12-2012

The use of mathematical and computational models to define the role of mutation and infection in colorectal cancer.

Chandler D. Gatenbee 1982-
University of Louisville

THE USE OF MATHEMATICAL AND COMPUTATIONAL MODELS TO DEFINE THE ROLE OF MUTATION AND INFECTION IN COLORECTAL CANCER

By

Chandler D. Gatenbee
B.A. University of Louisville, 2005

A Dissertation

Submitted to the Faculty of the

College Arts and Sciences of the University of Louisville

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

Department of Biology
University of Louisville
Louisville, KY

December 2012

THE USE OF MATHEMATICAL AND COMPUTATIONAL MODELS TO DEFINE THE ROLE OF MUTATION AND INFECTION IN COLORECTAL CANCER

CHANDLER D. GATENBEE

A Dissertation Approved on

November 19, 2012

by the following Dissertation Committee:

Dr. Paul Ewald

Dissertation Direction

Dr. Lee Dugatkin

Dr. Jennifer Mansfield-Jones

Dr. Fabian Crespo

Dr. Henry Harpending

DEDICATION

This dissertation is dedicated to my wife Amy, my son Rowan, and my parents Trudy and Doug. My family has always been there to support me when times got tough, helping me see the light at the end of the tunnel. Over the years they have made countless sacrifices to ensure that I had the opportunity to follow my research interests, no matter where they took me. I cannot thank my family enough, and I consider myself blessed.

ACKNOWLEDGEMENTS

I would like to thank the many professors who have supported and guided me throughout my education. In particular, Dr. Christopher Tillquist, Dr. Fabian Crespo, Dr. Henry Harpending, Dr. Alan Rogers, Dr. Lynn Jorde, Dr. Lee Dugatkin, and Dr. Mansfield-Jones have all played an instrumental role in my development from a wide-eyed undergraduate into a contributing member of the scientific community. I would especially like to thank my mentor, Dr. Paul Ewald, for helping me bring together my interests in human evolution, disease, infection, and mathematical modeling. I would also like to thank Dr. Ewald for giving me the opportunity to explore any and all of my research interests. Over my twelve year academic journey, these professors have helped me grown not only as a researcher, but as a person.

ABSTRACT

THE USE OF MATHEMATICAL AND COMPUTATIONAL MODELS TO DEFINE THE ROLE OF MUTATION AND INFECTION IN COLORECTAL CANCER

CHANDLER D. GATENBEE

November 19, 2012

Research over the past twenty five years has led to the development of the hypothesis that colorectal cancer is caused by the accumulation of mutations in tumor suppressor genes and proto-oncogenes. The last ten years has also revealed that the common JC Virus (JCV) is frequently found in colorectal tumors. This has led to the hypothesis that the virus, which is known to cause tumors in the lab, may play a role in colorectal cancer. However, the presence of JCV in colorectal tumors does not necessarily indicate a cause-effect relationship. Unlike *in vivo* and *in vitro* studies, mathematical and computational modeling provides an opportunity to evaluate the roles that mutation and infection play in colorectal tumorigenesis. Three probability models are developed to asses whether colorectal cancer can occur by mutation alone or if infection is required. Two models find that JCV is required for tumorigenesis, and that mutation alone is unable to generate any tumors. The third probability model finds the opposite; mutation is able to generate realistic numbers of colorectal cancer patients, while infection is not. All three models do indicate that selection for a stem cell mutation rate that is 100 times lower than transit cells provides protection from cancer, confirming the findings of other research groups. An agent based model is also developed to simulate many of the complexities that cannot be modeled in the probability

models. The results from the agent based model indicate that JCV exacerbates colorectal cancer and greatly increases the risk of developing cancer. It also finds that mutation alone is able to cause colorectal cancer, although not as frequently as JC virus associated cases. All together, these models indicate that both mutation and infection have the capacity to drive tumorigenesis, but that the presence of JC Virus increases the risk of developing colorectal cancer. This strongly suggests that the role of JCV in colorectal cancer deserves more attention. If future studies confirm these findings, it would indicate that the prevalence of colorectal cancer can be reduced by taking measures to prevent infection by JC Virus.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1 THE BARRIERS TO CANCER

1.1 OVERVIEW

Multicellular organisms have evolved several mechanisms to prevent an individual cell from dividing uncontrollably, a process that results in cancer and possibly death. These mechanisms include: tight regulation of the cell cycle; using apoptosis to kill cells that have accumulated too much damage; limiting the cell's maximum number of divisions; and keeping the cell anchored to the matrix of underlying tissue. It happens that interfering with these mechanisms can drive a cell to divide uncontrollably, leading to the formation of tumors. It is for this reason that these mechanisms have been identified as "cancer barriers", as when their presence prevents tumors from forming. These barriers can be removed via any combination of mutations (somatic or germline), non-infectious environmental carcinogens, and infection. The following chapter will review each barrier, the relevant signaling pathways, and provide examples of how each barrier can be removed.

1.2 WHAT IS CANCER?

Imagine if the cells in your body began to divide uncontrollably and without limit. Within a relatively short amount of time, the progeny of those cells would form large masses of tissue, called tumors, in and on your organs. Those organs would soon cease to function normally, and death would be imminent. It is for this reason that cells of multicellular organisms have evolved several sophisti-

cated mechanisms, herein referred to as cancer barriers, that regulate when and where cells can survive and divide. These barriers have been subject to positive selection, as without them the individual would never survive to reproduce. However, transforming from a normal cell into a cancer cell is not like flipping a switch, which occurs in an instant; instead, tumor progression is a multistep process that can take years to complete. Through careful examination of many types of cancers, researchers have been able to identify many of the common steps that occur during tumor progression, each of which is considered a hallmark of cancer [67, 50]. We will now briefly review how each barrier protects the individual, and how disruption of the barrier provides that cell with a selective advantage, driving it one step closer to evolving into a cancer cell.

1.2.1 *De-regulation of the Cell Cycle: Pro-growth and Anti-growth Barriers*

One of the largest barriers to cancer is the extremely tight regulation over the cell cycle, which determines when and where a cell can divide. The cell cycle consists of four distinct phases: G_1,S,G_2,M. When a cell is stimulated to divide it starts an intracellular signaling cascade that stimulates the formation of CDK:cyclin, which in turn catalyzes the phosphorylation of pRb. Hypophosphorylated pRb keeps the cell frozen in G_1 by binding to the E2F transcription factors [146]; phosphorylation deactivates the inhibitory properties of pRb, freeing the transcription factors and driving the cell into S phase.

DNA replication takes place during S phase. Afterwards, the cell makes sure that no significant DNA damage has occurred during replication, and if there is not any the cell continues on into the G_2 phase. G_2 is in turn followed by M phase, which consists of mitosis (division of DNA between daughter cells) and cytokinesis (division of cytoplasm and organelles between daughter cells). Afterwards, the cell moves back into G_1 . Once in G_1, the cell ensures that it received

the correct number of chromosomes, and if it did not it commits apoptosis (cell suicide).

The cell cycle is primarily regulated in two ways: 1) the cell is stimulated to divide in the presence of pro-growth signals; 2) the is cell forced to stop dividing in the presence of anti-growth signals. Pro-growth signals typically come in the form of growth factors. Growth factors are secreted by other cells, and when they bind to another cell's growth factor receptor, they stimulate that cell to divide by inactivating pRb. If a cell is not stimulated by enough growth factors it will remain in G_1; if in G_1 for a prolonged period of time, the cell is said be in a quiescent state termed G_0. Anti-growth signals come in the form of soluble signals and embedded signals, which can block cell division in one of two ways: either they force the cell into the G_0 (quiescent) state, or they force the cell to relinquish its ability to divide, usually due to terminal differentiation[50]. Normal cells have thus evolved pro-growth and anti-growth barriers to regulate cell division, helping prevent uncontrolled growth.

Cancer cells de-regulate the cell-cycle by removing the pro-growth and anti-growth barriers that normally determine when and where the cell can divide. The pro-growth barrier is frequently removed in one of three ways: 1) the cell starts producing its own growth factors (autocrine stimulation); 2) the cell produces growth factor receptors that are permanently activated; 3) the signaling pathway from the growth factor receptor is altered [67, 50]. No matter the path, the result is that the cell divides even when it is not externally stimulated by growth factors. Similarly, the anti-growth barrier can be removed in several different ways, although most of them converge on pRb [50]. The reason for this convergence is that if pRb is removed, the cell will always move from G_1 into S phase, preventing the cell from entering a quiescent state. Another common strategy for removing the anti-growth barrier is to avoid terminal differentiation, thus allowing the cell to continue dividing. A cancer cell might accomplish this by constitutively producing the Myc protein, which supplants Mad in the Mad:Max

complex, creating Myc:Max, a protein complex that that impairs differentiation and promotes growth [41].

While some have argued that the pro-growth barriers and anti-growth barriers are separate and distinct [50], others feel that the two should be combined into one barrier [38, 93]. The debate arises because the two barriers often converge on the same pathways, and both lead to a cell that constantly divides. For example, removal of pRb is said to remove the anti-growth barrier, yet up-regulation of CDK4 (which is in the same pathway as pRb) is said to remove the pro-growth barrier (reviewed in [93]).

Regardless of whether or not the two barriers are distinct, their removal increases their replication rate, resulting in increased fitness. This cell will then experience positive selection, as it can replicate more frequently, leading to an increase in its frequency throughout the population.

1.2.2 *Apoptosis Barrier*

Removal of the pro-growth and anti-growth barriers bring the cell very close to uncontrolled growth. However, there are several cell-cycle checkpoints that ensure the integrity of the cell's DNA; if the cell has accumulated too much damage (i.e. too many mutations, incorrect number of chromosomes, etc...), the p53 protein accumulates. Accumulation of p53 stimulates the transcription of p21, a Cyclin Dependent Kinase (CDK) inhibitor. When there is too much DNA damage, p53 indirectly halts the cell cycle , giving the cell time to repair the damage. If the damage cannot be repaired, p53 activates Puma (p53 up-regulated modulator of apoptosis), which binds to, and inhibits, the omnipresent anti-apoptotic Bcl-2 protein, increasing the permeability of the mitochondrial membrane, allowing for the secretion of cytochrome c. Cytochrome c then stimulates a caspase cascade that leads to cell suicide, a process termed apoptosis. Cells not only respond to internal DNA damage, but also to other stressors, such as infection or hypoxia

[67, 74]. Furthermore, cells can also undergo apoptosis in response to external signals, such as TNF-α or FASL, both of which might be secreted in response to infection.

Given the stopping power of apoptosis, a cell that successfully removes both pro-growth and anti-growth barriers still has a very good chance of being killed off by the accumulation of too many deleterious mutations, being infected, or experiencing hypoxia. Such a cell that is continuously dividing will almost inevitably acquire so many deleterious mutations that it undergoes apoptosis, thus successfully removing a pre-cancerous cell. However, as one might suspect, the apoptotic barrier is indeed removed in most cancers, as its removal provides the cell with a selective advantage by increasing its survival rate. In fact, inhibition of p53 by mutation alone is estimated to occur in 50% of cancers [51]. This statistic not only testifies to the critical role of p53, but also to the importance of apoptosis, which is able to eliminate unhealthy cells. If those cells cannot be eliminated, they remain free to divide, and thereby remove the remaining barriers.

1.2.3 *Replication Limit Barrier*

Removing the pro-growth, anti-growth, and apoptosis barriers should, theoretically, drive the cell to divide uncontrollably. However, it turns out that there is an another barrier that limits cellular replication, and this one does not depend upon cell-to-cell signaling. At the ends of each chromosome are several thousand 6bp repeats called telomeres. After each division, $\sim 50 - 100$bp of the telomeres are lost, due to the inability of DNA polymerase to completely replicate the 3′ ends of chromosomal DNA during cellular division [67]. Over enough replication cycles, the telomeres are lost, the chromosomes fuse, the cell experiences crisis and eventually is subject to apoptosis. In other words, the cell has its own internal mechanism to limit the number of divisions it can undergo, and thus potentially how long a cancer lineage would survive. However, $85 - 95\%$ of cancer cells

remove this barrier by up-regulating the expression of telomerase, an enzyme that adds the 6bp repeat back onto the chromosome's ends [123]; the remaining $5-15\%$ remove the barrier using a recombination-based inter-chromosomal exchange mechanism termed the Alternative Lengthening of Telomere (ALT) pathway [15]. Replacement of telomeres, and thus resetting of the cell's clock and removal of the replication limit barrier, gives the cell a selective advantage by giving it the potential to divide without limit.

1.2.4 *Angiogenesis: An Intermediate Event*

If a cell gains the ability to divide wherever it wants, without limit, it seems that cell would be able to create a massive tumor in a short amount of time. However, this is not necessarily the case. It has been demonstrated that tumors grown in absence of a blood supply, such as in the anterior chamber of the eye, only grow to $2-3$mm [67, 42]. Yet when these same cancer cells are placed in tissue with a proper blood supply they are able to rapidly generate large tumors. It seems that access to oxygen and nutrients, provided by the blood, are critical for large tumor growth. It appears the reason that tumor size is capped at $2-3$mm when in a vessel deprived environment is that the tumor cell outgrow the other cells in their micro-environment, including the oxygen supplying capillary Endothelial cell (EC)s [42, 136, 135]. As the tumor grows, the distance between the innermost tumor cells and the nearest capillary increases. Lack of oxygen (hypoxia) prevents those innermost tumor cells from replicating, as they are on the verge necrosis. The tumor initially manages to supply the hypoxic cells with oxygen first by co-opting the surrounding ECs. However, the inner-most tumor cells remain isolated from the co-opted blood vessels, and respond to their hypoxic condition by over-expressing compounds, such as Vascular Endothelial Growth Factor (VEGF) and fibrolast growth factor (FGF), that induce the production of new blood vessels from pre-existing blood vessels, a process called angiogenesis

[118]. In fact, the amount of VEGF produced is enhanced by hypoxia [62], as the most hypoxic tumor cells produce the most VEGF, creating a gradient of VEGF from the hypoxic tumor cell to the oxygen supplying EC [58, 73]. Taken together, these observations suggest that VEGF produced by hypoxic tumor cells will lead to the rapid extension of vessel tips, which will eventually "crawl" their way towards the hypoxic tumor cells. Once those cells receive the oxygen they crave, they are able to divide, and the tumor grows beyond 2-3mm.

While angiogenesis is primarily activated by pro-angiogenic molecules like VEGF and FGF, the process is also controlled by inhibitors, such as thromspondin. Therefore, the must be more activators than inhibitors for angiogenesis to be initiated [67]. As angiogenesis permits an increase in size of an already growing tumor, it generally considered an early to mid-stage event [50].

1.2.5 *Metastasis Barrier*

While cells dividing uncontrollably can lead to tumors, they may often be benign and removed surgically. If not removed early enough these benign tumors may begin to produce cells that invade new tissues, a process call metastasis. Unlike benign tumors these metastatic tumors are deadly, a fact illustrated by the observation that 90% of human cancer deaths are from metastases [131]. The acquisition of mobility is a complex process that is facilitated by angiogenesis. Part of the reason for this relationship is that during angiogenesis the dividing endothelial cells produce matrix metalloproteinases (MMPs), which break down both the extracellular matrix and the basal lamina, a process which creates an opening for mobile metastatic cancer cells to enter the blood supply [67]. A second reason is that VEGF also directly increases the permeability of the vascular wall by loosening cell-cell contacts, making it easier for mobile cancer cells to enter the blood stream (reviewed in Saharinen et al. [118]). Finally, the simple ex-

istence of more blood vessels increases the opportunities for mobile cancer cells to enter to blood stream and eventually invade new tissue.

As already noted, for a cell to gain entry into the blood stream it must first acquire mobility. All cells have the fundamental molecular tools for locomotion, but most are rendered immobile because they are bound to the extracellular matrix via integrins, and to their surrounding cells via Cell-Cell Adhesion Molecules (CAM)s , such as E-cadherin. Cancer cells, on the other hand, frequently have non-functional CAMs, allowing them to separate from their neighbors [67, 50]. Cancer cells also frequently have the ability to vary which integrins they express, giving them the ability to attach to whichever surface they may move to [50]. Together, loss of cell-cell adhesion and the an ability to bind to different surfaces provides cancer cells the opportunity to separate from the primary tumor mass and move throughout the environment. Thus, while CAMs and integins undoubtedly serve other purposes, they also serve as an effective barrier that protects the individual from having rogue cells wander throughout the body.

The next question is where should the mobile cancer cells go, and how do they get there? Chen et al. [21] have developed an agent based model to answer this very question. This model is built upon evolutionary dispersal theory, which suggests that when there is resource variability (oxygen, nutrients, etc..) in the tumor, such as when tumor-induced angiogenesis occurs, mobile cells are selected for, as they have the ability to move to areas of high resource concentrations, such as where new blood vessels have formed during angiogenesis. Mobile cells may move towards the underlying blood vessels by producing proteases, which degrade the basal lamina, giving the mobile cell the ability to burrow through the underlying tissue and enter the bloodstream.

Once in circulation the chances of cell survival are low. In an experiment cancer cells were radioactively labeled and injected into the bloodstream of lab animals, and after a few weeks only one in one thousand were still alive, indicating that

very few cancer cells can survive in the bloodstream (described in [67]). However, if that cell is able to survive the trip through the bloodstream and move into new tissue (metastasize), it will, at least initially, be privy to additional resources. If the other barriers have been removed, this metastatic cell will be able to divide indefinitely, leading to the formation of a potentially deadly metastatic tumor [50].

1.3 REMOVING THE CANCER BARRIERS

The causes of disease fall into one of three categories: genetic (mutation and methylation), non-infectious environmental, or infectious [23]. However, in many cases, diseases have primary causes and secondary causes; the disease cannot occur without the primary causes, and is exacerbated by the secondary cause(s) [38]. In much the same way, each cancer barrier can potentially be removed by mutation (somatic or germline), non-infectious environmental factors, or infectious agents [67]. As cancer is a multi-step process requiring the removal of several protective barriers, it may also be that for many cancers there are also primary and secondary causes. For example, cervical cancer may primarily caused by HPV infection, but the rate of tumor progression may be accelerated by inherited mutations in TNF-α [153]. The following section will review how each of the different causes can remove the barriers to cancer.

1.3.1 *Genetic Changes and Genomic Instability*

Mutations that remove cancer barriers can generally be divided into two categories: germline mutations and somatic mutations. A classic example of a germline mutation that can increase the risk of cancer is the inheritance of one defective or missing pRb gene, which can increase an individual's risk of developing retinoblastoma by 90% [67]. However, before retinoblastoma can actually develop,

the second copy of pRb must also be rendered defective. The dramatic increase in cancer risk by simply removing pRb illustrates the protein's key role in regulating division, and how its removal can lead to de-regulation of the cell cycle, allowing the cell to divide in the absence of growth factors [67].

Novel mutations can potentially occur every time a cell divides, as they result from replication errors. It has been estimated that the probability of nucleotide mis-incorporation is $\sim 10^{-6}$ per replication event, but proof-reading exonucleases and mismatch repair improves replication fidelity ~1000 fold[143], leading to a mutation rate of 10^{-9}, or one mis-incorporation for every 10^9 bp every cell generation [3].

This extremely low mutation rate led Hanahan and Weinberg [50] to state that "mutations are rare events, indeed so rare that the multiple mutations known to be present in tumor cell genomes are highly unlikely to occur within a human lifespan". This observation leads to two (not mutually exclusive) explanations for how cells acquire the ability cause cancer: 1) one or more of the other categories of barrier-removal is involved (i.e. non-infectious environmental, or infectious); 2) tumor cells have increased mutability, a phenomenon known as genomic instability.

There are three types of genomic instability: increased point mutation rates, Microsatellite Instability (MSI), and Chromosomal Instability (CIN). MSI may be caused by mutations in Mismatch Repair Enzymes (MMR), which can cause DNA polymerase to slip during replication of tandem repeats, resulting in the insertion/deletion of microsatellites [143]. CIN leads to alterations in large segments of chromosomes, including losses, gains, translocations, inversions, deletions, amplifications, and frequently aneuploidy [143]. Both MSI and CIN are observed in cancers, as MSI is frequently seen in individuals with hereditary non-polyposis colorectal cancer (HNPCC), while CIN is characteristic of most cancers and will be discussed more thoroughly [93].

Approximately seven genes have been associated with CIN and somatic mutation, which is believed to cause CIN through one of three pathways: chromosome segregation defects, telomere dysfunction, and dysregulation of DNA damage response [102]. Normally, the mitotic checkpoint ensures that chromosomes are segregated properly, but mutations in genes that regulate segregation can result in an unequal distribution of chromosomes, leading to aneudploidy [102]. Some of the frequently mutated genes are: mitotic arrest–deficient (MAD), budding uninhibited by benzimidazoles (BUB), anaphase-promoting complex/C (APC/C; not to be confused with adenomatous poplyposis coli (APC) discussed in Chapter 2)[102]. Additionally, an abnormal number of centrosomes can lead to formation of multiple spindle fibers during mitosis, which can also result in aneuploidy [102]. Telomere dysfunction can induce CIN when telomeres become extremely short, as the ends begin to fuse with neighboring chromosomes, resulting breakage-fusion-bridge cycles that can lead to dramatic genome reorganization[102]. However, if telomerase is up-regulated during later stages of cancer such CIN may cease to occur, and the reorganized tumor cell may gain immortality[102]. Finally, impaired DNA damage responses can also induce CIN. Normally, DNA damage responses protect the cell from exogenous and endogenous stresses by initiating signaling cascades that result in cell cycle arrest or apoptosis. If the DNA damage responses are impaired, the cell may accumulate large numbers of mutations, some of which may result in CIN. Many of the genes involved in these DNA damage responses are frequently mutated in cancers. Some of the more commonly mutated genes include: ataxia telangiectasia mutated (ATM), ataxia telangiectasia and Rad3-related (ATR) protein kinases, p53, BRCA1, and BRCA2.[102]. Of them all, p53 mutations are the most common [93].

CIN can dramatically increase the rate at which the above barriers are removed. If one allele of a gene is mutated (either through inheritance, novel mutation, or methylation) the cell is considered heterozygous, and may still function normally. However, if the second allele is also knocked out, possibly by

CIN, the heterozygous state is lost, a process known as Loss of Heterozygosity (LOH) [67]. LOH is far more common than mutation, occurring ~ 1 out of every 1000 cell divisions[67], and on average results in the loss of 25%-30% of all alleles in a tumor[102]. In fact, LOH is so common that it is considered the "hallmark" of CIN-positive tumors [67, 102]. LOH is believed to be caused by three different processes: mitotic non-disjunction, mitotic recombination, and gene conversion.[67, 102]. During mitotic non-disjunction, one chromatid fails to separate during mitosis, resulting in one cell that has 3 copies of a chromosome, and a second cell with only one chromosome. Thus, the cell with one chromosome would have lost its heterozygous state [67]. Mitotic recombination involves the exchange of DNA between homologous chromosomes, a process that generates diversity. However, such recombination can also result in a cell that is homozygous for an allele, should the swapped allele segregate with a homologous chromosome containing the same allele [67]. Gene conversion occurs when one of the two homologous chromosomes copies and inserts a segment of its DNA into the other homologous chromosome, resulting in a cell that has three copies of the same allele[67]. After segregation, one of the cells will become homozygous for that allele that was copied. LOH is so dangerous because if the allele that becomes homozygous is a defective tumor suppressor gene, a barrier to cancer will be removed and that cell will be one step closer to becoming a cancer cell.

In addition to mutation, genes may also be silenced via epigenetic changes (heritable changes not encoded in DNA) induced by methylation of a gene's promoter region [67]. Methylation occurs when methyl groups attach to the 5' position of a cytosine (C) nucleotide [67, 4]. In humans, the promoter region of DNA often contains unmethylated CpG islands. Transcription may be inhibited if these CpG islands become methylated, effectively "silencing" the gene [67, 4]. Such gene silencing via methylation is believed to be just as common as mutation,

and so it is possible that mutation could remove one allele while methylation silences the other [67, 4].

1.3.2 *Non-Infectious Environmental Carcinogens*

The second way in which a cell's cancer barriers might be removed is via exposure to various non-infectious environmental compounds. Examples of such chemical carcinogens include: polycyclic aromatic hydrocarbons (found in coal tars, soots, and oils); aromatic amines (found in dyes, tobacco smoke); N-nitroso compounds (some are present in cigarette smoke); alkylating agents (used in production of plastics, antifreeze, "mustard gas"); various inorganic substances (i.e. asbestos); and other natural products, such as aflatoxin, a carcinogen produced by the mold *Aspergillus* [67]. Many of these carcinogens are metabolized in the liver, where they become electrophillic and thus tend to bind to electron-rich DNA. The interaction between these carcinogens and DNA causes the DNA double helix to distort, resulting in an increased number of mutations during cell division[67]. Once the damaged DNA molecule has been replicated, it can be almost impossible for the cell to repair the damage, and so the mutation can be inherited by daughter cells. Once a mutation has occurred, the growth of those cells may initially be dependent upon promoting agent carcinogens. Over time, however, these cells may acquire additional mutations or epigenetic changes that allow them to divide in the absence of the promoting agent, leading to the evolution of self-sufficient cancer cell.

1.3.3 *Infection*

While it is not in the interest of a pathogen to induce cancer in its host, there are several reasons why one would expect that most chronic pathogens would evolve mechanisms that remove some of the cell's cancer barriers. De-regulation of the

cell cycle allows the intracellular pathogen to divide along with the cell while minimizing detection by the immune system; inhibiting apoptosis would allow the pathogen to survive infection-induced apoptosis; up-regulating telomerase increases the number of divisions the host cell and its pathogen can undergo; and removing metastasis barrier gives the pathogen the ability to move to different areas within the host, where it may have access to more resources or more easily get transmitted to other hosts [38]. The virus can increase its intra-host fitness by removing several barriers to cancer. However, a strain that frequently removed of all barriers would have a lower inter-host fitness, as its host would succumb to cancer soon after infection, limiting the number of possible transmissions. It may be that pathogens have to walk a fine line between these competing levels of selection, and that many have struck a balance between intra- and inter-host fitness. Evidence for this hypothesis comes from the observation that only a small proportion of individuals infected by oncogenic viruses actually develop cancer, suggesting that these viruses may only remove a few of the cell's cancer barriers[38].

Perhaps one of the best understood examples of a pathogen being the primary cause of cancer is that of human papilloma virus (HPV), the high risk strains (i.e. HPV 16, 18, and 31) of which are the agents behind cervical cancer. While HPV produces many different proteins during its life-cycle, only two appear to be required to transform a normal cell into a malignant cell [153]. These two proteins are E6 and E7, and each is quite efficient at removing various cancer barriers (reviewed in [153]). E6 activates the catalytic subunit of hTERT (human telomerase reverse transcriptase), thus removing the replication limit barrier . At the same time, E6 also removes the apoptosis barrier by binding to and degrading p53. Removal of apoptosis also results in the loss of the G_1 checkpoint, allowing the cell to divide even when there is DNA damage, a process which may induce chromosomal instability [153]. While E6 is able to remove apoptosis and the replication limit, E7 seems to play a key role in deregulating the cell cycle. E7 has the ability

to bind pRb, which frees the E2F transcription factors and drives the cell into S phase, thus removing the anti-growth barrier. Furthermore, E7 is able to bind to the CDK inhibitors p21 and p27, thereby increasing the levels of cyclins in the cell and driving it to divide, thus removing the pro-growth barrier[153]. Combined, these observations suggest that high risk HPV is able to de-regulate the cell cycle, up-regulate telomerase, inhibit apoptosis, and induce genomic instability. This only leaves the metastatic barrier remaining, which might be removed by one of the other categories of barrier removal (i.e. mutation or non-infectious environmental cause).

Hosts have evolved a complex set of mechanisms to protect against the damage caused by infection, some of which may result in cancer. However, in this evolutionary arms race pathogens frequently have an advantage, as they are able to evolve counter-strategies at a much faster rate due to their high replication rates and short generation times. zur Hausen [153] developed the concept of three Cellular Interfering Factor (CIF) pathways that the pathogen must overcome in order to drive the cell to become malignant. The existence of such pathways were "initially postulated to explain the restriction of tumor-virus gene expression in proliferating cells, and the long latency period [~20-30 years] between primary infection and the eventual emergence of invasive cancer" [153].

The first CIF pathway is CIF-I, which includes all pathways involved in pathogen recognition by the immune system. For example, T-cells have the ability to recognize HPV antigens presented on the surface of infected cells. It is in the interest of HPV to find a way to avoid elimination by the immune system. As it turns out, HPV has indeed evolved a counter-strategy, allowing it to evade detection by the immune system. The E5 protein of high-risk HPVs can down-regulate the expression of both MHC class I and MHC class II molecules, which present antigens to CD8 and CD_4 T-cells, respectively. This process is believed to delay early recognition by the immune system, although it may not be sufficient to permit persistent infection [153]. Some lucky HPVs might acquire the ability to

avoid detection during persistent infection by being fortunate enough to be in a cell that has also acquired mutations in human leukocyte antigen genes (HLA), which encode HLA proteins. Indeed, HLA mutations are found in 90% of cervical cancers [153]. It seems reasonable to assume that a cell which has increased genomic instability, possibly induced by HPV, would also be more likely to have a mutation in one of these HLA genes, as more mutation events means there is a greater chance the mutation will land in a particular locus.

The second CIF pathway, CIF-II, is the collection of mechanisms that inhibit the functioning of viral oncoproteins[153]. In the case of HPV, $p16^{INK4}$ appears to limit the effectiveness of E6, while $p14^{ARF}$ may be involved in moving E7 from the nucleoplasm to the nucleolus, thereby preventing E7-induced degradation of pRb [153]. However, this is only true in cells in which HPV can either express only E6 or E7, but not both. When both proteins are expressed, E6 blocks effects of $p14^{ARF}$, while E7 is able to circumvent the activity of $p16^{INK4}$. Together these proteins are therefore able to "help" each other, blocking the cell's inhibitory proteins, allowing HPV's oncoproteins to function.

The third CIF pathway is CIF-III, and includes all signals involved in paracrine control, particularly cytokines and chemokines. In particular, TNF-α (a cytokine that promotes inflammation and/or induces apoptosis) appears to limit the growth of HPV-immortalized cells, but not malignant HPV cervical cancer cells. This suggests that TNF-α is able to limit growth of most HPV infected cells, likely through the external stimulation of p53-independent apoptosis, and that something must happen in order for them to become malignant. This "something" might be mutations in the TNF-α gene, a hypothesis that is supported by the observation that many polymorphisms in the TNF-α promoter increase the risk of cervical cancer [153].

There has likely been great selection pressure on pathogens to evolve ways to circumvent these CIF barriers. Those strains that have such abilities will have a much greater fitness, as they would be able to survive longer and replicate more

frequently than strains that cannot overcome these barriers. In the case of HPV, it seems that the high-risk strains have evolved such strategies to overcome the three CIF pathways, a feat which not only increases the fitness of those strains, but simultaneously increases the risk of cervical cancer.

1.4 CANCER AS AN EVOLUTIONARY PROCESS

If a cell is able to remove the barriers to cancer, through any combination of the mechanisms described above, it will gain a selective advantage [84].De-regulation of the cell cycle gives the cell the ability to divide when and where others cannot; inhibition of apoptosis reduces the probability of cell death; removal of the replication limit allows the cell to divide more times than other cells; metastasis may give the cell the ability to escape a necrotic environment, moving to one that has more abundant resources. Given the selective advantage conferred on these cells, the removal of these barriers can be considered beneficial to the cell, but harmful to the host.

CHAPTER 2 MUTATION AND COLORECTAL CANCER

2.1 OVERVIEW

Colorectal cancer, the third most common cancer in men and women, is expected to kill 51,690 Americans in 2012. Due to its relevancy, researchers have spent the past 20 years trying to understand what drives colorectal tumorigenesis. Beginning in 1990, a hypothesis was put forth that specific mutations, commonly observed in colorectal tumor tissue, occur in a preferred order and are largely responsible for tumorigenesis. Subsequent studies have built upon this hypothesis, making it the most commonly accepted argument of colorectal cancer causation. However, the last ten years have revealed that a common infection, the JC polyomavirus, is frequently associated with human colorectal tumors. JCV expresses several viral oncoproteins that interfere with key cellular pathways, which is known to cause cancer in lab animals. Both observations have led many to further investigate the role of JCV in colorectal cancer.

2.2 STRUCTURE OF COLON AND CRYPTS

The colon is roughly organized as an outer layer of smooth muscle, a central layer of connective tissue, and an inner layer of absorptive epithelial lining. The structural subunit of the colon is the colon crypt, a collection of ~250 cells that penetrate into the underlying submucosa [127, 12]. Each of these colon crypts is sub-divided into three sections: the crypt base, the mid-crypt (a.k.a. the proliferative zone), and the upper crypt [12]. As the cells of the epithelial lining are

constantly shed into the lumen they must be replaced by cells generated from the $4-6$ pluripotent stem cells that reside at the crypt base [12, 105]. These stem cells are defined by several abilities that set them apart from other cells: stem cells remain undifferentiated; they are capable of proliferation and self-maintenance; they are pluripotent (i.e. they can produce many different kinds of cells); they are able to regenerate tissue after injury; they can divide indefinitely [12]. The cell cycle time of these stem cells has been measured to be between $12-32$ hours in mice, and is believed to be $4-8$ times longer in humans [104]. This means that the stem cell cycle time in humans could be between $2-10$ days; or, if the average cell cycle time is 22 hours, the average stem cell cycle time should be about five and a half days ($\frac{12+32}{2} \times 6$). Approximately 95% of the time colon crypt stem cells produce one daughter transit amplifying cell and one daughter stem cell, a process referred to as asymmetric division [79]. However, 5% of the time a stem cell may undergo symmetric division, producing either two daughter stem cells or two daughter transit cells [79]. If two stem cells are created, another stem cell is lost by differentiation, displacement, or apoptosis [12].

The transit cells produced by stem cells migrate upwards into the mid-crypt, where they gradually mature into one of four different cell types: absorptive colonocytes, mucus secreting goblet cells, and peptide producing endocrine cells [127]. As the transit cells migrate through the mid-crypt, they continue to replicate along the way; it has been estimated that 60% of the cells in the crypt are replicating, and most can be found in the bottom two-thirds of the crypt [103]. By the time the transit cells have differentiated, they have moved into the upper-crypt. As transit cells migrate towards the lumen they lose their ability to replicate, possibly because the upper-crypt lacks appropriate growth factors [103]. This limited replicative ability has been confirmed in studies demonstrating that cells in the upper crypt do not have the ability to regenerate a crypt after radiation injury [52]. Eventually, the differentiated cells reach the most superficial

part of the epithelial lining, where they undergo apoptosis and are shed into the lumen [127, 12].

It has been estimated that adding an additional stem cell to a crypt could create an additional $60 - 120$ cells in the crypt, possibly leading to dysplasia [103]. It is for this reason that the number of crypt stem cells is tightly regulated. It also appears stem cells cannot efficiently repairing DNA damage and often undergo p53 mediated apoptosis [12]. These lost stem cells can then be replaced by symmetric division. Not only does this process help regulate the number of stem cells in the crypt, it also prevents the accumulation of carcinogenic mutations [12].

If a mutation provides a stem cell with an advantage or is neutral (possibly because it is a recessive mutation), that mutation may spread through the crypt via a niche succession, a process which appears to be somewhat stochastic and dependent upon symmetric division [56, 12]. The process of niche succession begins when a mutated stem cell produces two daughter stem cells during symmetric division. Afterwards, the mutation is present in each of the two stem cells, as well as each stem cell's progeny. As this process is repeated, a mutation may come to be present in all cells in the crypt, a phenomenon that has been estimated to occur every 8.2 years in humans [63]. Niche succession might be accelerated if the mutation occurs in a tumor suppressor gene or oncogene, which can increase the replication and survival rate of the cell [56].

It appears that a mutation might not only be able to spread within a crypt, but it might also be able to spread between crypts. If a mutation occurs in a gene that regulates apoptosis, such as p53 or Bcl-2, the apoptotic regulation of stem cell numbers is lost, leading to an excess of stem cells [12]. Should too many stem cells accumulate, the crypt will respond by bifurcating, thus distributing the number of stem cells between the two crypts, a process termed crypt fission [12]. Thus, any mutations in that first crypt will now be in two crypts. One can imagine how this can allow a mutation to spread throughout the colon. However, it appears that crypt fission is a relatively rare event in humans, and has been

estimated that there can be 30 years between crypt fission events suggesting that a single mutation may spread across some, but not all, crypts [56].

