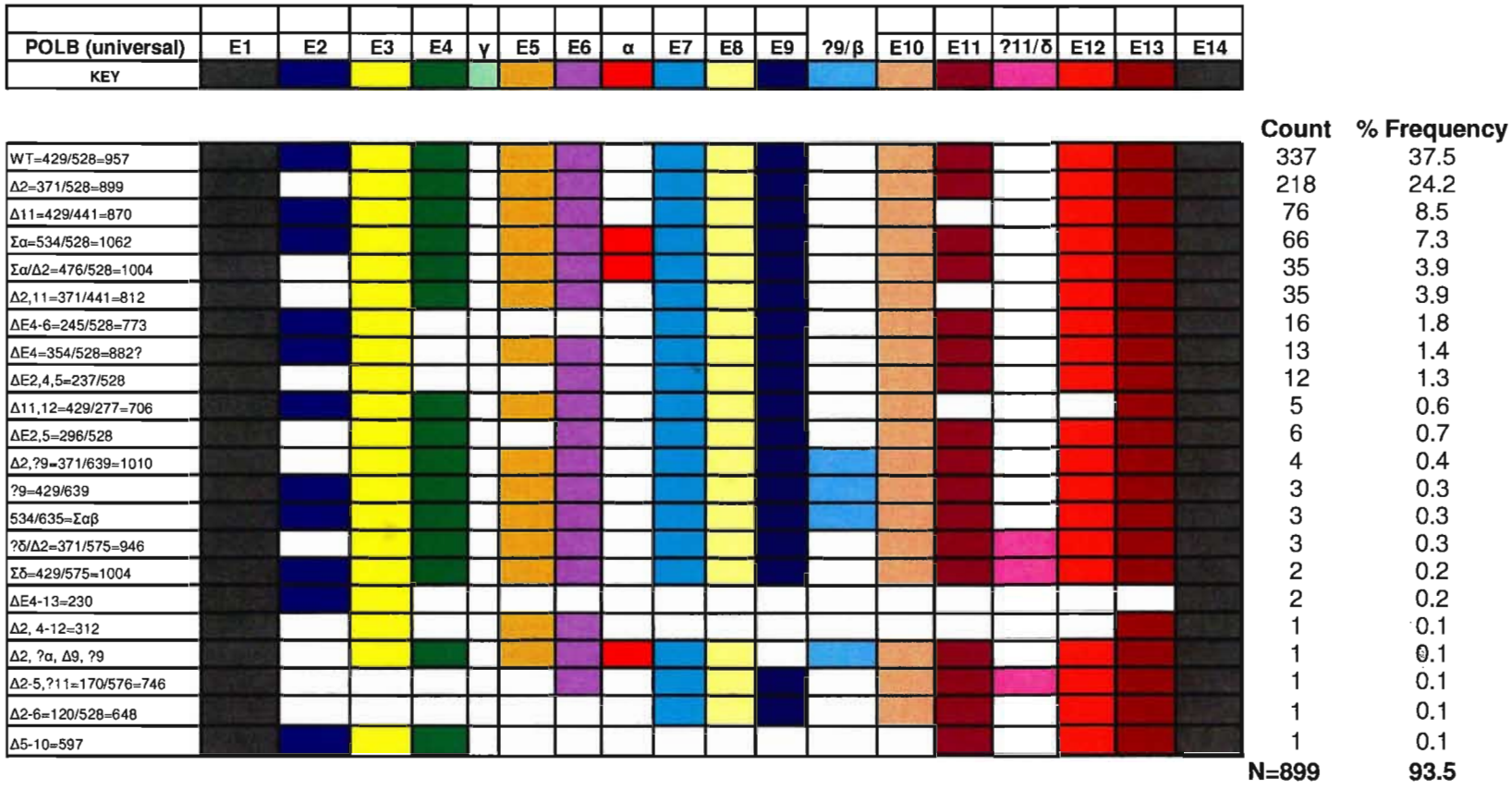
MOTIF 5'→ 3'	LOCATION	PAIRING MOTIF 5'→ 3'	LOCATION
caGTGAGCTGAGatc	intron 2	atCTCAGCTCACg	intron 9
caGTGAGCTGAGatc	intron 2	tcCTCAGCTCAAg	intron 11* (1st)
caGTGAGCTGAGat(t)	intron 11* (2nd)		

4.3.7. Location of Sequence Motifs in *POLB* Introns Adjacent to Frequently Spliced Exons in Colon Cancer

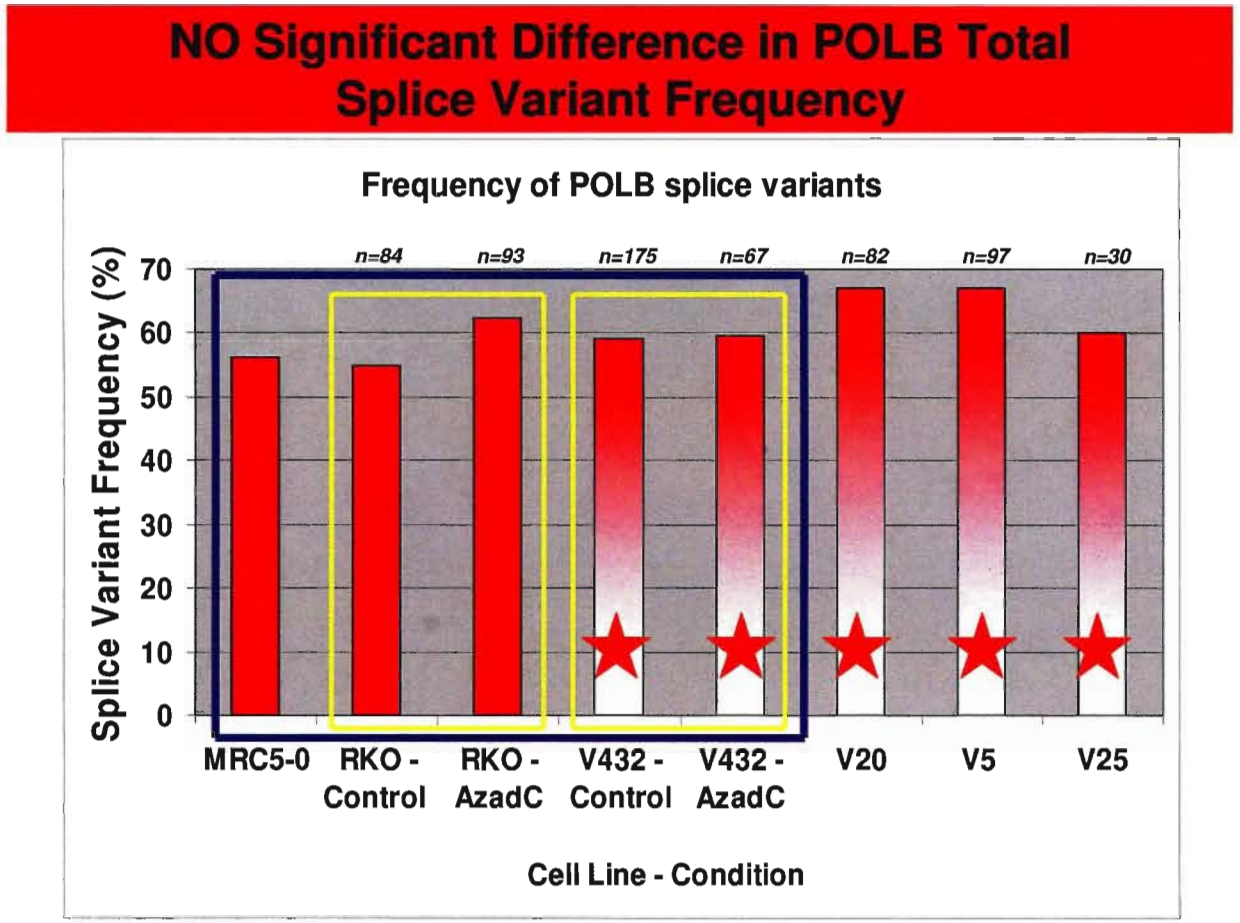
When the splicing profile of *POLB* was examined in the colon cancer cell lines before and after AzadC treatment, significant differences in E2 skipping and E11 skipping were observed and intron retention in I9 and I11 were seen that had not been observed in previous studies in MRC5. The frequencies of *POLB* mRNA transcript types identified in the colon cancer cell lines are presented in tabular form as Supplementary Table S4.0 (grey shading denotes definitive sequence confirmed identification has been obtained). A splicing key summarizing *POLB* splicing from all colon cancer cell lines combined is included as Supplementary Figure S4.0. A comparison of the WT frequency of POLB across cell lines is included as Supplementary Figure S4.1. A comparison graph of the main POLB SVs and/or exon skipping events of interest is included as Supplementary Figures S4.2-S4.3. While it is not clear from the results that these events represent global drug effects (changes in E2 and E11 skipping occur after oxidative stress in MRC5 and may be part of a general stress response), the position of the pairing motifs correspond to introns adjacent to frequently skipped exons for both *POLB* and *hMLH1* in the colon cancer cell lines examined and as such these motifs and how they contribute to splicing outcomes warrant further investigation.