2.3 MUTATION MODEL: THE CANONICAL VIEW OF COLORECTAL CANCER

Although the incidence of colorectal cancer has been decreasing over the past two decades, it remains the third most common cancer in the men and women [1]. In 2012, it is estimated that there will be 143,000 new cases of, and 51,690 deaths from, colorectal cancer [1]. It is also estimated that 50% of the entire Western population will develop a colorectal tumor by the time they are 70 years old, and that 10% of those will develop into malignant tumors [64].

Beginning in the 1990s, researches began to put together a theory of colorectal cancer that remains strong to this day [56]. At its heart, this theory argues that colorectal cancer is initiated by a mutation in a single gene, and that tumorigenesis progresses by the sequential accumulation of other specific mutations. The process of accumulation is believed to be accelerated by genomic instability [64, 39].

Many of the important genes involved in colorectal cancer have been identified by studying two heritable forms of the disease, Family Adenomatous Polyposis (FAP) and Hereditary Nonpolyposis Colorectal Cancer (HNPCC) [39]. FAP is a dominantly inherited disease that affects ~1 out of 7,000 individuals (or $< 1\%$ of all colorectal cancer cases), while HNPCC accounts for $2-4\%$ of all colorectal cancers [64]. In both diseases the median age of developing cancer is 42, while the median age of sporadic colon cancer is 67 [64]. In the case of FAP, researchers have identified five genes that are commonly mutated in colorectal cancer: Adenomatous Polyposis Coli (APC), Kristen Rat Sarcoma Virus (KRAS), DCC, SMAD, and p53 [39, 28, 64]. It was also discovered that mutations in Mismatch Repair Enzymes (MMR), such as MHS2, MLH1, MSH6, and PMS2, are common in HNPCC patients [64, 4]. It has been estimated that these MMR muta-

tions lead to a $2-3$ fold increase in the mutation rate, making it more likely that a tumor-suppressor is knocked out or an oncogene activated, leading to tumor formation [64]. Together, these observations have led researches to conclude that colorectal cancer develops via the sequential accumulation of specific mutations, a process which is often accelerated by genomic instability [39, 64]. Furthermore, it has been argued that these mutations may spread throughout the crypt via niche succession, leading to monoclonal conversion of the crypt [39, 56]. The following sections will review the role of each gene, the proposed timing of each mutation, and how it drives tumor progression.

2.3.1 APC Mutation Required for Formation of Aberrant Crypt Foci and Early Adenomas

APC regulates the amounts of β-catenin, the central protein in the Wnt pathway, which is involved in activating cellular proliferation during development[67]. Under normal conditions β-catenin is degraded by a multi-protein destruction complex that contains APC, Glycogen Synthase Kinase 3 (GSK3), and axin. When this destruction complex is assembled, GSK3 catalyzes the phosphorylation of β-catenin, marking β-catenin for destruction by proteasomes, leading to a low concentration of β-catenin within the cell [67]. However, if an extracellular Wnt protein binds to a transmembrane Wnt receptor, the destruction complex is prevented from forming, leading to an accumulation of cytoplasmic β-catenin [67]. β-catenin accumulates, moves to the nucleus, binds to and activates several transcription factors, such as T-cell Factor (TCF) [83]. The TCF:β-catenin complex then activates transcription of **myc**, preventing cellular differentiation and allowing the cell to divide [142]. Interestingly, β-catenin also plays a major role in colon crypt organization, as its expression is down-regulated in the mid-crypt (due to a decrease in Wnt signals), resulting in cell cycle arrest and differentiation in the

mid-crypt [142]. Finally, given that other catenins bind to cadherins, it has been suggested that β-catenin might also be involved in cellular adhesion [64, 40].

It has been discovered that FAP patients have a deletion in one of the two copies of chromosome 5q [39]. Further research has revealed this deletion is in the APC gene, and that this deletion results in a truncated and non-functional APC protein [64, 134]. Truncated APC losses its ability to form the destruction complex, leading to the accumulation of β-catenin, preventing differentiation and driving the cell to divide [67], essentially giving the mutated cells a stem cell phenotype [4]. It is believed FAP patients, being $APC^{+/-}$, have an increased rate of crypt fission, allowing the mutation to spread within the colon [56]. However, the true effects of losing the tumor suppressor abilities of APC are not felt until both alleles are lost via a mutation or methylation of the normal APC allele [56, 37]. When APC is completely lost, it is believed that transit cells acquire the ability to divide in the mid-crypt [56]. These $APC^{-/-}$ these cells will then out-replicate their heterozygous neighbors, leading to monoclonal conversion of the crypt [56]. The combination of monoclonal conversion and crypt fission is believed to lead to the formation of dysplastic Aberrant Crypt Foci (ACF) (a.k.a. microadenomas), which are microscopic collections of abnormal crypts [56]. Formation of ACF is also considered the earliest stage of colorectal cancer, and may lead to the formation of early adenomas (benign epithelial tumors) [39, 64].

The removal of APC is often considered the initiating event of colorectal cancers because it is frequently found in ACFs as well as 78% of adenomas [56, 60]. Furthermore, APC is lost in 80% of sporadic colorectal cancers, again suggesting that loss of APC is a common mode of initiating colorectal tumorigenesis [56]. However, it has also been discovered that 15% of colorectal cancers do not have APC mutations, an observation that suggests other mutations are capable of initiating tumorigenesis [129]. Gain of function mutations in β-catenin are also commonly found in colorectal tumors, occurring in ~50% of colorectal tumors with wild-type APC [130], and thus may account for some of the 15% of cases

without APC mutations. However, it has also been demonstrated that adenomas with altered β-catenin are less likely to progress to malignancy than adenomas with APC mutations (β-catenin mutations were found in 12.5% of small adenomas, but only in 1.4% of malignant tumors), indicating that APC and β-catenin mutations do not have the same effect [119]. Even so, the frequency of APC and β-catenin mutations indicates that the Wnt pathway plays a central role in colorectal tumorigenesis.

2.3.2 *KRAS Mutations Drive Early Adenoma to Intermediate Adenoma*

The Ras-MAPK pathway is involved in stimulating cell division in the presence of growth factors [67]. The signaling cascade is initiated when a growth factor, such as PDGF or EGF, binds to a growth factor receptor. This binding activates Ras via phosphorylation, initiating a cascade of intracellular protein kinases (Raf, MEK, and MAPK) [67]. These kinases in turn activate the production of nuclear transcription factors (Ets, Jun, Fos, Myc, and E2F), resulting in the synthesis of cyclins and CDK molecules that phosphorylate pRb, driving the cell into S phase, resulting in division [67].

Mutations in KRAS (the gene that encodes the Ras protein), which are found in 40% of sporadic colorectal cancers, can produce a hyperactive form of Ras which drives the cell to divide even in the absence of growth factors [67, 4]. KRAS is thus considered an oncogene because mutating it results in a protein that forces cell division. There are several reasons why it believed KRAS mutations are responsible for driving an early adenoma to late adenoma [64]. First, KRAS mutations are found in 50% of adenomas greater than 1cm, but only in 10% of adenomas less than 1cm. This observation suggests that knocking out Ras might be required for the tumor to grow more than 1cm, after the formation of ACF [39]. Second, it has been discovered that while cells with only KRAS mutations are hyperproliferating, they do not result in the formation of ACF [60], suggesting

that KRAS mutations are able to accelerate tumorigenesis, but not sufficient to initiate it.

2.3.3 *18q Deletions May Drive Intermediate Adenoma Into a Late Adenoma*

18q deletions are the second most common region lost, and are found in 70% of colorectal carcinomas and 50% of late adenomas, suggesting that this deletion drives the formation of late adenomas [39]. There is debate over which genes are responsible for this shift, although recent candidates include SMAD proteins, DCC , and *cables* [102].

SMAD is a protein involved in the TGF-β (transforming growth factor) pathway. The binding of TGF-β to a TGF-β receptor (TGFR) triggers the phosphorylation of SMAD [67]. Once activated, SMAD moves the nucleus, where it initiates transcription of the cell-cycle inhibiting proteins p21 and p15 (recall that these proteins inhibit Cyclin Dependent Kinase (CDK)s, preventing the formation of CDK:cyclin complexes, which are required for division)[67]. Thus, TGF-β and SMAD normally inhibit cellular division. However, SMAD mutations can prevent the transcription of p21 and p15, leaving the cell free to divide even in the presence of anti-growth signals, such as TGF-β. It has been estimated that 30% of colorectal cancers have SMAD mutations, and it is believed that these mutations drive intermediate adenomas into late adenomas by increasing the rate of cellular proliferation [28].

The second most commonly lost region in FAP patients occurs on chromosome 18q, a region that contains DCC [39, 4]. It was initially believed that DCC is involved in cellular adhesion, as it has significant homology to adhesion molecules [39]. However, more recent research indicates that DCC is involved in cell growth, particularly that of axons [4, 83]. It is believed that DCC inhibits cell growth when it is not bound to its ligand, netrin-1 [4]. Mutations in DCC are believed to prevent the binding of DCC to netrin-1, a process that results in abnormal cell

survival [4]. Even so, recent studies have found that DCC is lost in only about 6% of colorectal tumors, suggesting that it does not play a major role in most colorectal cancers [102].

Cables is a linker protein that increases the tyrosine phosphorylation of CDKs by non-receptor tyrosine kinases such as Src, Abl, and Wee1 [102]. These tyrosine kinases inactivate CDKs by dual phosphorylation at the N-terminal Thr-Tyr sequence in CDKs, inhibiting cell cycle progression [83] . Thus, loss of *cables*, through the 18q deletion, can decrease the concentrations of active tyrosine kinases, increasing the amount of active CDKs, driving the cell to divide. It has be estimated that loss of *cables* occurs in $6-70$ of sporadic colorectal cancers, and that the other allele might be inactivated by hypermethylation [100].

2.3.4 *Loss of p53 Drives Late Adenoma Into a Carcinoma*

Researchers have discovered that FAP patients often have deletion in the small arm of chromosome 17 (i.e. 17p), which contains p53 [39]. Complete loss of p53, through loss of 17p in one chromosome and mutation in the other p53 allele, can prevent the cell from committing apoptosis, even in the presence of significant DNA damage. Even so, it appears that loss of p53, which occurs in 80% of colorectal cancers, is a fairly late event in colorectal tumorigenesis [64, 5]. Loss of p53 is found in 75% of colorectal carcinomas, but is rare in adenomas, suggesting that inhibition of apoptosis is required for an adenoma to develop into a carcinoma (a malignant epithelial tumor) [64, 39]. Furthermore, patients with inherited p53 mutations are not at a higher risk of developing colorectal cancer [46], again suggesting that loss of p53 is not is sufficient for tumor initiation, but is required for transformation from benign tumor to a malignant tumor.

2.3.5 *Other Mutations Required for Metastasis: Possible Role for the PI3K Pathway*

The above research suggests that APC or β-catenin mutations initiate carcinogenesis, and that subsequent mutations in KRAS, SMAD, DCC, and p53 increase cellular proliferation and survival, driving ACF to develop into adenomas and eventually carcinomas, the aptly named adenoma-carcinoma sequence. The final step towards metastasis is believed to occur via the accumulation of a variety of other mutations [56, 64]. One pathway implicated in the metastasis of colorectal cancer cells is the PI3K-Akt pathway. This pathway is turned on when PI3K becomes activated by Ras (which is activated in the presence of growth factors) [83]. Activated PI3K catalyzes the addition of a phosphate group to PIP_2, converting it to PIP_3 [67]. PIP_3 in turn recruits kinases which phosphorylate and activate Akt and Rho/Rac/Cdc42 [67, 71, 83]. Akt is able to phosphorylate and inactivate several cell cycle inhibitor proteins (p21,p27,MYT1,GSK3, and FOXO) and pro-apoptotic proteins (Bad, FasL,caspase 8, and FOXO), thus encouraging division and discouraging apoptosis [83]. Activation of Rho/Rac/Cdc42 changes actin and myosin, altering the shape of a cell and inducing mobility, resulting in creeping "ameboid" movement, which could play an important role in invasion and metastasis [83].

Even though Ras plays a key role in activating the PI3K-Akt pathway, KRAS mutations may not necessarily result in the activation of Akt or Rho/Rac/Cdc42. This is because there is a regulatory protein, PTEN, in the PI3K-Akt pathway [67]. PTEN removes a phosphate from PIP_3, converting it back to PIP_2, thus preventing the activation of Akt and Rho/Rac/Cdc42 [67]. In normal cells, concentrations of PTEN are high in the absence of growth factors, and so Akt and Rho/Rac/Cdc42 remain inactive[67].

Recent research has revealed that PTEN is silenced via promoter methylation in 82% of Indian patients with sporadic metastatic colorectal cancer [114], and that PI3KCA activating mutations are found in 32% of colorectal cancers [120].

Furthermore, PI3KCA mutations are only found in 2% of pre-malignant tumors, but in 32% of colorectal tumors [120]. Taken together, this evidence suggesting that PI3KCA mutations and/or PTEN silencing arises late in tumorigenesis, just before invasion and metastasis [120].

2.3.6 *COX-2 and Angiogenesis*

It has been discovered that COX-2 is over-expressed in 43% of adenomas and 86% of carcinomas [32]. It is believed that over-expression of COX-2 results in an increased production of prostoglandin E2, which regulates proliferation, survival, migration, and invasion of colorectal tumors [48]. It is also believed that over-expression of COX-2 induces the production of VEGF and FGF, both of which are involved in angiogenesis [44]. COX-2's role in angiogenesis is supported by the observation that homozygous deletion of COX-2 impairs the growth of tumors and reduces tumor vascularity [148]. Thus, the over-expression of COX-2 may accelerate the rate of angiogenesis, allowing the tumor to grow beyond 2mm.

2.3.7 *How Many Mutations?*

It has been estimated that the human genome contains more than 100 tumor suppressor genes and oncogenes [67]. However, the above data suggests that only a handful of mutations are frequently found in colorectal cancer, both inherited and sporadic. Given that at least $4-5$ mutations are required for carcinoma formation, and at least one for metastasis, it has been estimated that a metastatic tumor may develop after the accumulation of a minimum of $6-7$ independent mutations [39, 64]. More recent genome wide sequencing studies have discovered that colorectal tumors have an average of 80 mutations, but estimate that less than 15 of these actually drive tumorigenesis [149, 72].

2.3.8 *Genomic Instability And Tumorigenesis*

As noted in Chapter 1, it is argued that, given the low human mutation rate ($\sim 10^{-9}$ per cell generation), it is unlikely that all of the genes required for tumorigenesis can be knocked out within a human lifetime. As such, genomic instability is often invoked to explain how all of the required genes could be "hit" by mutation [78, 77, 50]. Through the study of HNPCC and FAP patients it has been discovered that patients with colorectal cancer do indeed exhibit genomic instability.

Microsatellite instability (MSI) has been called the "hallmark" of HNPCC [4]. Defective MMR enzymes, particularly MLH1, MSH2, MSH6, and PMS2, are common in these patients. The loss of DNA replication fidelity (i.e. insertion or deletion of microsatellites), due to defects in MMR genes, have been reported to increase the mutation rate of HNPCC patients by $2-3$ orders of magnitude [7, 124, 36]. Microsatellite Instability (MSI) does not appear to be limited to HNPCC patients either, as it is found in $13-20\%$ of sporadic colorectal cancers [4, 64]. MLH1 appears to be the primary MMR gene affected, as it is methylated in 80% of sporadic colorectal cancers with MSI [4]. Tumors with MSI also often have frameshift mutations in the microsatellite region of the TGFR gene, making the cell immune to the growth-suppressing effects of TGF-β[64]. MSI tumors also sometimes have defects in HLA genes, possibly resulting defective MHC proteins that allow the tumor cell to evade elimination by the immune system [14].

While MSI is characteristic of HNPCC, Chromosomal Instability (CIN) is regarded as the hallmark of sporadic cancers, as it is observed in $65-70\%$ of such cases [102]. It is also observed that these CIN tumors do not usually have MSI or higher mutation rates [7, 36], leading some to suggest that tumors only require one type of genomic instability [64]. Even though CIN tumors do not have a higher mutation rate, they do exhibit losses or gains of entire chromosomes and

a high frequency of Loss of Heterozygosity (LOH) , which can lead to complete deactivation of tumor suppressor genes [102].

The three common methods of generating CIN involve chromosome segregation defects, telomere dysregulation, and defects in DNA damage responses [102]. In the case of colorectal cancer, the genes involved in chromosomal segregation defects are kinetochore proteins involved in the spindle checkpoint (i.e. hZw10, hZwilch/ FLJ10036, and hRod/KNT), as well as Ding, a protein that is essential for proper chromosome disjunction [145]. Plk1, which regulates entry into mitosis and centrosome duplication, is mutated in up to 63% of colorectal tumors [102]. It has also been reported that $77-90\%$ of colon cancer cells have shorter telomeres than the normal surrounding cells, suggesting that telomere dysfunction might also be playing a role in generating CIN in colorectal cancer cells (reviewed in [102]). Only p53 has been directly implicated DNA damage response defects in colorectal cancer, and may play a permissive role for developing CIN, likely by letting CIN cells survive despite having severe genomic abnormalities [102].

Some research suggests that APC may also be involved in generating CIN. This suggestion comes from the observation that APC also plays a role in cytoskeletal regulation, and it has the ability to bind spindle microtubules and centrosomes [102]. However, further investigation has revealed that the genomic instability in mice with these APC mutations is quite different that that observed in actual tumors [102]. This finding suggests that APC mutations can cause genomic rearrangements, but that these are not consistent with the CIN observed in tumors, leading Pino and Chung to conclude that the role of APC in CIN is "provocative but incompletely defined".

Given that CIN is so common in colorectal cancers, the next question is whether or not CIN initiates tumorigenesis, or simply exacerbates it. Several studies have demonstrated that CIN does indeed occur very early in the adenoma-carcinoma sequence, as CIN is frequently observed in adenomas [125]. Once such study

found that polyps less than 2mm exhibited CIN on chromosomes 5q, 1p, 8p, 15q, and 18q, regions that contain many of the genes frequently mutated/lost in colorectal cancer. Mathematical modeling also suggests that CIN initiates colorectal cancer, although it is difficult to find experimental evidence to support this model [85, 97, 125]. Furthermore, currently there id no data that directly connects CIN to the acquisition of specific mutations frequently observed in colorectal cancer, making it difficult to prove that CIN initiates tumorigenesis[102]. Thus, the debate over whether CIN initiates tumorigenesis or simply accelerates it via LOH continues [102, 126].

While many consider genomic instability a requirement of tumorigenesis, others believe that cancer can develop without such instability [9]. These authors support their argument by noting that not all colorectal cancers exhibit CIN, and that MMR mutations generally occur after APC mutations. Furthermore, these authors argue that selection can drive the mutation to spread within the crypt, and that selection would thus override mutation as the primary evolutionary force driving tumorigenesis [9]. These authors also note that mutations in critical genes, such as APC and p53, are not truly recessive, and would provide a selective an advantage after a single mutation. Thus, an alternative hypothesis might be that MMR and CIN do not play a critical role in tumorigenesis, but are simply the result of mutations that provide the cell with a selective advantage. This hypothesis is also in line with the observation that the aneuploidy, a result of CIN , can sometimes inhibit tumor progression [102]. This might be interpreted as CIN actually providing the cell with a selective disadvantage, leading to the selection of cells that do not exhibit CIN.

2.3.9 *Cancer Stem Cells*

Given that transit cells are rapidly sloughed off, it is frequently argued that colon crypt stem cells accumulate the mutations necessary to convert them into cancer

stem cells [4, 102]. Cancer stem cells, which make up $0.25\% - 2.5\%$ of the cells in a tumor, are defined as cells that have the ability to self-renew, perpetuate themselves for long periods, maintain the ability generate a variety of differentiated cells, and have the ability to generate tumors when transplanted into other tissue [102]. However, there is now some evidence suggesting that transit cells may undergo mutation and selection that enables them to linger in the crypt, giving them time to accumulate the extra mutations required for tumorigenesis [56, 70]). While there is no direct evidence to support this, such a process could explain why colorectal tumors are often composed of differentiated cells [56].

CHAPTER 3 JC VIRUS AND COLORECTAL CANCER

3.1 OVERVIEW

The last ten years have revealed that a common infection, the JC polyomavirus, is frequently associated with human colorectal tumors. JCV expresses several viral oncoproteins that interfere with key cellular pathways, and is known to cause cancer in lab animals. Both observations have led many to investigate if JCV is involved in colorectal tumorigenesis. The results from these studies have generated a body of intriguing evidence implicating a role for JCV in colorectal tumorigenesis.

3.2 POLYOMAVIRUSES AND TUMORS

The past $10-15$ years have revealed that a common infection, JC Virus (JCV), has tumorigenic potential and is frequently associated with a variety of tumors, including colorectal tumors[29]. Such oncogenic potential likely results from JCV's ability to interfere with many of the same pathways that are disrupted in the mutation hypothesis (see Chapter 2). JCV, belongs to the family of polyomaviruses, which, prior to 2000, were grouped with the papillomaviruses (such as HPV) under the family of papoviruses [47]. Polyomaviruses are named for their well known oncogenic abilities; *poly* is Greek for many, while *oma* is Greek for tumors, together meaning "many tumors" [82]. There are five human polyomaviruses: JC Virus (JCV) , BK Virus (BKV) (both discovered in 1971), Karolinska Institute virus (KIV), Washington University virus (WUV) (both discovered in 2007), and

Merkel Cell polyomavirus (MCV; discovered in 2008) [47]. JCV, BKV, KIV and WUV are all closely related to the non-human primate polyomavirus simian virus 40 (SV40), whose gene products are frequently used in the lab to induce tumors, illuminating which pathways are often dysregulated during tumorigenesis [47]. MCV, a close relative of JCV, is linked to the rare but aggressive Merkel Cell skin cancer (MCC) (reviewed in [47]). Some of the evidence involved in this revelation includes the discovery that MCC patients have MCV titers that are 59 times higher than controls [101, 139]; MCV is found in Merkel cell tumors, with an average copy number of 5.2 MCVs per tumor cell [47]; and MCV's interacts with several key proteins, such as pRb, Hsc70, and PP2A [47]. Interestingly, even though MCV is the cause of the rare MCC, it is a common virus, as 88% of adults without MCC are seropositive for the virus [101].It is also widespread throughout the human body [47]. The explanation for this pattern is that MCV is only reactivated in the elderly or immunocompromised individuals, and that Merkel cells may be especially susceptible to transformation by MCV [47]. As discussed below, this is interesting because JCV, and MCV share many characteristics and epidemiological patterns.

While JCV is associated with many tumors, it is most commonly known as being the etiologic agent behind the fatal Progressive Multifocal Leukoencephalopathy (PML). JCV is able to infect the oligodendrocytes of the brain. If reactivation occurs, the virus becomes lytic, leading to demyelination and cytolytic destruction of the oligodenroglia, resulting in PML [69, 82]. It is believed that immunosuppression is primarily responsible for the reactivation of JCV. This is supported by evidence that $5\% - 8\%$ of AIDS patients develop PML, as they are severely immunocompromised [6]. Finally, with regards to JCV's association to tumors, it is interesting to note that the unusual astrocytes associated with PML are indistinguishable from tumor cells in high-grade glial neoplasia [47].

JCV is a very common DNA virus, occurring in 45%-80.5% of adults [144, 18]. JCV is such a common virus that it can be found in all human populations [133], suggesting that it has been with humans throughout our evolution, lending itself to studies of human migration patterns [66]. However, as might be expected, there are several different types and subtypes of JCV found in different regions of the world, each identified by polymorphisms in their IG region: A (EU) is almost exclusively found in Europe, B (Af2, Af3, B2, MY, SC, B1, CY) is most common in Africa and Asia, while C (Af1) is only found in a few regions in African (subtypes found in parenthesis) [49, 133]. The prevalence of JCV also varies around the world: JCV DNA shed in urine samples was, on average, found in 13.8% of samples in Europe, 11.85% of samples in Asia, and 8.9% of samples in Africa [133].

Seroprevalence studies for JCV antibodies show that seroprevalence increases with age. In a study of 2,435 individuals in England and Wales, ranging in ages from 1-69 years old, Knowles et al. [68] found that 11% of children under 5 years of age are seropositive for JCV's VP1 capsid protein, but that prevalence rises throughout life, reaching 50% in the 60-69 year old age group [1]. In a similar study, Viscidi et al. [144] examined the serum of 947 individuals attending out-patient clinics in Rome, ranging in age from 1-93 years old. Like Knowles et al. [68], the authors found that JCV seroprevalence increases with age: the seroprevalence of individuals 10 years of age was only 9.5%, but jumped to 50% in individuals 10-20 years old [144] . Seroprevalence reaches 68.8% by the time individuals are 40-49, and maxes out at 80.5% in individuals older than 70 years of age [144] . Finally, Egli et al. [33] examined 400 blood donors in Switzerland for the presence

1 Seroprevalence values tend to be higher than prevalence values detected from urine samples, as described above: In the study conducted by Egli et al, the seroprevalence for JCV was 58%, while JCV DNA was only found in 19% of urine samples [33]

of JCV antibodies, finding that 58% of individuals had IgG antibodies for JCV but no IgM antibodies [33] .

The above studies reveal several aspects of JCV. First, the increasing seroprevalence and spike between the ages of 10-20 suggests that JCV is most frequently acquired during childhood and adolescence. Second, the absence of IgM (the first antibody produced during novel infection) and presence of IgG (which is characteristic of persistent infections) indicates that JCV is a chronic infection [90]. This conclusion is buttressed by the observation that the same strain of JCV can be found an individual's urine sample, taken several years apart from one another, indicating the individual is persistently shedding the same strain, as opposed to being reinfected by a different strain [65].

The above prevalence patterns have led researchers to conclude that transmission of JCV requires close contact, and that JCV is likely transmitted among family members [68]. This conclusion is supported by a study finding that, in Tokyo, there is no evidence for different JCV genotypes spreading between the local American and native Japanese populations [61]. Since JCV is frequently found in urban sewage samples, many authors have concluded that the virus is spread by consuming contaminated water and/or food, or by coming into contact with contaminated surfaces (i.e. clothes, countertops, eating utensils, etc...) (reviewed in [47]). Once ingested, JCV infects the tonsillar tissue, where it is frequently detected [47, 87]. However, JCV is also frequently found in the bone marrow and B lymphocytes, which may help the virus spread to other tissues (i.e. colon, brain, etc...) via the circulatory system [47, 86]. It is believed that JCV eventually infects the kidneys, where it establishes a persistent infection, leading to the frequent shedding of virus via the urine [47].

The circular dsDNA genome of JCV is 5130bp long and is encased within a non-enveloped 72 pentamer capsid [47]. JCV's genome is evenly divided into early and late regions, each with lengths of 2.4 kb and 2.3 kb, respectively [31]. The early and late regions are separated by a Non-Coding Regulatory Region (NCRR) that contains the origin of replication and transcriptional control elements, and usually contains two 98bp tandem repeats that serve as enhancers [47, 31]. The early region encodes the Large T Antigen (T-ag) (a.k.a LT), the Small T Antigen (t-ag) (a.k.a. ST), as well as several T' antigens (T'165, T'136, and T'135) which are expressed from alternately spliced early transcripts [141]. Its noteworthy that these early proteins are named tumor antigens because they were originally detected using antibodies from animals with tumors [47]. The late region encodes the viral capsid proteins VP1, VP2, and VP3, as well as agnoprotein [47].

It is believed that JCV gains entry into the target cell by using its VP1 protein to bind the cell's GT1b 5HT2AR serotonin receptor [34]. After attachment, JCV enters the cell via clathrin-dependent endocytosis, a process in which the virus is internalized through the inward budding of the plasma membrane, forming clathrin coated pits [47, 82]. JCV is then delivered to endosomes and caveosomes, facilitating the movement of the virus to the endoplasmic reticulum, where viral un-coating is occurs [82]. Finally, JCV is translocated to the nucleus [82].

Once inside the nucleus, the transcription factors AP1, NF-1, NF-κB, NFAT, and YB-1 bind to JCV's promoter region, leading to the transcription of early region mRNAs, and eventually translation in the cytoplasm [82]. After translation, T-ag accumulates to high concentrations and initiates cellular division by inhibiting pRb (see below for more details). T-ag next binds to the origin of replication in the NCRR , unwinds the viral DNA, and hijacks the hosts DNA polymerase to replicate the viral DNA [31, 82]. Eventually, T-ag suppresses early gene transcription and initiates transcription of the late viral genes, agnoprotein, VP1, VP2,

and VP3 [82]. Agnoprotein is believed to associate with T-ag help regulate viral replication [116, 59], while the latter three proteins are assembled together in the cytoplam and translocated to the nucleus, where viral encapsidation takes place [47].

What happens after virion assembly appears to vary. Usually polyomaviruses spread from cell to cell by lysing their host cell [47]. However, electron microscopy studies have demonstrated that virions can be secreted from the plasma membrane of intact cells, suggesting that lysis is not always required for cell to cell transmission [57, 22]. Furthermore, transformed cells may have some viral DNA integrated into its host genome (not all of the genome has to be integrated because viral replication is not required to sustain the tumor) [111]. In fact, such integration may drive transformation because the absence of viral replication and lysis, but expression of T-ag , can promote cell proliferation and tumorigenesis [29]. With regard to colon cells, infected cells start to lose viral DNA soon after infection, and is only detectable up to 21 days after infection, suggesting that infected colon cells are more susceptible to transformation than lysis [111].

After successful infection, JCV remains with the host for the remainder of their life. This is supported by the observation that individuals that are positive for JCV excrete the same strain, have low levels of IgM and high levels of T-ag IgM antibodies (see Section 3.3). During this period, healthy immunocompetent individuals do not exhibit any specific symptoms even though they have low levels of viral gene expression and sporadic reactivation, a phenomenon that is observed in 0.5-20% of individuals [31]. However, if the individual becomes immunocompromised JCV can become completely reactivated, leading to an increase in virus titers and disease [31, 47].

3.5 JCV ONCOPROTEINS AND THE HOST PROTEINS THEY MANIPULATE

JCV increases its fitness by manipulating many host proteins involved in the cell cycle [47], leading to an increased number of replications for both JCV and the infected cell. Furthermore, JCV has the ability to increase its survival by inhibiting innate immune signaling [47]. As might be expected, these processes increase the probability of oncogenic transformation. Finally, it is noteworthy that many of the proteins JCV interacts with are the same proteins that are mutated in the mutation model. The following section will review the plethora of proteins JCV interacts with.

3.5.1 *Large T-Antigen*

The large T antigen of JCV plays a critical role in driving viral replication and transforming cells [47]. Lab experiments demonstrate that infected cells expressing T-ag often become immortalized, can escape contact inhibition, and exhibit anchorage-independent growth (reviewed in [47]). However, T-ag alone is sufficient to induce immortalization; only the combination of T-ag and hTERT (the active unit of telomerase) is sufficient to bypass senescence and cell crisis, resulting in an immortalized cell (reviewed in [47]). JCV's T-ag induces these phenotypic changes by interacting with several proteins, including β-catenin, pRb, p53, p300/CBP, IRS-1, Nbs1, and Bub1.

3.5.1.1 *β-catenin*

Several studies have demonstrated that T-ag is able to bind β-catenin [111, 35, 45]. In particular, it has been determined that T-ag residues from 412-688 of T-ag bind the 695-781 residues (C-terminal) of β-catenin [45]. Experiments show that expression of T-ag increases the level of β-catenin within the cell [45]. T-ag is also able to stabilize β-catenin, possibly by preventing β-catenin from binding to

the APC/CK1/GSK-3β destruction complex [45]. Subsequent experiments have shown that cells expressing T-ag also have increased levels of T-ag and β-catenin in the nucleus, while cells not expressing T-ag only have β-catenin in the cytoplasm [45]. All together, these experiments demonstrate that T-ag is able transport β-catenin into the nucleus [45]. Once in the nucleus, β-catenin increases the transcription of Myc, driving cell division [35, 67, 45]. The interaction of T-ag and β-catenin is significant because the interferes with the Wnt pathway much like Adenomatous Polyposis Coli (APC) mutations. As it has been argued that APC mutations initiate tumorigenesis, the finding that T-ag stabilizes β-catenin, the target of APC and the center of the Wnt pathway, suggests this may be a mechanism by which T-ag can initiate tumorigenesis.

3.5.1.2 *pRb*

Like other tumor viruses, JCV's T-ag can interact with pRb [47, 91]. In fact, JCV not only interacts with pRb itself, but also with other retinoblastoma proteins, p103 and p107[47]. It has been determined that the N-terminal domain of T-ag (which contains the LXCXE motif) is responsible for binding to the pRb family of proteins [54]. Such binding of the pRb family proteins disrupts replication inhibition, releasing E2F proteins, driving the cell to divide [47]. Furthermore, mice with knocked out pRb, p107 and p103 are unable to halt the cell cycle in G1, even in the presence of limited resources, contact with other cells (i.e. loss of contact inhibition), and DNA damage [26, 117]. Thus, via its interaction with pRb proteins, T-ag is capable of driving the cell into S phase, even under conditions when replication would be normally prevented.

3.5.1.3 *p53*

Most viruses, including JCV, are able to inactivate the pro-apoptotic p53 protein. T-ag accomplishes this by binding to p53's core DNA binding domain, thereby inhibiting p53's ability to act as a transcription factor [47, 122]. T-ag and p53

expression are positively associated, while T-ag and p21 are inversely related, suggesting that inhibition of p53 also decreases the amount of p21, thus independently allowing for phosphorylation of pRb, driving the cell to divide [96]. The power of the relationship between pRb and p53 is increased by the fact that T-ag's inhibition of pRb drives the cell to divide in the presence of DNA damage, and inactivation of p53 prevents apoptosis or senescence in the presence of such damage [47]. Another important role of p53 is its ability to inhibit angiogenesis, especially in tumors [83]. Thus, inactivation of p53 may drive the cell to divide, avoid apoptosis, and possibly prevent inhibition of angiogenesis. Given that p53 is mutated or deleted in 50% of human cancers [53], including colon cancer, inactivation of p53 by T-ag is important in that it provides an alternative mechanism by which this crucial tumor suppressor can be removed.

3.5.1.4 *p300/CBP*

While most studies suggest that T-ag inactivates p53, there are handful indicating that T-ag may also stabilize p53 (reviewed in [47]). While the exact results are unclear, it is hypothesized that such stabilization of p53 may T-ag the ability to interact with p300/CBP [13]. CBP/p300 are proteins that act as adapters or co-activators by using their acetyltransferase activity [47]. However, it turns out that T-ag cannot bind p300/CBP in the absence of p53 [75]. Thus, T-ag may stabilize p53 so as to gain access to p300/CBP [13]. It is argued that T-ag binding to p300/CBP, using p53 as an adaptor, can result in the production of Myc, driving the cell to divide [128]. It has also been suggested that T-ag uses p300/CBP to activate the promotors that are normally inhibited by pRb, such as E2F, again initiating replication [47]. Although how exactly T-ag acts on p300/CBP is unclear [47], one might speculate that T-ag's interaction with p300/CBP inhibits apoptosis. Normally, p300/CBP acetylates p53, which is accompanied by the removal of phosphates in the regulatory region of p53 [83]. The Ser residue in the transactivating region of p53 is then free to be phosphorylated, a process that can trigger

apoptosis [83]. *If* T-ag inactivates p300/CBP (again, this is unknown), the protein would not able to acetylate p53, leaving the phosphates in the regulatory region intact, thus preventing apoptosis.

3.5.1.5 *IRS-1*

Insulin receptor substrate-1 (IRS-1) is docking protein normally found in association with an insulin growth factor 1 receptor (IGF-1R) and the plasma membrane [83, 47]. However, T-ag has the ability to translocate IRS-1 from the cytoplasm to the nucleus [106], resulting in cell division, inhibition of apoptosis, and induction of genomic instability. IRS-1 initiates division and inhibits apoptosis via its activation of the PI3K/Akt pathway, which down-regulates cell cycle inhibitors (p21,p27,MyT1,GSK3,and FOXO) and pro-apoptotic molecules (Bad, FasL,caspase 8, and FOXO) [83, 47]. Furthermore, nuclear IRS-1 has been found bound to β-catenin, resulting in increased transcription of Myc and cyclin D, thereby increasing cell growth [20] . Nuclear IRS-1 has also been found in a complex with Rad51, which is the main enzymatic component of homologous recombination directed DNA repair (HRR) [137]. It appears that T-ag, through its interaction with IRS-1, impairs HRR, which might be compensated for by an increase in non-homologous end joining (NHEJ), an alternative DNA repair process [110]. However, this compensatory action of NHEJ is associated the accumulation of spontaneous mutations [140], increasing genomic instability.