Supplementary Table S4.0: Frequency (%) of *POLB* mRNA Types in Colon Cancer Cell Lines

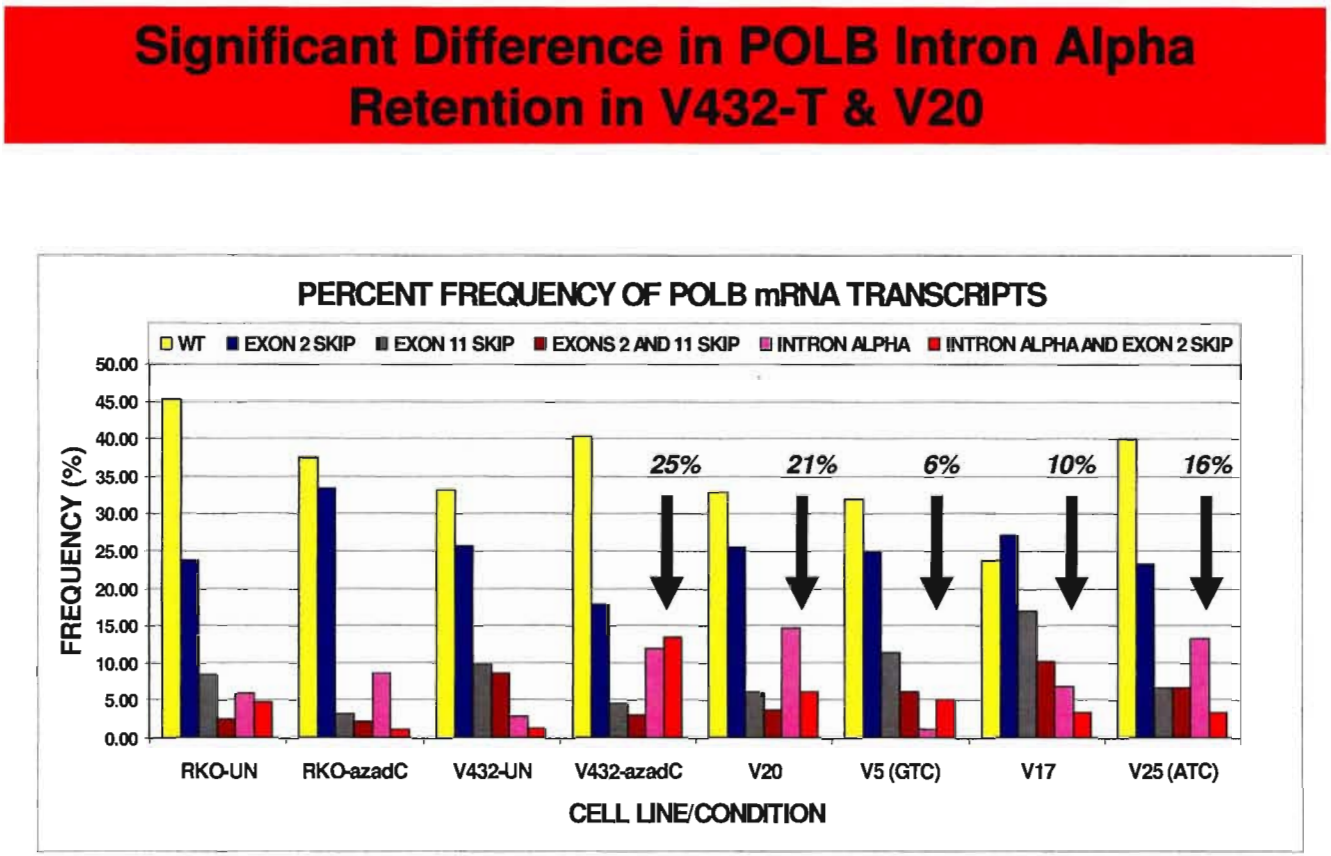
POLB: % Frequency universal primers	SW480 CONTROL	SW480 azadC	RKO CONTROL	RKO azadC	V432 CONTROL	V432 azadC	V20	V5	V25	V17
VARIANT	**N=112**	**N=122**	**N=84**	**N=93**	**N=153**	**N=67**	**N=82**	**N=97**	**N=30**	**N=59**
WT=429/528=957	45.5	43.4	45.2	37.6	32.0	40.3	32.9	32.0	40.0	23.7
Σα/Δ2=476/528=1004	3.6	2.5	4.8	1.1	0.7	13.4	6.1	5.2	3.3	3.4
Σα=534/528=1062	5.4	10.7	6.0	8.6	3.3	11.9	14.6	1.0	13.3	6.8
Σδ=429/575=1004	0.0	0.0	1.2	0.0	0.0	0.0	1.2	0.0	0.0	0.0
∑δ/Δ2=371/575=946	0.0	0.0	0.0	1.1	1.3	0.0	0.0	0.0	0.0	0.0
Δ2=371/528=899	17.0	23.8	23.8	33.3	25.5	17.9	25.6	24.7	23.3	27.1
Δ11=429/441=870	8.9	8.2	8.3	3.2	9.8	4.5	6.1	11.3	6.7	16.9
Δ2,11=371/441=812	0.0	0.0	2.4	2.2	7.8	3.0	3.7	6.2	6.7	10.2
ΔE4-6=245/528=773	0.0	0.0	4.8	3.2	0.0	0.0	1.2	5.2	3.3	3.4
ΔE4=354/528=882?	0.0	0.0	2.4	0.0	3.9	0.0	0.0	4.1	3.3	0.0
Δ2, 4-12=312	0.0	0.0	1.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
?220/560=780?	0.0	0.0	0.0	0.0	0.0	1.5	0.0	0.0	0.0	0.0
371/237?=608?	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.0	0.0	0.0
?187/441=628?	0.0	0.0	0.0	0.0	0.0	1.5	2.4	0.0	0.0	0.0
ΔE4-13=230	0.0	0.0	0.0	0.0	0.0	1.5	0.0	1.0	0.0	0.0
?429/370=799?	0.9	0.0	0.0	0.0	0.7	0.0	1.2	0.0	0.0	0.0
387/528=915?	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0
?800?	0.0	0.0	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0
534/635=Σαβ	0.9	0.0	0.0	0.0	0.7	0.0	1.2	0.0	0.0	0.0
?354/429or441?	0.9	0.8	0.0	0.0	0.0	0.0	1.2	0.0	0.0	0.0
?435? Uncut	0.9	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	1.7
??250/371??	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.7
?? 429 or 441 & 237? ??	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
?187/528?	0.0	0.0	0.0	0.0	0.7	0.0	0.0	2.1	0.0	0.0
ΔE2,5=296/528	1.8	0.8	0.0	0.0	1.3	0.0	0.0	1.0	0.0	0.0
?187?	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
?372? Uncut	0.9	0.0	0.0	0.0	0.0	0.0	0.0	3.1	0.0	1.7
?178/370?	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
200/150?	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
ΔE2,4,5=237/528	1.8	1.6	0.0	0.0	5.2	0.0	0.0	0.0	0.0	0.0
Δ2, ∑?	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
800 uncut	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
900 uncut	1.8	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
Δ2-6=120/528=648	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Δ5-10=597	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Δ2, ∑α, Δ9, ∑9	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
∑9=429/639	2.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
380/441	2.7	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Δ11,12=429/277=706	0.9	2.5	0.0	0.0	0.7	0.0	0.0	0.0	0.0	0.0
250?	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
429/420?	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Δ2,∑9=371/639=1010	0.0	0.8	0.0	0.0	2.0	0.0	0.0	0.0	0.0	0.0
Δ2-5,∑11=170/576=746	0.0	0.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
?237/323=560?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
?237/441=678?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
?902 OR wt?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
441/354?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
429/237?	0.0	0.0	0.0	1.1	0.7	0.0	0.0	0.0	0.0	3.4
400/528?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
371/323=694=e2, 12-13?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
?122/528=650?	0.0	0.0	0.0	1.1	0.0	0.0	0.0	0.0	0.0	0.0
429/354?	0.0	0.0	0.0	1.1	0.0	3.0	0.0	0.0	0.0	0.0