3.5.1.6 *Nbs1*

Nbs1 is a component of the MRN (Mre11, Rad50, Nbs1) complex, and plays an important in DNA repair and detection of double strand breaks (reviewed in [47]). It is believed that T-ag binds to Nbs1, and may result in chromosomal instability, although the particular result of the interaction is largely unknown [47]. This hypothesis comes from the observation that Nbs1 is mutated in the Nijmengen breakage syndrome, which is associated with CIN and an increased

risk of cancer [47]. Finally, it has also been suggested that LT's binding to Nbs1 may allow for increased replication of JCV DNA [150].

3.5.1.7 *Bub1*

Bub1 is a mitotic checkpoint kinase and is critical in maintaining genomic integrity [25]. T-ag is able to bind Bub1, leading to a compromised spindle checkpoint [25]. Mice with reduced expression of Bub1 have increased tumorigenesis and aneuploidy (reviewed in [47]), suggesting that T-ag's binding to Bub1 may induce the Chromosomal Instability (CIN) that is so characteristic of colorectal tumor cells.

3.5.2 *Small t-Antigen*

While T-ag is primarily a nuclear protein, t-ag is found in both the nucleus and cytoplasm [47]. Like its big brother, t-ag has the ability to induce cellular proliferation, even in the absence of T-ag [47]. Microarray analyses have shown that t-ag can alter many genes involved in proliferation, apoptosis, integrin signaling, and immune responses [88]. Most of these alterations can be traced to t-ag's ability to bind and inactivate the serine–threonine protein phosphatase PP2A [47]. PP2A inactivation leads to stabilization of Myc and has a similar effect as PI3K, leading to increased rate of division, inhibition of apoptosis, and possibly increased mobility [152, 151]. That t-ag has a similar effect as PI3K is significant because PI3K is normally activated by Ras, and so may have a similar effect as mutated Kristen Rat Sarcoma Virus (KRAS).

t-ag's inhibition of PP2A also leads to the activation of several kinases, including MAPK, Akt, and PKCζ (reviewed in 47). Activation of MAPK (which normally requires growth factors) initiates the production or activation of several transcription factors (i.e. Ets, Jun, Fos, Myc, and E2F) that drive the cell to divide [67]. As discussed in Section 2.3.5, activation of Akt results in the inhibition of

apoptosis. PKCζ is involved in activating NFκB, a protein that increases inflammation, stimulates cell division, and inhibits apoptosis [67, 83]. It is noteworthy that several other viruses associated with tumors, either directly or indirectly (i.e. hepatitis C virus, herpes simplex virus, HIV, and human T-cell leukemia virus), also up-regulate NFκB [83].

3.5.3 *Agnoprotein*

Agnoprotein is produced late in the viral lifecycle and is primarily found in the cytoplasm [27]. In addition to its regulatory role in viral transcription and translation, Agnoprotein has the ability to circumvent the cell cycle checkpoint, resulting in an accumulation of DNA damage [27]. While some of this many be the result of Agnoprotin's ability to bind T-ag and p53, the primary mechanism by which Agnoprotein generates genomic instability is through its binding of Ku70 and Ku80, DNA repair proteins involved in non homologous end joining (NHEJ) DNA double strand break repair [27]. This has been demonstrated in an experiment in which cells expressing agnoprotein were treated with cisplatin, a DNA damaging agent. It was found that cells expressing Agnoprotein had significantly lower levels of Ku70 and Ku80 than controls (which did not express agnoprotein), resulting in aneuploidy [27]. The authors concluded that Angoprotein's localization of Ku70 to the perinuclear region permitted evasion of the cell cycle checkpoint, leading to the accumulation of DNA damage and CIN [27].

3.6 IN THE LAB: JCV AND TUMORIGENESIS

Given JCV's interaction with several key tumor suppressors and DNA repair proteins, one might expect that JCV will have the ability to induce tumors. Indeed, lab experiments have demonstrated that JCV is capable of transforming cells in culture as well as in laboratory animals. Transgenic experiments involve

the insertion of an exogenous gene into the genome of a living organism. It has been demonstrated that the expression of T-ag in intestinal enterocytes results in hyperplasia, and eventually dysplasia (reviewed in 47). Furthermore, transgenic mice that express T-ag and t-ag develop adrenal neuroblastomas, neuroectodermal tumors, pituitary adenomas, and malignant peripheral nerve sheath tumors (MPNST) [82, 109]. Finally, it is noteworthy that T-ag positive cells eventually lose expression of T-ag but maintain their transformed phenotype, suggesting JCV may induce tumorigenesis by some sort of "Hit and Run" mechanism [109].

JCV not only has the ability to induce tumors in transgenic models, but injection of JCV is oncogenic in several animals, including hamsters, rats, and non-human primates [82]. JCV infection of newborn Syrian hamsters results in the development of several different tumors, including medulloblastomas, primitive neuroectodermal tumors, astrocytomas, glioblastoma multiforme, and peripheral neuroblastomas [109]. It has also been demonstrated that hamsters innoculated with the Mad-1 strain of JCV develop medulloblastomas, while those infected with the Mad-4 strain develop pineocytomas and medulloblastomas, demonstrating that different strains can cause tumors in different cell types [98]. In the case of rats, injection of the JCV Tokyo-1 strain into the brain results in undifferentiated neuroectodermal tumors in 75% of infected animals, some of which remain oncogenic when transplanted into other rats [109]. Finally, owl monkeys and squirrel monkeys infected with JCV develop astrocytomas, glioblastomas, and neuroblastomas by 16–24 months of age [109].

All of the above animals are non-permissive for JCV infection, which may make them more susceptible to transformation, presumably because they integrate JCV DNA into their genome and are thus unable to lyse the cell, decreasing their intra-host transmission [111]. However, an equally important finding is that colonic cells infected with JCV start to lose JCV DNA 14-21 days after infection, suggesting that colonic cells are also non-permissive to JCV infection and thus more susceptible to transformation [111].

The above lab experiments demonstrate that JCV has the capacity to be tumorigenic in lab animals. However, there is considerable debate over whether or not JCV is involved in human cancers. Even so, there is intriguing evidence that JCV could indeed play such a role. Part of the argument comes from the observation that 70% of colorectal cancers are caused by "chance and environment" (mutation and environment), while 5% are from inherited mutations [9]. Similarly, it has also been estimated that ~25% of colorectal cancers result from multifactorial contributions of different risk factors [9]. While the authors argue that these 25% of cases occur as the result of inheriting many rare dominant alleles that have low penetrance, but together increase the risk of colorectal cancer [9], an alternative hypothesis might be that infection is one of those critical environmental factors that accounts for increased risk to cancer. The hypothesis is based upon the observation that JCV is frequently associated with many cancers, including human brain tumors, lung cancer, esophageal cancer, gastric cancer, (reviewed in [76]), and at least five independent laboratories (and several studies conducted by each lab) have detected both JCV DNA and T-ag in colorectal tumors [69, 35, 55, 76, 96].

In 1999, the Laghi laboratory used semi-quantitative PCR to detect the presence of LT in the mucosa of colorectal tumors as well as adjacent tissue [69]. The authors detected T-ag DNA in 89% of 25 healthy colorectal cells, 25 colorectal cancer cells, and 4 cancers, indicating that JCV is present in both healthy and malignant tissue. However, subsequent semi-quantitative PCR revealed that the JCV viral load is ten times higher in cancer tissue that in the adjacent healthy tissue, suggesting that JCV is more active in tumor cells. However, the viral load in these tumor cells is only 0.1 JCV viral copies per human genome [69]. Rollison [115] argues that there should be at least 1 viral copy per human genome, and so the results of Laghi et al. [69] do not indicate that JCV is driving tumorigenesis.

Two follow-up studies by Ricciardiello et al. were conducted to shed more light on the association between JCV and colorectal tumors. Like Laghi et al. [69], the authors used JCV T-ag specific PCR primers to detect T-ag in 81.2% of normal healthy colorectal tissues (the use of JCV specific primers is important because it allowed the authors to avoid amplifying other polyomavirus T-ag sequences)[112]. Further investigation revealed that only the Mad-1 strain of JCV , which is characterized by two 98bp deletions in the NCRR, is found in healthy and malignant colorectal tissue [112]. This is significant because it indicates that the circulating archetype strain is unable to infect colorectal tissue, which was confirmed by the finding that the archetype strain was absent in all samples [112]. The authors offer several hypothesis about why only the Mad-1 strain is found in colorectal tissue: 1) genomic rearrangements resulting in the Mad-1 strain might occur in non-lymphoid tissue, and then the Mad-1 strain uses lymphocytes to infect the colon; 2) genomic rearrangements may occur in the colon, giving the Mad-1 strain the ability to proliferate in colorectal tissue; and 3) Mad-1 may be a circulating strain that has the ability to infect colorectal tissue [112]. The following year, 2001, further revealed that a variant of the Mad-1 strain, which lacks one of the 98bp repeats, is found exclusively in colorectal tumors, but is absent in the adjacent healthy tissue [113]. Given that higher viral loads are found in colorectal tumors, the authors suggest that the Δ98 Mad-1 strain is more efficient at proliferating in colorectal tissue. The authors also suggest that this strain of JCV might be involved in the generation of CIN [113]. Using the information available, the authors hypothesize that transformation by the Δ98 Mad-1 JCV strain might occur through two mechanisms: 1) Mad-1 has a selective advantage in colorectal tissue, but some impairment of the immune system might select for the Δ98 variant, which has the ability to transform cells; 2) Mad-1 integrates into the human genome, and pre-existing genomic instability results in the Δ98 variant that is capable of transforming cells [113].

In 2002, the Enam lab used PCR, microdissection, and immunohistochemistry to detect JCV DNA and proteins in 27 colonic tumors [35]. The authors detected early region DNA in 81.5% of samples, Agnoprotein DNA in 59.2% of samples, and VP1 DNA in 14.8 % of samples[35]. Immunohistochemisty, which detects proteins using specific antibodies, found the expression of T-ag in 62.9% of samples, Agnoprotein 44.4% in samples, but no VP1 protein in any samples [35]. This is significant, because the lack of VP1 protein suggests that **JCV!** (**JCV!**) is unable to replicate productively in these tumor cells. Subsequent laser capture microdissection, which is capable of isolating specific cells, was conducted on normal mucosa, precancerous adenomas, and invasive adenocarcinomas so as to verify the presence JCV DNA and proteins in these tissues. Gene amplification revealed that early JCV DNA and T-ag protein are found in both precancerous adenomas and invasive adenocarcinomas, while only a "weak signal" of JCV DNA is found in the adjacent healthy tissue [35]. These results suggest that JCV is only found a few healthy tissues, but at higher concentrations in colorectal tumors, where T-ag is able to interact with key host proteins (i.e. p53 and pRb) [35]. Furthermore, the presence of T-ag and JCV DNA in pre-cancerous adenomas and invasive adenocarcinomas suggests that JCV could potentially be involved in initiating tumorigenesis, again by dysregulating key pathways such as the Wnt pathway, apoptosis, and cell cycle regulation.

In a similar study, Hori et al. used nested PCR, Southern Blot, and immunohistochemistry to detect the presence of T-ag, Agnoprotein, and VP proteins in 23 colorectal adenomas and 20 healthy colonic mucosa from Japan. The authors detected T-ag in 26.1% of colorectal cancers, 4.8% of adenomas, and in 0% of healthy colonic mucosa [55]. Consistent with the findings of Enam et al. [35], VP1 was not detected in any samples, but unlike Enam et al. [35], the authors were unable to detect Agnoprotein in any samples [55]. The absence of VP1, which indicates JCV is not actively replicating, suggests that JCV may integrate early

DNA in the the host genome, and that subsequent expression of T-ag is involved in tumorigenesis [55].

That same year, 2005, Theodoropoulos et al. [138] used PCR to detect the presence of JCV DNA in Greek adenomas and adenocarcinomas, and real-time PCR to determine the levels of expression. Similar to previous studies, PCR detected JCV DNA in 61% of adenocarcinomas and 60% of adenomas[138]. Also like Laghi et al. [69], real-time PCR detected a viral load of 9×10^3 to 20×10^3 copies/µg DNA in adenocarcinomas and adenomas, but only 50-450 copies/µg DNA in healthy tissue [138]. The finding that JCV viral load is much higher in cancer tissue suggests that the higher concentration of JCV increases the risk of cancer. The authors conclude that JCV is likely to be involved in initiating tumorigenesis, possibly by inducing chromosomal instability [138].

In 2008, Lin et al. [76] also used PCR and immunohistochemistry to detect JCV DNA, T-ag , and VP1 in 22 colrectal tumors from Taiwanese patients . Similar to previous studies, T-ag was detected in 63.6% of colorectal cancer tissues but not in adjacent healthy tissue [76]. Again, VP1 was not detected in any tissue [76], suggesting that JCV integrates into the host genome.

Many of the studies above studies might be criticized because they lack large sample sizes. However, in 2009 Nosho et al. [96] conducted a large scale study of 766 colorectal cancer samples. The authors used immunohistochemistry to detect levels of p53, p21, β-catenin, COX2, Cyclin D1, and JCV T-ag, as well as whole-genome amplification to detect Loss of Heterozygosity (LOH) in the regions frequently associated with colorectal cancer (i.e. 2p, 5q, 17q, and 18q) [96]. The results show that expression of T-ag is positively associated with expression p53 ($p < 0.0001$), nuclear β-catenin ($p < 0.006$), COX-2 ($p = 0.02$), and loss of p21 ($p < 0.0001$) [96]. The positive association of T-ag and p53, accompanied with loss of p21 (which is activated by p53), strongly suggests that T-ag is able to dysregulate the p53 pathway [96]. T-ag's positive association with nuclear β-catenin reinforces the argument that JCV is able to translocate β-catenin to the nucleus,

thereby disrupting the Wnt pathway. The positive association between T-ag and COX-2 suggests that JCV is able to induce angiogenesis. However, expression of T-ag is not associated with alterations in Ras, PIK3CA, BRAF or cyclin D1 [96]. The authors also found that T-ag is over-expressed in 35% of colorectal cancers, again suggesting that T-ag plays a key role in tumorigenesis. Finally, the authors discovered that T-ag expression is significantly associated with CIN , which was defined as LOH in chromosomes 2p, 5q, 17q, and 18q [96]. This is noteworthy because these are the same regions frequently lost in the mutation hypothesis. While T-ag expression is not significantly associated with patient survival, the authors conclude that T-ag likely contributes to CIN and dysregulation of the p53 pathway, the combined effects of which may result in the uncontrolled proliferation of cancer cells [96].

In 2010 Del Valle and Khalili [29] examined 50 commercially available colorectal samples for immunoreactivity to T-ag. Thirty four percent of those samples were positive for T-ag, and of those 88% were also positive for Agnoprotein, while none expressed VP1 [29]. These results are significant because they are the first to indicate that JCV T-ag can be found in commercially available tissue arrays Del Valle and Khalili [29], suggesting that JCV may be responsible for their transformation. Also, like many previous studies, the absence of VP1 indicates that JCV is incapable of productively infecting tumor cells, but that those cells retain the ability to express T-ag and Angoprotein, promoting cell proliferation and tumor formation [29].

That same year Niv et al. [95] determined JCV titers (using anti-bodies to VP1) in patients undergoing colonoscopy, some of whom had colorectal cancer and others who were healthy . This is an important study because it was the first study to directly compare JCV titers in colorectal cancer patients versus healthy patients (other studies compared tumor tissue to adjacent normal tissue). While the sample size was fairly small (7 adenomas and 11 tumors), the authors observed statistically significant higher titers of JCV in patients with advanced ade-

nomas and tumors, compared to healthy individuals [115]. While the authors found no correlation between T-ag expression and JCV seroreactivity, it was discovered that JCV antibody levels are higher in individuals with more advanced disease, suggesting that immunosuppression and/or JCV reactivation is involved in disease progression [115].

While the above nine studies suggest that JCV is at least present in colorectal tumors, and may drive tumorigenesis, there are a handful of studies that were unable to corroborate those results. A study by Losa et al. [80] was only able to detect JCV DNA in one out of 100 colorectal tumors. Similarly, in 2004, Newcomb et al. [94] screened 45 healthy donors and 233 colorectal cancer patients for JCV DNA. The authors were unable to detect JCV in any of colorectal tumor samples [94]. However, Rollison [115] has noted that these conflicting results are likely to due to differences in assay sensitivity, possible contamination, and differences in JCV prevalence in the populations examined . Indeed, Newcomb et al. [94] have been criticized for using novel primers (i.e. those not used in the positive studies), as well as for not using any positive controls to verify that the primers worked in their formalin-fixed paraffin-embedded samples, which are notoriously difficult to work with as formalin fixation breaks DNA [10].

In addition to being associated with colorectal tumors, JCV has also been found to induce chromosomal instability CIN , something many believe initiates colorectal tumorigenesis. Ricciardiello et al. [111] have demonstrated that T-ag alone is able to induce CIN. To do this, they used RKO cells, which are a line of diploid colon cancer cells that express wild-type p53, β-catenin, and APC. They transfected the RKO cells with Mad-1 and the Δ98bp strain. Within seven days the authors observed CIN, which was characterized by chromosomal breakages, dicentric chromosomes, and aneuploidy [111]. The controls used in the study showed no such CIN [111]. The authors concluded that T-ag's binding of p53 and β-catenin are sufficient to induce CIN [111]. It also seems likely that T-ag's interaction with Nbs1, Bub1, IRS-1, and Agnoprotein's interaction with

Ku70, would also contribute to genomic instability. The observations made by Ricciardiello et al. [111] are also consistent with the study by Nosho et al. [96], who also found that T-ag expression is significantly associated with LOH , and is defined as LOH in chromosomes 2p, 5q, 17q, and 18q. Ricciardiello et al. [111] also observed that cells started to lose viral DNA soon after transfection, and was only detectable by PCR 14-21 days after initiation of the experiment [111]. Given these results, the authors proposed the following "hit and run" scenario for JCV-associated colorectal carcinogenesis: 1) after integration of viral DNA into the host genome (a phenomenon common in polyomavirus transformed cells), expression of the early genes (particularly T-ag) induces CIN, forcing most cells might enter crisis, while fortunate few increase their fitness by removing key tumor suppressors genes via CIN ; 2) those cells retain their transformed phenotype, but continue to lose JCV DNA, reducing the amount of CIN due to the loss of T-ag; eventually the transformed cells that completely lose expression of T-ag have the highest fitness, as they re-acquire genomic stability while retaining their ability to divide without limit, leading to tumor formation [111].

Taken together, the above studies paint the following picture of how JCV might induce tumorigenesis: 1) JCV is ingested by consuming contaminated water, and establishes a persistent infection in the kidneys; 2) JCV infects lymphocytes, and if the Mad-1 strain has evolved (either in the kidney, lymphocytes, or colon cells) JCV is acquires the ability to infect colon cells; 3) once the Mad-1 strain infects the colon, it integrates into the genome, preventing productive infection and thus expression of VP1; 4) if the Δ98 Mad-1 strain evolves, possibly due to genomic instability, T-ag is expressed at high levels, dysregulating the cell cycle and inhibiting apoptosis, driving the cell to divide uncontrollably and inducing CIN; 5) eventually the cell loses key tumor suppressor genes, along with expression of T-ag , resulting in a T-ag independent transformed cell. An alternative model is that JCV Mad-1 is able to infect colon cells, and reactivation, due to some sort of

immunosuppresion caused by mutations in the Cellular Interfering Factor (CIF) barriers, gives JCV the ability to deregulate the cell cycle and inhibit apoptosis.

CHAPTER 4 A NEED FOR MODELING

The reviews in Chapters 2 and 3 reveal that there are two models of colorectal tumorigenesis, the Mutation model and the Infection model. At its core, the Mutation model proposes that key genes are preferentially removed in the following order: Adenomatous Polyposis Coli (APC), Kristen Rat Sarcoma Virus (KRAS), 18q, p53, and perhaps PI3K. One might conclude that the Mutation model hypothesizes that the cancer barriers are removed the the following order: anti-growth (APC mutations prevent differentiation), pro-growth (KRAS and SMAD mutations allow the cell to divide in the absence of growth factors), apoptosis (inactivation of p53 inhibits apoptosis), metastasis (PI3K mutations activate Rho/Rac/Cdc42).

The Infection model hypothesizes that JC Virus (JCV) plays a role in tumorigenesis by interacting with many of the same or similar proteins involved in the Mutation model: Large T Antigen (T-ag) transports β-catenin to the nucleus, which has similar effect as mutating APC; Small T Antigen (t-ag)'s interaction with PP2A activates PI3K,much like mutations in KRAS; JCV's inhibition of p53 prevents transcription of p21, allowing division to occur in the presence of anti-growth signals, thus having the same effect of mutating SMAD; inhibition of p53 also prevents apoptosis, just like in the Mutation model. Thus, the Infection model hypothesizes that JCV is able to simultaneously remove the several cancer barriers: T-ag's interaction with β-catenin and pRb inhibits differentiation and promotes proliferation, allowing the cell to divide in the presence of anti-growth signals; the pro-growth barrier is removed by T-ag 's interaction with IRS-1, p300/CBP, and t-ag's interaction with PP2A, allowing the cell to divide in the absence of pro-growth signals; the apoptosis barrier is removed by T-ag's

interaction with p53, IRS-1, and t-ag's interaction with PP2A; the metastasis barrier might be removed by t-ag's interaction with PP2A, which activates PI3K, resulting cytoskeletal changes and increased mobility.

There are two hypotheses of JCV's role in colorectal tumorigenesis. The Hit and Run model posits that the Δ98 Mad-1 strain integrates into the host genome and removes key tumor suppressor genes via T-ag Chromosomal Instability (CIN). T-ag expression is eventually lost, leading the re-acquisition of genomic stability and maintenance of the transformed phenotype.

The Reactivation model hypothesizes that JCV becomes latent after infection, but is reactivated if some sort of immunosuppression occurs when mutation compromises the Cellular Interfering Factor (CIF) barriers. Once reactivated, JCV expresses its oncoproteins, removing the pro-growth, anti-growth, and apoptosis barriers.

Despite the evidence presented Chapter 3, the Infection model is contentious because it is difficult to determine role of JCV in colorectal caner. A primary reason for this is that JCV is so prevalent that it is not entirely surprising that JCV can be found in tumors [82]. It could simply be that JCV latently infects healthy cells and but remains detectable in the tissue after tumor formation. Furthermore, just because JCV can cause tumors in non-human hosts does not necessarily mean it will cause tumors in humans, as JCV may only cause tumors in non-permissive hosts. Even so, the criteria frequently used to establish a cause-effect relationship between infection and cancer includes the detection of the viral genome and/or its products in tumor tissue but its absence in healthy tissue, and a molecular basis for virally induced oncogenesis, and a consistency of the association [99]. JCV meets these requirements: Δ98 Mad-1 is found exclusively in tumors but is absent in adjacent healthy tissue; JCV produces several proteins that interfere with pathways traditionally associated with colorectal tumorigenesis; at least nine studies from five independent laboratories have demonstrated an association between JCV and colorectal cancer.

While controversial, the consistent association between JCV and colorectal cancer, along with JCV's ability to interfere with several tumor suppressors and induce CIN, strongly suggests that JCV should increase the risk of colorectal cancer. However, is this increase in risk negligible, moderate, or significant? If JCV does significantly increase the risk of colorectal cancer, it would be worthwhile to develop a vaccine, which could help prevent colorectal cancer, as opposed to treating it.

Lab studies are not be ideal for determining how much JCV increase the risk of colorectal cancer, as it takes decades for the disease manifest itself, much longer than the lifespan of most lab animals. Animal studies are also not ideal because they are non-permissive hosts, which may make them more likely to develop tumors. Furthermore, JCV's high prevalence in the human population makes it difficult to use population based studies to determine how much JCV increases risk. Mathematical and computer modeling, on the other hand, can help determine if JCV has the potential to increase the risk of colorectal cancer. The use of such models allows one to simulate how JCV interacts with cells over a human lifetime. One can also remove infection from the model to estimate the prevalence of colorectal cancer in the absence of JCV, thereby simulating a population in which the prevalence of JCV is zero. Such a simulation helps determine if the prevalence of JCV and colorectal cancer are related. Again, neither of these conclusions can come from population and lab based studies.

Three models have developed to estimate whether or not JCV is involved in colorectal tumorigenesis. The first is a probability model that determines the age-specific probability of developing colorectal cancer by mutation or infection. Although this is a simple model, it sheds light on whether the mutation or infection is the primary driver of colorectal cancer.

The second model is a separate probability model that estimates the age at which colorectal cancer develops under the infection and mutation models. This

too is a simple model, but it provides an independent estimate of which force, infection or mutation, plays the most important role in colorectal tumorigenesis.

The third model is a more complex agent based model (ABM) that simulates the behavior of cells and their interaction with one another. In this model, tumors "emerge" from changes in cellular behavior induced by mutation and/or JCV. Due to the nature of ABMs, this model is able to capture more of the complexities of tumorigenesis.

The results from these models will shed light on the drivers of colorectal tumorigenesis. Is mutation or infection the primary driver of tumorigenesis? Is mutation alone sufficient to drive tumorigenesis, and if so, which barriers provide the most protection? If infection is involved, what is its role and how does it increase the risk of colorectal cancer? The answer to these questions and others may be useful in understanding the drivers of colorectal tumorigenesis, and how those drivers might be blocked so as to reduce the prevalence of colorectal cancer.

CHAPTER 5 PROBABILITY MODELS

5.1 OVERVIEW

Two different hypothesis of colorectal tumorigenesis have been put forth: the first argues that mutation drives tumorigenesis, while the second hypothesizes that JC Virus (JCV) infection plays an important role in tumorigenesis. While much evidence suggests that JCV at least has the potential to be oncogenic, there is relatively little evidence about how oncogenic the virus actually is in humans. Much of the difficulty in assessing the impact of JCV is that fact that it is so common, and so it is not surprising that it is associated with various cancers. However, while lab and epidemiological studies may not be able to asses the risk of JCV infection has on colorectal cancer, mathematical models may be able to. A probability model developed by Calabrese and Shibata [17] can be modified to determine the probabilities of developing colorectal cancer, with and without infection. These models can be further modified to account for genomic instability[1].

5.2 ORIGINAL CALABRESE MODEL [?]

In 2010, Calabrese and Shibata [17] developed a simple heuristic probability model that calculates the age-specific cumulative probability that mutation will remove all cancer barriers, leading to colorectal tumorigenesis. In this model,

1 Please see A or a complete description of the R code used to run this model

PARAMETERS	
$\mu = 3 \times 1000\text{bp} \times 10^{-9} = 3 \times 10^{-6}$	mutation rate, 3 genes, 1000bp per gene
d = number of divisions in d days	number of divisions for a given age
$k = 6$	number of barriers to cancer
$N_0 = 15\,(10^6)$	number of intestinal crypts
$m = 8$	number of stem cells per crypt

Table 5.1: Original Calabrese Model Parameters [17]

Algorithm 5.1 Probability of Colorectal Cancer Developing by Mutation, per 100,000 individuals [17]

$$P_M = 1 - \left(1 - \left(1 - (1-\mu)^d\right)^k\right)^{Nm} \times 100{,}000$$

there are six cancer barriers, which are derived from the paper by Hanahan and Weinberg [50]: pro-growth, anti-growth, apoptosis, angiogenesis, replication limit, and metastasis barriers. In this model, there are five parameters: d, the number of stem cell divisions; $m \times N$, where m = the number of stem cells in each crypt, and N =the number of colon crypts, yielding the total number of stem cells in the colon; k, the number of critical rate-limiting pathway driver mutations (i.e. the number of mutations required to remove all barriers to cancer); and μ, the mutation rate [17]. The values used in the model are found in Table 5.1.

The logic behind the probability model of Calabrese and Shibata [17] is as follows: $1 - \mu$ is the probability that there is not a mutation in a gene , so $(1-\mu)^d$ is the probability that there is not a mutation in the gene after d stem cell divisions. Similarly, $\left(1 - (1-\mu)^d\right)^k$ is the probability that k barriers are *not* knocked by mutation after d divisions. Finally, $\left(1 - \left(1 - (1-\mu)^d\right)^k\right)^{Nm}$ is the probability that k barriers are not knocked out in N stem cells, in each of m colon crypts, divide d times. Therefore, the cumulative probability of oncogenesis by mutation can be defined as the probability that k barriers **are** knocked out after Nm stem cells divided d times, which is summed up in Algorithm 5.1.

Inserting the parameter values in Table 5.1 allows one to calculate the cumulative probability (i.e. prevalence) of colorectal cancer for each age group by using

different values of d. For example, the predicted prevalence of colorectal cancer in individuals 70 years or less is $P_{M70} = 1 - \left(1 - \left(1 - \left(1 - 3 \times 10^{-6}\right)^{6387.5}\right)^{6}\right)^{8 \times 1.5 \times 10^{7}} = 0.00559$, while the prevalence of colon cancer in individuals 65 or less is calculated to be $P_{M65} = 0.0036$. Incidence is the number of new cases between age groups, and so one can convert prevalence to incidence by determining the difference in prevalence between any two age groups. For example, the incidence of colorectal cancer in 70-75 year old individuals would be $I_{M70} = 0.00559 - 0.0036 = 0.00199$. Using these incidence value, one can then predict the incidence of colorectal cancer per 100,000 individuals. For example, the incidence of colorectal cancer in individuals 70-75 years old, per 100,00 individuals, would be $0.00199 \times 100000 = 199$ individuals [17]. Incidence values can be calculated for each age group and compared to observed incidence values of colorectal cancer [2], providing a sense of how well the model predicts colorectal cancer incidence. The results of such a comparison, using the probability model and parameter values of Calabrese and Shibata [17], is found in Figure 5.1.

5.3 INFECTION MODEL DERIVED FROM THE CALABRESE MODEL

Calabrese and Shibata's original Mutation model can be modified to determine what the incidence of colorectal cancer would be if JCV is involved in tumorigenesis. By assuming that JCV removes three protective cancer barriers (pro-growth, anti-growth, and apoptosis), one can change the number of barriers mutations must remove from $k = 6$ to $k = 3$. The prevalence of JCV must also be accounted for, as not everyone in the population is infected. By finding the slope of regression line for observed JCV seroprevalence [68], $\bar{R}$, one has an estimate of, on average, how many more individuals are infected by JCV every five years. Figure

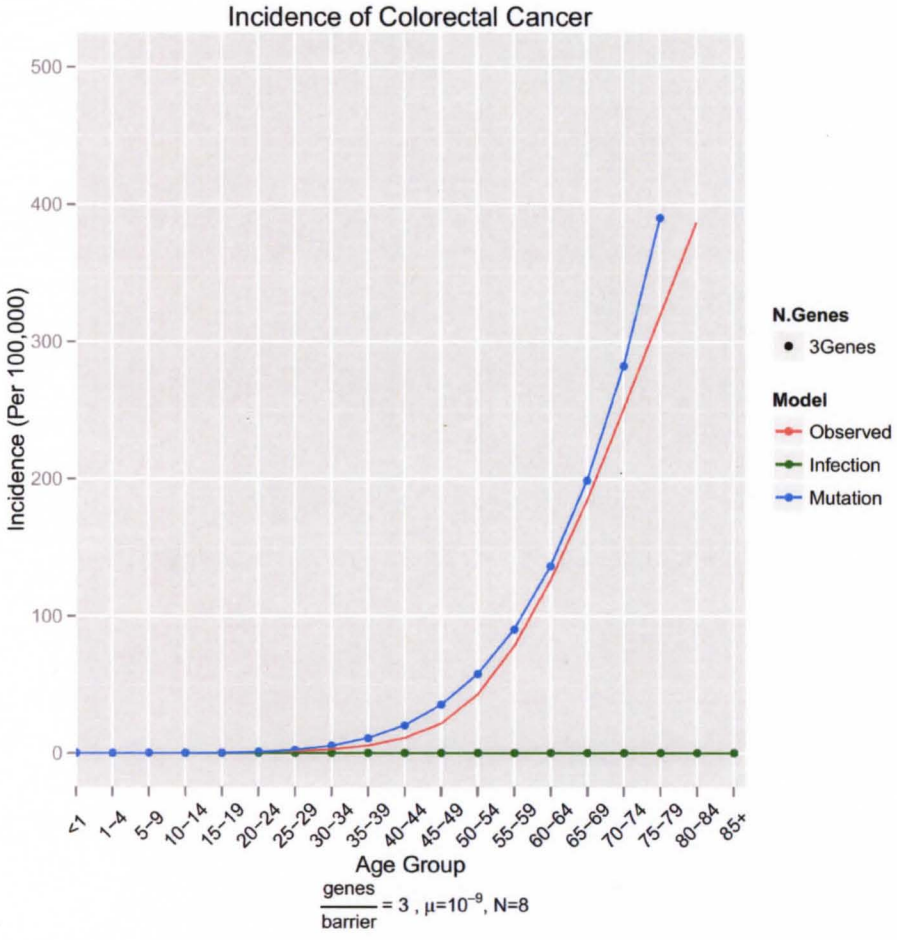


Figure 5.1: Modeled Incidence, Mutation

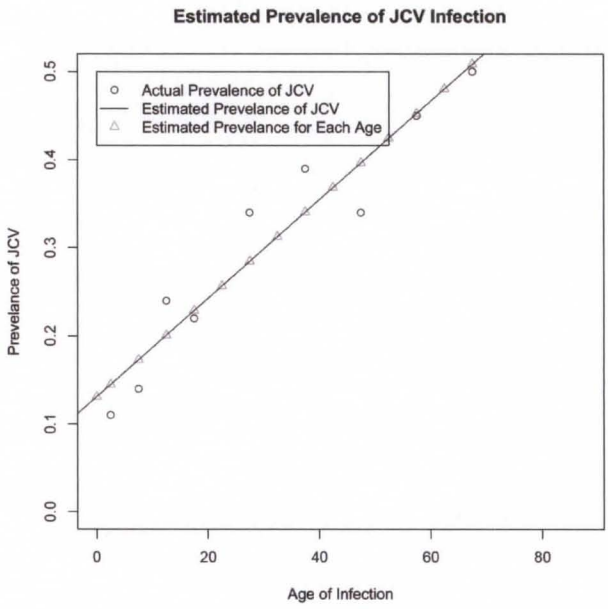


Figure 5.2: Prevalence of JCV

5.2 illustrates this calculation, which finds that $\bar{R} = 0.04205513$. By multiplying $\bar{R}$ by the prevalence of colorectal cancer, one can thus estimate what fraction of the age group is infected with JCV and has colorectal cancer. A third modification has to do with the observation that not all stem cells are infected by JCV. Del Valle et al. found JCV T-Ag in 95% of CD133+/CD44+ rat mesnchymal stem cells (rMSC), suggesting that JCV has the potential to infect most, but not all, stem cells. Assuming that JCV is also able to infect a similar number of colorectal stem cells, the total number of cells under consideration in the Infection model becomes $Nm = 8 \times 15000000 \times 0.95$.

To estimate the probability of infection driven colorectal tumorigenesis it is necessary to sum across all combinations of current age and age of infection. The reason for this is that the probability of cancer in each age group depends up how long the individual has been infected. For example, an individual who is 70 could have been infected at age 15 or age 65, but the former individual would have a greater probability of developing cancer than the latter individual. Thus, to determine the prevalence of cancer for all 70 year old individuals, one must

Algorithm 5.2 Prevalence of Colorectal Cancer Developing by Infection for each age group, per 100,000 individuals.

$$P_{Id} = \sum_{j}^{d} \left[1 - \left(1 - \left(1 - (1-\mu)^{d-j} \right)^{k-3} \right)^{Nm \times 0.95} \right] \bar{R} \times 100,000$$

sum across all differences in current age, d, and the age of infection, j. Summing across all ages of infection thus provides the prevalence of colorectal cancer in that age group. Making these changes to Calabrese and Shibata [17]'s mutation model yields the infection model found in Algorithm 5.2.

Iterating Algorithm 5.2 across all age groups generates modeled prevalence, which can be converted to incidence and compared to both Calabrese and Shibata's model and the observed incidence. The modeled incidence values can be found in Figure 5.3.