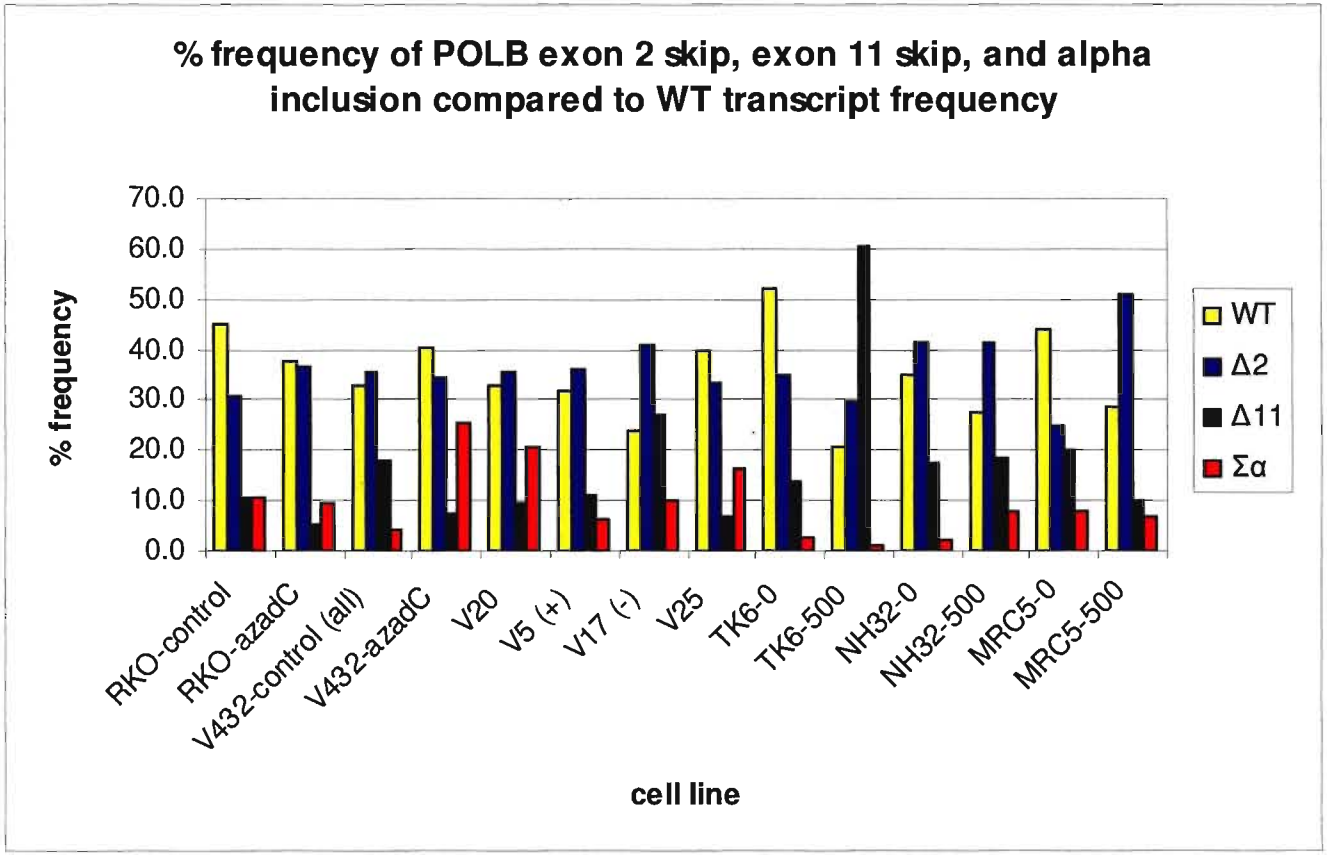
Supplementary Figure S4.0. *POLB* mRNA transcript types and frequencies identified in all colon cancer cell lines combined. Overall, 841/899 (93.5%) of *POLB* transcripts classified have been identified. Exon skipping and intron retention (with or without exon skipping) are observed and 13.1% of *POLB* transcripts classified in this study contain intron retention events.



Supplementary Figure S4.1: The *POLB* WT frequency is similar in MRC5-0 (baseline) and in the colon cancer cell lines both before and after treatment with AzadC.



Supplementary Figure S4.2: *POLB* exon alpha inclusion is up-regulated in both transiently and stably demethylated VACO cell lines but not in the untreated parental cell line suggesting AzadC treatment brought about a global change that altered the signaling pathway regulating the exon. Impairment of MapK signaling may result in an increase in exon alpha inclusion suggesting one possibility of AzadC treatment may be to restore MapK signaling (BRAF related?) in these cell lines. Exon alpha ends in the same TGCA(G) motif that appears to be associated with *hMLH1* splicing in response to MRC5-250HP and certain colon cancer C1 SVs.



Supplementary Figure S4.3: Comparison of *POLB* key splicing events (irrespective of transcript type of origin) across various conditions. TK6, NH32, and MRC5 data is from Disher and Skandalis (2007).