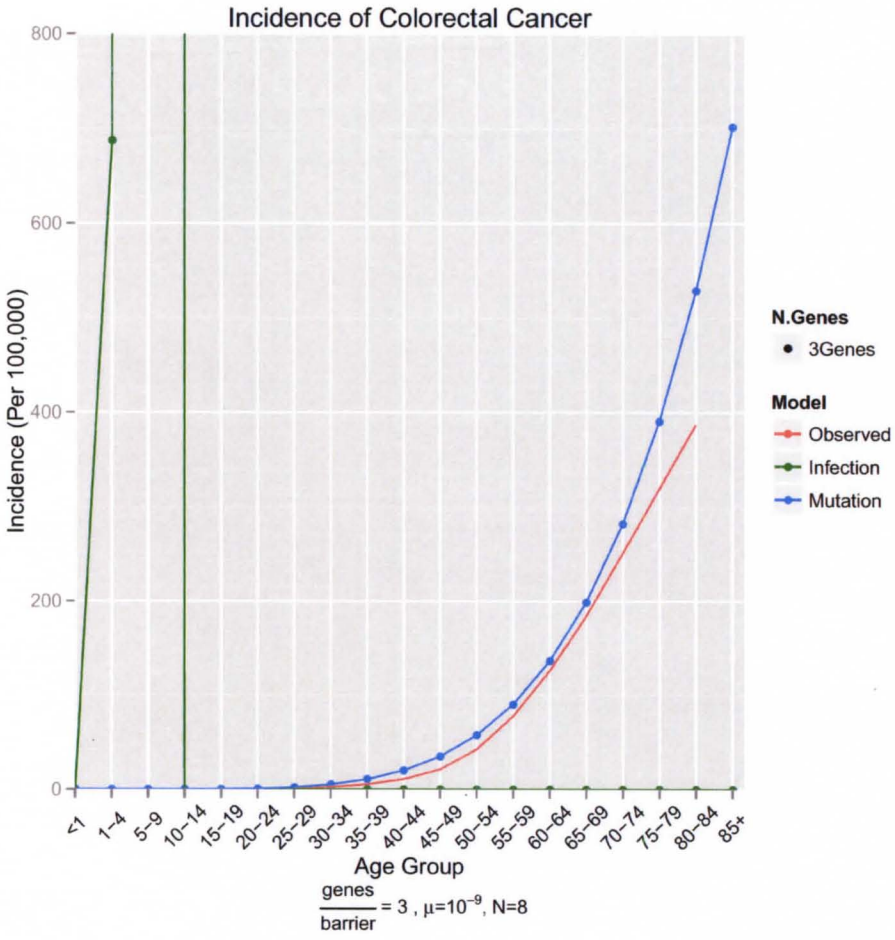


Figure 5.3: Incidence, COP

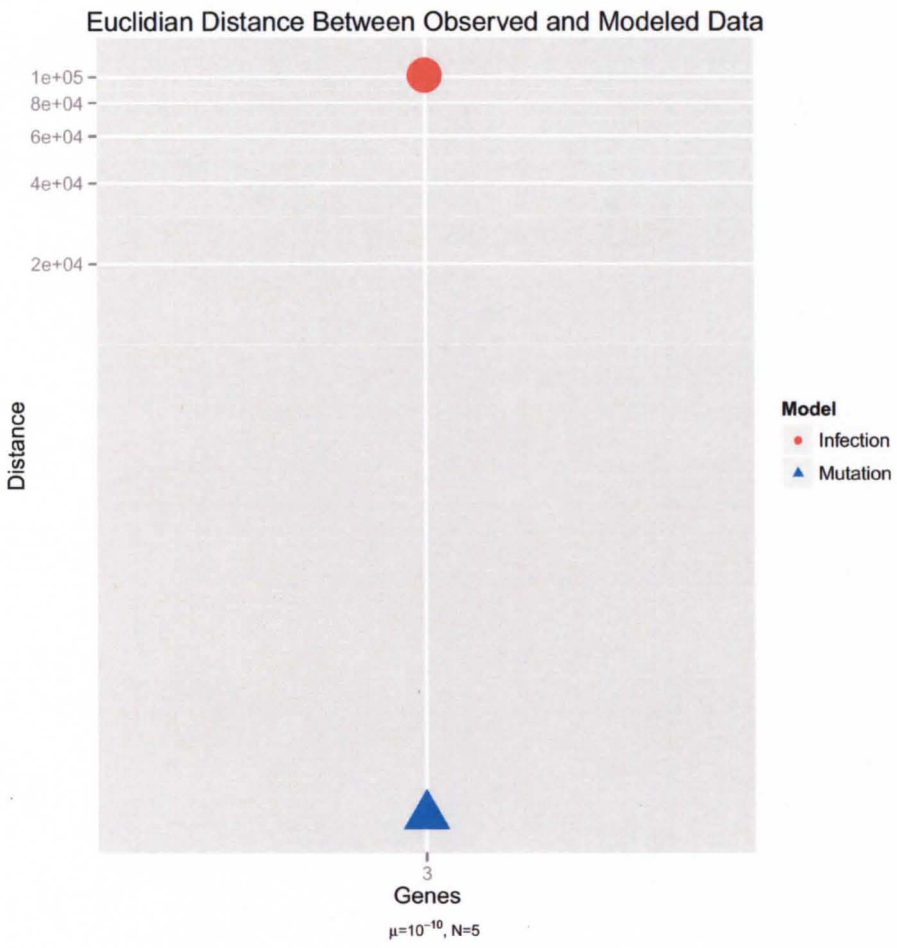


Figure 5.4: Euclidian Distance Between Modeled Incidence and Observed Incidence, Using Original Parameter Values

Every individual in the Infection model develops colorectal cancer by the time they are 10 years old, which is far from reality. Euclidian distances between the modeled incidence and observed incidence provide a way to determine which model's results are closest to the real thing. As expected, given the results in Figure5.3, the Euclidian distances found in Figure 5.4 illustrate that Calabrese and Shibata's original Mutation model is much better fit, while the Infection model is nowhere close to the observed incidence. Given these results, one can conclude that, as modeled, infection cannot be involved in tumorigenesis, and that mutation is the primary cause of colorectal cancer.

PARAMETERS	
$\mu = 10^{-10}$ or $\mu = 10^{-11}$	stem cell point mutation rate
d = number of divisions in d days	number of divisions for a given age
$k = 3, 6, 9, 12, 16$	number of barriers to cancer
$N_0 = 15\,(10^6)$	number of intestinal crypts
$m = 5$	number of stem cells per crypt

Table 5.2: New Parameter Values

In future discussions, these two models will be referred to as the Calabrese model with Original Parameters (COP) models.

5.4 CALABRESE MODELS WITH NEW PARAMETERS (CNP)

While the mutation model developed by Calabrese and Shibata [17] provides an exceptional fit to the observed incidence, it makes several assumptions that may not be valid. The first assumption is that the stem cell mutation rate is the same as the transit cell mutation rate. However, several authors suggest that stem cells have mutation rates that are 10-100 times lower than normal cells [19, 43, 16], precisely to avoid accumulating oncogenic mutations over a lifetime. In fact, Frank et al. estimate that the mutation rate of stem cells may be several orders of magnitude lower than that of somatic cells, somewhere between 10^{-10} and 10^{-11} mutations per base pair, assuming the average gene is 1000bp long. Second, Calabrese and Shibata [17] use eight stem cells per crypt, while Bjerknes and Cheng [8] estimate that there are only 4-6 stem cells per crypt. Third, it has been estimated that there could be up to 100 genes involved in tumorigenesis[67]. If there are six barriers there could be up to 16 genes per barrier.The following probability models thus set the stem cell point mutation rate at either 10^{-10}or 10^{-11}, with number of stem cells per crypt at five, and have 3,6,9,12, or 16 genes per barrier. A summary of these new parameter values can be found in Table Table 5.2.

The incidence values generated from using Algorithms 5.1 and 5.2 with new parameter values are found in Figure 5.5, while the Euclidian distances are found in Figure 5.6.

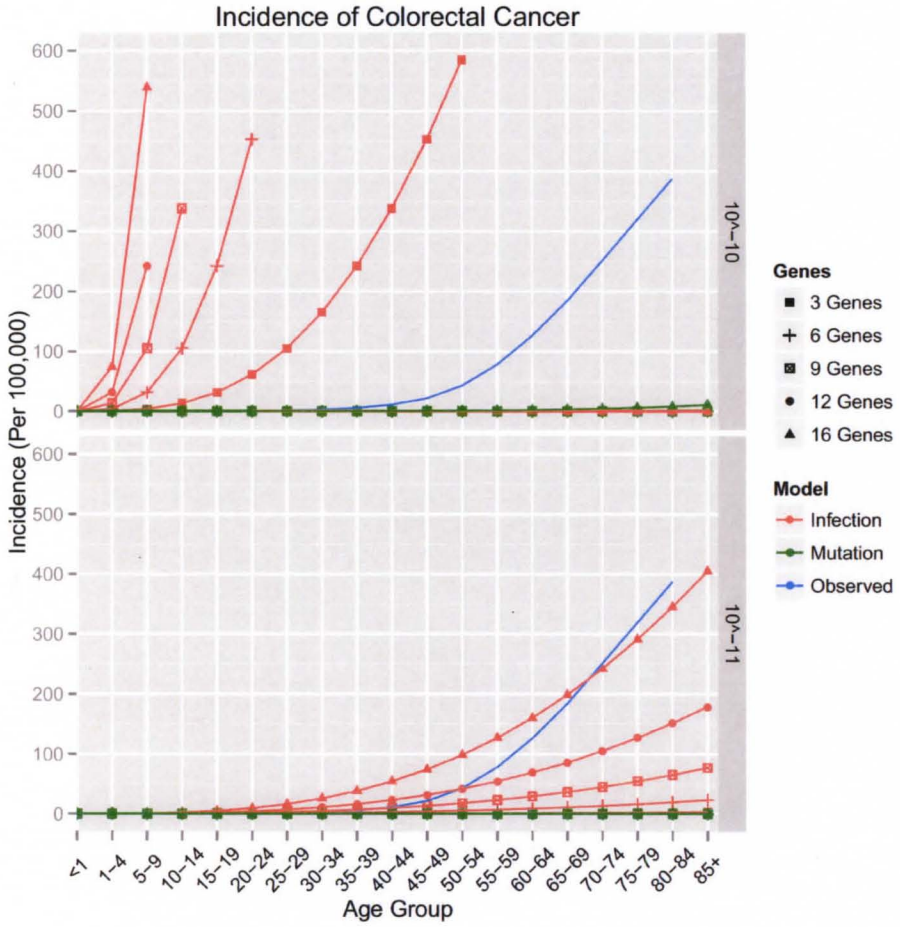


Figure 5.5: Incidence, CNP

The results in Figures 5.5 and 5.6 yield several conclusions. The first is that, given the modeled incidence values, mutation is not able to drive tumorigenesis, which is in contrast the COP models. This is true across all models, indicating that the difference between the Original parameter and New parameter models is due to the lower stem cell mutation rate, whether it be $\mu = 10^{-10}$ or $\mu = 10^{-11}$. This in turn reveals that a lower stem cell mutation rate does protect the cell from accumulating too many oncogenic mutations. Third, the Euclidian distances

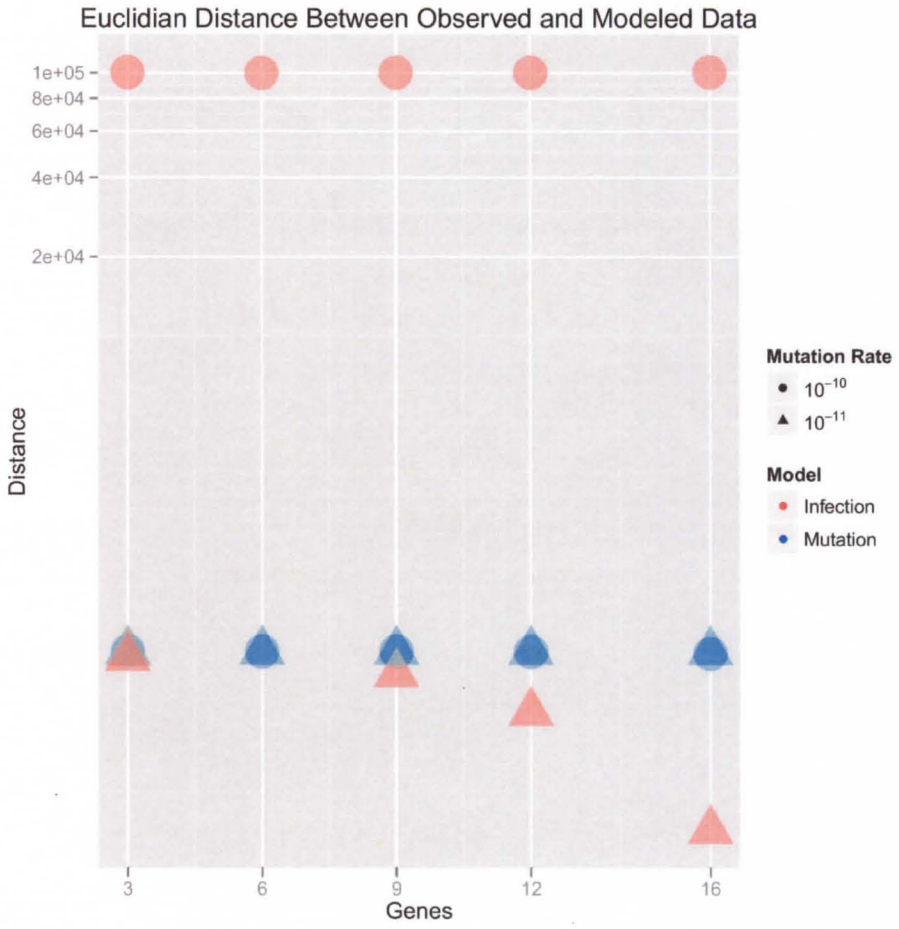


Figure 5.6: Euclidian Distances between Models With New Parameter Values

suggest that infection is only able to produce realistic incidence values if the stem cell point mutation rate is $\mu = 10^{-11}$, as at $\mu = 10^{-10}$ colorectal cancer occurs far too early and frequently. Finally, the Euclidian distances suggest that the model with 16 genes and a stem cell mutation rate of $\mu = 10^{-11}$ best replicates the observed incidence. All together, the results from this model suggest that infection is the primary cause of colorectal cancer, as mutation alone is unable to generate any cases of cancer. Finally, mutation is required to remove the last three barriers, and so may be considered a secondary cause.

The Calabrese model with New Parameters (CNP) models suggest that JCV infection is absolutely required for colorectal tumorigenesis, indicating that it is the primary cause of cancer. However, one can easily argue that the models are far too simple, and that they lack one of the driving forces of tumorigenesis, genomic instability. A second assumption that CNP make is that all stem cells have an equal chance of removing all of the barriers. However, it is more likely that the population of cells that removes the first barrier is the most likely to remove the second. And that second, smaller, population is the most likely to remove the third. Thus, the population of cells that are most likely to remove all of the barriers decrease over time. The following family of models incorporates this decreasing population size and genomic instability by making a few more modifications to the Calabrese models.

The decreases population size is accounted for by leaving behind the cells that do not remove that first barrier, so that $N_1 < N_0$. Similarly, the population of N_1 of cells that remove the second barrier, N_2, is even closer to metastasis, and all other cells in N_1 are ignored, meaning that $N_2 < N_1 < N_0$. Over time the population of pre-metastatic cells gradually decreases in a step-wise fashion until the final population of cells that need to remove the last barrier is determined.

The second modification is based on the assumption that genomic instability doubles the mutation rate every time a barrier is removed. This value is based on the observation that Hepatitis C Virus induces a mutator phenotype, increasing the mutation rate 5-10 fold [81]. Spreading this increase across six barriers means that the mutation rate can double every time a barrier is removed. Over six barriers, this means that $\mu_0 = 10^{-10}$,$\mu_1 = 2 \times 10^{-10}$,$\mu_2 = 4 \times 10^{-10}$,$\mu_3 = 8 \times 10^{-10}$, $\mu_4 = 16 \times 10^{-10}$, and $\mu_5 = 32 \times 10^{-10}$ (assuming a stem cell point mutation rate of $\mu = 10^{-10}$

Algorithm 5.3 Probability Stems Cell Removes One Barrier, Mutation Model

$$P_M(N,\mu) = 1 - \left(1 - \left(1 - (1-\mu)^{\frac{d}{6}}\right)^{k=1}\right)^N$$

Algorithm 5.4 Probability Stem Cell Removes One Barrier, Infection Model

$$P_I(N,\mu) = 1 - \left(1 - \left(1 - (1-\mu)^{\frac{d}{3}}\right)^{k=1}\right)^N$$

These modifications are incorporated into the Genomic Instability Mutation Model (GIM) in the following manner. The probability that one stem cell knocks down the first barrier, $P_M(1,\mu_0)$, is found in Algorithm 5.3. The expected total number of cells knock down the first barrier can be calculated as $N_1 = P_M(1,\mu_0)N_0$, where $N_0 = 5 \times 15(10^6)$, the total number of stem cells in the colon. Similarly, $N_2 = P_M(1,\mu_1)N_1$. Over time, the population of cells deceases as fewer and fewer cells have removed each barrier. This process is repeated for each barrier until the population size of cells that have removed the first five barriers is determined. Using this population size, N_5, and the the highest mutation rate μ_5, $P_{Md_c} = P_M(N_5,\mu_5) \times 100{,}000$ is the prevalence of colorectal cancer in age group d, per 100,000 individuals.

The Genomic Instability Infection Model (GII) is modified in a similar fashion, yielding Algorithm 5.4, which finds the probability one infected cell will remove one barrier. Note that in the GII model mutation only has to remove three barriers, as JCV has already removed other three. JCV has also induced genomic instability, and so the initial mutation rate is μ_3. As in the GIM model, the population size of cells that removed the first barrier is $N_1 = P_I(1,\mu_3)N_0$. In the GII model, $N_2 = P_I(N_1,\mu_4)$ is the population size of cells that have removed five barriers. Thus, the probability that an individual has developed cancer at age d is $P_{Id}(N_2,\mu_5)$. As in the COP and CNP infection models, the prevalence of each age group is determined by summing all ages of infection for each age group; $P_{Id} = \sum_j^d \left[P_{Id}(N_2,\mu_5) - P_{Ij}(N_2,\mu_5)\right]\bar{R} \times 100{,}000$, where d= current age, and j=

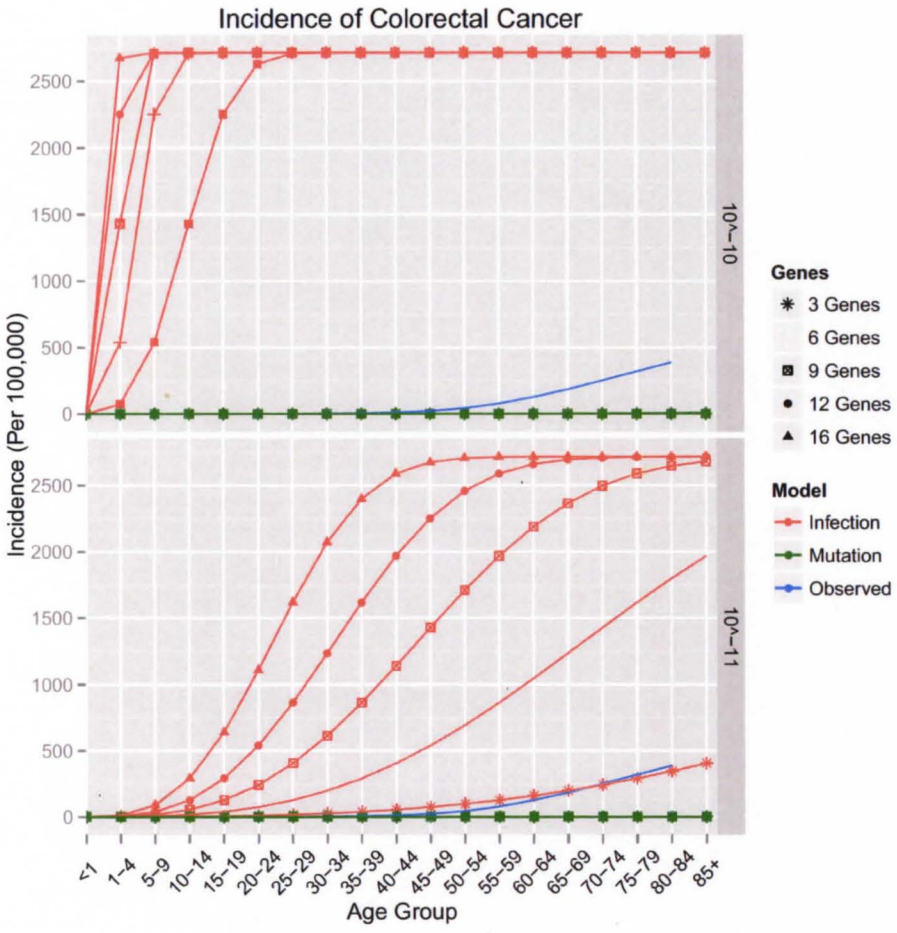


Figure 5.7: Incidence, Genomic Instability

age of infection by JCV. Also like the COP and CNP models, the initial population size of cells used in the infection model is 95% of the cells in the mutation model. Once prevalence is calculated it is converted to incidence as described above.

There are two sub-models for each model, and 5 different genes barrier in each sub-model, creating a total of ten sub-models. The first set of sub-models uses a stem cell point mutation rate of $\mu = 10^{-10}$, and uses 3,6,9,12,or 16 genes. The second set of sub-models uses the same collection of genes, but has a stem cell point mutation rate of $\mu = 10^{-11}$. The results of each model are found in Figure 5.7. The Euclidian distances of all models are found in Figure 5.8.

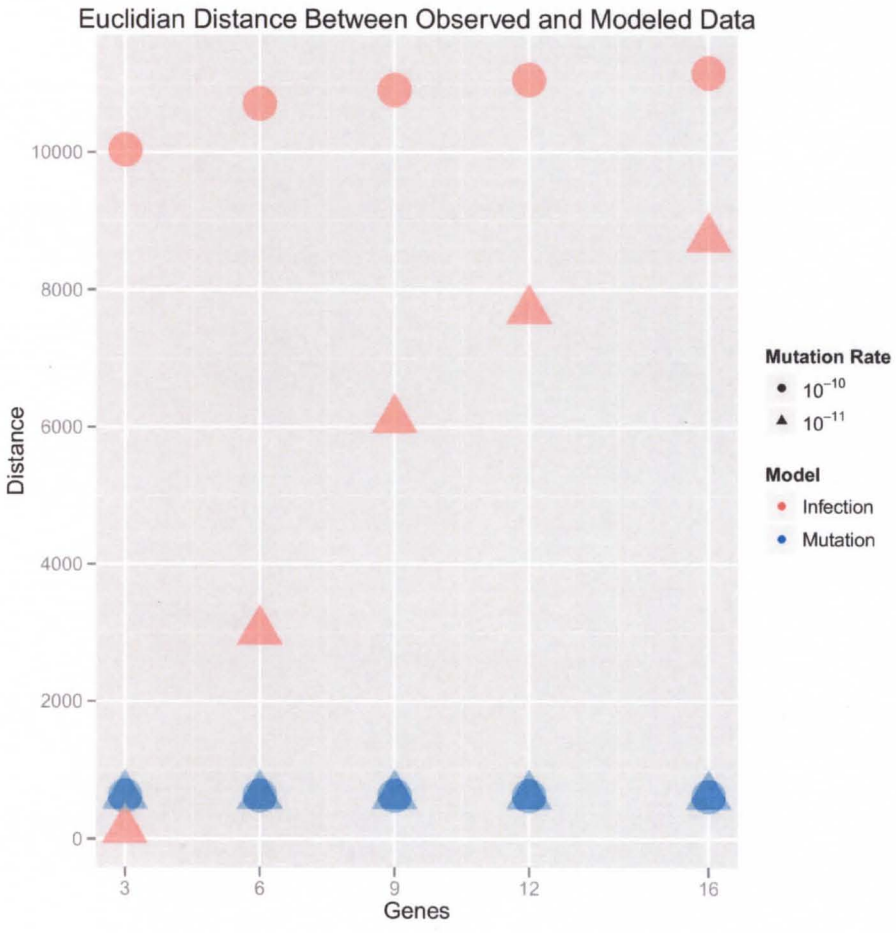


Figure 5.8: Euclidian Distance, Genomic Instability

The results of the Genomic Instability Models largely agree with those of the CNP models. Both sets of models find that mutation is unable to drive tumorigenesis, while infection is only able to generate realistic incidence values if the stem cell point mutation rate is $\mu = 10^{-11}$. Again, these results suggest that infection is the primary cause, as mutation cannot generate realistic prevalence values. However, colorectal cancer cannot occur without mutation removing the last three barriers, and so mutation may be considered a secondary cause. The results of these models only differ in that the CNP models find that sixteen genes provide the best fit, while the GII models indicates that three genes per barrier fits best.

5.6 CONCLUSIONS

The above results strongly suggest that infection plays a significant role in the development of colorectal cancer. JCV should thus be considered a major risk factor. These results are consistent across all incarnations of the model (except for the original), suggesting that they are robust. The results also show that infection always results in a higher incidence of colorectal cancer than mutation, again indicating that it increases the risk of cancer. However, this result is not surprising given that the probability of removing three barriers will always be higher than the probability of removing six. What is surprising is that mutation is unable to generate any incidence values, no matter the mutation rate. This is in contrast to the COP models, where $\mu = 10^{-9}$, which suggests that lowering the stem cell mutation rate by one order of magnitude is sufficient to protect against cancer. However, at $\mu = 10^{-10}$ infection causes colorectal cancer far to early and frequently. JCV is a very common infection, yet colorectal cancer is relatively rare, so this result may indicate that the stem cell mutation rate is $\mu = 10^{-11}$ so as to protect against infection induced cancers. This value is consistent with the estimation that the stem cell mutation rate could be 10-100 times lower than other cells [43].

All in all, these models require infection for tumorigenesis, and so JCV can be interpreted as being the primary cause of colorectal cancer. Mutation plays an important secondary role by removing those barriers that infection does not. Thus, these results suggest that colorectal cancer is a multi-factorial disease.

CHAPTER 6 | GEOMETRIC MODEL

6.1 OVERVIEW

A fourth probability model is developed to better understand the drivers of colorectal cancer. This model is not built around the work of Calabrese and Shibata [17], but is based upon the Geometric distribution, which calculates the probability of a certain number of failed trials before the first success. This algorithm simulates the minimal number of divisions that need to occur for cancer to develop. This model uses the same assumptions of the probability models, and is applied to both Mutation and Infection model. Unlike the other probability models, the Geometric model finds that the Mutation model best replicates the prevalence of colorectal cancer[1].

6.2 THE GEOMETRIC MODEL

The Geometric distribution is defined as $\Pr(X = k) = (1-p)^{k-1} p$. Given that p is the probability of success, and k is the number of Bernoulli trials, the Geometric distribution can be interpreted as follows: $1-p$ is the probability of failure, so $(1-p)^{k-1}$ is the probability of failure before the final k^{th} trial, and so $(1-p)^{k-1} p$ is the probability that there are $k-1$ failures and success on the final k^{th} trial. In the case of this model, each stem cell division is considered a Bernoulli trial, where the probability of removing a barrier (i.e. a successful event) is μ. The

1 Please see B or a complete description of the R code used to run this model

Geometric distribution can therefore be used to determine how many stem cell divisions are required for each barrier to be removed by mutation.

6.3 IMPLEMENTATION

The program **R** (v 2.11.1; [107]) has the function **rgeom(n,prob)**, which randomly generates n independent observations of how many trials occur before a successful event, given that the probability of success is p. Each of the n observations can be thought of as a cell lineage, and the deviates produced by **R** are interpreted as the number of divisions required for a barrier to be removed by mutation.

In the Mutation model, b_1 =**rgeom(n=**5×15000000**, prob=**$3 \times 1000 \times 10^{-10}$**)** creates a vector containing 7.5×10^7 elements. Each element represents the number of years it takes for a barrier to be removed by mutation in each of the 7.5×10^7 stem cells, given that the probability of removing a barrier in one of three 1000bp genes is 10^{-7}. After the creating of b_1, a new vector, b_2, is created in a similar fashion, except that the mutation rate is doubled, thus taking into account genomic instability. The process is repeated for vectors b_3, b_4, b_5, and b_6, doubling the mutation rate each time a vector is created, so that the mutation rate used to create b_6 is $\mu = 2^5 \times 3 \times 1000 \times 10^{-10}$. After their creation, all vectors are added together into the vector T, each element of which now represents the total number of years it takes for all six barriers to be removed in each of the 7.5×10^7 stem cells. The minimum number of years it takes a lineage to remove all six barriers is recorded, as it represents the first lineage to initiate tumorigenesis within the individual. This process is repeated $1,000$ times, so as to simulate colorectal tumorigenesis in $1,000$ individuals. Unlike the probability models, every individual will develop colorectal cancer at some point, and so the Geometric model only produces colorectal cancer patients.

The Infection model assumes that JCV is able to remove three barriers and generate genomic instability. Thus, the only difference between the Infection and

Mutation model is that the mutation rate used to create b_1 is in the Infection Model is $\mu_3 = 2^3 \times 3 \times 1000 \times 10^{-10}$ (assuming there are 3 genes per barrier, and the stem cell point mutation rate is $\mu = 10^{-10}$). Similarly, the mutation rate used to generate b_2 is twice that used to create b_1, and b_3 is twice that used to create b_2. Finally, as in the other probability models, it is assumed that JCV infects 95% of cells, Otherwise, the two models and their implementation are identical.

Each model is run using either a mutation rate of 10^{-10} or 10^{-11}, and either 3,6,9,12, or 16 genes per barrier. All results for each combination of genes and mutation rates are collected and binned into age groups.

6.4 RESULTS

The Geometric models produce the number of new patients in each age group, and so may be considered incidence values. Since this model is stochastic it is unlikely it will produce cancer patients in all age groups. For example, it may produce one or two patients that are 72 and 74, but none that are 76, leaving the $75-79$ year old age bin empty. This is in contrast to the probability models in Chapter 5, which produce the cumulative distribution function calculating the probability of cancer in each age group. Therefore, as the results in Figure 6.1 suggest, modeled incidence values for the Geometric models are slightly deceptive, as the incidence goes up and down, simply because some age groups do not have any patients in them. Prevalence may provide a better picture of the results, as it is the total number of individuals that have colorectal cancer at that age, regardless of when they developed colorectal cancer. For example, the prevalence of colorectal cancer in individuals 70-74 includes everyone that developed cancer at 30,40,50, etc.... If the stochastic model does not generate an individual for a given age group the prevalence remains the same as the prevalence in the previous age group, and so does not dip up and down like incidence does. Finally, prevalence provides more accurate distance measurements, as modeled

AGE	MORTALITY
< 20	0%
$20-34$	0.6%
$35-44$	2.5%
$45-54$	8.6%
$55-64$	16.5%
$65-74$	22%
$75-84$	29%
$85+$	20.8%

Table 6.1: Mortality From Colorectal Cancer, $2005-2009$ [2]

incidence values can skyrocket and then crash to zero. If the observed incidence is also close to zero for the age group the model will have a low distance score, despite its poor reproduction of the observed prevalence. Prevalence avoids this because when the modeled data maxes out, it stays there, and so no calculations are biased by having the modeled data return to zero.

Incidence is converted to prevalence for a given age group by summing how many individuals have cancer in all previous age groups. However, some patients die from cancer before they reach that age group, and so cancer mortality should be to be taken into account. This can be accomplished by multiplying the prevalence in each age group by 1−age specific mortality. This is done before adding that age group to the next age group, and thus removes the individuals that died from cancer in that age group. Mortality rates of colorectal cancer can be found in Table 6.1. After taking mortality into account, prevalence can be calculated. These Geometric models's prevalence values are found in Figure 6.2, while the Euclidian distances for each model are found in Figure 6.3.

6.5 CONCLUSIONS

Consistent with findings of the probability models, the results of the Geometric model strongly indicate that infection increases the risk of cancer. As modeled,

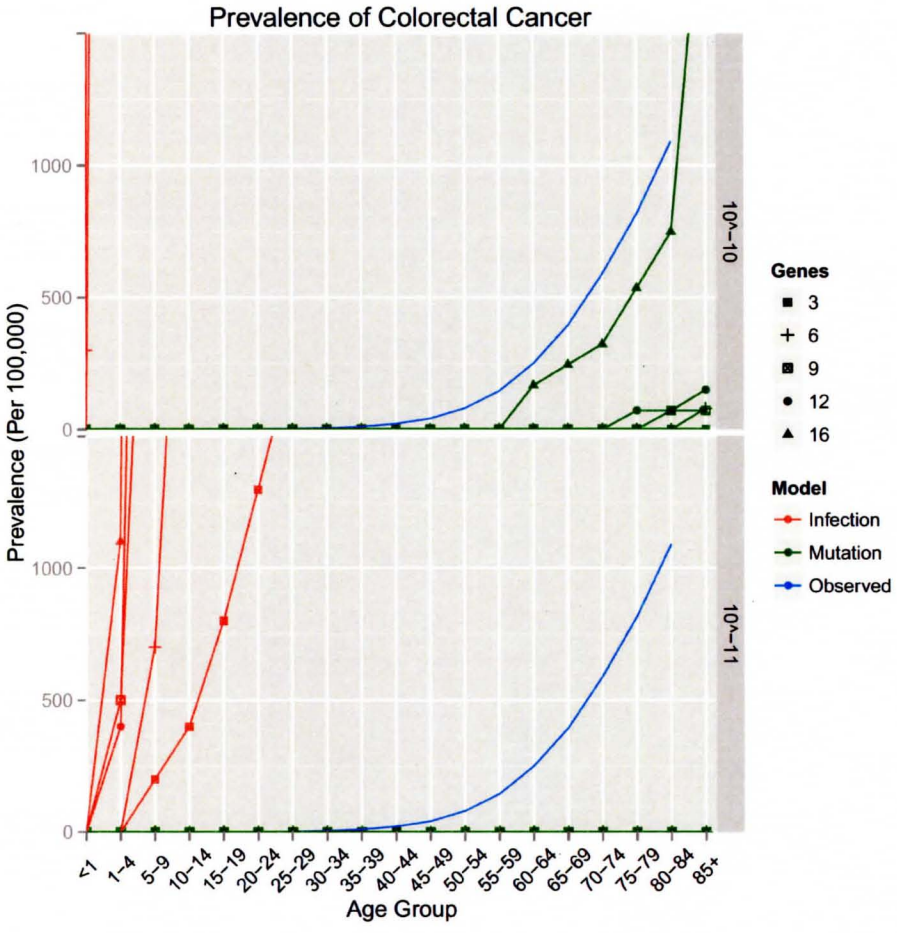


Figure 6.2: Modeled Prevalence,Geometric Model

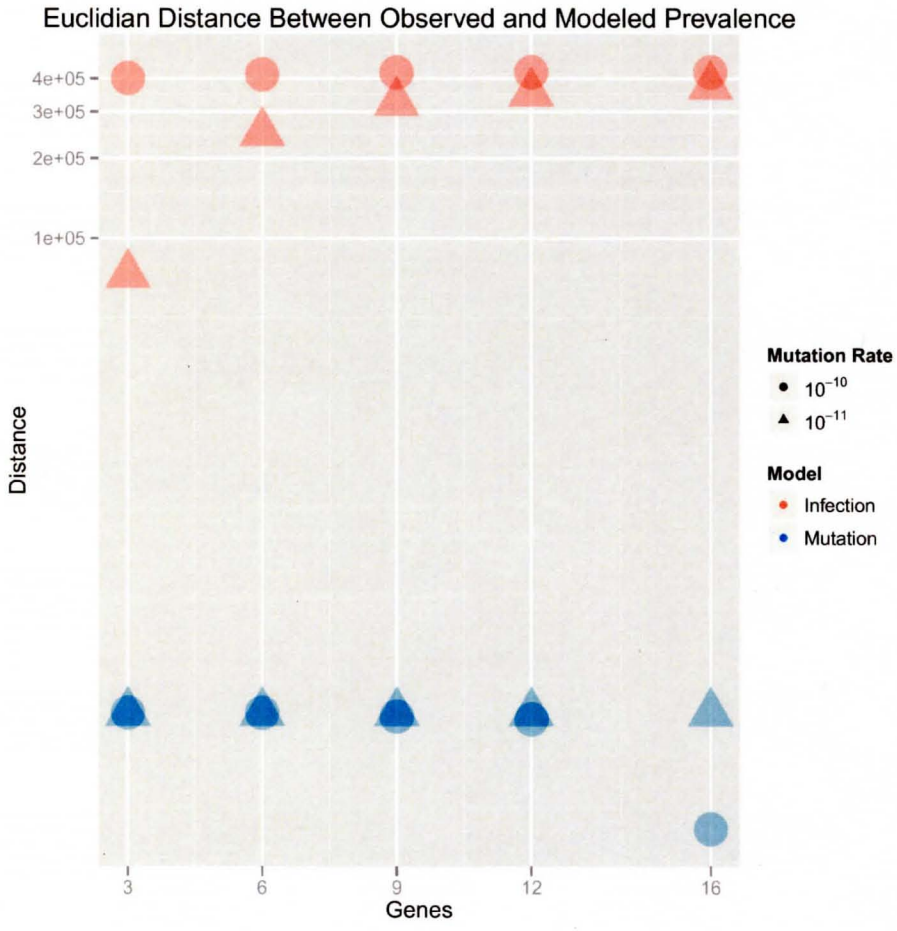


Figure 6.3: Euclidian Distance, Geometric Model

the Infection model does not fit well with the observed data, no matter if the mutation rate is $\mu = 10^{-10}$ or 10^{-11}. However, closeness of fit is highly dependent on the parameters of the model that have been altered and presumably on parameters that haven't been altered. Therefore, lack of fit should not indicate that infection is not playing a role. If nothing else, these results illustrate that the presence of infection dramatically increases the risk of cancer.