Project Conclusion

There are several interesting findings that emerge from this thesis that warrant further investigation. Firstly, characterization of the *hMLH1* splicing profile using full-length mRNA transcripts reveals extensive splicing in all cell lines, dual splice site use, and potential U2-independent splicing (a novel category of splice variants), although the authenticity of the splice variants requires verification. Secondly, the splicing profile of *hMLH1* is sensitive to oxidative stress. The pattern of changes appears to be driven, at least in part, by known splicing factors including SRp55, SRp40, SF2/ASF, TRA2-BETA and/or FOX-2 that may be regulating the other splicing factors. These possibilities warrant further investigation and the splicing patterns identified here can guide those investigations. Consistent with this idea, a key regulatory motif, TGCA, was identified in certain alternatively spliced *hMLH1* exons. The motif is also found in *POLB* exon alpha and preliminary findings suggest TGCA use may be regulated by common factors in *POLB* and *hMLH1*. *POLB* exon alpha is known to be up-regulated by MAPK knockdown implicating disfunction of this pathway in the colon cancer cell lines. This finding is consistent with *BRAF* mutations that characterize the SNPCC cell lines examined however the functional significance of the observed splicing events, if any, currently remains unknown and should be determined in future studies. Thirdly, no *hMLH1* intron retention was observed in MRC5, however exon border extension is observed in the colon cancer cell lines. Moreover, novel *POLB* intron retention events (rare) are observed in the colon cancer cell lines. These should be examined systematically to determine whether they are cancer specific and have any diagnostic potential. Fourthly, there appears to be a drug effect of AzadC on the splicing profile of transiently compared to stably demethylated colon cancer cell lines that warrants further investigation because of the clinical implications in cancer treatment. Finally, a curvilinear relationship was observed between *hMLH1* promoter methylation level and transcript diversity. Both high and low levels of methylation appear to be associated with low diversity, whereas intermediate levels of methylation result in higher diversity. There was no significant difference between the frequencies of the CT among the VACO

cell lines irrespective of transient or stable demethylation. However, RKO (transiently demethylated) exhibited a significantly higher frequency of CT compared to VACO cell lines pointing to cell line specific differences and underscoring the importance of comparing additional cell lines to confirm the results. Despite this limitation, when all cell lines were examined together, a negative correlation was observed between frequency of *hMLH1* exon 17 skipping and the level of methylation in the core promoter (region C). Region C has been shown to be important for silencing of *hMLH1* and contains several known transcription factor binding sites that have been shown to influence *hMLH1* expression. This raises the possibility that differential promoter occupation may be contributing to the observed splicing pattern and warrants further investigation.

References

Andersson, R., Enroth, S., Rada-Iglesias, A., et al. 2009. Nucleosomes are well positioned in exons and carry characteristic histone modifications. *Genome Research* **19:** 1732-1741.

Ang, P.W., Loh, M., Liem, N., et al. 2010. Comprehensive profiling of DNA methylation in colorectal cancer reveals subgroups with distinct clinicopathological and molecular features. *BMC Cancer* **10:** 227-234.

Auclair, J., Buisine, M.P., Navarro, C., et al. 2006. Systematic mRNA analysis for the effect of *MLH1* and *MSH2* missense and silent mutations on aberrant splicing. *Human Mutation* **27(2):** 145-154.

Baker, S.G., and Kramer, B. 2007. Paradoxes in carcinogenesis: New opportunities for research directions. *BMC Cancer* **7:** 151-156.

Barash, Y., Blencowe, B., and Frey, B. 2010. Model-based detection of alternative splicing signals. *Bioinformatics* **26:** 325-333.

Beard, W.A., and Wilson, S.H. 2000. Structural design of a eukaryotic DNA repair polymerase: DNA polymerase β. *Mutation Research* **460:** 231-244.

Beard, W.A., and Wilson, S.H. 1998. Structural insights into DNA polymerase β fidelity: Hold tight if you want it right. *Chemistry & Biology* **5(1):** R7-R13.

Bentley, D.L. 2005. Rules of engagement: co-transcriptional recruitment of pre-mRNA processing factors. *Current Opinion in Cell Biology* **17:** 251-256.

Berget, S.M. 1995. Exon recognition in vertebrate splicing. *J. Biol. Chem.* **270:** 2411-2414.

Bettstetter, M., Dechant, S., Ruemmele, P., et al. 2007. Distinction of hereditary nonpolyposis coloretal cancer and sporadic microsatellite-unstable colorectal cancer through quantification of MLH1 methylation by real-time PCR. *Clinical Cancer Research* **13:** 3221-3228.

Black, D.L. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annual Review of Biochemistry* **72:** 291-336.

Blasi, M.F., Ventura, I., Aquilina, G., et al. 2006. A human cell-based assay to evaluate the effects of alterations in the *MLH1* mismatch repair gene. *Cancer Research* **66(18):** 9036-9044.

Bronner CE, Baker SM, Morrison PT, et al. 1994. Mutation in the DNA mismatch repair gene homologue hMLH1 is associated with hereditary non-polyposis colon cancer. *Nature* **368(6468):** 258-261.

Burnette, J.M., Miyamoto-Sato, E., Schaub, M.A., Conklin, J., and Lopez, J.A. 2005. Subdivision of large introns in Drosophila by recursive splicing at nonexonic elements. *Genetics* **170:** 661-674.

Caceres, J.F., and Kornblihtt, A.R. 2002. Alternative splicing: multiple control mechanisms and involvement in human disease. *Trends in Genetics* **18(4):** 186-193.

Canadian Cancer Society's Steering Committee on Cancer Statistics. (2011, May). *Canadian Cancer Statistics 2011.* Toronto, ON: Canadian Cancer Society.

Cartegni, L., Wang, J., Zhu, Z., Zhang, M.Q., and Krainer, A.R. 2003. ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research* **31(13):** 3568-3571.

Cejka, P., Stojic, L., Mojas, N., et al. 2003. Methylation-induced G_2/M arrest requires a full complement of the mismatch repair protein hMLH1. *The EMBO Journal* **22(9):** 2245-2254.

Chandradas, S., Deikus, G., Tardos, J.G., Bogdanov, V.Y. 2009. Antagonistic roles of four SR proteins in the biosynthesis of alternatively spliced tissue factor transcripts in monocytic cells. *Journal of Leukocyte Biology* **87:** 147-152.

Charbonnier, F., Martin, C., Scotte, M., et al. 1995. Alternative splicing of *MLH1* messenger RNA in human normal cells. *Cancer Research* **55:** 1839-1841.

Chen, J., and Sadowski, I. 2005. Identification of the mismatch repair genes PMS2 and MLH1 as p53 target genes using serial analysis of binding elements. *PNAS* **102:** 4813-4818.

Christmann, M., Tomicic, M.T., Roos, W.P., and Kaina, B. 2003. Mechanisms of human DNA repair: an update. *Toxicology* **193:** 3-34.

Colussi, C., Parlanti, E., Degan, P., et al. 2002. The mammalian mismatch repair pathway removes DNA 8-oxoGMP incorporated from the oxidized dNTP pool. *Current Biology* **12:** 912-918.

Cramer, P., Pesce, C.G., Baralle, F.E., and Kornblihtt, A.R. 1997. Functional association between promoter structure and transcript alternative splicing. *PNAS* **94:** 11456-11460.

Cramer, P., Caceres, J.F., Cazalla, D., et al. 1999. Coupling of transcription with alternative splicing: RNA Pol II promoters modulate SF2/ASF and 9G8 effects on an exonic splicing enhancer. *Molecular Cell* **4:** 251-258.

Culler, S.J., Hoff, K.G., Voelker, R.B., et al. 2010. Functional selection and systematic analysis of intronic splicing elements identify active sequence motifs and associated splicing factors. *Nucleic Acids Research* **38(15):** 5152-5165.

Cunningham, J.M., Christensen, E.R., Tester, D.J., et al. 1998. Hypermethylation of the hMLH1 promoter in colon cancer with microsatellite instability. *Cancer Research* **58:** 3455-3460.

Damianov, A., and Black, L. 2010. Autoregulation of Fox protein expression to produce dominant negative splicing factors. *RNA* **16:** 405-416.

Daniov, G.L., Prasad, R., Wilson, S.H., Bohr, V.A. 1999. Role of DNA polymerase β in the excision step long patch mammalian base excision repair. *Journal of Biological Chemistry* **274:** 13741-13743.

de la Mata, M., Alonso, C.R., Kadener, S., et al. 2003. A slow RNA polymerase II affects alternative splicing *in vivo*. *Molecular Cell* **12(2):** 525-532.

Deng, G., Bell, I., Crawley, S., et al. 2004a. BRAF mutation is frequently present in sporadic colorectal cancer with methylated *hMLH1*, but not in hereditary nonpolyposis colorectal cancer. *Clinical Cancer Research* **10:** 191-195.