Unlike the Calabrese probability models, the results from the Geometric model indicate that mutation can drive of tumorigenesis, as the modeled and observed prevalence values are close. However, this is only the case when $\mu = 10^{-10}$, and when there are 16 genes per barrier. At $\mu = 10^{-11}$ the mutation model fails to produce any colorectal cancer patients, which is in agreement with the Calabrese model with New Parameters (CNP) and Genomic Instability models from Chapter 5. This indicates that a lower stem cell mutation rate does protect against cancer, but only if it is 100 times lower than normal. It thus seems that either mutation or infection can be the driver of tumorigenesis, but that which one fits best depends on the mutation rate. These mixed results indicate that further investigation is required.

CHAPTER 7 AGENT BASED MODEL

7.1 OVERVIEW

The family of models described in the previous chapters suggest that both mutation and JCV are capable of initiating colorectal tumorigenesis. However, those models also make many simplifying assumptions that do not capture the complex process of tumorigenesis. An agent based model is developed to address these complexities, with the aim of attaining a more accurate picture of mutation and infection's role in colorectal tumorigenesis. In this model, each mutation and/or viral oncoprotein generates a behavioral change in the cell, and the accumulation of these phenotypic changes can result in the emergence of a metastatic tumor. Agent based models also record a great deal of data which can be used to determine not only which factors increase the risk of colorectal cancer, but also how each does so.

7.2 NEED FOR AN AGENT BASED MODEL

The models described in Chapter 5 and Chapter 6 demonstrate both infection and mutation are able to drive tumorigenesis. However, the argument could be made that these models are too simple, and ignore many important observations on colon tissue dynamics, colorectal tumorigenesis, and JCV's lifecycle. For example, symmetric division is not modeled, which is a significant omission as it is proposed to be a mechanism involved in spreading a mutation throughout the crypt (i.e. monoclonal conversion). Second, the models assume that a single mu-

tation completely removes each barrier, and yet it is known that a single mutation in one tumor suppressor allele (i.e. p53, pRb, APC, etc...) does provide a selective advantage, but that removal of the other allele is required to gain the full selective advantage[9, 67]. Third, the probability model treats all barriers as equal in the sense they simply bring the cell one step closer to metastasis. However, the removal of each barrier provides the cell with a particular selective advantage that allow the genotype to increase in frequency within the population. Fourth, the previous models assume a constant population size, which is very unrealistic, as the definition of cancer is uncontrolled growth. This is a particularly poor assumption, as the more cells there are, the greater the probability that one of them will acquire all of the mutations required for tumorigenesis. Finally, the earlier models assume JCV is active immediately after infection, and remains active throughout the host's life. However, in appears that JCV instead becomes latent upon infecting colon cells, and some sort of reactivation is required for the virus to express its oncoproteins. Such reactivation may occur either due to immunosuppression, due to mutations Cellular Interfering Factor (CIF)-II genes, or by a "hit and run" mechanism.

It is not possible to model these complex processes with a simple probability model, and so an Agent Based Model (ABM) is developed. ABMs are ideal for such complex processes, as they allow each cell's behavior to change over time, either in response to internal and/or external changes, in this case mutation and infection. All of the cells then interact with each other, and combinations of different behaviors can lead to the emergence of different patterns, in this case tumor formation. Thus, this ABM seeks to address the simplifying assumptions of the probability models by modeling how mutation and infection affect the behavior of cells, and how the combination of these behaviors can result in the formation of a metastatic tumor.

7.3 OVERVIEW OF MODELS

There are three families of ABMs, each of which is programmed using NetLogo version 5.0.3[1] [147]. Each of following models are run using 3, 6,9,12,16 genes per barrier, with the average gene length being 1,000bp. Thus, there is a total of 25 models, each of which is run 1,000 times.

7.3.1 *Mutation Model*

The first model is the Mutation Model. In this model, mutation is the only way that the protective cancer barriers can be removed, even if the individual is infected. As such, the Mutation model assumes that JCV has no role in tumorigenesis. If a single mutation lands in a cancer barrier gene, the mutant phenotype is expressed 50% of the time. The beneficial phenotype will always be expressed if a second mutation lands in that same gene.

7.3.2 *Infection Models*

In the Infection models JCV randomly infects one cell every year. Once that cell is infected, there is a 2% chance that the infection will spread to its neighbor cell. If the infection fails to spread to all cells, JCV does not establish a chronic infection and will attempt to infect the individual the following year. The infection is considered chronic if JCV successfully spreads to all cells. The parameters used in the Infection models are calibrated so as to represent the observed seroprevalence of JCV (see Figure 7.1).

1 Please see the ODD in Appendix C for a complete description of the code used in this ABM

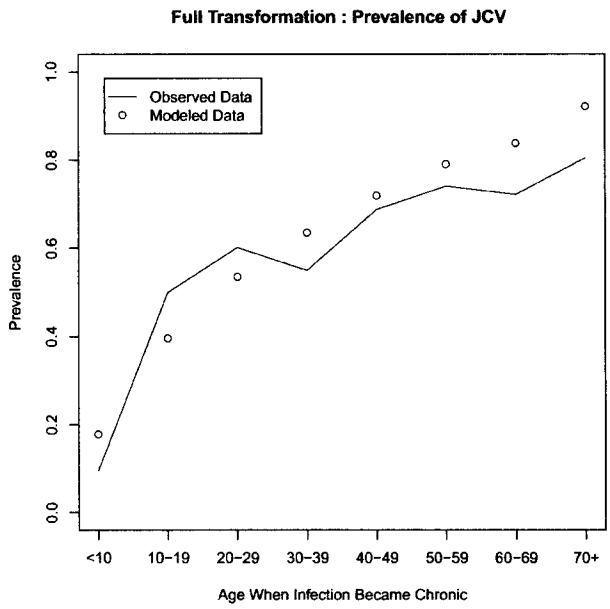


Figure 7.1: Modeled Prevalence of JCV

Each of the following infection models have two sub-models. In each case, the first model, referred to as the Full Model, has active JCV completely remove each barrier, while in the second model, termed the Partial Model, JCV partially removes each barrier. In the case of the Partial models, if there is one pre-existing mutation, and JCV partially removes that barrier, the entire barrier is be removed. The situation is the same if JCV partially removes a barrier and then mutation finishes the task.

7.3.2.1 *Reactivation Model*

The first infection model is the Reactivation model. This model hypothesizes that JCV has no effect until the individual becomes immunocompromised. For an individual to become immunocompromised they must acquire two mutations in the CIF-II genes. Recall from Chapter 1 that the CIF-II barrier is involved in inhibiting the function of viral oncoproteins[153]. If the CIF-II. barrier is removed, simulating immunosuppression, JCV is reactivated and removes the pro-growth, anti-growth, and apoptosis barriers (see Chapter 4 for a review of how JCV inter-

feres with these barriers). If there is only one mutation in a CIF-II gene, JCV only has an effect 50% of the time. This is meant to simulate that the individual is not completely immunocompromised, and so retains defenses against JCV. However, if there are two mutations in the CIF-II genes, JCV is always active.

7.3.2.2 *Hit and Run Model*

The Hit and Run model is the second infection model, and is based upon the research conducted by Ricciardiello et al. [112, 113] (see Chapter 3 for a review of their findings). The Hit and Run model hypothesizes that genomic rearrangements in JCV's Non-Coding Regulatory Region (NCRR) lead to the development of the Mad-1 strain, which may use lymphocytes to infect the colon. Alternatively, these genomic rearrangements may occur in JCV that is already in the colon. Either way, the Mad-1 strain develops into the Mad-1 $\Delta 98$ strain when there is a second deletion in the NCRR. This event changes the cell's phenotype in two ways: 1) the apoptosis, pro-growth and anti-growth barriers are removed; 2) the virus induces genomic instability, increasing the mutation rate 7.5 fold. The increase in virally induced genomic instability comes from the observation that Hepatitis C Virus induces a mutator phenotype which increases the mutation rate 5-10 fold [81], and so 7.5 is somewhere between the two. Unlike the Reactivation model, JCV Mad-1 $\Delta 98$ is only active for $14 - 21$ days, after which JCV's oncoproteins cease to be expressed, and the cell returns to its previous phenotype.

7.4 MODELING THE STRUCTURE OF A COLON CRYPT

Figure 7.2 is the world of the ABM. The box on the bottom left is the colon crypt, which is divided into the inner crypt (pink) and outer crypt (yellow). The tissues are laid out in squares so as to represent a columnar crypt that has been laid flat, so that the crypt base is in the center of the square. The black area

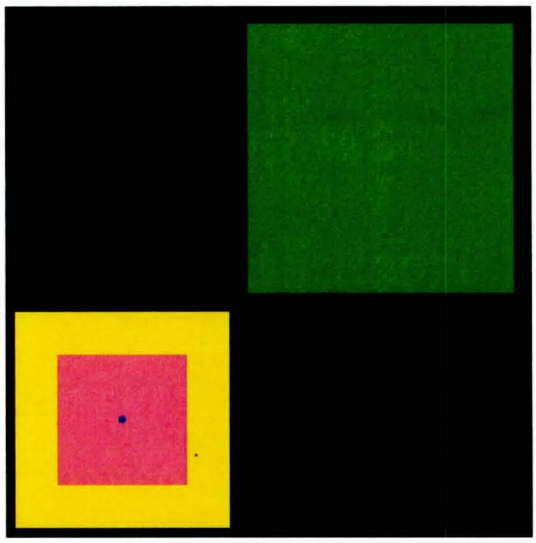


Figure 7.2: Modeled Colon Crypt

outside of the crypt represents the lumen. The green square is the metastatic tissue, which could hypothetically be anywhere in the body. At the center of the crypt are five stem cells (blue). Each patch in each tissue represents one cell from the underlying tissue.

Everyday (i.e each tick) each patch, which assumed to contain underlying blood vessels, supplies 0.25 units of oxygen to the colon crypt. Each patch then diffuses 100% of its oxygen to its neighboring patches, 20 times a day, simulating constant oxygen diffusion [21]. The oxygen is thus dispersed throughout the crypt where it is consumed by cells. Since oxygen dynamics determine cell division and movement, all oxygen related parameters are calibrated so that each crypt contains an average of 250-300 cells, and so that each stem cell divides every four to five days [127, 12, 104, 103].

7.5 WILD-TYPE BEHAVIOR

7.5.1 *Stem Cells*

At the center of the crypt (i.e. the base of the crypt) are five stem cells laid out in a circle [12, 103, 105]. Each of these stem cells is able to divide either asymmetrically or symmetrically. Which type of division occurs is determined using a Bernoulli distribution, where the probability of successful asymmetric division is $p = 0.95$ [79]. If symmetric division occurs, the most fit stem cell (i.e. the stem cell with the fewest deleterious mutations) produces two daughter stem cells, one of which replaces the least fit stem cell (i.e. that stem cell with the most deleterious mutations). This is modeled because it is assumed that the number of stem cells is tightly regulated, and that the most fit stem cell has the greatest probability of surviving while the least fit stem cell is most likely to die. After this selection has occurred, the remaining stem cells divide asymmetrically. When asymmetric division occurs, each stem cell produces one daughter transit cell. This model assumes that stem cell division is prevented by contact inhibition. Thus, a stem cell will only divide if there is an empty patch within its cone of vision, which has an angle of 180° and radius of 1.5 patch units.

During division each stem cell acquires various mutations that are inherited by their daughter cells (see section 7.6 for details on how mutation is modeled). Finally the ABM assumes that the crypt can determine if it has too few stem cells, and will respond by having the most-fit stem cell hatch one daughter stem cell. The only time this will occur is if one stem cell acquires a metastatic mutation, giving it the ability to roam throughout the crypt. Again, this is modeled based on the assumption that the number of stem cells in a crypt is tightly regulated.

7.5.2 *Transit Cells*

Wild-type transit cells produced by stem cells will only divide if there is an empty patch within their cone of vision, again simulating contact inhibition. However, transit cells must also meet further requirements to divide. They must be in the inner crypt, where they are not fully differentiated; they have enough oxygen to divide; and they have telomeres remaining. If a transit cell meets these requirements, divides, undergoes mutation, produces one daughter cell, divides its oxygen equally between itself and the daughter cell, looses one unit of its telomere, metabolizes oxygen, and randomly moves to one of the empty patches in its cone of vision, where it consumes oxygen. If the transit cell meets all of the criteria except for having enough oxygen to divide, the cell will simply move to one of the empty patches, metabolizing oxygen in the process, and then consume more oxygen on the patch it now occupies. If there is not enough oxygen in the patch the cell moves to, it will consume half of what oxygen is available. When a cell reaches the outer crypt it will follow the same rules of movement and oxygen consumption, but will not be able to divide because the cell has differentiated. Finally, transit cells are shed once they reach the lumen. Transit cells can also die if they lose all of their telomeres or have a low relative fitness when resources become scarce during population growth (see subsection 7.8.2 for details).

7.6 MODELING MUTATION

7.6.1 *Genome Regions*

It is estimated that the human genome contains ~7×10^9bp and 70,000 genes, each of which has an average length of 1000bp (reviewed in [92]). Assuming there are six barriers to cancer, and three genes per barrier, one can estimate that there are $6 \times 3 \times 1000 = 18{,}000$ bp that if mutated or targeted by JCV will

damage a cancer barrier. Interfering with these barriers is beneficial to the cell, as it either increases the rate of replication or probability of survival (see section 7.7 for details). Therefore, any mutation that lands in these 18,000bp is considered a beneficial mutation.

Assuming that mutations in the remaining genes are deleterious, one can estimate that mutations in any of the $(70,000 - 3 \times 6) \times 1000 = 69982000$bp will result in a mutation that will decrease a cell's fitness. The ABM also assumes there are seven different different genomic instability genes that are involved in Chromosomal Instability (CIN) [102]. Thus, there are $7 \times 1000 = 7,000$bp that, if mutated, will increase the mutation rate. Finally, assuming all other mutations are neutral, one can estimate that any mutation in the $7 \times 10^9 - (70000 \times 1000) = 6.93 \times 10^9$bp will have no effect on the cell's fitness.

7.6.2 *Generating Mutations*

7.6.2.1 *Host Mutations*

The stem cell mutation rate used in these models was originally $\mu = 10^{-10}$, and the transit cell mutation $\mu = 10^{-9}$, but only 1 out of 25,000 runs developed a metastatic tumor. This hardly provides any information, so the stem cell mutation rate is set to $\mu = 10^{-9}$, while the transit cell mutation rate is $\mu = 10^{-8}$. While these mutation rates are higher than observed values, they generate plenty of results and maintain the hypothesis that the stem cell mutation rate is lower than the transit cell mutation rate.

Using these mutation rates, one can calculate the expected number of mutations in each genome region during each division: $18,000 \times 10^{-8} = 1.8 \times 10^{-4}$ and $18,000 \times 10^{-9} = 1.8 \times 10^{-5}$ beneficial mutations in transit cells and stem cells, respectively; $69982000 \times 10^{-8} = 0.69982$ and $69982000 \times 10^{-9} = 0.069982$ deleterious mutations in transit cells and stem cells, respectively; $7000 \times 10^{-8} = 7 \times 10^{-5}$ and $7000 \times 10^{-9} = 7 \times 10^{-6}$ genomic instability mutations in transit and stem

cells, respectively; and $6.93 \times 10^9 \times 10^{-8} = 69.3$ and $6.93 \times 10^9 \times 10^{-9} = 6.93$ neutral mutations in transit cells and stem cells, respectively.

Randomly drawing numbers from a Poisson distribution, which predicts the number of successes in a fixed interval, allows one to randomly assign how many of each type of mutation occurs in each genome region during each cell division. For example, drawing a random number from a Poisson distribution with $\lambda = 69.3$ will determine how many neutral mutations land in a transit cell during division; a random number from a Poisson distribution with $\lambda = 0.069$ will determine how many deleterious mutations occur in a stem cell division.

If a mutation lands in a beneficial region, a random number is drawn from the Uniform distribution with a range of 1 to 6, one number for each cancer barrier (pro-growth, anti-growth, apoptosis, replication limit, metastasis, and CIF-II). There is an equal probability that each number will be chosen from the Uniform distribution, so using this method allows one to randomly select which barrier is mutated.

Each model is run using different numbers of genes per barrier, using either 3,6,9,12,or 16 genes per barrier. Therefore, as the number of genes per barrier increases, so does the probability of a beneficial mutation. For example, if there are 16 genes per barrier, there are $6 \times 16 \times 1000 = 96000$bp that can be mutated to increase the cells fitness. The length of the other gene regions will change accordingly as well.

Finally, it is likely the case that more than one mutation is required to completely remove a barrier. In this model, and like Knudson's two hit hypothesis, it is assumed that two events are required to completely remove a barrier, either by mutation or infection. However, one mutation still has an affect on the phenotype, and the barrier can be considered partially removed. If a cell has one mutation, the probability that the mutant phenotype will be expressed is 50%, which is modeled using a Bernoulli distribution where the probability of success is 0.5. Thus, everyday there is a 50% change that the mutant phenotype will be

expressed. If a cell has two mutations in a barrier, it is completely removed and the mutant phenotype is expressed 100% of the time.

7.6.2.2 *JCV Mutations*

Activation of JCV in the Hit and Run model requires mutations in the NCRR region, and these mutations are modeled in a similar fashion to host mutations. The NCRR region is 430bp in length and the mutation rate of JCV is estimated to between 7.8×10^{-4} and 4×10^{-6} per site per year [31, 47, 121]. This suggests that the average mutation rate of JCV is 1.074×10^{-6}bp/site/day. Thus, the probability that the JCV strain within single infected cell will acquire one NCRR on mutation on any given day is $1.074 \times 10^{-6} \times 430 = 4.6182 \times 10^{-4}$. Therefore, using the Poisson distribution with $\lambda = 4.618 \times 10^{-4}$ will generate how many NCRR mutations occur. Finally, NCRR mutations can only occur if the host cell is dividing.

7.6.3 *Genomic Instability*

Since genomic instability, particularly CIN, is believed to play a role in tumorigenesis, it is modeled as well. Every time a mutation lands in one of the seven genomic instability genes, the mutation rate increases linearly by a factor of two. For example, if there is one genomic instability mutation, the new mutation rate is double the original; if there are four genomic instability mutations the new mutation rate is eight times the original mutation rate; and so on until the new mutation rate is fourteen times the original mutation rate.

The JCV Mad-1 Δ98 strain is able to induce genomic instability in the Hit and Run model. This virally induced genomic instability is combined with existing genomic instability, so if the cell has 3 genomic instability mutations and the JCV Mad-1 Δ98 phenotype, the mutation rate will be increased $6 \times 7.5 = 45$ fold.

However, once JCV is lost, the mutation rate returns to six times the original mutation rate.

7.7 MUTANT BEHAVIOR

7.7.1 *Stem Cells*

Mutations that land in beneficial genes provide that cell with a selective advantage. However, the particular advantage depends on what type of cell mutates. If a stem cell acquires one mutation in its metastasis barrier it gains the ability to move around the crypt, moving to the neighbor with the most oxygen. Movement to the patch with the most oxygen is modeled because it has been proposed as the mechanism that selects for mobile metastatic cells [21]. If the mobile stem cell accumulates a second mutation in a metastasis gene, and is next to a blood vessel, it will try to invade the metastatic tissue (these blood vessels are produced during angiogenesis, see Section 7.8.1). However, the probability that the stem cell will survive the bloodstream and successfully invade the metastatic tissue is only 1/1000 [67]. If the stem cell does successfully invade the metastatic tissue it will go through four rounds of symmetric division to produce five metastatic stem cells, each of which continues to produce transit cells.

If the pro-growth barrier of a stem cell is disrupted, either by mutation or interaction with viral oncoproteins, it gains the ability to divide even if there is not an empty patch within its cone of vision, thus simulating the loss of contact inhibition. If both the metastasis and pro-growth barriers are removed, the mobile stem cell is able to move around the crypt and always divide in the inner crypt, but never in the outer crypt.

If a mobile stem cell has the anti-growth barrier removed, it is able to divide in the outer crypt, but only if there is an empty patch within its cone of vision. If this

mobile stem cell has both the anti-growth and pro-growth barriers removed, that stem cell will be able to move around the crypt and divide anywhere, everyday.

Stem cells that have removed the apoptosis barrier are exempt from the fitness search conducted during symmetric division, even if they have the most deleterious mutations. This behavior means that a stem cell with the most deleterious mutations, and possibly the most beneficial mutations too, will survive, spreading their mutations throughout the population. As stem cells are considered immortal, they do not have a replication limit barrier. Finally, any infected stem cell will reactivate JCV if the CIF barrier is removed by mutation, thus simulating reactivation by immunosuppression.

7.7.2 *Transit Cells*

Transit cells that acquire one metastasis mutation change their behavior from randomly moving to a patch in their cone of vision to moving to the neighbor patch that has the most oxygen. However, they follow all other wild-type division rules, and thus will only divide if they are in the inner crypt and there is an empty patch in their cone of vision. If the transit cell accumulates a second metastatic mutation, and is next to a blood vessel, it will attempt to invade the metastatic tissue, again with the probability of success being 1/1000. If the metastatic transit cell is able to invade the metastatic tissue it will follow the same rules of division that it followed in the crypt, except that there are not regions in which the transit cell cannot divide.

If a transit cell removes the pro-growth barrier, it will always divide in the inner crypt, regardless of whether or not their is an empty patch within its cone of vision. However, it will remain unable to divide in the outer crypt. If the cell has both the metastasis and pro-growth barriers removed it will still divide anywhere in the inner crypt, but will chose to move to the neighboring patch with the most oxygen.

Transit cells having the anti-growth barrier removed gain the ability to divide in the outer crypt, so long as there is an empty patch in its cone of vision. However, if the cell also has pro-growth mutations they will acquire the ability to divide even when there are no empty patches, meaning that they can divide every day, anywhere in the crypt. If these cells have the anti-growth, pro-growth, and metastasis barriers removed they will be able to divide anywhere, everyday, and will move to the neighboring patch with the most oxygen.

If a transit cell has the replication limit barrier removed, it will stop losing its telomeres during each division. If this mutation occurs in concert with the pro-growth and anti-growth mutations, the cell will divide anywhere, everyday, and without limit. This cell will only stop dividing is if it is shed into the lumen, or has low fitness and cannot survive when resources become scarce during population growth (see 7.8.2). However, if the cell has the apoptosis barrier removed it will always survive when resources are scarce, even it has the lowest fitness.

7.8 TUMOR EMERGENCE

7.8.1 *Angiogenesis*

Unlike the probability models, this ABM does not assume that angiogenesis is the result of mutation and barrier removal. Instead, this model assumes that angiogenesis naturally emerges when the tissue becomes hypoxic. This tends to occur when there are too many cells for the amount of oxygen being produced by a normal crypt. The body thus responds by producing new blood vessels to supply the growing tissue with the oxygen it needs.

Angiogenesis is modeled by asking hypoxic patches that do not have any vessels with five patches to create a new blood vessels. Any other hypoxic patches within a radius of five patches secrete VEGF molecules, which migrate towards the closest vessel. Once the VEGF molecule is within 0.5 patch units of the blood

vessel, it stimulates the expansion of the existing blood vessel. The result is the gradual growth of new blood vessels, winding their way to the most hypoxic areas of the tissue.

Each of the new vessels is assigned a random lifespan that can be as high as 250 days. Each vessel adds oxygen to each patch, and there can be up to three vessels on a single patch. The increased amount of oxygen supplied by angiogenesis increases the number of cells that the tissue can support, allowing the population to increase in size.

The increase in population size is an important component of the model because, the more cells there are, the greater the chance that at least one will remove all of the barriers to cancer. Also, because angiogenesis tends to create areas with higher concentrations of oxygen, metastatic cells migrate towards the blood vessels, where they may attempt to invade the metastatic tissue. This behavior thus replicates the close relationship between angiogenesis and metastasis.

7.8.2 *Population Cap*

There are two population size limits in this model. The first is a limit of 300 cells in the colon crypt. The model assumes that the crypt only produces enough resources to support it's normal number of cells, ~250-300 cells. If the population rises above 300 cells, resources become scarce and only the most fit cells survive while the least fit cells die off. Least fit cells are defined as the cells that have the lowest amount of oxygen and the most deleterious mutations. However, any cells that inhibit apoptosis are not included in this fitness search, and so there is selection for apoptosis mutations. If enough cells accumulate the mutation, the population will grow beyond 300 cells.

Due to limitations in computing power, a second population limit has to be set. If this limit were not in place the large population sizes would slow the simulation to a crawl, making it difficult to complete 1000 runs of each model.

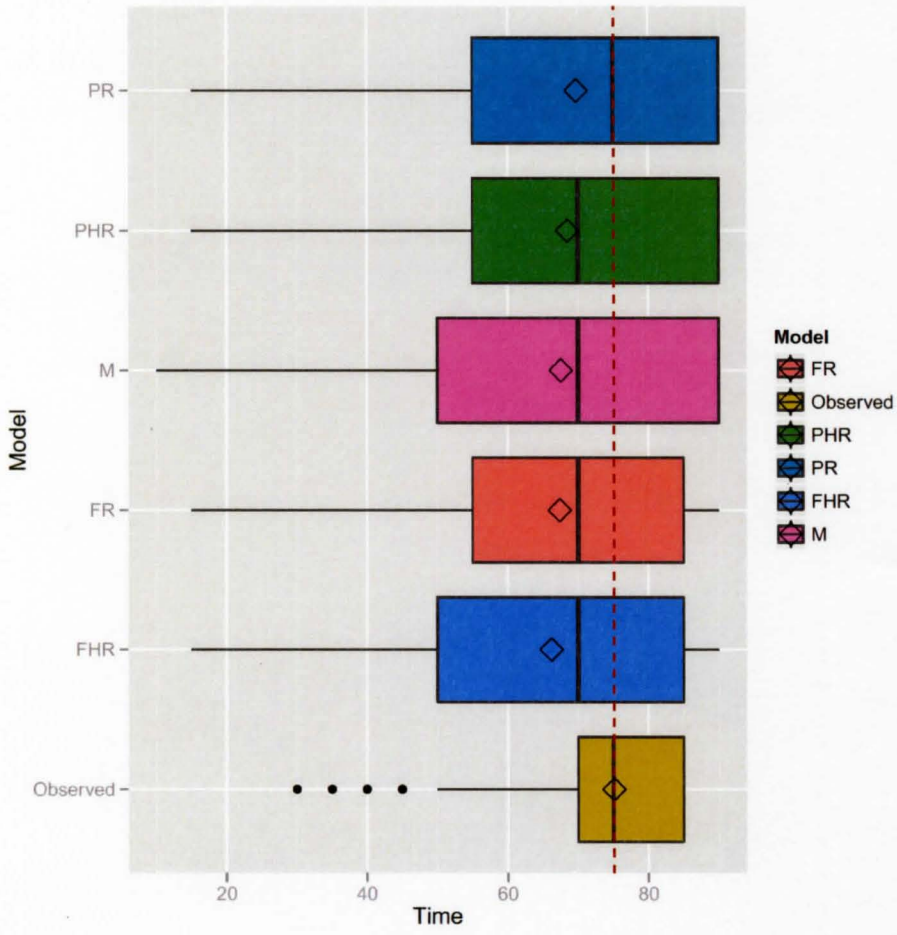
This second population cap is set at 5000 cells, which is more than a sixteen fold increase in population size. If this population cap is reached, cells are randomly selected for death until the population returns to 5000 cells. No cells are excluded from this search, so the probability of being killed is the same for all cells.

7.8.3 *Tumor Formation*

Due to the population cap and limits in computing power, it is not possible to diagnose tumor formation by tissue size. For example, a polyp forms when the population reaches a size equivalent to a sphere with a diameter of 2cm, but this is hundreds of thousands of cells and exceeds the amount of available computing power. Due to this limitation, tumors are diagnosed by the presence of cells that have removed most barriers. A colon tumor is considered to have formed when at least one cell in the population has completely removed the pro-growth, anti-growth, apoptosis, replication limit barriers.

A metastatic tumor forms when at least one cell successfully invades the metastatic tissue (which is only possible if the metastasis barrier is removed), and has all of the other barriers removed. In the case of stem cells, all barrier except the replication limit barrier must be removed. If such a metastatic tumor forms, the age of metastatic tumor formation is recorded and the run is stopped. The run is only stopped when metastatic tumors form because the statistics for colorectal tumor prevalence are for metastatic tumors [2], thus facilitating comparisons between modeled and observed data. The only other way a run is ended is if an individual reaches 100 years of age and has not developed a metastatic tumor.

7.9.1 *Average Age of Tumor Formation*

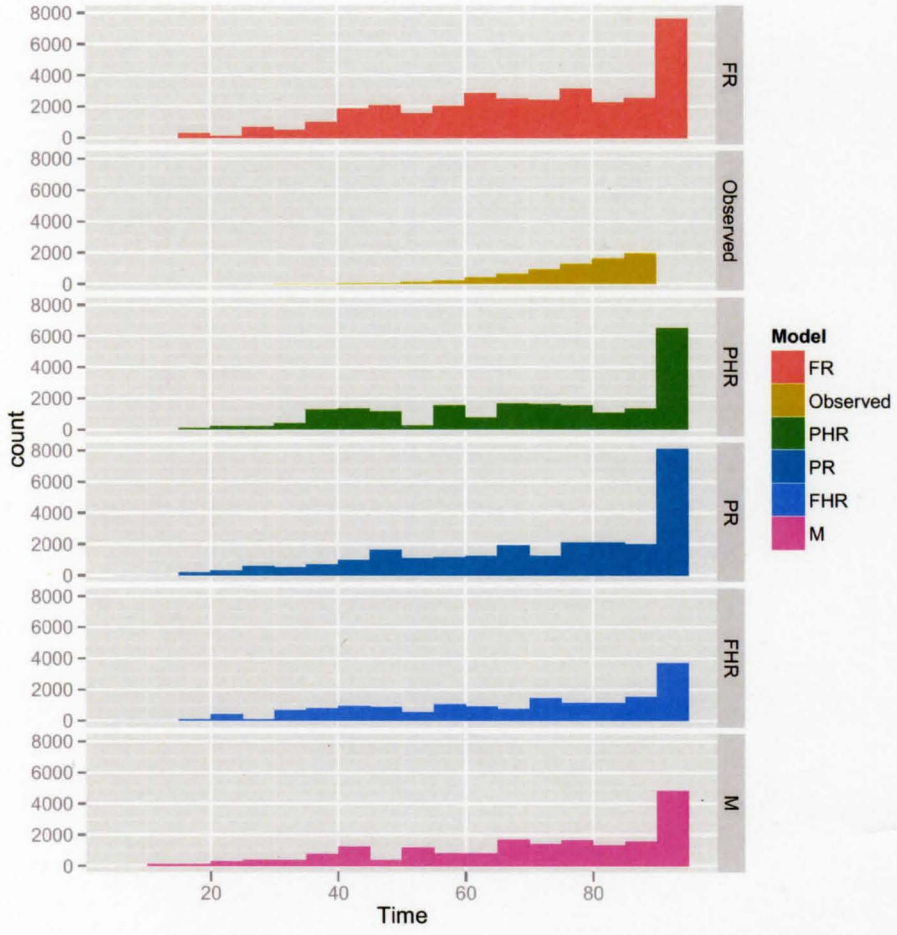


FR=Full Reactivation, PHR=Partial Hit and Run, PR=Partial Reactivation, FHR=Full Hit and Run, M=Mutation

$\diamond$ =mean, | =median, left bound=lower quartile, right bound=upper quartile

Figure 7.3: Age Distribution of Colorectal Cancer by Model

Figure 7.3 illustrates that all models tend to cause cancer primarily between the ages of 50-85. While spread of observed and modeled data are quite different, this finding is consistent with the observation that the average age of colorectal cancer is 77.



FR=Full Reactivation, PHR=Partial Hit and Run, PR=Partial Reactivation, FHR=Full Hit and Run, M=Mutation

◇ =mean, | =median, left bound=lower quartile, right bound=upper quartile

Figure 7.4: Age Distribution of Colorectal Cancer by Model

The age distribution of colorectal cancer in Figure 7.4 illustrates that the data are not normally distributed, and thus ANOVA cannot be used to compare the models. However, the Kruskal-Wallis one-way analysis of variance is able to compare multiple datasets that are not normally distributed. This test finds that the probability of the models being from the same underlying distribution is $p \lll 0.5$. This in turn suggests that the process underlying each model do significantly affect the age distribution of colorectal cancer.

The primary goal of this paper is determine the role of mutation and infection in colorectal cancer. Thus, it is useful to know if the infection models are different than the mutation model. Using the non-parametric two-sided Kolmogorov-Smirnov test on each infection model versus the mutation model reveals all infection models are significantly different than the Mutation model.

7.9.2 *Initiators of Tumorigenesis*

Genomic instability is often argued to be the driver of tumorigenesis, and active JCV would seem to be a key driver if infection plays a role in tumorigenesis. The ABM records all events and when they take place, providing the opportunity to examine which events most frequently initiate tumorigenesis. As Figure 7.5 illustrates, the first event is not necessarily the initiating event, as there is frequently a long time lag between it and the next event. Infection also cannot be considered an initiating event because latent JCV does not change the behavior of the cell. Therefore, the initiating event is here defined as the event that has the shortest time period between it and the next event. This definition is adopted because this is the event that accelerates, or at least jumpstarts, tumorigenesis, as the following events occur within a shorter time span than before. While this definition is not perfect, it at least provides some insight into which events accelerate the accumulation of beneficial mutations. Finally, the mutations of the parental stem cell could not be recorded, so it is not possible to determine exactly which barriers were removed in stem cells and which were removed in daughter transit cells.

Figure 7.6 illustrates how many times each barrier removal event initiated tumorigenesis. Inhibition of apoptosis and up-regulation of telomerase are the most frequent initiators, a result that is consistent with observation that p53 is removed in 80% of colorectal cancers, while telomerase is up-regulated in 85-95% of cancer cells [5, 123].

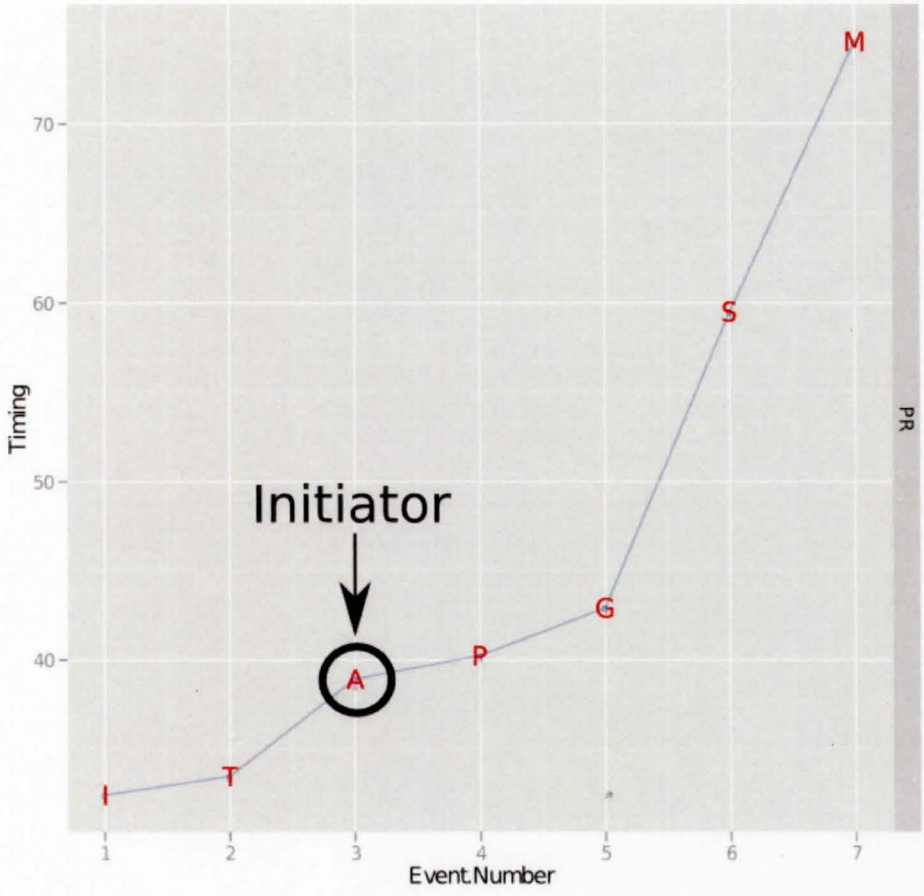
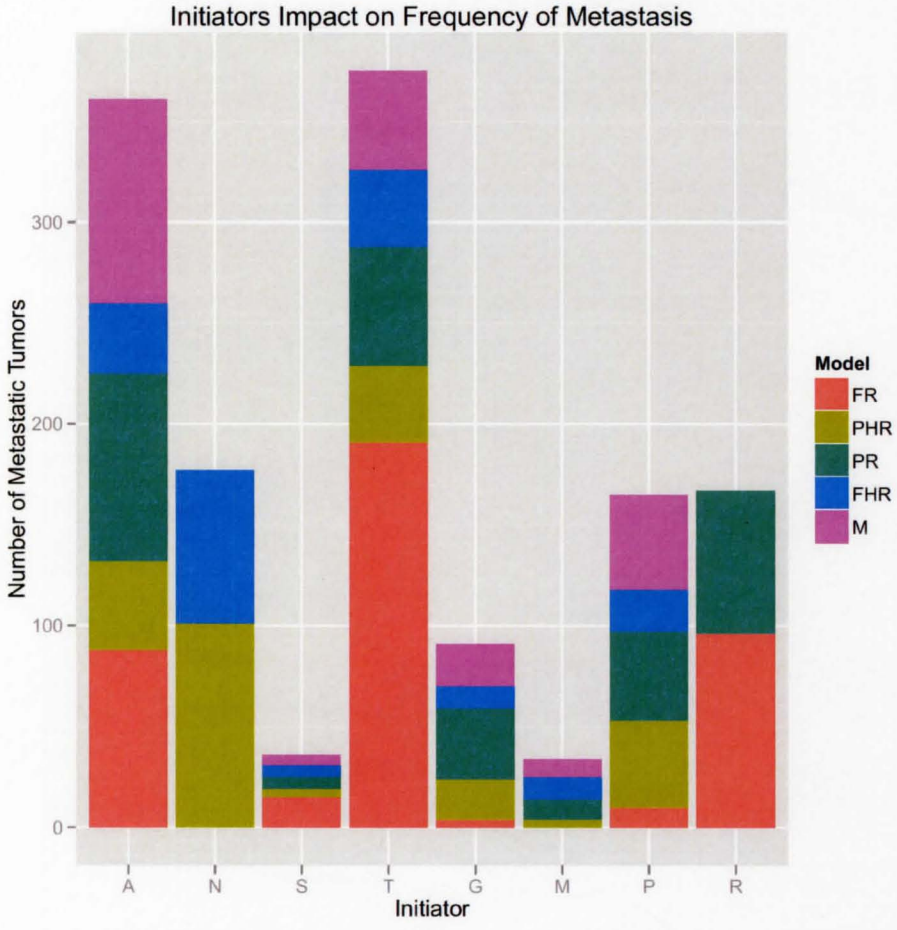


Figure 7.5: Example of Initiating Event

Inhibition of apoptosis may frequently be an initiating event in stem cells because it allows the stem cell to avoid being replaced during symmetric division, even if it has the most deleterious mutations. This increases the longevity of the stem cell, which may provide it with more time to accumulate additional mutations. However, up-regulation of telomerase does not affect the stem cell since it already has the ability to divide indefinitely.

Unlike stem cells, transit cells would benefit from up-regulating telomerase. However, even if a transit cell has the potential to divide indefinitely it will still be shed or die from accumulating too many deleterious mutations. Therefore, it is not immediately clear why up-regulation of telomerase is so frequently an initiating event, even in transit cells. Inspecting the mutations that precede the up-



Models: FR=Full Reactivation, FHR=Full Hit and Run, PR=Partial Reactivation, PHR=Partial Hit and Run, M= Mutation. Events:A=Apoptosis, N=Hit and Run, S=Genomic Instability,T=Telomerase,G=Anti-growth,M=Metastasis,P=Pro-growth,R=Reactivation

Figure 7.6: Initiating Events

regulation of telomerase sheds light on why this event may initiate tumorigenesis. Figure 7.7 shows that the complete removal of apoptosis and up-regulation of telomerase is preceded by a single metastasis mutation. This mutation gives the cell the ability to move to the patch with the most oxygen, even if those areas are lower in the crypt. The transit cell then moves throughout the crypt, looking for resources, and thus avoids being shed. Subsequent up-regulation of telomerase and inhibition of apoptosis provide the cell with the ability to divide indefinitely and avoid apoptosis no matter how many deleterious mutations that cell has. All three mutations together allow the cell to replicate without limit, but never die (apoptosis and metastasis). Such an immortal transit cell would have plenty of time to remove the remaining barriers, and given their higher mutation rate they may be able to do so at an accelerated pace. This hypothesis is supported by the research of Lamprecht and Lipkin [70], who presented evidence that transit cells can acquire mutations that allow them to remain in the crypt.

It may also be that up-regulation of telomerase is frequently an initiator in the infection models because, while the cells are immune to apoptosis, and can divide anywhere and everyday, they still have limited replicative potential. Without up-regulating telomerase, the ability to divide so frequently may backfire on the cells, as they can soon lose their telomeres. However, up-regulating telomerase via mutation gives that cell the ability to survive, and divide anywhere, everyday, and without limit. This hypothesis is consistent with the observation that cells expressing Large T Antigen (T-ag)often become immortalized, can escape contact inhibition, and exhibit anchorage-independent growth, but only if they also up-regulate hTERT (reviewed in [47]). FInally, this phenotype may provide the cell with the more opportunities to accumulate additional mutations, as it is constantly dividing.

Reactivation of JCV by the removal of the CIF-II barrier is also common. Such reactivation is accompanied by inhibition of apoptosis, and removal of the pro- and anti-growth barriers. Even if no other barriers are removed, this combination

of events allows the cell evade apoptosis and replicate more frequently. Since mutation occurs during every division, reactivation gives the cell the opportunity to accumulate more beneficial mutations at a faster rate. A similar situation occurs when JCV hits and runs, except that removal of the barriers and increase in mutation rate only lasts for $14-21$ days. Even though this is a short period of time, this event (N) can either mutate other tumor suppressor or oncogenes, or initiate genomic instability by mutation genes involved in CIN. This latter event would be particularly important, as it allow a mutator phenotype to last after Δ98 Mad-1 is lost [109].

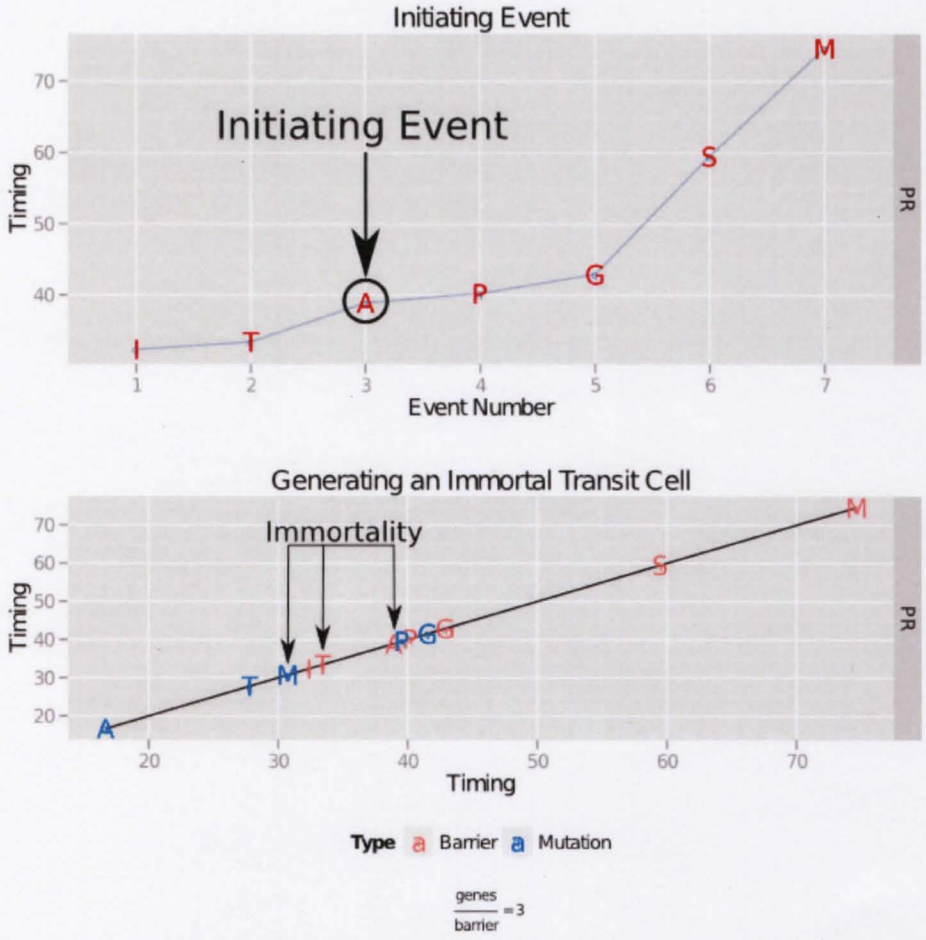


Figure 7.7: Creating an Immortal Cell

7.9.3 *Role of Genomic Instability*

While **CIN!** (**CIN!**) is found in 65-70% of sporadic cancers, there is debate over whether or not it induces or exacerbate tumorigenesis [4, 102]. As before, genomic instability is considered to initiate tumorigenesis if the following events occur rapidly. However, it is considered an exacerbating event if it occurs within an individual but does not initiate tumorigenesis. Finally, if genomic instability never occurs it does not play a role in colorectal cancer. Defining the role of genomic instability in this manner produces the results found in Figure 7.8.

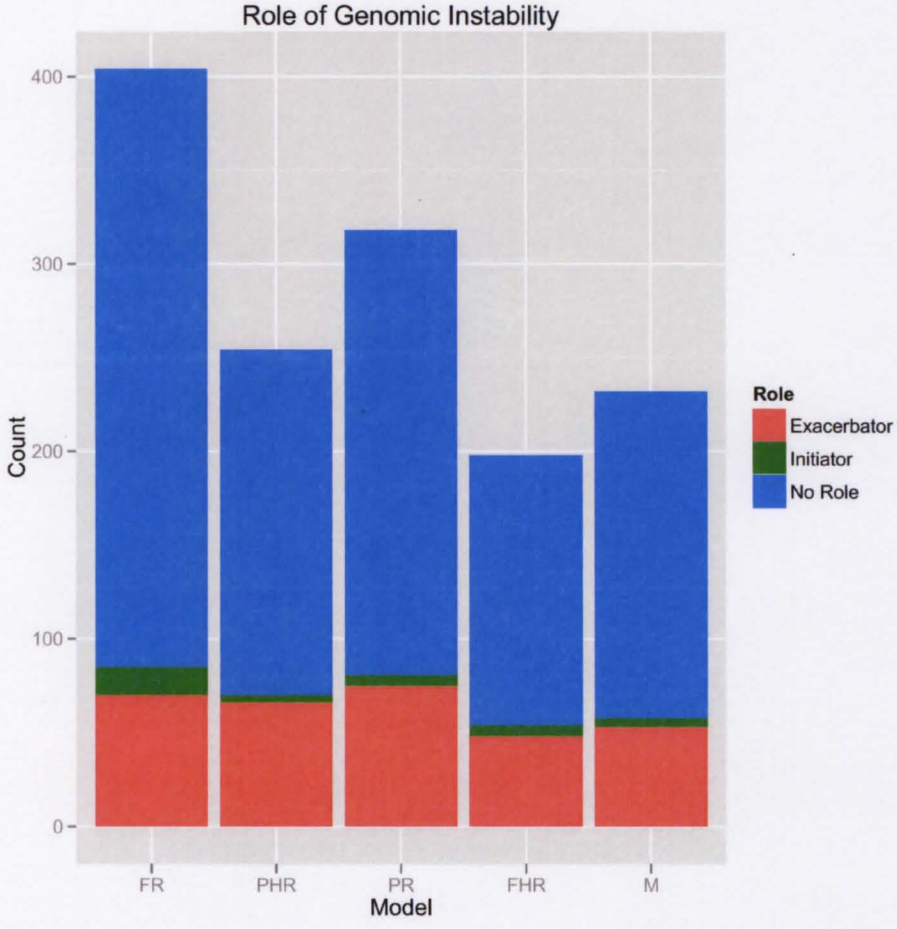


Figure 7.8: Creating an Immortal Cell

These results suggest that when present, genomic instability generally exacerbates tumorigenesis. Figure 7.8 also illustrates that, as modeled, genomic instability does not often occur in colorectal cancer. This is in contrast to laboratory studies, which find CIN occurs in 65-70% of sporadic cancers, including colorectal cancer [102]. This indicates that the ABM does not accurately genomic instability. This could for several reasons: there need to be more genomic instability genes; genomic instability needs to occur when certain barriers are removed, such as apoptosis; genomic instability needs to increase the mutation rate more than modeled. However, these modifications may not change the most of the conclusions because each model takes advantage of genomic instability. Both Mutation and Reactivation models can mutate genomic instability genes, increasing the chances another barrier is removed. JCV-induced genomic instability can increase the cell's mutation rate by mutating the hosts own genomic instability genes. Thus, the overall conclusions may remain similar, but there would like be more tumors in each model. Even so, any future in carnations of this model could make the above modifications.

7.9.4 *The Role of Infection*

If infection plays a role in colorectal tumorigenesis it would be useful to understand how JCV increases the risk of cancer. Does JCV frequently initiate the process, or does it exacerbate it? If either are true, this knowledge could be used to prevent or treat colorectal cancer. Here, an initiating role for JCV is when the virus' reactivation, either by hit and run or removal of CIF-II, is also the initiating event. JCV is considered to be an exacerbator when activation occurs but is not the initiating event. JCV has no role if it never becomes active, and mutation removes all barriers. Non infected individuals are those who were generated during an infection model, but were never infected by JCV. Using these definitions,

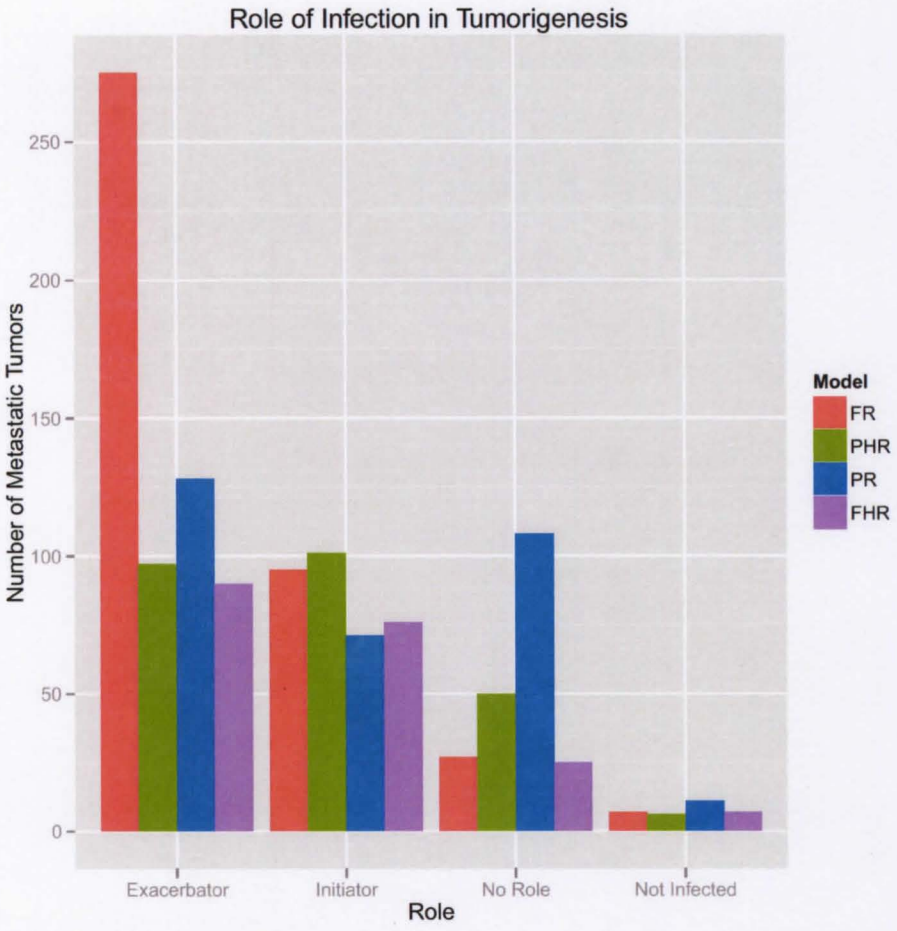


Figure 7.9: Role of Infection in Colorectal Cancer

the distribution of roles JCV plays in colorectal tumorigenesis can be found in 7.9.

These results indicate that, on average, JCV exacerbates colorectal cancer. This is likely because immunosuppression, via removal of CIF-II, occurs at 59.2 years (SD=20.2 years) in the Full Hit and Run model, while in the Partial Reactivation model activation occurs at 62.9 years(SD=19.8years). Similarly, the average age of the first hit and run event is 57 years and 59 years in the Full and Partial Hit and Run models, respectively. Thus, JCV may primarily act as an exacerbator because the individual has already accumulated several mutations, and active

JCV is able to remove the rest, either by forcing the cell to divide (Reactivation models) and/or inducing genomic instability (Hit and Run models).

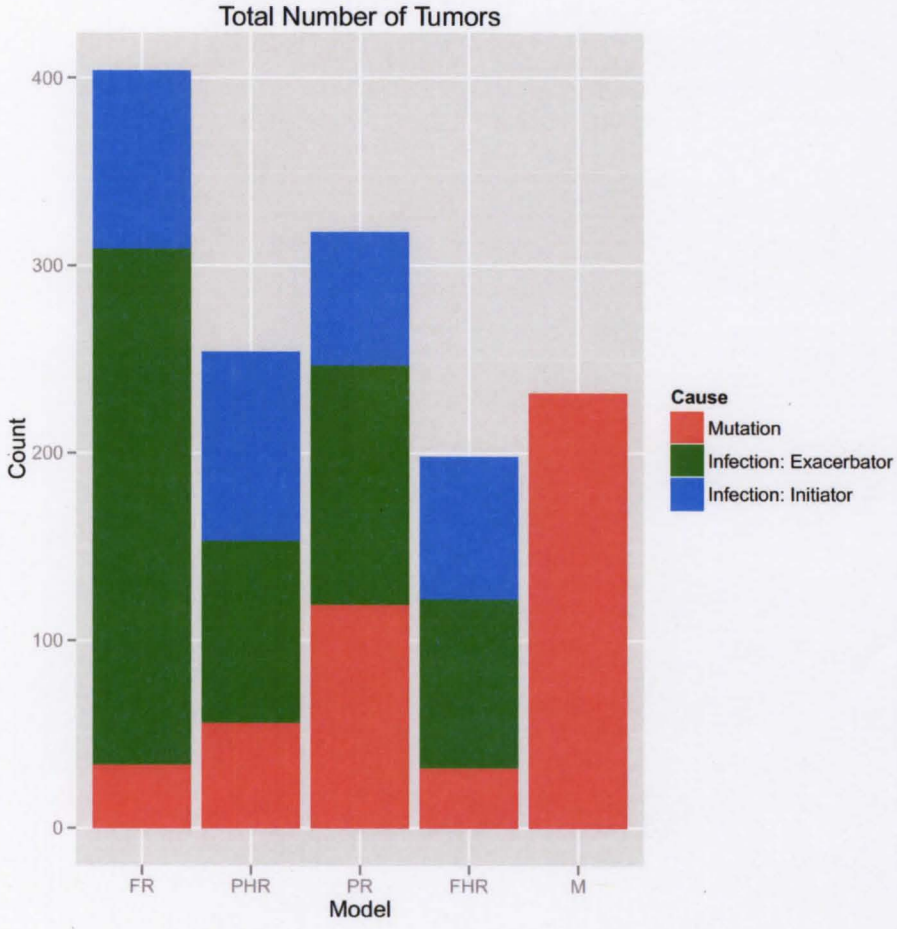
JCV is also frequently an initiator of colorectal cancer. If reactivation of JCV occurs early enough, it would be able to keep the cell alive and dividing constantly, increasing the chances of future mutations. However, this seems to occur far less often than when JCV exacerbates colorectal cancer. This likely because the probability of completely removing the CIF-II barrier, inducing immunosuppression, is only $(16000 \times 10^{-8})^2 = 2.56 \times 10^{-8}$, at most. This event is so unlikely that it wouldn't occur in most individuals, and if it did they would most likely be older, and thus more likely to already have accumulated initiating mutations.

The probability of JCV evolving into the $\Delta 98$ Mad-1 phenotype and hitting and running is $(1.074 * 10^{-6} * 430\text{bp})^2 = 2.132777 \times 10^{-7}$. Again, this is a fairly rare event, and tends to occur in older individuals. However, if JCV did hit and run earlier, it would likely initiate cancer by inducing genomic instability.

7.9.5 *Number of Tumors Formed*

Another question one might ask is which model generates the most tumors. This question can be answered by summing how many of the 25,000 individuals developed cancer, and binning by model. One can further determine whether mutation or infection are responsible for tumorigenesis. Mutation is considered the cause if a tumor formed when JCV is not present or played no role. Infection is considered the cause when it either exacerbates or initiates tumorigenesis. The results can be found in Figure 7.10.

An important conclusion that can be drawn from Figure 7.10 is that mutation is able to cause colorectal cancer in the absence of JCV. This is true across all models. Since mutation is required for tumorigenesis, and infection is not, mutation can be considered the primary cause of colorectal cancer.



FR=Full Reactivation, PHR=Partial Hit and Run, PR=Partial Reactivation, FHR=Full Hit and Run, M=Mutation

Figure 7.10: Number of Tumors by Model

A second conclusion one can draw from Figure 7.10 is that, in the worlds where infection can cause cancer it does so as much or more often than mutation. This finding indicates that even though infection is not required for tumorigenesis, JCV does play in important role in colorectal cancer.

7.10 also illustrates how much exacerbation by JCV increases the risk of cancer. Three of the four infection models generate more tumors than mutation, indicating that active JCV increases the risk of colorectal cancer. In all cases JCV's primary role is that of an exacerbator. This adds weight to the hypothesis that activation of JCV occurs after an individual accumulates several mutations, and that JCV is able increase the chances all barriers are removed by its ability inhibit apoptosis and the pro- and anti-growth barriers. In other words, JCV is able to complete the process of tumorigenesis in individuals that would otherwise not have developed colorectal cancer. Combined with the finding that JCV activates at ~60 years of age, this suggests that JCV should cause more tumors in older individuals than the mutation model does. As Figure 7.4 illustrates, this is the case.

Figure 7.10 also illustrates that the Reactivation models generate the most tumors, suggesting that JCV is most tumorigenic when the host is immunocompromised. This is likely because the formation of a cancer stem cell only requires that the cell already removed three barriers, and so JCV only needs to remove the other three. This is particularly true for the Full Reactivation model because JCV completely removes those three barriers. This is in contrast to the Partial Reactivation model, as the cell must acquire additional mutations to completely remove the apoptosis, anti-growth, and pro-growth barriers. This hypothesis is consistent with the finding that the Full Reactivation model generates many more tumors than the Partial Reactivation model.

A surprising finding is that the Full Hit and Run model generates the fewest tumors. A possible explanation has to do with that fact activation of Δ98 Mad-1 JCV generates genomic instability and completely removes the apoptosis, pro-

growth, and ant-growth barriers. Genomic instability will make the cell generate more mutations than normal, with more being deleterious than beneficial. Removal of the apoptosis, pro-growth, and ant-growth barriers allows the cells to divide anywhere and everywhere without being killed. Thus, this cell will produce a large volume of daughter cells with the same phenotype. Together this means that the population will grow rapidly, and that the cells driving the growth will have an increased number of deleterious mutations. Angiogenesis is a fairly slow process, as it takes time for the vessels to spread throughout the crypt, and so it may not be able to provide the rapidly growing population with the oxygen it needs. This not a problem when Δ98 Mad-1 is active, but as soon as it deactivates all of the cells that did not already have apoptosis removed will likely die because they have too many deleterious mutations. Therefore, the only cells that will survive are those that either already inhibited apoptosis, or had JCV-induced genomic instability remove the barrier. This would remove most of the cells that were hit and run by JCV, almost making the impact of hit and run minimal. This hypothesis seems to match up with the finding that the Full Hit and Run and Mutation models generate a similar number of tumors.

This scenario does is not necessarily true for the Partial Hit and Run model. In this model, while there is JCV-induced genomic instability, the pro-growth, anti-growth, and apoptosis barriers are only removed half of the time. This will results in a smaller rate of population growth for two reasons: 1), the population does not grow as rapidly because the pro- and anti-growth barriers are only removed half of the time, and so the cell cannot divide every time and everywhere; 2) cells that accumulate large numbers deleterious mutations can be die when apoptosis is not being inhibited by Δ98 Mad-1 JCV. The decreased rate of population growth means that, compared to the Full Hit and Run model, there will be more resources available to cells that survive. Some of these cells would have acquired beneficial mutations and would have more resources than cells in the

Full Hit and Run model, so they would be more likely to survive and accumulate more mutations that result in the formation of a metastatic tumor.

Finally, the above scenario also does not apply to the Reactivation models because after JCV becomes active apoptosis is inhibited for the remainder of that cell's life, and thus will not be killed when resources become scarce. This is particularly true for the Full Reactivation model, which may also help explain why it causes the most tumors.

7.9.6 *Metastatic Cell Type*

It is generally believed that stem cells are the cells which metastasize. The reasoning behind this is that stem cells have a long lifespan, giving them plenty of time to accumulate all of the necessary mutations. This is in contrast to transit cells, whose existence is fleeting, theoretically preventing them from acquiring all mutations needed for metastasis. To test this hypothesis, the ABM records which cell type, transit or stem, metastasizes. The results can be found in Figure 7.11

It seems reasonable to assume that because transit cells inherit their parental stem cell's mutations, they may only have to acquire one more mutation to have the opportunity to metastasize. Furthermore, there is only one of these parental stem cells, while over the course of several years that stem cell will produce hundreds or thousands of daughter cells, any of which can acquire that last mutation. That transit cells have a higher mutation rate makes this even more likely. It is also quite possible that the transit cell inherits the stem cell mutations that confer immortality. This long lifespan, combined with a higher mutation rate, gives these transit cells many opportunities to accumulate beneficial mutations, more so than stem cells, which have a lower mutation rate. This hypothesis is consistent with the evidence that transit cells may undergo mutation and selection that enables them to linger in the crypt, giving them time to accumulate the extra mutations required for tumorigenesis [56, 70].

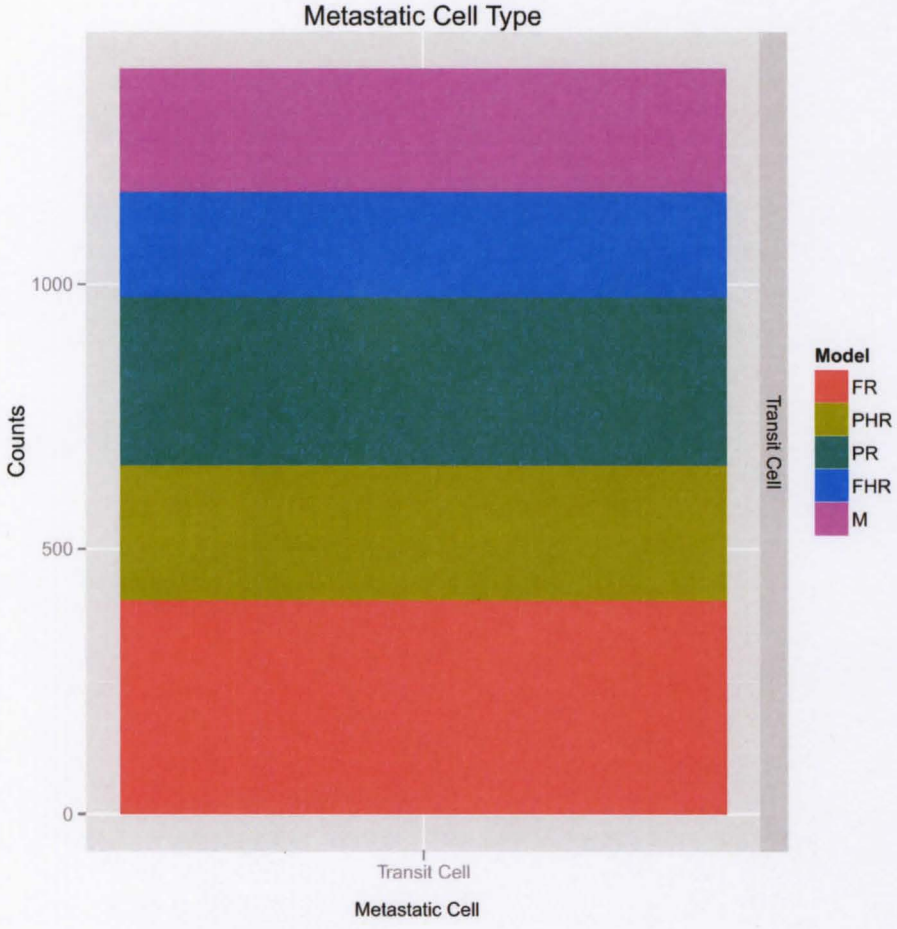


Figure 7.11: Metastatic Cell Type

7.9.7 *Prevalence*

The world of the ABM can be interpreted in one of two ways.The first interpretation is that the world represents one crypt in a single individual, and each run represents 1 out of 1,000 individuals. The second interpretation is that all runs represent 1,000 crypts in a single individual. Either situation is far from realistic, as it has been estimated that there are 1.5×10^7 colon crypts in an individual. That prevalence can only be calculated by assuming that each run simulates the events in a single crypt means that modeled prevalence values must be interpreted with caution. Even so, modeled prevalence values may reveal the age distribution one might expect given each model's hypothesis, and thus which model is most likely to be realistic. Each of these models were tested using 3, 6,9,12, or 16 genes, which will also shed some light on the number of genes per barrier. The modeled prevalence values are found in Figure 7.12, while the Euclidian distance between the observed prevalence and modeled prevalence is found in Figure 7.13.

The modeled prevalence values suggest that the Full Reactivation model with six genes per barrier best replicates the observed prevalence, supporting the hypothesis that infection plays an important role in tumorigenesis. Both the Mutation and Full Hit and Run models with nine genes per barrier are not too far behind. These results are consistent with the finding that all models are able to generate tumors.

7.10 CONCLUSIONS

The findings presented above reveal that mutation is the primary cause of colorectal cancer, as it is able to generate a large number of tumors without JCV. Mutation is also critical in tumorigenesis, not only because it must remove the barriers infection cannot, but because mutations that up-regulate telomerase and inhibit apoptosis frequently initiate tumorigenesis.

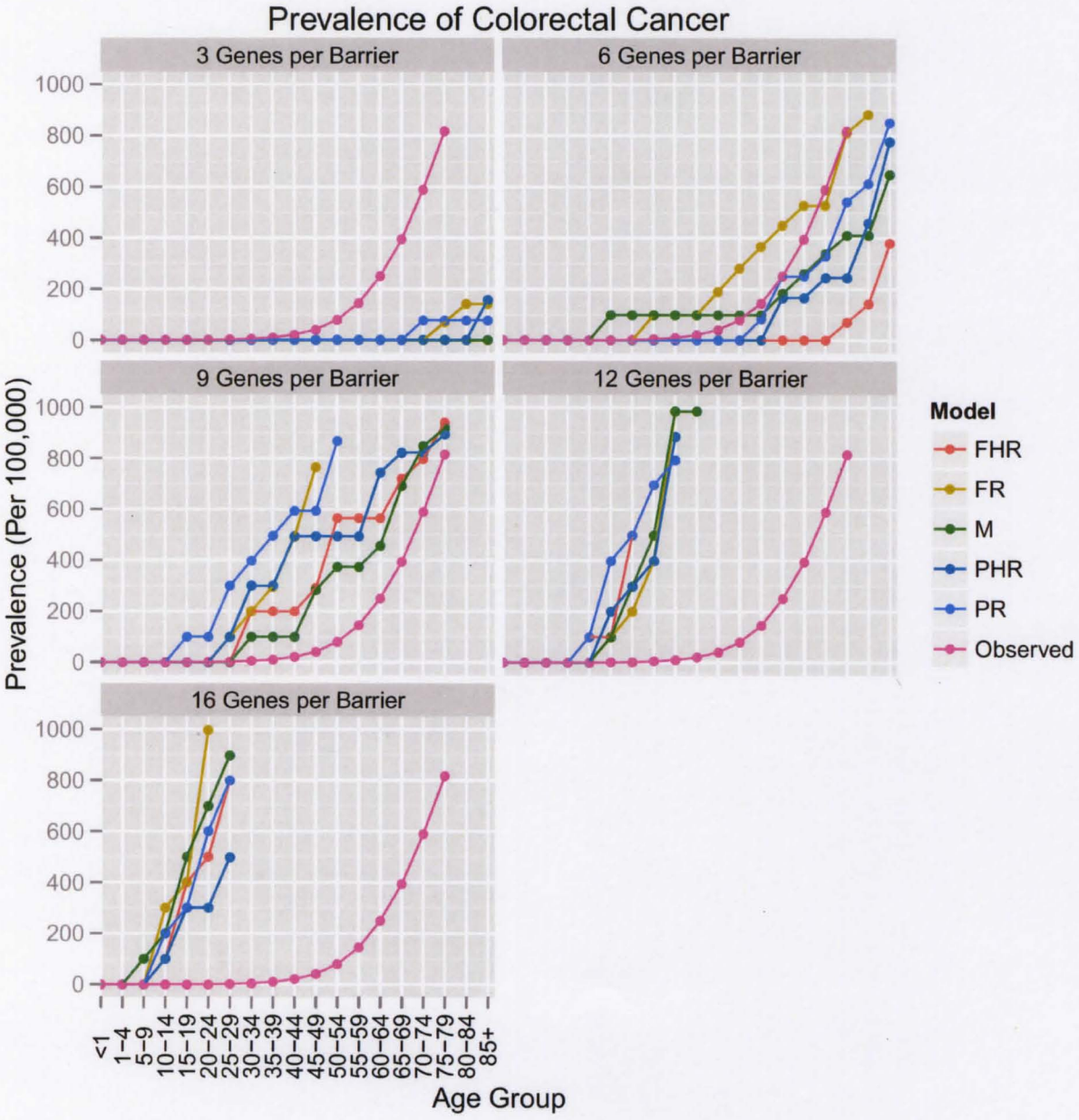


Figure 7.12: Modeled Prevalence, ABM

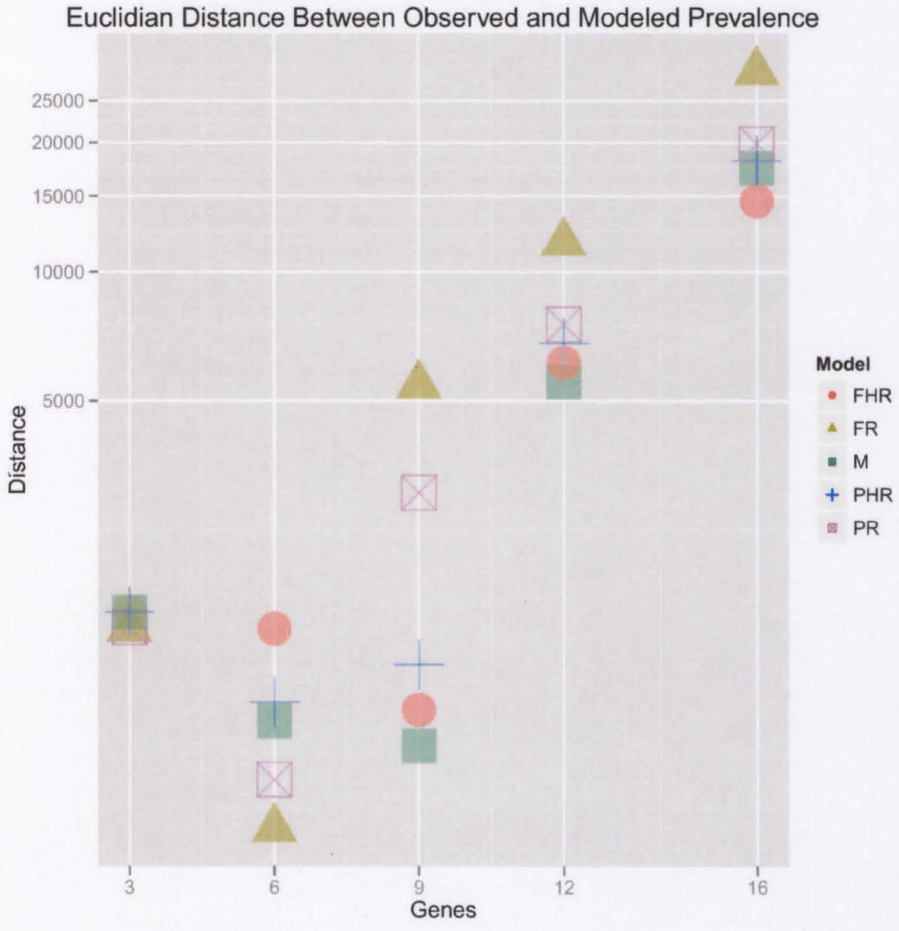


Figure 7.13: Euclidian Distances, ABM

While mutation is the primary cause of colorectal cancer, infection plays an important secondary role, usually exacerbating cancer. The Reactivation models, which posit that JCV is activated by mild immunosuppression, is most frequently an exacerbator, but is highly tumorigenic. This is likely because immunosuppression occurs later in life, after tumorigenesis is already initiated. Once activated, JCV removes three important barriers that keep itself and the cell replicating and immune to apoptosis. Together, these processes will keep the cell alive long enough to accumulate additional mutations, thus increasing the chances that a cancer stem cell will evolve.