Deng, G., Song, G., Pong, E., et al. 2004b. Promoter methylation inhibits *APC* gene expression by causing changes in chromatin conformation and interfering with the binding of transcription factor CCAAT-binding factor. *Cancer Research* **64:** 2692-2698.

Deng, G., Peng, E., Gum, J., et al. 2002. Methylation of hMLH1 promoter correlates with the gene silencing with a region-specific manner in colorectal cancer. *British Journal of Cancer* **86:** 574-579.

Deng, G., Chen, A., Pong, E., and Kim, Y.S. 2001. Methylation in hMLH1 promoter interferes with its binding to transcription factor CBF and inhibits gene expression. *Oncogene* **20:** 7120-7127.

Deng, G., Chen, A., Hong, J., et al. 1999. Methylation of CpG in a small region of the hMLH1 promtoer invariably correlates with the absence of gene expression. *Cancer Research* **59:** 2029-2033.

Desmet, F.O., Hamroun, D., Lalande, M., Collod-Beroud, G., Claustres, M., and Beroud, C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals.

Nucleic Acids Research **37(9):** e67 doi:10.1093/nar/gkp215 (Freeware available at: www.umd.be/HSF/)

Disher, K., and Skandalis, A. 2007. Evidence of the modulation of mRNA splicing fidelity in humans by oxidative stress and p53. *Genome* **50(10):** 946-953.

Dreumont, N., Bourgeois, C.F., Lejeune, F., et al. 2010. Human RBMY regulates germline-specific splicing events by modulating the function of serine/arginine rich proteins 9G8 and TRA2-β. *Journal of Cell Science* **123:** 40-50.

Drummond, J.T., Anthoney, A., Brown, R., and Modrich, P. 1996. Cisplatin and adriamycin resistance are associated with MutLalpha and mismatch repair deficiency in an ovarian tumor cell line. *Journal of Biological Chemistry* **271:** 19645-19648.

Ehrlich, M., Turner, J., Gibbs, P., et al. 2008. Cytosine methylation profiling of cancer cell lines. *PNAS* **105(12):** 4844-4849.

Eshelman, J.R., Lang, E.Z., Bowerfind, G.K., et al. 1995. Increased mutation rate at the hprt locus accompanies microsatellite instability in colon cancer. *Oncogene* **10(1):** 33-37.

Fackental, J.D., and Godley, L.A. 2008. Aberrant RNA splicing and its functional consequences in cancer cells. *Disease Models & Mechanisms* **1:** 37-42.

Fededa, J.P., Petrillo, E., Gelfand, M.S., et al. 2005. A polar mechanism coordinates different regions of alternative splicing within a single gene. *Molecular Cell* **19:** 393-404.
Filippov, V., Filippova, M., Duerksen-Hughes, P.J. 2007. The early response to DNA damage can lead to activation of alternative splicing activity resulting in CD44 splicing changes. *Cancer Research* **67(16):** 7621-7630.

Filippov, V., Schmidt, E.L., Filippova, M., Duerksen-Hughes, P.J., 2008. Splicing and splice factor SRp55 participate in the response to DNA damage by changing isoform ratios of target genes. *Gene* **420(1):** 34-41.

Fink, S., and Cookson, B. 2005. Apoptosis, pyroptosis, and necrosis: mechanistic description of dead and dying eukaryotic cells. *Infection and Immunity* **73(4):** 1907-1916.

Fink, D., Zheng, H., Nebel, S., et al. 1997. *In vitro* and *in vivo* resistance to cisplatin in cells that have lost DNA mismatch repair. *Cancer Research* **57:** 1841-1845.

Fink, D., Nebel, S., Norris, P.S., et al. 1998. Enrichment for DNA mismatch repair-deficient cells during treatment with cisplatin. *International Journal of Cancer* **77:** 741-746.

Fortini, P., Pascucci, B., Parlanti, E., Sobol, R.W., Wilson, S.H. 1998. Different DNA polymerases are involved in the short- and long-patch base excision repair in mammalian cells. *Biochemistry* **37:** 3575-3580.

Friedman, B.A., Stadler, M.B., Shomron, N., Ding, Y., Burge, C.B. 2008. Ab initio identification of functionally interacting pairs of cis-regulatory elements. *Genome Resesarch* **18:** 1643-1651.

Frosina, G., Fortini, P., Rossi, O., et al. 1996. Two pathways for base excision repair in mammaliean cells. *Journal of Biological Chemistry* **271:** 9573-9578.

Fukumura, K., et al. 2009. U1-independent pre-mRNA splicing contributes to the regulation of alternative splicing. *Nucleic Acids Research* **37(6):** 1907-1914.

Galluzzi, L., and Kroemer, G. 2008. Necroptosis: A specialized pathway of programmed necrosis. *Cell* **135:** 1161-1163.

Gardina, P.J., Clark, T.A., Shimada, B., et al. 2006. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7:**325.

Geiszt, M., Lekstrom, K., Leto, T.L. 2004. Analysis of mRNA transcripts from the NAD(P)H oxidase 1 (*Nox1*) Gene. *Journal of Biological Chemistry* **279(49):** 516661-516668.

Genuardi, M., Anti, M., Capozzi, E., et al. 1998. *MLH1* and *MSH2* constitutional mutations in colorectal cancer families not meeting the standard criteria for hereditary nonpolyposis colorectal cancer. *International Journal of Cancer* **75:** 835-839.

Ghazi, S., von Holst, S., Picelli, S., et al. 2010. Colorectal cancer susceptibility loci in a population-based study. *American Journal of Pathology* **177(6):** 2689.

Ghigna, C., Valacca, C., and Biamonti, G. 2008. Alternative splicing and tumor progression. *Current Genomics* **9:** 556-570.

Glaab, W.E., Tindall, K.R., Skopek, T.R. 1999. Specificity of mutations induced by methyl methanesulfonate in mismatch repair-deficient human cell lines. *Mutation Research* **427(2):** 67-78.

Glaab, W.E., Risinger, J.I., Umar, A., et al. 1998a. Resistance to 6-thioguanine in mismatch repair-deficient human cancer cell lines correlates with an increase in induced mutations at the HPRT locus. *Carcinogenesis* **19:** 1931-1937.

Glaab, W.E., Risinger, J.I., Umar, A., et al. 1998b. Cellular resistance and hypermutability in mismatch repair-deficient human cancer cell lines following treatment with methyl methanesulfonate. *Mutation Research* **398(1-2):** 197-207.

Hagemann, S., Heil, O., Lyko, F., and Brueckner, B. 2011. Azacytidine and decitabine induce gene-specific and non-random DNA demethylation in human cancer cell lines. *PLoS ONE* **6(3):** e17388.

Hardmann, R.A., Afshari, C.A., and Barret, J.C. 2001. Involvement of mammalian MLH1 in the apoptotic response to peroxide-induced oxidative stress. *Cancer Research* **61:** 1392-1397.

Helleman, J., van Staveren, I.L., Dinjens, W.N.M., et al. 2006. Mismatch repair and treatment resistance in ovarian cancer. *BMC Cancer* **6:** 201-210.

Herman, J.G., Umar, A., Polyak, K., et al. 1998. Incidence and functional consequences of MLH1 promoter hypermethylation in colorectal carcinoma. *PNAS* **95:** 6870-6875.

Hitchens, M.P., and Ward, R.L. 2009. Constitutional (germline) MLH1 epimutation as an aetiological mechanism for hereditary non-polyposis colorectal cancer. *Journal of Medical Genetics* **46:** 793-802.

Hodges, C., Bintu, L., Lubkaowska, L., Kashlev, M., and Bustamante, C. 2009. Nucleorsomal fluctuations govern the transcription dynamics of RNA polymerase II. *Science* **325:** 626-628.