The Hit and Run models are also most frequently exacerbators, but they are less tumorigenic. The transient nature of hit and run means that the cell will only have genomic instability and the transformed phenotype for a short period of time, after which it returns to its previous state. The cell will only become more carcinogenic if it removes other barriers while the Δ98 Mad-1 phenotype is present. JCV-induced genomic instability may mutate the host's genomic instability genes, maintaing a mutator phenotype after Δ98 Mad-1 is lost, thereby increasing the chances additional mutations accumulate [109]. This won't happen with every cell, which may explain why the hit and run mechanism is less tumorigenic.

While JCV most often exacerbates tumorigenesis, it also frequently initiates the process. Activation of JCV by immunosuppression may be able to induce cancer by forcing the cell to divide, increasing the chance of mutation and possible barrier removal. Since the cell is constantly dividing, the likelihood of a beneficial mutation occurring after reactivation is higher than if JCV is not present, which may be how JCV initiates tumorigenesis in the Reactivation model. If a hit and run event occurs early, JCV-induced genomic instability could initiate tumorigenesis by increasing the chances that a mutation in a beneficial gene soon occurs, thus initiating tumorigenesis.

Another important finding is that in the infection models mutation is responsible for fewer tumors than in the model where mutation is the only cause of colorectal cancer. This may be because individuals have already acquired mutations in the cancer barriers prior to activation of JCV, which usually occurs around 60 years of age. JCV can remove the rest of the barrier either by keeping the cell alive long enough for it to acquire additional mutations and/or by generating genomic instability. The finding that, more often than not, JCV exacerbates tumorigenesis is consistent with this hypothesis.

It is difficult to say which infection model is most accurate, as the Hit and Run models have the most experimental support, while the Full Reactivation model best replicates the observed data. However, in all cases mutation is the primary cause of colorectal cancer, but JCV plays an important role by exacerbating, and less frequently initiating, colorectal cancer. Given that mutation and infection play key roles in tumorigenesis, colorectal cancer can be considered a multifactorial disease.

CHAPTER 8 CONCLUSIONS

Taken together, the models presented herein indicate that both mutation and infection play important roles in colorectal tumorigenesis. Each of the probability models find that mutation is insufficient to drive colorectal cancer, no matter if the stem cell point mutation rate is $\mu = 10^{-10}$ or $\mu = 10^{-11}$. Conversely, the infection model is able to generate realistic incidence values when $\mu = 10^{-11}$. These results suggest that infection is the primary cause of colorectal cancer, as tumorigenesis will not occur in the absence of JCV. However, the Geometric model comes to the opposite conclusion, as it finds that mutation alone is able to drive colorectal cancer. Where both models agree is in finding that the presence of infection dramatically increases the risk of colorectal cancer.

The ABM appears to resolve the conflicting results of the probability and Geometric models. This collection of models finds that both mutation and infection are able to drive tumorigenesis, and that all models can generate realistic prevalence values. Since colorectal cancer does not absolutely require infection, JCV cannot be considered the primary cause. Despite its secondary role, the presence of active JCV increases the risk of cancer, as it increases the number of tumors and initiates or exacerbates tumorigenesis more often that it plays no role. The finding that both mutation and infection play important roles in colorectal tumorigenesis is also consistent with the estimate that ~25% of colorectal cancers result from multifactorial contributions of different risk factors [9]. While the authors argue that these 25% of cases occur as the result of inheriting many rare dominant alleles that have low penetrance, but together increase the risk of colorectal cancer, the results presented here suggest the alternative hypothesis

that infection is one of those critical environmental factors that accounts for an increased risk of cancer.

Colorectal cancers are usually divided into two categories, but these results suggest that it should be divided into three. This first is all colorectal cancers caused by germline mutations, primarily Family Adenomatous Polyposis (FAP) and Hereditary Nonpolyposis Colorectal Cancer (HNPCC). This category contains the fewest individuals, likely because the high-risk alleles reduce the individual's fitness, and have decreased in frequency due to negative selection. The second category is colorectal cancer caused soley by somatic mutations. As modeled, this is the second largest category. The individuals in this category likely develop colorectal cancer through the accumulation of the mutations described in Chapter 2, namely Adenomatous Polyposis Coli (APC) , Kristen Rat Sarcoma Virus (KRAS), SMAD, and p53.

The third and new category is JCV associated colorectal cancers. As modeled, this is the largest category of colorectal cancer cases. Individuals in this category may frequently have somatic mutations in APC , KRAS, SMAD, and/or p53, predisposing them to colorectal cancer. In the absence of JCV, these individuals may simply develop benign colorectal tumors, which are present in 50% of the population [64]. Activation of JCV, either by immuosuppression or the evolution of Δ98 Mad-1, typically occurs at age 60, after the somatic mutations have occurred. This event may transform the tumor from benign to malignant, causing tumors in individuals that would not have developed malignant colorectal cancer without JCV.

In addition to shedding light on the drivers of tumorigenesis, these models add weight to the hypothesis that natural selection has favored a lower stem cell mutation rate. While evolution has the power to select against the germline mutations that increase the risk of FAP and HNPCC, it cannot directly select against somatic mutations. However, by selecting for a lower mutation rate in stem cells, evolution can decrease the frequency of colorectal cancer. Both the

probability and Geometric models confirm this hypothesis, as mutation is unable to generate cancer patients when the stem cell mutation rate is 100 times lower than the transit cell rate. This finding is in agreement with the work of Frank et al. [43], who also concluded that a stem cell mutation rate that is 100 times lower provides sufficient protection against cancer.

Unfortunately, not much can be said about the number of genes per barrier, as the number that fits best is different across each set of models. The primary conclusion that can be made is that the more genes that are involved in a pathway, the more likely that pathway is to be disturbed.

The results from these models suggest colorectal cancer is a multifactorial disease. Mutation is required for tumorigenesis, but infection by JCV increases the risk of colorectal cancer, either by initiating or exacerbating colorectal cancer. In combination with the findings of numerous studies that demonstrate JCV's oncogenic potential and frequent presence in colorectal tumors, the results presented here suggest that JCV's role in colorectal cancer deserves more attention. A good place for future studies to start might be to determine titers of Mad-1 JCV in colon tumors of cancer patients throughout their treatment. If it is found that individuals with higher titers of JCV Mad-1 are at higher risk of colorectal cancer, it would reinforce the hypothesis that JCV is an important risk factor for colorectal cancer. If further studies confirm these findings there would be good reason to develop a vaccine against JCV. While vaccination would not eradicate colorectal cancer, as mutation can still drive tumorigenesis in the absence of infection, it would reduce the number of colorectal cancer cases. These results, therefore, are encouraging, as they present the possibility of decreasing the prevalence of colorectal by reducing the rate of infection by JCV.

REFERENCES

[1] Cancer facts and figures 2012. Technical report, American Cancer Society, Atlanta, GA, 2012. (Cited on page 21.)

[2] Surveillance, epidemiology, and end results (seer) program. Technical report, National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, November 2012. URL http://seer.cancer.gov/. (Cited on pages 60, 78, 98, 148, 159, and 191.)

[3] RJ Albertini, JA Nicklas, JP O'Neill, and SH Robison. In vivo somatic mutations in humans: Measurement and analysis. *Genetics*, 24:305–326, 01 1990. doi: 10.1146/annurev.ge.24.120190.001513. URL http://pubget.com/paper/2088171. (Cited on page 10.)

[4] Tannaz Armaghany, Jon D. Wilson, Quyen Chu, and Glenn Mills. Genetic alterations in colorectal cancer. *Gastrointest Cancer Research*, 5(1): 19–27, Jan-Feb 2012. URL http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3348713/. (Cited on pages 12, 13, 21, 23, 24, 25, 26, 29, 32, and 106.)

[5] SJ Baker, AC Preisinger, JM Jessup, C Paraskeva, S Markowitz, JK Willson, S Hamilton, and B Vogelstein. p53 gene mutations occur in combination with 17p allelic deletions as late events in colorectal tumorigenesis. *Cancer Res*, 50(23):7717–7722, 12 1990. URL http://pubget.com/paper/2253215. (Cited on pages 26 and 101.)

[6] JR Berger and M Concha. Progressive multifocal leukoencephalopathy: The evolution of a disease once considered rare. *J Neurovirol*, 1(1):5–18, 03 1995.

doi: 10.3109/13550289509111006. URL http://pubget.com/paper/9222338. (Cited on page 34.)

[7] NP Bhattacharyya, A Skandalis, A Ganesh, J Groden, and M Meuth. Mutator phenotypes in human colorectal carcinoma cell lines. *Proc Natl Acad Sci U S A*, 91(14):6319–6323, 07 1994. URL http://pubget.com/paper/8022779. (Cited on page 29.)

[8] M. Bjerknes and H. Cheng. Clonal analysis of mouse intestinal epithelial progenitors. *Gastroenterology*, 116(1):7–14, January 1999. (Cited on page 66.)

[9] Walter F. Bodmer. Cancer genetics: Colorectal cancer as a model. *J Hum Genet.*, 51(5):391–396, 2006. (Cited on pages 31, 46, 84, and 120.)

[10] C. Richard Boland. Evidence for an association between jc virus and colorectal neoplasia. *Cancer Epidemiol Biomarkers Prev*, 13:2285–2286, 2004. (Cited on page 51.)

[11] Bruce M. Boman and Emina Huang. Human colon cancer stem cells: A new paradigm in gastrointestinal oncology. *Journal of Clinical Oncology*, 26 (17):2828–2838, 06 2008. (Cited on page 208.)

[12] Catherine Booth and Christopher S. Potten. Gut instincts: thoughts on intestinal epithelial stem cells. *Journal of Clinical Investigation*, 105(11):1493, 2000. (Cited on pages 18, 19, 20, 88, 89, 216, 226, 228, and 240.)

[13] DR Borger and JA DeCaprio. Targeting of p300/creb binding protein coactivators by simian virus 40 is mediated through p53. *J Virol*, 80 (9):4292–4303, 05 2006. doi: 10.1128/JVI.80.9.4292-4303.2006. URL http://pubget.com/paper/16611888. (Cited on page 41.)

[14] P Branch, DC Bicknell, A Rowan, W F Bodmer, and P Karran. Immune surveillance in colorectal carcinoma. *Nat Genet*, 9(3):231–232, 03 1995. doi: 10.1038/ng0395-231. URL http://pubget.com/paper/7773283. (Cited on page 29.)

[15] Tracy M Bryan and Thomas R Cech. Telomerase and the maintenance of chromosome ends. *Current Opinion in Cell Biology*, 11 (3):318–324, 6 1999. URL http://www.sciencedirect.com/science/article/B6VRW-3XJSRG0-6/2/0a23d7047aeb2c38ae6e0cdf80f0d26e. (Cited on page 6.)

[16] J. Cairns. Mutation selection and the natural history of cancer. *Nature*, 255: 197–200, 1975. (Cited on pages 66 and 217.)

[17] Peter Calabrese and Darryl Shibata. A simple algebraic cancer equation: calculating how cancers may arise with normal mutation rates. *BMC Cancer*, 10(1), 2010. URL http://www.biomedcentral.com/1471-2407/10/3. (Cited on pages vii, 58, 59, 60, 63, 65, 66, 75, 148, 149, 150, 153, 155, 156, 162, and 217.)

[18] JJ Carter, KG Paulson, GC Wipf, D Miranda, MM Madeleine, LG Johnson, BD Lemos, S Lee, AH Warcola, JG Iyer, P Nghiem, and DA Galloway. Association of merkel cell polyomavirus-specific antibodies with merkel cell carcinoma. *J Natl Cancer Inst*, 101(21):1510–1522, 11 2009. doi: 10.1093/jnci/djp332. URL http://pubget.com/paper/19776382. (Cited on page 35.)

[19] Rachel B. Cervantes, James R. Stringer, Changshun Shao, Jay A. Tischfield, and Peter J. Stambrook. Embryonic stem cells and somatic cells differ in mutation frequency and type. *Proceedings of the National Academy of Sciences*, 99(6):3586–3590, 03 2002. (Cited on pages 66 and 217.)

[20] J Chen, A Wu, H Sun, R Drakas, C Garofalo, S Cascio, E Surmacz, and R Baserga. Functional significance of type 1 insulin-like growth factor-mediated nuclear translocation of the insulin receptor substrate-1 and beta-catenin. *J Biol Chem*, 280(33):29912–29920, 08 2005. URL http://pubget.com/paper/15967802. (Cited on page 42.)

[21] Jun Chen, Kathleen Sprouffske, Qihong Huang, and Carlo C. Maley. Solving the puzzle of metastasis: The evolution of cell migration in neoplasms. *PLoS ONE*, 6(4):e17933, 04 2011. (Cited on pages 8, 88, 94, 239, and 240.)

[22] ET Clayson, LV Brando, and RW Compans. Release of simian virus 40 virions from epithelial cells is polarized and occurs without cell lysis. *J Virol*, 63(5):2278–2288, 05 1989. URL `http://pubget.com/paper/2539518`. (Cited on page 38.)

[23] Gregory M. Cochran, Paul W. Ewald, and Kyle D Cochran. Infectious causation of disease: An evolutionary perspective. *Perspectives in Biology and Medicine*, 43(3):406–448, 2000. (Cited on page 9.)

[24] Tatiana R Coelho, Luis Almeida, and Pedro A Lazo. Jc virus in the pathogenesis of colorectal cancer, an etiological agent or another component in a multistep process? *Virology Journa*, 7(42), 2010. (Cited on page 208.)

[25] M Cotsiki, RL Lock, Y Cheng, GL Williams, J Zhao, D Perera, R Freire, A Entwistle, EA Golemis, TM Roberts, PS Jat, and OV Gjoerup. Simian virus 40 large t antigen targets the spindle assembly checkpoint protein bub1. *Proc Natl Acad Sci U S A*, 101(4):947–952, 01 2004. doi: 10.2307/3148614. URL `http://pubget.com/paper/14732683`. (Cited on page 43.)

[26] JH Dannenberg, A van Rossum, L Schuijff, and H te Riele. Ablation of the retinoblastoma gene family deregulates g(1) control causing immortalization and increased cell turnover under growth-restricting conditions. *Genes Dev*, 14(23):3051–3064, 12 2000. URL `http://pubget.com/paper/11114893`. (Cited on page 40.)

[27] Armine Darbinyan, Khwaja M. Siddiqui, Dorota Slonina, Nune Darbinian, Shohreh Amini, Martyn K. White, and Kamel Khalili. Role of jc virus agnoprotein in dna repair. *The Journal of Virology*, 78(16):8593–8600, 8 2004.

URL http://jvi.asm.org/cgi/content/abstract/78/16/8593. (Cited on page 44.)

[28] R. Justin Davies, Richard Miller, and Nicholas Coleman. Colorectal cancer screening: prospects for molecular stool analysis. *Nature Reviews Cancer*, 5(3):199–209, 03 2005. URL http://search.ebscohost.com/login.aspx?direct=true&db=aph&AN=16298348&site=ehost-live. (Cited on pages 21 and 25.)

[29] Luis Del Valle and Kamel Khalili. Detection of human polyomavirus proteins, t-antigen and agnoprotein, in human tumor tissue arrays. *Journal of Medical Virology*, 82(5):806–811, 2010. doi: 10.1002/jmv.21514. URL http://dx.doi.org/10.1002/jmv.21514. (Cited on pages 33, 38, and 50.)

[30] Luis Del Valle, Sergio Piña-Oviedo, Georgina Perez-Liz, Brian J. Augelli, S. Ausim Azizi, Kamel Khalili, Jennifer Gordon, and Barbara Krynska. Bone marrow-derived mesenchymal stem cells undergo jcv t-antigen mediated transformation and generate tumors with neuroectodermal characteristics. *Cancer Biology and Therapy*, 9(4):286–294, February 2010. (Cited on page 62.)

[31] S. Eash, K. Manley, M. Gasparovic, W. Querbes, and W. Atwood. The human polyomaviruses. *Cellular and Molecular Life Sciences*, 63(7):865–876, 2006-04-07. URL http://dx.doi.org/10.1007/s00018-005-5454-z. (Cited on pages 37, 38, 93, and 218.)

[32] CE Eberhart, RJ Coffey, A Radhika, FM Giardiello, S Ferrenbach, and RN DuBois. Up-regulation of cyclooxygenase 2 gene expression in human colorectal adenomas and adenocarcinomas. *Gastroenterology*, 107(4): 1183–1188, 10 1994. URL http://pubget.com/paper/7926468. (Cited on page 28.)

[33] A Egli, L Infanti, A Dumoulin, A Buser, J Samaridis, C Stebler, R Gosert, and HH Hirsch. Prevalence of polyomavirus bk and jc infection and replication in 400 healthy blood donors. *J Infect Dis*, 199(6):837–846, 03 2009. doi: 10.1086/597126. URL http://pubget.com/paper/19434930. (Cited on pages 35 and 36.)

[34] GF Elphick, W Querbes, JA Jordan, GV Gee, S Eash, K Manley, A Dugan, M Stanifer, A Bhatnagar, WK Kroeze, BL Roth, and WJ Atwood. The human polyomavirus, jcv, uses serotonin receptors to infect cells. *Science*, 306(5700): 1380–1383, 11 2004. doi: 10.1126/science.1103492. URL http://pubget.com/paper/15550673. (Cited on page 37.)

[35] Sahnila Enam, Luis Del Valle, Cesar Lara, Dai-Di Gan, Carlos Ortiz-Hidalgo, Juan P. Palazzo, and Kamel Khalili. Association of human polyomavirus jcv with colon cancer: Evidence for interaction of viral t-antigen and {beta}-catenin. *Cancer Research*, 62(23):7093–7101, 12 2002. URL http://cancerres.aacrjournals.org/cgi/content/abstract/62/23/7093. (Cited on pages 39, 40, 46, 48, and 208.)

[36] JR Eshleman, EZ Lang, GK Bowerfind, R Parsons, B Vogelstein, JK Willson, ML Veigl, WD Sedwick, and SD Markowitz. Increased mutation rate at the hprt locus accompanies microsatellite instability in colon cancer. *Oncogene*, 10(1):33–37, 01 1995. URL http://pubget.com/paper/7824277. (Cited on page 29.)

[37] M Esteller, A Sparks, M Toyota, M Sanchez-Cespedes, G Capella, MA Peinado, S Gonzalez, G Tarafa, D Sidransky, SJ Meltzer, SB Baylin, and JG Herman. Analysis of adenomatous polyposis coli promoter hypermethylation in human cancer. *Cancer Res*, 60(16):4366–4371, 08 2000. URL http://pubget.com/paper/10969779. (Cited on page 23.)

[38] Paul Ewald. *Advances in Parasitology*, volume 68, chapter An Evolutionary Perspective on Parasitism as a Cause of Cancer, pages 21–43. Elsevier, 2009. (Cited on pages 4, 9, and 14.)

[39] Eric R. Fearon and Bert Vogelstein. A genetic model for colorectal tumorigenesis. *Cell*, 61(5):759–767, 6 1990. URL http://www.sciencedirect.com/science/article/pii/0092867490090186I. (Cited on pages 21, 22, 23, 24, 25, 26, 28, and 209.)

[40] Riccardo Fodde, Ron Smits, and Hans Clevers. Apc, signal transduction and genetic instability in colorectal cancer. *Nat Rev Cancer*, 1(1):55–67, 2001. (Cited on page 23.)

[41] KP Foley and RN Eisenman. Two mad tails: what the recent knockouts of mad1 and mxi1 tell us about the myc/max/mad network. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1423(3):0–0, 05 1999. URL http://pubget.com/paper/10382539. (Cited on page 4.)

[42] Judah Folkman. Anti-angiogenesis: new concept for therapy of solid tumors. *Ann Surg*, 175(3):409–416, 03 1972. URL http://pubget.com/paper/5077799. (Cited on page 6.)

[43] SA Frank, Y Iwasa, and Nowak MA. Patterns of cell division and the risk of cancer. *Genetics*, 163(4):1527–32., April 2003. (Cited on pages 66, 73, 122, and 217.)

[44] R Fukuda, B Kelly, and GL Semenza. Vascular endothelial growth factor gene expression in colon cancer cells exposed to prostaglandin e2 is mediated by hypoxia-inducible factor 1. *Cancer Res*, 63(9):2330–2334, 05 2003. URL http://pubget.com/paper/12727858. (Cited on page 28.)

[45] DD Gan and K Khalili. Interaction between jcv large t-antigen and beta-catenin. *Oncogene*, 23(2):483–490, 01 2004. doi: 10.1038/sj.onc.1207018. URL http://pubget.com/paper/14724577. (Cited on pages 39 and 40.)

[46] JE Garber, AM Goldstein, AF Kantor, MG Dreyfus, JF Fraumeni, and FP Li. Follow-up study of twenty-four families with li-fraumeni syndrome. *Cancer Res*, 51(22):6094–6097, 11 1991. URL http://pubget.com/paper/1933872. (Cited on page 26.)

[47] O Gjoerup and Y. Chang. Update on human polyomaviruses and cancer. *Advanced Cacner Research*, 106:1–51, 2010. (Cited on pages 33, 34, 36, 37, 38, 39, 40, 41, 42, 43, 45, 93, 104, and 218.)

[48] A Greenhough, HJ Smartt, AE Moore, HR Roberts, AC Williams, C Paraskeva, and A Kaidi. The cox-2/pge2 pathway: Key roles in the hallmarks of cancer and adaptation to the tumour microenvironment. *Carcinogenesis*, 30(3):377–386, 03 2009. doi: 10.1093/carcin/bgp014. URL http://pubget.com/paper/19136477. (Cited on page 28.)

[49] J Guo, T Kitamura, H Ebihara, C Sugimoto, T Kunitake, J Takehisa, YQ Na, MN Al-Ahdal, A Hallin, K Kawabe, F Taguchi, and Y Yogo. Geographical distribution of the human polyomavirus jc virus types a and b and isolation of a new type from ghana. *J Gen Virol*, 77(5):919–927, 05 1996. doi: 10.1099/0022-1317-77-5-919. URL http://pubget.com/paper/8609488. (Cited on page 35.)

[50] Douglas Hanahan and Robert A. Weinberg. The hallmarks of cancer. *Cell*, 100(1):57–70, 1 2000. URL http://www.sciencedirect.com/science/article/B6WSN-4195FC1-5/2/aef1d48431eadea4567b697b1fee0514. (Cited on pages 2, 3, 4, 7, 8, 9, 10, 29, 59, and 209.)

[51] Curtis C. Harris. p53 tumor suppressor gene: from the basic research laboratory to the clinic–an abridged historical perspective. *Carcinogenesis*, 17(6):1187–1198, 06 1996. doi: 10.1093/carcin/17.6.1187. URL http://pubget.com/paper/pgtmp_3ce2d76f724dc3a7815e59fe79498f19. (Cited on page 5.)

[52] JH Hendry, CS Potten, A Ghafoor, JV Moore, SA Roberts, and PC Williams. The response of murine intestinal crypts to short-range promethium-147 beta irradiation: Deductions concerning clonogenic cell numbers and positions. *Radiat Res*, 118(2):364–374, 05 1989. URL http://pubget.com/paper/2727264. (Cited on page 19.)

[53] M Hollstein, K Rice, MS Greenblatt, T Soussi, R Fuchs, T Sørlie, E Hovig, B Smith-Sørensen, R Montesano, and CC Harris. Database of p53 gene somatic mutations in human tumors and cell lines. *Nucleic Acids Res*, 22 (17):3551–3555, 09 1994. URL http://pubget.com/paper/7937055. (Cited on page 41.)

[54] PS Holman, OV Gjoerup, T Davin, and BS Schaffhausen. Characterization of an immortalizing n-terminal domain of polyomavirus large t antigen. *J Virol*, 68(2):668–673, 02 1994. URL http://pubget.com/paper/8289370. (Cited on page 40.)

[55] Ryouta Hori, Yoshihiro Murai, Kouichi Tsuneyama, Hekmat Abdel-Aziz, Kazuhiro Nomoto, Hiroyuki Takahashi, Chun-mei Cheng, Tomohiko Kuchina, Brian Harman, and Yasuo Takano. Detection of jc virus dna sequences in colorectal cancers in japan. *Virchows Archiv*, 447(4):723–730, 10 2005. URL http://dx.doi.org/10.1007/s00428-005-0014-3. (Cited on pages 46, 48, and 49.)

[56] Adam Humphries and Nicholas A. Wright. Colonic crypt organization and tumorigenesis. *Nat Rev Cancer*, 8(6):415–424, 2008. (Cited on pages 20, 21, 22, 23, 27, 32, and 113.)

[57] Michael J. Imperiale. *The Human Polyomaviruses: An Overview*, pages 53–71. John Wiley & Sons, Inc., 2002. ISBN 9780471221944. doi: 10.1002/0471221945.ch5. URL http://dx.doi.org/10.1002/0471221945.ch5. (Cited on page 38.)

[58] Lars Jakobsson, Claudio A. Franco, Katie Bentley, Russell T. Collins, Bas Ponsioen, Irene M. Aspalter, Ian Rosewell, Marta Busse, Gavin Thurston, Alexander Medvinsky, Stefan Schulte-Merker, and Holger Gerhardt. Endothelial cells dynamically compete for the tip cell position during angiogenic sprouting. *Nat Cell Biol*, 12(10):943–953, 10 2010. URL http://dx.doi.org/10.1038/ncb2103. (Cited on page 7.)

[59] G Jay, S Nomura, CW Anderson, and G Khoury. Identification of the sv40 agnogene product: A dna binding protein. *Nature*, 291(5813):346–349, 05 1981. doi: 10.1038/291346a0. URL http://pubget.com/paper/6262654. (Cited on page 38.)

[60] J Jen, SM Powell, N Papadopoulos, KJ Smith, SR Hamilton, B Vogelstein, and KW Kinzler. Molecular determinants of dysplasia in colorectal lesions. *Cancer Res*, 54(21):5523–5526, 11 1994. URL http://pubget.com/paper/7923189. (Cited on pages 23 and 24.)

[61] A Kato, T Kitamura, C Sugimoto, Y Ogawa, K Nakazato, K Nagashima, WW Hall, K Kawabe, and Y Yogo. Lack of evidence for the transmission of jc polyomavirus between human populations. *Arch Virol*, 142(5):875–882, 01 1997. URL http://pubget.com/paper/9191854. (Cited on page 36.)

[62] Robert S. Kerbel. Tumor angiogenesis: past, present and the near future. *CARCINOGENESIS(LONDON)*, 21(3):505–515, 03 2000. URL http://pubget.com/paper/10688871. (Cited on page 7.)

[63] KM Kim and D Shibata. Methylation reveals a niche: Stem cell succession in human colon crypts. *Oncogene*, 21(35):5441–5449, 08 2002. doi: 10.1038/sj.onc.1205604. URL http://pubget.com/paper/12154406. (Cited on page 20.)

[64] Kenneth W. Kinzler and Bert Vogelstein. Lessons from hereditary colorectal cancer. *Cell*, 87(2):159–170, 10 1996. URL

http://www.sciencedirect.com/science/article/B6WSN-41BD859-3/2/84f3700dd0de458e8dbde8640ca115e1. (Cited on pages 21, 22, 23, 24, 26, 27, 28, 29, and 121.)

[65] T Kitamura, C Sugimoto, A Kato, H Ebihara, M Suzuki, F Taguchi, K Kawabe, and Y Yogo. Persistent jc virus (jcv) infection is demonstrated by continuous shedding of the same jcv strains. *Journal Clinical Microbiology*, 35:1255–1257, 1997. (Cited on page 36.)

[66] Andrew Kitchen, Michael M. Miyamoto, and Connie J. Mulligan. Utility of dna viruses for studying human host history: Case study of jc virus. *Molecular Phylogenetics and Evolution*, 46 (2):673–682, 2 2008. URL http://www.sciencedirect.com/science/article/B6WNH-4PP775S-2/2/51782c18d02b573fa52dc9833d32e0e3. (Cited on page 35.)

[67] Lewis J Kleinsmith. *Principles of Cancer Biology*. Pearson Benjamin Cummings, 2006. (Cited on pages 2, 3, 5, 6, 7, 8, 9, 10, 12, 13, 22, 23, 24, 25, 27, 28, 40, 43, 44, 66, 84, 94, 209, 217, and 240.)

[68] Wendy A. Knowles, Pam Pipkin, Nick Andrews, Andrew Vyse, Philip Minor, David W.G. Brown, and Elizabeth Miller. Population-based study of antibody to the human polyomaviruses bkv and jcv and the simian polyomavirus sv40. *Journal of Medical Virology*, 71(1):115–123, 2003. (Cited on pages 35, 36, 60, 148, 150, 151, 152, 221, and 223.)

[69] Luigi Laghi, Ann E. Randolph, D. P. Chauhan, Giancarlo Marra, Eugene O. Major, James V. Neel, and C. Richard Boland. Jc virus dna is present in the mucosa of the human colon and in colorectal cancers. *Proceedings of the National Academy of Sciences of the United States of America*, 96(13):7484–7489, 06 1999. URL http://www.pnas.org/content/96/13/7484.abstract. (Cited on pages 34, 46, 47, 49, and 208.)

[70] Sergio A. Lamprecht and Martin Lipkin. Migrating colonic crypt epithelial cells: Primary targets for transformation. *Carcinogenesis*, 23(11):1777–1780, 11 2002. doi: 10.1093/carcin/23.11.1777. URL http://pubget.com/paper/pgtmp_dcd1e4c19cc4334959d10e9a8425bff5. (Cited on pages 32, 104, and 113.)

[71] L Q Le and L F Parada. Tumor microenvironment and neurofibromatosis type i: connecting the gaps. *Oncogene*, 26(32):4609–4616, 2007. URL http://dx.doi.org/10.1038/sj.onc.1210261. (Cited on page 27.)

[72] RJ Leary, JC Lin, J Cummins, S Boca, LD Wood, DW Parsons, S Jones, T Sjöblom, BH Park, R Parsons, J Willis, D Dawson, JK Willson, T Nikolskaya, Y Nikolsky, L Kopelovich, N Papadopoulos, LA Pennacchio, TL Wang, SD Markowitz, G Parmigiani, KW Kinzler, B Vogelstein, and VE Velculescu. Integrated analysis of homozygous deletions, focal amplifications, and sequence alterations in breast and colorectal cancers. *Proc Natl Acad Sci U S A*, 105(42):16224–16229, 10 2008. doi: 10.1073/pnas.0808041105. URL http://pubget.com/paper/18852474. (Cited on page 28.)

[73] Sunyoung Lee, Shahla M. Jilani, Ganka V. Nikolova, Darren Carpizo, and M. Luisa Iruela-Arispe. Processing of vegf-a by matrix metalloproteinases regulates bioavailability and vascular patterning in tumors. *The Journal of Cell Biology*, 169(4):681–691, 05 2005. (Cited on page 7.)

[74] A.J. Levine. p53, the cellular gatekeeper for growth and division. *Cell*, 88 (3):9–9, 02 1997. doi: 10.1016/S0092-8674(00)81871-1. URL http://pubget.com/paper/9039259. (Cited on page 5.)

[75] NL Lill, MJ Tevethia, R Eckner, DM Livingston, and N Modjtahedi. p300 family members associate with the carboxyl terminus of simian virus 40 large tumor antigen. *J Virol*, 71(1):129–137, 01 1997. URL http://pubget.com/paper/8985331. (Cited on page 41.)

[76] Paul Yann Lin, Chiung-Yau Fung, Fang-Pei Chang, Wen-Shih Huang, Wen-Cheng Chen, Jeng-Yi Wang, and Deching Chang. Prevalence and genotype identification of human jc virus in colon cancer in taiwan. *Journal of Medical Virology*, 80(10):1828–1834, 2008. (Cited on pages 46 and 49.)

[77] LA Loeb, CF Springgate, and N Battula. Errors in dna replication as a basis of malignant changes. *Cancer Res*, 34(9):2311–2321, 09 1974. URL http://pubget.com/paper/4136142. (Cited on page 29.)

[78] Lawrence A. Loeb. A mutator phenotype in cancer. *Cancer Research*, 61: 3230–3239, 2001. (Cited on page 29.)

[79] Markus Loeffler, Andreas Birke, Douglas Winton, and Christopher Potten. Somatic mutation, monoclonality and stochastic models of stem cell organization in the intestinal crypt. *Journal of Theoretical Biology*, 160(4):471–491, 2 1993. URL http://www.sciencedirect.com/science/article/pii/S0022519383710313. (Cited on pages 19, 89, 216, and 226.)

[80] J. Hernández Losa, V. Fernandez-Soria, C. Parada, R. Sanchez-Prieto, S. Ramón Y Cajal, C. G. Fedele, and A. Tenorio. Jc virus and human colon carcinoma: An intriguing and inconclusive association. *Gastroenterology*, 124(1):268–269, 01 2003. URL http://linkinghub.elsevier.com/retrieve/pii/S0016508503700390?showall=true. (Cited on page 51.)

[81] K Machida, KT Cheng, VM Sung, S Shimodaira, KL Lindsay, AM Levine, MY Lai, and MM Lai. Hepatitis c virus induces a mutator phenotype: Enhanced mutations of immunoglobulin and protooncogenes. *Proc Natl Acad Sci U S A*, 101(12):4262–4267, 03 2004. doi: 10.1073/pnas.0303971101. URL http://pubget.com/paper/14999097. (Cited on pages 69, 87, 218, and 224.)

[82] Melissa S. Maginnis and Walter J. Atwood. Jc virus: An oncogenic virus in animals and humans? *Seminars in Cancer Biology*, 19(4):261–269, 8 2009. URL http://www.sciencedirect.com/science/

article/B6WWY-4VP6688-1/2/de81a328aebcac9156dc39cad9a4680a. (Cited on pages 33, 34, 37, 38, 45, 55, 208, and 222.)