Horowitz, D.S. and Krainer, A.R. 1994. Mechanisms for selecting 5' splice sites in mammalian pre-mRNA splicing. *Trends in Genetics* **10:** 100-106.

Hui, L., Zhang, X., Wu, X., Lin, Z., Wang, Q., Li, Y., and Hu, G. 2004. Identification of alternatively spliced mRNA variants related to cancers by genome-wide ESTs alignment. *Oncogene* **23:** 3013-3023.

Johnson et al. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. *Science* **302:** 2141-2144.

Jurica, M.S., and Moore, M.J. 2003. Pre-mRNA splicing: awash in a sea of proteins. *Molecular Cell* **12:** 5-14, July, 2003.

Kadener, S., Cramer, P., Nogues, G., et al. 2001. Antagonistic effects of T-Ag and VP16 reveal a role for RNA pol II elongation on alternative splicing. *EMBO Journal* **20:** 5759-5768.

Kadyrov, F.A., Dzantiev, L., Constantin, N., and Modrich, P. 2006. Endonucleolytic function of MutLα in human mismatch repair. *Cell* **126:** 297-308.

Kalnina, Z., Zayakin, P., Silina, K., and Line, A. 2005. Alterations of pre-mRNA splicing in cancer. *Genes, Chromosomes & Cancer* **42:** 342-357.
Karni, R., de Stanchina, E., Lowe, S.W., Sinha, R., Mu, D., and Krainer, A.R. 2007. The gene encoding the splicing factor SF2/ASF is a proto-oncogene. *Nature Structural & Molecular Biology* **14:** 185-193.

Ke, S., and Chasin, L.A. 2010. Intronic motif pairs cooperate across exons to promote pre-mRNA splicing. *Genome Biology* **11(8):** R84 doi:10.1186/gb-2010-11-8-r84

Kim, E., Magen, A., and Ast, G. 2007. Different levels of alternative splicing among eukaryotes. *Nucleic Acids Research* **35:** 125-131.

Kim, H., Jen, J., Vogelstein, B., and Hamilton, S.R. 1994. Clinical and pathological characteristics of sporadic colorectal carcinomas with DNA replication errors in microsatellite sequences. *American Journal of Pathology* **145(1):** 148-156.

Kireeva, M.L., Hancock, B., Cremona, G.H., et al. 2005. Nature of the nucleosomal barrier to RNA polymerase II. *Molecular Cell* **18:** 97-108.

Kolasinska-Zwierz, P., Down, T., Latorre, I., et al. 2009. Differential chromatin marking of introns and expressed exons by H2K36me3. *Nature Genetics* **41:** 376-381.

Kondo, E., Horii, A., and Fukushige, S. 2001. The interacting domains of three MutL heterodimers in man: hMLH1 interacts with 39 homologous amino acid residues within hMLH3, hPMS1, and hPMS2. *Nucelic Acids Research* **29(8):** 1695-1702.

Kornblihtt, A.R., de la Mata, M., Fededa, J.P., Muñoz, M.J., and Nogués, G. 2004. Multiple links between transcription and splicing. *RNA* **10:** 1489-1498.

Kornblihtt, A.R. 2007. Coupling transcription and alternative splicing. *Advances in Experimental Medicine and Biology* **623:** 175-189.

Kornblihtt, A.R., Schor, I.E., Allo, M., and Blencowe, B.J. 2009. When chromatin meets splicing. *Nature Structural and Molecular Biology* **16(9):** 902-903.

Kosinski, J., Hinrichsen, I., Bujnicki, J.M., Friehoff, P., and Plotz, G. 2010. Identification of Lynch Syndrome mutations in the MLH1-PMS2 interface that disturb dimerization and mismatch repair. *Human Mutation* **31(8):** 975-982.

Kramer, A. 1996. The structure and function of proteins involved in mammalian pre-mRNA splicing. *Annual Review of Biochemistry* **65:** 367-409.

Kültz, D. 2005. Molecular and evolutionary basis of the cellular stress response. *Annual Review of Physiology* **67:** 225-257.

Kunkel, T.A., and Erie, D.A. 2005. DNA mismatch repair. *Annual Review of Biochemistry* **74:** 681-710.

Kuroyanagi, H. 2009. Fox-1 family of RNA-binding proteins. *Cellular and Molecular Life Sciences.* **66:** 3895-3907.

Kyng, K.J., and Bohr, V.A. 2005. Gene expression and DNA repair in progeroid syndromes and human aging. *Ageing Research Reviews* **4:** 579-602.
Ladd, A.N., and Cooper, T.A. 2002. Finding signals that regulate alternative splicing in the post-genomic era. *Genome Biology* **3(11):** reviews0008.1-0008.16.

Levenson, J.M., and Sweatt, J.D. 2005. Epigenetic mechanisms in memory formation. *Nature Reviews Neuroscience* **6(2):** 108-118.

Li, G.M. 2008. Mechanisms and functions of DNA mismatch repair. *Cell Research* **18:** 85-98.

Li, X., Wang, J., and Manley, J.L. 2005. Loss of splicing factor SF2/ASF induces G2 cell cycle arrest and apoptosis, but inhibits internucleosomal DNA fragmentation. *Genes & Development* **19:** 2705-2714.

Lin, C.F., Mount, S.M., Jarmolowski, A., Makalowski, W. 2010. Evolutionary dynamics of U12-type spliceosomal introns. *BMC Evolutionary Biology* **10:** 47-56.

Lin, J.C., Jeong, S., Liang, G., et al. 2007. Role of nucleosomal occupancy in the epigenetic silencing of the *MLH1* CpG island. *Cancer Cell* **12:** 432-444.

Lindahl, T. 1993. Instability and decay of the primary structure of DNA. *Nature* **362:** 709-715.

Loeb, K.R., and Loeb, L.A. 2000. Significance of multiple mutations in cancer. *Carcinogenesis* **21(3):** 379-385.

Long, J.C., and Caceres, J.F. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochemical Journal.* **417:** 15-27.

Loughrey, M.B., Waring, P.M., Trivett, T.A., et al. 2007. Incorporation of somatic BRAF mutation testing into an algorithm for the investigation of hereditary non-polyposis colorectal cancer. *Familial Cancer* **6(3):** 301-310

Luger, K., Mader, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. 1997. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389:** 251-260.

Luo, Y., Lin, F., and Lin, W. 2004. ATM-mediated stabilization of hMutL DNA mismatch repair proteins augments p53 activation during DNA damage. *Molecular and Cellular Biology* **24(14):** 6430-6444.

Lyko, F., and Brown, R. 2005. DNA methyltransferase inhibitors and the development of epigenetic cancer therapies. *Journal of the National Cancer Institute* **97(20):** 1498-1506.

Lynch, H.T., and de la Chapelle, A. 1999. Genetic susceptibility to non-polyposis colorectal cancer. *Journal of Medical Genetics* **36:** 801-818.

Martin, S.A., McCabe, N., Mullarkey, M., et al. 2010. DNA Polymerases as potential therapeutic targets for cancers deficient in the DNA mismatch repair proteins MSH2 or MLH1. *Cancer Cell* **17(3-3):** 235-248.

Martinez-Conteras, R., Fisette, J., Nasim, F.H., et al. 2006. Intronic binding sites for hnRNP A/B hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS Biology* **4(2):** e21 doi:10.1371/journal.pbio.0040021.

Mastrocola, A.S., and Heinen, C.D. 2010. Nuclear reorganization of DNA mismatch repair proteins in response to DNA damage. *DNA Repair (Amst)* **9(2):** 120-133.

Mendillo, M.L., Hargreaves, V.V., Jamison, J.W., et al. 2010. A conserved MutS homolog connector domain interface interacts with MutL homologs. *PNAS* **106(52):** 22223-22228.

Merdzhanova, G., Edmond, V., De Seranno, S., et al. 2008. E2F1 controls alternative splicing pattern of genes involved in apoptosis through upregulation of the splicing factor SC35. *Cell Death & Differentiation* **15:** 1815-1823.

Michor, F., Iwases, J., and Nowak, A. 2004. Dynamics of cancer progression. *Nature Reviews Cancer* **4:** 197-205.

Mihaylova, V.T., Bindra, R.S., Yuan, J., et al. 2003. Decreased expression of DNA mismatch repair protein gene Mlh1 under hypoxic stress in mammalian cells. *Molecular and Cellular Biology* **23(9):** 3265-3273.

Morak, M., Schackert, H.K., Rahner, N., et al. 2008. Further evidence for heritability of an epimutation in one of 12 cases with *MLH1* promoter methylation in blood cells clinically displaying HNPCC. *European Journal of Human Genetics* **16:** 804-811.

Mossman, D., Kim, K., and Scott, R.J. 2010. Demethylation by 5-aza-2'deoxyctytidine in colorectal cancer cells targets genomic DNA whilst promoter CpG island methylation persists. *BMC Cancer* **10:** 366-375.

Nasim, M.T., Chernova, T.K., Chowdhury, H.M., Yue, B., and Eperon, I.C. 2003. hnRNP G and Tra2β: opposite effects on splicing matched by antagonism in RNA binding. *Human Molecular Genetics* **12(11):** 1337-1348.

National Cancer Institute at the National Institutes of Health. 2011. Colon and Rectal Cancer. <http://www.cancer.gov/cancer topics/types/colon-and-rectal> and <http://www.cancer.gov/clinicaltrials/search/results?protocolsearchid=6317779&vers=1> (last accessed August, 12, 2011)

Nilsen, H., and Krokan, H.E. 2001. Base excision repair in a network of defence and tolerance. *Carcinogenesis* **22(7):** 987-998.

Nogués, G., Kadener, S., Cramer, P., et al. 2002. Transcriptional activators differ in their abilities to control alternative splicing. *Journal of Biological Chemistry* **277:** 43110-43114.

Novoyatleva, T., Heinrich, B., Tang, Y., Benderska, N., and Butchbach, M.E. 2008. Protein phosphatase 1 binds to the RNA recognition motif of several splicing factors and regulates alternative pre-mRNA processing. *Human Molecular Genetics* **17(1):** 52-70.

O'Brien, V., and Brown, R. 2006. Signalling cell cycle arrest and cell death through the MMR system. *Carcinogenesis* **27(4):** 682-692.

Olmos, Y. Brosens, J.J., and Lam, E.W.-F. 2011. Interplay between SIRT proteins and tumor suppressor transcription factors in chemotherapeutic resistance of cancer. *Drug Resistance Updates* **14:** 35-44.

Pagenstecher, C., Ihner, M., Friedl, W., et al. 2006. Aberrant splicing in *MLH1* and *MSH2* due to exonic and intronic variants. *Human Genetics* **119:** 9-22.

Palmirotta, R., Verì, M.C., Curia, M.C., et al. 1998. Transcripts with splicings of exons 15 and 16 of the hMLH1 gene in normal lymphocytes: implications in RNA-based mutation screening of hereditary non-polyposis colon cancer. *European Journal of Cancer* **34(6):** 927-930.

Papadopoulos, N., Nicolaides, N.C., Wei, Y.F., et al. 1994. Mutation of a mutL homolog in hereditary colon cancer. *Science* **263:** 1625-1629.

Papouli, E., Cejka, P., and Jiricny, J. 2004. Dependence of the cytotoxicity of DNA-damaging agents on the mismatch repair status of human cells. *Cancer Research* **64:** 3391-3394.

Patel, A.A., and Steitz, J.A. 2003. Splicing double: insights from the second spliceosome. *Nature Reviews Molecular Cell Biology* **4:** 960-970.

Peasland, A., Matheson, E., Hall, A., and Irving, J. 2010. Alternative splicing of hMLH1 in childhood acute lymphoblastic leukaemia and characterization of the Δ9/10 isoform as a dominant negative species. *Leukemia Research* **34(3):** 322-327.

Pitsikas, P., Lee, D., and Rainbow, A.J. 2007. Reduced host cell reactivation of oxidative DNA damage in human cells deficient in the mismatch repair gene hMSH2. *Mutagenesis* **22(3):** 235-243.

Ponta, H., Sherman, L., and Herrlich, P.A. CD44: from adhesion molecules to signaling regulators. *Nature Reviews Molecular Cell Biology* **4:** 33-45.

Ponthier, J.L., Schluepen, C., Chen, W., et al. 2006. Fox-2 splicing factor binds to a conserved intron motif to promote inclusion of protein 4.1R alternative exon 16. *Journal of Biological Chemistry* **281:** 12468-12474.

Prasad, R., Dianov, G.L., Bohr, V.A., and Wilson, S.H. 2000. FEN1 stimulation of DNA polymerase β mediates an excision step in mammalian long patch base excision repair. *Journal of Biological Chemistry* **275:** 4460-4466.

Quan, L., Stassen, A.P.M., Ruivenkamp, C.A.L., et al. 2011. Most lung and colon cancer susceptibility genes are pair-wise linked in mice, humans and rats. *PLoS ONE* **6(2):** e14727.

Rajendran, R., Garva, R., Krstic-Demonacos, M., and Demonacos, C. 2011. Sirtuins: molocular traffic lights in the crossroad of oxidative stress, chromatin remodeling and transcription. *Journal of Biomedicine and Biotechnology* (accepted May 30, 2011, provisional PDF available at: http://www.hindawi.com/journals/jbb/aip/368276/).

Robson-Dixon, N.D., and Garcia-Blanco, M.A. 2004. MAZ elements alter transcription elongation and silencing of the fibroblast growth factor receptor 2 exon IIIb. *Journal of Biological Chemistry* **279(8):** 29075-29084.

Roca, X., Sachidanandam R., and Krainer, A.R. 2003. Intrinsic differences between authentic and cryptic 5' splice sites. *Nucleic Acids Research* **31:** 6321-6333.

Roca, X., Sachidanandam R., and Krainer, A.R. 2005. Determinants of the inherent strength of human 5' splice sites. *RNA* **11:**683-698.

Roche Applied Science. 2011. Apoptosis, Cell Death, and Cell Proliferation Manual. <http://www.roche-applied-science.com/PROD_INF/MANUALS/CELL_MAN/apoptosis_003_004.pdf> (accessed August 15, 2011)

Russo, M.T., Blasi, M.G., Chiera, F., et al. 2004. The oxidized deoxynucleoside triphosphate pool is a significant contributor to genetic instbility in mismatch reapir-deficient cells. *Molecular Cell Biology* **24:** 465-74.

Schwartz, S., Meshorer, E., and Ast, G., 2009. Chromatin organization marks exon-intron structure. *Nature Structural & Molecular Biology* **16:**990-995.

Schwerk, C., and Schulze-Osthoff, K. 2005. Regulation of apoptosis by alternative pre-mRNA splicing. *Molecular Cell* **9:** 1-13.

Senapathy, P., Shapiro, M.B., and Harris, N.L. 1990. Splice junctions, branch point sites, and exons: Sequence statistics, identification, and applications to genome project. *Methods in Enzymology* **183:** 253-278.

Shapiro, M.B. and Senapathy, P. 1987. RNA splice junctions of different classes of eukaryotes: Sequence statistics and functional implications in gene expression. *Nucleic Acids Research* **15:** 7155-7174.

Shepard, P.J., and Hertel, K.J. 2009. The SR protein family. *Genome Biology* **10:** 242.

Shukla, S., and Fisher, A. 2008. Tra2β as a novel mediator of vascular smooth muscle diversification. *Circulation Research* **103:** 485-492.

Simmons, S.O., Fan, C., and Ramabhadran, R. 2009. Review: Cellular stress response pathway system as a sentinel ensemble in toxicological screening. *Toxicological Sciences* **111(2):** 202-225.

Skandalis, A. and Uribe, E. 2004. A survey of splice variants of the human hypoxanthine phosphoribosyl transferase and DNA polymerase beta genes: products of alternative or aberrant splicing? *Nucleic Acids Research* **32(22):** 6557-6564.

Sorek, R., Lev-Maor, G., Reznik, M., Dagan, T., Belinky, F., Graur, D., and Ast, G. 2004. Minimal conditions for exonization of intronic sequences: 5' splice site formation in Alu exons. *Molecular Cell* **14:** 221-231.

Stamm, S., Ben-Ari, S., Rafalska, I., et al. 2005. Function of alternative splicing. *Gene* **344:** 1-20.

Stoilov et al. 2004. Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA. *Human Molecular Genetics* **13(5):** 509-524.

Sugnet, C.W., Kent, W.J., Ares Jr., M., and Haussler, D. 2004. Transcriptome and genome conservation of alternative splicing events in humans and mice. *Pacific Symposium on Biocomputing* **9:** 66-77.

Takeo, K., Kawai, T., Nishida, K., et al. 2009. Oxidative stress-induced alternative splicing of transformer 2β (*SFRS10*) and *CD44* pre-mRNAs in gastric epithelial cells. *AJP – Cell Physiology* **297:** C330-C338.

Thanaraj, T.A., and Clark, F. 2001. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Research* **29(12):** 2581-2593.

Thannickal, V.J., and Fanburg, B.L. 2000. Reactive oxygen species in cell signaling. *AJP – Lung Cellular and Molecular Physiology* **279:** L1005-L1028.

Tilger, H., Nikolaou, C., Althammer, S., et al. 2009. Nucleosome positioning as a determinant of exon recognition. *Nature Structural & Molecular Biology* **16:** 996-1001.

Tournier, I., Venzain, M., Martins, A., et al. 2008. A large fraction of unclassified variants of the mismatch repair genes *MLH1* and *MSH2* is associated with splicing defects. *Human Mutation* **29(12):** 1412-1424.

Tsuda, K., Someya, T., Kuwasako, K., et al. 2010. Structural basis for the dual RNA-recognition modes of human TRA2-β RRM. *Nucleic Acids Research* **1-16:** doi: 0.1093.

Tsukamoto, Y., Matsuo, N., Ozawa, K., et al. 2001. Expression of a novel RNA-splicing factor, RA301/Tra2β, in vascular lesions and its role in smooth muscle cell proliferation. *American Journal of Pathology* **158:(5):** 1685-1694.

Veigl, M.L., Kasturi, L., Olechnowicz, J., et al. 1998. Biallelic inactivation of hMLH1 by epigenetic gene silencing, a novel mechanism causing human MSI cancers. *PNAS* **95:** 8698-8702.

Venables, J.P., Bourgeois, C.F., Dalgliesh, C., et al. 2005. Up-regulation of the ubiquitous alternative splicing factor Tra2β causes inclusion of a germ cell-specific exon. *Human Molecular Genetics* **14(16):** 2289-2303.

Wan, P.T.C., Garnett, M.J., Roe, S.M., et al. 2004. Mechanism of activation of the RAF-ERK signaling pathway by oncogenic mutations of B-RAF. *Cell* **116:** 855-867.

Wang, B., Xiao, Z., and Ren, E.C. 2009. Redefining the p53 response element. *PNAS* **106(34):** 14373-14378.

Wang, H.C., Su, Y.R., Han, K.J., et al. 2004. Multiple variants and a differential splicing pattern of kinectin in human hepatocellular carcinoma. *Biochemistry and Cell Biology* **82:** 321-327.

Wang, J., Smith, P.J., Krainer, A.R., Zhang, M.Q. 2005. Distribution of SR protein exonic splicing motifs in human protein-coding genes. *Nucleic Acids Research* **33(16):** 5053-5062.

Wang, X., Wang, K., Radovich, M., et al. 2009. Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. *BMC Genomics* **10(Suppl 1):** S4. (doi: 10.1186/1471-2164-10-S1-S4)

Wang, Z., Lo, H.S., Yang, H., et al. 2003. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. *Cancer Research* **63:** 655-657.

Warnick, C.T., Dabbas, B., Ilstrup, S.J., Ford, C.D., Strait, K.A. Cell type-dependent regulation of *hMLH1* promoter activity is influenced by the presence of multiple redundant evidence. *Molecular Cancer Research* **1(8):** 610-618.

Watermann, D.O., Tang, Y., zur Hausen, A., et al. 2006. Splicing factor TRA2-BETA1 is specifically induced in breast cancer and regulates alternative splicing of the *CD44* gene. *Cancer Research* **66:(9):** 4774-4780.

Wilson, S.H. 1998. Mammalian base excision repair and DNA polymerase β. *Mutation Research* **407(3):** 203-215.

World Health Organization. 2011. Cancer Fact Sheet N°297 February 2011. <http://www.who.int/mediacentre/factsheets/fs297/en/> (accessed August, 15, 2011)

Wu, Q., and Krainer, A.R. 1999. AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Molecular and Cellular Biology* **19(5):** 3225-3226.

Xiao, X., Wang, Z., Jang, M., Burge, C.B. 2007. Coevolutionary networks of splicing cis-regulatory elements. *PNAS* **104(47):** 18583-18588.

Xie, H., Zhu, W.Y., Wasserman, A., Grebinskiy, V., Olson, A., and Mintz, L. 2002. Computational analysis of alternative splicing using EST tissue information. *Genomics* **80:** 326-330.

Xu, D. and Mattox, M. 2006. Identification of a splicing enhancer in MLH1 using COMPARE a new assay for determination of relative RNA splicing efficiencies. *Human Molecular Genetics* **15(2):** 329-336.

Xu, Q., Modrek, B., and Lee, C. 2002. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research* **30:** 3754-3766.

Xu, Q. and Lee, C. 2003. Discovery of novel splice forms and functional analysis of cancer-specific alternative splicing in human expressed sequences. *Nucleic Acids Research* **31:** 5635-5643.

Yan, J., and Marr, T.G. 2005. Computational analysis of 3'-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome* **15:** 369-375.

Yeo, G.W., Coufal, N.G., Liang, T.Y., et al. 2009. An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology* **16(2):** 130-137.

Yeo, G., Nostrand, E., and Liang, T. 2007. Discovery and analysis of evolutionarily conserved intronic splicing regulatory elements. *PLoS Genetics* **3(5):** 815-829.

Yeo, G., Nostrand, E.L., Liang, T.Y. 2007b. Discovery and analysis of evolutionalrily conserved intronic splicing regulatory elements. *PLoS Genetics* **3(5):** e85.

Yeo, G. and Burge, C.B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology* **11:** 377-394.

Yu, J., Mallon, M.A., Zhang, W., et al. 2006. DNA repair pathway profiling and microsatellite instability in colorectal cancer. *Clinical Cancer Research* **12:** 5104-5111.

Zhang, C., Hastings, M.L., Krainer, A.R., and Zhang, M.Q. 2007. Dual-specificity splice sites function alternatively as 5' and 3' splice sites. *PNAS* **104(38):** 15028-15033.

Zhou, H., and Lou, H. 2008. Repression of prespliceosome complex formation at two distinct steps by Fox-1/Fox-2 proteins. *Molecular and Cellular Biology* **28(17):** 5507-5516.

Zhou, H., Baraniak, A.P., and Lou, H. 2007. Role for Fox-1/Fox-2 in mediating the neuronal pathway of calcitonin/calcitonin gene-related peptide alternative RNA processing. *Molecular and Cellular Biology* **27:** 830-841.