[83] Friedrich Marks, Ursula Klingmüller, and Karin Müller-Decker. *Cellular signal processing : an introduction to the molecular mechanisms of signal transduction*. Garland Science, 1 edition, 2009. (Cited on pages 22, 25, 26, 27, 41, 42, and 44.)

[84] Lauren M. F. Merlo, John W. Pepper, Brian J. Reid, and Carlo C. Maley. Cancer as an evolutionary and ecological process. *Nat Rev Cancer*, 6(12): 924–935, 12 2006. URL http://dx.doi.org/10.1038/nrc2013. (Cited on page 17.)

[85] Franziska Michor, Yoh Iwasa, Christoph Lengauer, and Martin A. Nowak. Dynamics of colorectal cancer. *Seminars in Cancer Biology*, 15(6):484–493, 12 2005. URL http://www.sciencedirect.com/science/article/B6WWY-4GSBGTG-2/2/ca68f1a168e092dfcc877ed18adffb74. (Cited on pages 31, 208, and 209.)

[86] MC Monaco and EO Shin, J amd Major. Jc virus infection in cells from lymphoid tissue. *Developments In Biological Standardization*, 94:115–122, 1998. (Cited on page 36.)

[87] MC Monaco, PN Jensen, J Hou, LC Durham, and EQ Major. Detection of jc virus dna in human tonsil tissue: evidence for site of initial viral infection. *Journal of Virology*, 72:9918–9923, 1998. (Cited on page 36.)

[88] CS Moreno, S Ramachandran, DG Ashby, N Laycock, CA Plattner, W Chen, WC Hahn, and DC Pallas. Signaling and transcriptional changes critical for transformation of human cells by simian virus 40 small tumor antigen or protein phosphatase 2a b56gamma knockdown. *Cancer Res*, 64(19):6978–6988, 10 2004. doi: 10.1158/0008-5472.CAN-04-1150. URL http://pubget.com/paper/15466190. (Cited on page 43.)

[89] Yoshihiro Murai, Hua-Chuan Zheng, Hekmat Osman Abdel Aziz, Hong Mei, Tomohiko Kutsuna, Yuko Nakanishi, Koichi Tsuneyama, and Yasuo Takano. High jc virus load in gastric cancer and adjacent non-cancerous mucosa. *Cancer Science*, 98(1):25–31, 2007. (Cited on page 208.)

[90] Kenneth Murphy, Paul Travers, and Mark Walport. *Immunobiology*. Garland Science, seventh edition, 2007. (Cited on page 36.)

[91] Dyson N, Bernards R, Friend SH, Gooding LR, Hassell JA, Major EO, Pipas JM, Vandyke T, and Harlow E. Large t antigens of many polyomaviruses are able to form complexes with the retinoblastoma protein. *j virol*, 64(3): 1353–1356, 03 1990. URL `http://pubget.com/paper/2154613`. (Cited on page 40.)

[92] Michael W. Nachman and Susan L. Crowell. Estimate of the mutation rate per nucleotide in humans. *Genetics*, 156(1):297–304, 09 2000. URL `http://www.genetics.org/content/156/1/297.abstract`. (Cited on pages 90, 217, and 228.)

[93] Simona Negrini, Vassilis G. Gorgoulis, and Thanos D. Halazonetis. Genomic instability - an evolving hallmark of cancer. *Nat Rev Mol Cell Biol*, 11 (3):220–228, 2010. (Cited on pages 4, 10, and 11.)

[94] Polly A. Newcomb, Angela C. Bush, Gerald L. Stoner, Johanna W. Lampe, John D. Potter, and Jeannette Bigler. No evidence of an association of jc virus and colon neoplasia. *Cancer Epidemiology Biomarkers & Prevention*, 13 (4):662–666, 04 2004. URL `http://cebp.aacrjournals.org/content/13/4/662.abstract`. (Cited on page 51.)

[95] Yaron MD Niv, Alex Vilkin, and Zohar Levi. Patients with sporadic colorectal cancer or advanced adenomatous polyp have elevated anti-jc virus antibody titer in comparison with healthy controls: A cross-sectional study.

Journal of Clinical Gastroenterology, 44(7):489–494, August 2010. (Cited on page 50.)

[96] K Nosho, K Shima, S Kure, N Irahara, Y Baba, L Chen, GJ Kirkner, CS Fuchs, and S Ogino. Jc virus t-antigen in colorectal cancer is associated with p53 expression and chromosomal instability, independent of cpg island methylator phenotype. *Neoplasia (New York, N.Y.)*, 11(1), 01 2009. URL http://ukpmc.ac.uk/abstract/MED/19107235. (Cited on pages 41, 46, 49, 50, and 52.)

[97] Martin A. Nowak, Natalia L. Komarova, Anirvan Sengupta, Prasad V. Jallepalli, Ie-Ming Shih, Bert Vogelstein, and Christoph Lengauer. The role of chromosomal instability in tumor initiation. *Proceedings of the National Academy of Sciences of the United States of America*, 99(25):16226–16231, 12 2002. URL http://www.pnas.org/content/99/25/16226.abstract. (Cited on page 31.)

[98] BL Padgett, DL Walker, GM ZuRhein, and JN Varakis. Differential neurooncogenicity of strains of jc virus, a human polyoma virus, in newborn syrian hamsters. *Cancer Res*, 37(3):718–720, 03 1977. URL http://pubget.com/paper/189911. (Cited on page 45.)

[99] JS Pagano, M Blaser, MA Buendia, B Damania, K Khalili, N Raab-Traub, and B Roizman. Infectious agents and cancer: Criteria for a causal relation. *Semin Cancer Biol*, 14(6):19–19, 12 2004. doi: 10.1016/j.semcancer.2004.06.009. URL http://pubget.com/paper/15489139. (Cited on page 55.)

[100] Y Park do, H Sakamoto, SD Kirley, S Ogino, T Kawasaki, E Kwon, M Mino-Kenudson, GY Lauwers, DC Chung, BR Rueda, and LR Zukerberg. The cables gene on chromosome 18q is silenced by promoter hypermethylation and allelic loss in human colorectal cancer. *Am J Pathol*, 171(5):11–11, 11 2007. doi: 10.2353/ajpath.2007.070331. URL http://pubget.com/paper/17982127. (Cited on page 26.)

[101] DV Pastrana, YL Tolstov, JC Becker, PS Moore, Y Chang, and CB Buck. Quantitation of human seroresponsiveness to merkel cell polyomavirus. *PLoS Pathog*, 5(9):e1000578, 09 2009. doi: 10.1371/journal.ppat.1000578. URL http://pubget.com/paper/19750217. (Cited on page 34.)

[102] Maria S. Pino and Daniel C. Chung. The chromosomal instability pathway in colon cancer. *Gastroenterology*, 138(6):2059–2072, 5 2010. URL http://www.sciencedirect.com/science/article/B6WFX-4YXNHS1-9/2/72f0fa707ac19868522e67af3da1b730. (Cited on pages 11, 12, 25, 26, 29, 30, 31, 32, 91, 106, and 107.)

[103] Christopher S. Potten, Catherine Booth, and D. Mark Pritchard. The intestinal epithelial stem cell: the mucosal governor. *International Journal of Experimental Pathology*, 78(4):219–243, 1997. (Cited on pages 19, 20, 88, 89, and 240.)

[104] C.S. Potten. The significance of spontaneous and induced apoptosis in the gastrointestinal tract of mice. *Cancer Metastasis Rev*, 11(2):179–195, 09 1992. URL http://pubget.com/paper/1394796. (Cited on pages 19 and 88.)

[105] CS Potten and M Loeffler. Stem cells: Attributes, cycles, spirals, pitfalls and uncertainties. lessons for and from the crypt. *Development*, 110(4):1001–1020, 12 1990. URL http://pubget.com/paper/2100251. (Cited on pages 19 and 89.)

[106] M Prisco, F Santini, R Baffa, M Liu, R Drakas, A Wu, and R Baserga. Nuclear translocation of insulin receptor substrate-1 by the simian virus 40 t antigen and the activated type 1 insulin-like growth factor receptor. *J Biol Chem*, 277(35):32078–32085, 08 2002. doi: 10.1074/jbc.M204658200. URL http://pubget.com/paper/12063262. (Cited on page 42.)

[107] R Development Core Team. R: A language and environment for statistical computing, 2010. URL http://www.R-project.org. (Cited on page 76.)

[108] Steven F Railsback and Volker Grimm. *Agent-Based and Individual-Based Modeling: A Pracitcal Introduction*. Princeton, 2011. (Cited on page 230.)

[109] Krzysztof Reiss and Kamel Khalili. Viruses and cancer: Lessons from the human polyomavirus, jcv. *Oncogene*, 22(42):6517–6523, 2003. (Cited on pages 45, 105, and 118.)

[110] Krzysztof Reiss, Luis Del Valle, Adam Lassak, and Joanna Trojanek. Nuclear irs-1 and cancer. *Journal of Cellular Physiology*, 227(8):2992–3000, 2012. doi: 10.1002/jcp.24019. URL http://dx.doi.org/10.1002/jcp.24019. (Cited on page 42.)

[111] L Ricciardiello, M Baglioni, C Giovannini, M Pariali, G Cenacchi, A Ripalti, MP Landini, H Sawa, K Nagashima, RJ Frisque, A Goel, CR Boland, M Tognon, E Roda, and F Bazzoli. Induction of chromosomal instability in colonic cells by the human polyomavirus jc virus. *Cancer Research*, 63(21): 7256–7262, 11 2003. URL http://pubget.com/paper/14612521. (Cited on pages 38, 39, 45, 51, 52, and 224.)

[112] Luigi Ricciardiello, Luigi Laghi, Pradeep Ramamirtham, Christina L. Chang, Dong K. Chang, Ann E. Randolph, and C. Richard Boland. Jc virus dna sequences are frequently present in the human upper and lower gastrointestinal tract. *Gastroenterology*, 119(5):1228–1235, 11 2000. URL http://www.sciencedirect.com/science/article/B6WFX-45V7W23-4N/2/7003ec2bbb05d6936287ef43d373d719. (Cited on pages 47 and 87.)

[113] Luigi Ricciardiello, Dong K. Chang, Luigi Laghi, Ajay Goel, Christina L. Chang, and C. Richard Boland. Mad-1 is the exclusive jc virus strain present in the human colon, and its transcriptional control region has a deleted 98-base-pair sequence in colon cancer tissues. *The Journal of Virology*, 75(4):1996–2001, 2 2001. URL http://jvi.asm.org/cgi/content/abstract/75/4/1996. (Cited on pages 47, 87, and 223.)

[114] M. Rizvi, Asgar Ali, Syed Mehdi, Sundeep Saluja, and Pramod Mishra. Association of epigenetic alteration in pten gene with colorectal cancer progression among indian population. *International Journal of Colorectal Disease*, pages 1–2. doi: 10.1007/s00384-012-1482-y. URL http://dx.doi.org/10.1007/s00384-012-1482-y. (Cited on page 27.)

[115] Dana E. Rollison. Jc virus infection a cause of colorectal cancer? *Journal of Clinical Gastroenterology*, 44(7), 2010. URL http://journals.lww.com/jcge/Fulltext/2010/08000/JC_Virus_Infection_A_Cause_of_Colorectal_Cancer_.7.aspx. (Cited on pages 46 and 51.)

[116] M Safak, R Barrucco, A Darbinyan, Y Okada, K Nagashima, and K Khalili. Interaction of jc virus agno protein with t antigen modulates transcription and replication of the viral genome in glial cells. *J Virol*, 75(3):1476–1486, 02 2001. doi: 10.1128/JVI.75.3.1476-1486.2001. URL http://pubget.com/paper/11152520. (Cited on page 38.)

[117] J Sage, GJ Mulligan, LD Attardi, A Miller, S Chen, B Williams, E Theodorou, and T Jacks. Targeted disruption of the three rb-related genes leads to loss of g(1) control and immortalization. *Genes Dev*, 14(23):3037–3050, 12 2000. URL http://pubget.com/paper/11114892. (Cited on page 40.)

[118] Pipsa Saharinen, Lauri Eklund, Kristina Pulkki, Petri Bono, and Kari Alitalo. Vegf and angiopoietin signaling in tumor angiogenesis and metastasis. *Trends in Molecular Medicine*, In Press, Corrected Proof:–, 2011. URL http://www.sciencedirect.com/science/article/B6W7J-52KW30J-1/2/f0e669c7ab650328a4af431735ff8c07. (Cited on page 7.)

[119] WS Samowitz, MD Powers, LN Spirio, F Nollet, F van Roy, and ML Slattery. Beta-catenin mutations are more frequent in small colorectal adenomas than in larger adenomas and invasive carcinomas. *Cancer Res*, 59(7):

1442–1444, 04 1999. URL http://pubget.com/paper/10197610. (Cited on page 24.)

[120] Yardena Samuels, Zhenghe Wang, Alberto Bardelli, Natalie Silliman, Janine Ptak, Steve Szabo, Hai Yan, Adi Gazdar, Steven M. Powell, Gregory J. Riggins, James K. V. Willson, Sanford Markowitz, Kenneth W. Kinzler, Bert Vogelstein, and Victor E. Velculescu. High frequency of mutations of the pik3ca gene in human cancers. *Science*, 304(5670):554–554, 04 2004. URL http://www.sciencemag.org/content/304/5670/554.short. (Cited on pages 27 and 28.)

[121] Laura A. Shackelton, Andrew Rambaut andOliver G. Pybus, and Edward C. Holmes. Jc virus evolution and its association with human populations. *Journal of Virology*, 80(20):9928–9933, October 2006. (Cited on page 93.)

[122] AK Sharma and G Kumar. A 53 kda protein binds to the negative regulatory region of jc virus early promoter. *FEBS Lett*, 281(1-2):272–274, 04 1991. URL http://pubget.com/paper/1849841. (Cited on page 40.)

[123] JW Shay and S Bacchetti. A survey of telomerase activity in human cancer. *Eur J Cancer*, 33(5):787–791, 04 1997. doi: 10.1016/S0959-8049(97)00062-2. URL http://pubget.com/paper/9282118. (Cited on pages 6 and 101.)

[124] D Shibata, MA Peinado, Y Ionov, S Malkhosyan, and M Perucho. Genomic instability in repeated sequences is an early somatic event in colorectal tumorigenesis that persists after transformation. *Nat Genet*, 6(3):273–281, 03 1994. doi: 10.1038/ng0394-273. URL http://pubget.com/paper/8012390. (Cited on page 29.)

[125] IM Shih, W Zhou, SN Goodman, C Lengauer, KW Kinzler, and B Vogelstein. Evidence that genetic instability occurs at an early stage of colorectal tumorigenesis. *Cancer Res*, 61(3):818–822, 02 2001. URL http://pubget.com/paper/11221861. (Cited on pages 30 and 31.)

[126] OM Sieber, K Heinimann, and IP Tomlinson. Genomic instability–the engine of tumorigenesis? *Nat Rev Cancer*, 3(9):701–708, 09 2003. doi: 10.1038/nrc1170. URL http://pubget.com/paper/12951589. (Cited on page 31.)

[127] Benjamin D. Simons and Hans Clevers. Stem cell self-renewal in intestinal crypt. *Experimental Cell Research*, 317(19):2719–2724, 11 2011. URL http://www.sciencedirect.com/science/article/pii/S0014482711002862. (Cited on pages 18, 19, 20, and 88.)

[128] G Singhal, RK Kadeppagari, N Sankar, and B Thimmapaya. Simian virus 40 large t overcomes p300 repression of c-myc. *Virology*, 377(2):6–6, 08 2008. doi: 10.1016/j.virol.2008.04.042. URL http://pubget.com/paper/18570961. (Cited on page 41.)

[129] KJ Smith, KA Johnson, TM Bryan, DE Hill, S Markowitz, JK Willson, C Paraskeva, GM Petersen, SR Hamilton, and B Vogelstein. The apc gene product in normal and tumor cells. *Proc Natl Acad Sci U S A*, 90(7):2846–2850, 04 1993. doi: 10.2307/2361635. URL http://pubget.com/paper/8385345. (Cited on page 23.)

[130] AB Sparks, PJ Morin, B Vogelstein, and KW Kinzler. Mutational analysis of the apc/b-catenin/tcf pathway in colorectal cancer. *Cancer Research*, 58: 1130–4, 1998. (Cited on page 23.)

[131] M.B. Sporn. The war on cancer. *Lancet*, (347):1377–1381, 1996. (Cited on page 7.)

[132] Kathleen Sprouffske, John W Pepper, and Carlo C Maley. Accurate reconstruction of the temporal order of mutations in neoplastic progression. *Cancer Prevention Research*, 2011. (Cited on page 209.)

[133] Chie Sugimoto, Tadaichi Kitamura, Jing Guo, Mohammed N. Al-Ahdal, Sergei N. Shchelkunov, Berta Otova, Paul Ondrejkai, Jean-Yves Chol-

let, Sayda El-Safi, Mohamed Ettayebi, Gerard Gresenguet, Tanil Kocagoz, Sanong Chaiyarasameeii, Kyaw Zin Thant, Soe Thein, Kyaw Moe, Nobuyoshi Kobayashi, Fumiaki Taguchi, and Yoshiaki Yogo. Typing of urinary jc virus dna offers a novel means of tracing human migrations. *Proceedings of the National Academy Sciences*, 94:9191–9196, 1997. doi: 10.1073/pnas.94.17.9191. URL http://ci.nii.ac.jp/naid/80009851357/en/. (Cited on page 35.)

[134] T Takayama, M Ohi, T Hayashi, K Miyanishi, A Nobuoka, T Nakajima, T Satoh, R Takimoto, J Kato, S Sakamaki, and Y Niitsu. Analysis of k-ras, apc, and beta-catenin in aberrant crypt foci in sporadic adenoma, cancer, and familial adenomatous polyposis. *Gastroenterology*, 121(3):599–611, 09 2001. doi: 10.1053/gast.2001.27203. URL http://pubget.com/paper/11522744. (Cited on page 23.)

[135] Ian F. Tannock. Population kinetics of carcinoma cells, capillary endothelial cells, and fibroblasts in a transplanted mouse mammary tumor. *Cancer Research*, 30(10):2470–2476, 10 1970. URL http://cancerres.aacrjournals.org/content/30/10/2470. (Cited on page 6.)

[136] IF Tannock. The relation between cell proliferation and the vascular system in a transplanted mouse mammary tumour. *British Journal of Cancer*, 22(2):258–73, June 1968. (Cited on page 6.)

[137] J Thacker. The role of homologous recombination processes in the repair of severe forms of dna damage in mammalian cells. *Biochimie*, 81(1-2):77–85, 01 1999. URL http://pubget.com/paper/10214913. (Cited on page 42.)

[138] George Theodoropoulos, Dimitris Panoussopoulos, Ioannis Papaconstantinou, Maria Gazouli, Marina Perdiki, John Bramis, and Andreas Ch. Lazaris. Assessment of jc polyoma virus in colon neoplasms. *Diseases of the Colon & Rectum*, 48(1):86–91, 2005-01-01. doi: 10.1007/s10350-004-0737-2. URL http://dx.doi.org/10.1007/s10350-004-0737-2. (Cited on page 49.)

[139] YL Tolstov, DV Pastrana, H Feng, JC Becker, FJ Jenkins, S Moschos, Y Chang, CB Buck, and PS Moore. Human merkel cell polyomavirus infection ii. mcv is a common human infection that can be detected by conformational capsid epitope immunoassays. *Int J Cancer*, 125(6):1250–1256, 09 2009. doi: 10.1002/ijc.24509. URL http://pubget.com/paper/19499548. (Cited on page 34.)

[140] J Trojanek, S Croul, T Ho, JY Wang, A Darbinyan, M Nowicki, L Del Valle, T Skorski, K Khalili, and K Reiss. T-antigen of the human polyomavirus jc attenuates faithful dna repair by forcing nuclear interaction between irs-1 and rad51. *J Cell Physiol*, 206(1):35–46, 01 2006. doi: 10.1002/jcp.20425. URL http://pubget.com/paper/15965906. (Cited on page 42.)

[141] PW Trowbridge and RJ Frisque. Identification of three new jc virus proteins generated by alternative splicing of the early viral mrna. *J Neurovirol*, 1(2): 195–206, 06 1995. doi: 10.3109/13550289509113966. URL http://pubget.com/paper/9222358. (Cited on page 37.)

[142] van de Wetering M., Sancho E., Verweij C., de Lau W., Oving I., Hurlstone A., van der Horn K., Batlle E., Coudreuse D., Haramis A. P., Tjon-Pon-Fong M., Moerer P., van den Born M., Soete G., Pals S., Eilers M., Medema R., and Clevers H. The beta-catenin/tcf-4 complex imposes a crypt progenitor phenotype on colorectal cancer cells. *Cell*, 111(2):241–250, 2002. URL http://www.ingentaconnect.com/content/els/00928674/2002/00000111/00000002/art01014. (Cited on pages 22 and 23.)

[143] Ranga N. Venkatesan, Jason H. Bielas, and Lawrence A. Loeb. Generation of mutator mutants during carcinogenesis. *DNA Repair*, 5 (3):294–302, 3 2006. URL http://www.sciencedirect.com/science/article/B6X17-4HTBM2C-1/2/2d35172dae9e297f43e86f9e4ece8d47. (Cited on pages 10 and 217.)

[144] Raphael P. Viscidi, Dana E. Rollison, Vernon K. Sondak, Barbara Silver, Jane L. Messina, Anna R. Giuliano, William Fulp, Abidemi Ajidahun, and Daniela Rivanera. Age-specific seroprevalence of merkel cell polyomavirus, bk virus, and jc virus. *Clinical and Vaccine Immunology*, 18(10):1737–1743, 10 2011. (Cited on pages 35, 221, and 223.)

[145] Z Wang, JM Cummins, D Shen, DP Cahill, PV Jallepalli, TL Wang, DW Parsons, G Traverso, M Awad, N Silliman, J Ptak, S Szabo, JK Willson, SD Markowitz, ML Goldberg, R Karess, KW Kinzler, B Vogelstein, VE Velculescu, and C Lengauer. Three classes of genes mutated in colorectal cancers with chromosomal instability. *Cancer Res*, 64(9):2998–3001, 05 2004. URL `http://pubget.com/paper/15126332`. (Cited on page 30.)

[146] Robert A. Weinberg. The retinoblastoma protein and cell cycle the retinoblastoma protein and cell cycle control. *Cell*, 81:323–330, 1995. (Cited on page 2.)

[147] U. Wilensky. *NetLogo*. Center for Connected Learning and Computer-Based Modeling,Northwestern University, Evanston, IL., 1999. URL `http://ccl.northwestern.edu/netlogo/`. (Cited on page 85.)

[148] CS Williams, M Tsujii, J Reese, SK Dey, and RN DuBois. Host cyclooxygenase-2 modulates carcinoma growth. *J Clin Invest*, 105(11):1589–1594, 06 2000. doi: 10.1172/JCI9621. URL `http://pubget.com/paper/10841517`. (Cited on page 28.)

[149] LD Wood, DW Parsons, S Jones, J Lin, T Sjöblom, RJ Leary, D Shen, SM Boca, T Barber, J Ptak, N Silliman, S Szabo, Z Dezso, V Ustyanksky, T Nikolskaya, Y Nikolsky, R Karchin, PA Wilson, JS Kaminker, Z Zhang, R Croshaw, J Willis, D Dawson, M Shipitsin, JK Willson, S Sukumar, K Polyak, BH Park, CL Pethiyagoda, PV Pant, DG Ballinger, AB Sparks, J Hartigan, DR Smith, E Suh, N Papadopoulos, P Buckhaults,

SD Markowitz, G Parmigiani, KW Kinzler, VE Velculescu, and B Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, 11 2007. doi: 10.1126/science.1145720. URL http://pubget.com/paper/17932254. (Cited on page 28.)

[150] X Wu, D Avni, T Chiba, F Yan, Q Zhao, Y Lin, H Heng, and D Livingston. Sv40 t antigen interacts with nbs1 to disrupt dna replication control. *Genes Dev*, 18(11):1305–1316, 06 2004. doi: 10.1101/gad.1182804. URL http://pubget.com/paper/15175262. (Cited on page 43.)

[151] E Yeh, M Cunningham, H Arnold, D Chasse, T Monteith, G Ivaldi, WC Hahn, PT Stukenberg, S Shenolikar, T Uchida, CM Counter, JR Nevins, AR Means, and R Sears. A signalling pathway controlling c-myc degradation that impacts oncogenic transformation of human cells. *Nat Cell Biol*, 6(4):308–318, 04 2004. doi: 10.1038/ncb1110. URL http://pubget.com/paper/15048125. (Cited on page 43.)

[152] J Zhao, OV Gjoerup, RR Subramanian, Y Cheng, W Chen, TM Roberts, and WC Hahn. Human mammary epithelial cell transformation through the activation of phosphatidylinositol 3-kinase. *Cancer Cell*, 3(5):13–13, 05 2003. doi: 10.1016/S1535-6108(03)00088-6. URL http://pubget.com/paper/12781366. (Cited on page 43.)

[153] Harald zur Hausen. *Infections Causing Human Cancer*. Wiley-Blackwell, 2011. (Cited on pages 9, 14, 15, 16, 86, 209, and 222.)

APPENDIX A R CODE FOR PROBABILITY MODELS

A.1 ESTIMATING PREVALENCE OF COLON CANCER: CONSTANT MUTATION, CONSTANT STEM CELL POPULATION

The following code creates two different models of colon cancer development, but in each case the mutation rate and stem cell numbers remain constant. The first model argues that mutation drives the development of colon cancer, while the second hypothesizes that infection initiates the oncogenic process. Each model is also executed using the parameters from Calabrese and Shibata [17] as well as more recent parameter estimates[1].

A.1.1 *Creating the Initial Dataframe*

A.1.1.1 *Import Observational Data*

The first step of the program is to import actual data of colorectal cancer (CRC), using the SEER database SEE [2]. This was accomplished by downloading the data and creating new vectors (CRC_2003_2007 and CRC_2000_2006) containing the incidence data. The data used in Calabrese and Shibata [17] was found in the appendix of their paper. The JCV prevalence data was taken from Table 1 of Knowles et al. [68]. Note that from ages 0-19, prevalence was recorded every 5 years, but from ages 20-69 prevalence was recorded only every 10 years. As the CRC data are recorded every 5 years, the first five years in each age group

1 The parameters values used in this example are different than the ones used to generate the results in Chapter 5. Here, genomic instability is 1.5, while in the actual models it is 2. Gene length here is 1500bp, while in the models it is 1000bp.

over 20 years old was assigned an NA value. For example, prevalence of JCV at 30-34 is considered "NA", while at the ages of 35-39 JCV prevalence is 0.39 (i.e. prevalence is not divided across the ten years).

Plots found on the SEER website and Calabrese and Shibata [17] plot incidence data for the median age of an age group. For example, the incidence of CRC in the age group of 30-35 is plotted at age 32.5. As such, the vector AverageAgeCancer was created to contain these median ages. The code used to accomplish these tasks can be found below:

```
Seer0307<-read.csv("/Users/chandlergatenbee/Documents/UofL/CancerProbPaper/Observed
    Data/CRC03-07.csv") #Colon cancer data from SEER 2003-2007

CRC_2003_2007<-Seer0307[,2] #Slice CRC incidence values from table

Seer0006<-read.csv("/Users/chandlergatenbee/Documents/UofL/CancerProbPaper/Observed
    Data/CRC00-06.csv") #Colon cancer data from SEER 2000-2006

CRC_2000_2006<-Seer0006[,2] #Slice CRC incidence values from table

CalabreseData<-c(0, 0, 0, 0.07, 0.18, 0.47, 1.46, 2.82, 5.59, 11.14, 21.59, 42.72,
    77.94, 125.98, 184.04, 250.96, 319.14, 387.22, NA) #Actual data used in
    Calabrese 2010; from SEER 1992-1999

CalabreseAgeRange<-seq(0,90,5) #Age Values used by Calabrese to calculate
    probability of cancer

SeerAges<-Seer0307[,1] #Age categories used by SEER

JCVPrevalence<-c(NA,0.11,0.14, 0.24, 0.22, NA, 0.34, NA, 0.39, NA, 0.34, NA, 0.45,
    NA, 0.5, NA, NA, NA, NA) #JCV Prevalence by age From Knowles 2003

AverageAgeCancer<-rep(0,length(CalabreseAgeRange))
        for(i in 2:length(CalabreseAgeRange)){
                AverageAgeCancer[i]<-mean(CalabreseAgeRange[(i-1):i])}
```

A.1.1.2 *Create Empty Dataframes that will be filled in with results*

Two sets of vectors were created to capture the results from the analysis, the first set using old parameter values, the second set using new paramter values: "ProbInfection", "Incidence_Infection", "ProbMutation", and "Incidence_Mutation" were used to collect the results created using the parameters found in Calabrese and Shibata [17]; "ProbInfectionNew", "Incidence_InfectionNew", "ProbMutationNew", and "Incidence_MutationNew" were used to collect the results created using new parameter values. The code used to create these vectors can be found below:

```
        #Empty vectors to be used with parameters from Calabrese
ProbInfection<-rep(0,length(CalabreseAgeRange))
Incidence_Infection<-rep(0,length(CalabreseAgeRange))
ProbMutation<-rep(0,length(CalabreseAgeRange))
Incidence_Mutation<-rep(0,length(CalabreseAgeRange))


        #Empty vectors to be used with New Parameters
ProbInfectionNew<-rep(0,length(CalabreseAgeRange))
Incidence_InfectionNew<-rep(0,length(CalabreseAgeRange))
ProbMutationNew<-rep(0,length(CalabreseAgeRange))
Incidence_MutationNew<-rep(0,length(CalabreseAgeRange))
```

A.1.1.3 *Estimating Incidence of JCV Infection*

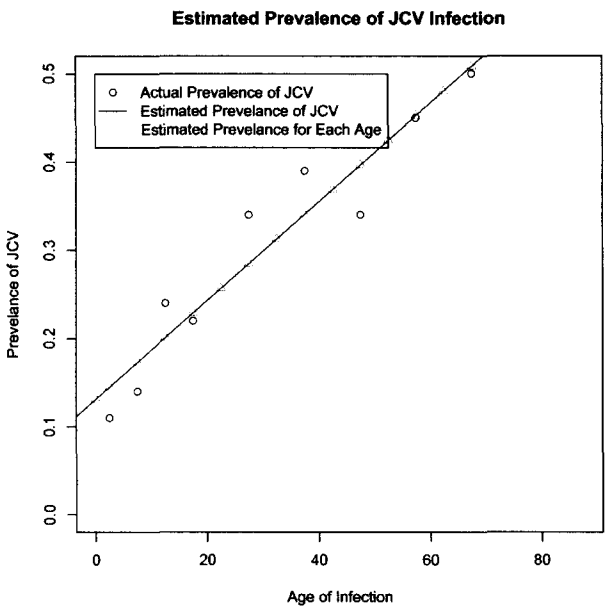
JCV prevalence data from Knowles et al. [68] only contains estimates for ages 1-4, 5-9, 10-14, 15-19, 20-29, 30-39, 40-49, 50-59, 60-69, while the CRC prevalence data is available for ages 0-85+, in increments of five years. Thus, linear interpolation of the regression line was performed on the Knowles et al. [68] data set, so as to fill in all missing values as well as to estimate JCV prevalence in individuals

over 70 years old. The estimated prevalence values were then stored in the vector "JCVEstimatedPrev". The code used was as follows:

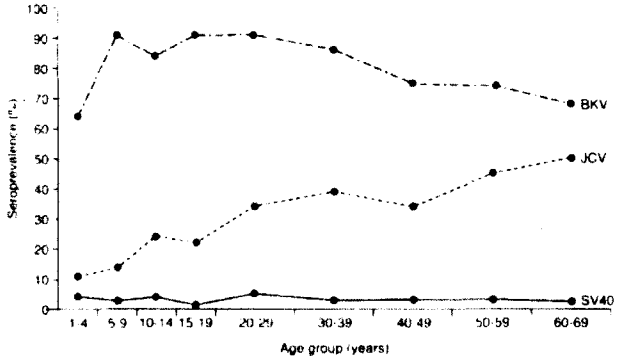
```
RegPrev<-lm(JCVPrevalence~AverageAgeCancer)
JCVEstimatedPrev<-coef(RegPrev)[2]*AverageAgeCancer+coef(RegPrev)[1]
```

These estimated values were then plotted against the actual values from Knowles et al. [68] to ensure that the prevalence pattern remained consistent. The following code generated the plot which can be compared to the plot from Knowles et al. [68] (Figure A.1):

```
xrangelm<-AverageAgeCancer
yrangelm<-seq(0,0.5,0.5/(length(xrangelm)-1))
plot(xrangelm,yrangelm,xlab="Age of Infection",ylab="Prevelance of JCV", main="
    Estimated Prevalence of JCV Infection",type="n")
points(AverageAgeCancer,JCVPrevalence,pch=1,col=1) #Observed Prevalence
points(AverageAgeCancer,JCVEstimatedPrev,pch=2,col=2) #Estimated Prevalence
abline(RegPrev)
legend(0,0.5,c("Actual Prevalence of JCV","Estimated Prevelance of JCV","Estimated
    Prevelance for Each Age"),lty=c(NA,1,NA),pch=c(1,NA,2),col=c(1,1,2))
```



(a) Estimated and Observed Prevalence of JCV



(b) Observed Prevalence of JCV from Knowles et al. [68]

Figure A.1

Finally, the model developed herein requires JCV incidence, however Knowles et al. [68] only provides JCV prevalence. Thus, JCV incidence was estimated by finding the average change in JCV prevalence, the results being stored in the vector "JCVIncidence". These values were then averaged, yielding an average incidence of 0.02716503 every five years (Note: the first incidence value was not included as it is NA):

```
JCVIncidence<-rep(NA,length(AverageAgeCancer))
for(i in 2:length(JCVEstimatedPrev)){
```

```
                JCVIncidence[i]<-JCVEstimatedPrev[i]-JCVEstimatedPrev[(i-1)] }


AvgJCVIncidence<-mean(JCVIncidence[-1])
```

A.1.1.4 *Creating the Dataframe*

All vectors were compiled into a dataframe labeled "CRC" using the code below. The resulting dataframe can be found in Figure A.2.

```
CRC<-data.frame(SeerAges, CalabreseAgeRange, AverageAgeCancer, JCVPrevalence,
    JCVEstimatedPrev, JCVIncidence, CRC_2000_2006, CRC_2003_2007, CalabreseData,
    ProbMutation, ProbMutationNew, Incidence_Mutation, Incidence_MutationNew,
    ProbInfection, ProbInfectionNew, Incidence_Infection, Incidence_InfectionNew)
```

> CRC

	SeerAges	CalabreseAgeRange	AverageAgeCancer	JCVPrevalence	JCVEstimatedPrev	JCVIncidence	CRC_2000_2006	CRC_2003_2007	CalabreseData	ProbMutation	ProbMutationNew	Incidence_Mutation	Incidence_MutationNew	ProbInfection	ProbInfectionNew	Incidence_Infection	Incidence_InfectionNew
1	<1	0	0.0	NA	0.1310294	NA	0.00	0.0	0.00	0	0	0	0	0	0	0	0
2	1-4	5	2.5	0.11	0.1450000	0.01397059	0.00	0.0	0.00	0	0	0	0	0	0	0	0
3	5-9	10	7.5	0.14	0.1729412	0.02794118	0.00	0.0	0.00	0	0	0	0	0	0	0	0
4	10-14	15	12.5	0.24	0.2088824	0.02794118	0.06	0.0	0.07	0	0	0	0	0	0	0	0
5	15-19	20	17.5	0.22	0.2288235	0.02794118	0.26	0.2	0.18	0	0	0	0	0	0	0	0
6	20-24	25	22.5	NA	0.2567647	0.02794118	0.81	0.6	0.47	0	0	0	0	0	0	0	0
7	25-29	30	27.5	0.34	0.2847059	0.02794118	1.86	1.2	1.46	0	0	0	0	0	0	0	0
8	30-34	35	32.5	NA	0.3126471	0.02794118	4.06	2.7	2.82	0	0	0	0	0	0	0	0
9	35-39	40	37.5	0.39	0.3405882	0.02794118	7.74	4.6	5.59	0	0	0	0	0	0	0	0
10	40-44	45	42.5	NA	0.3685294	0.02794118	14.67	9.2	11.14	0	0	0	0	0	0	0	0
11	45-49	50	47.5	0.34	0.3964706	0.02794118	27.65	16.9	21.59	0	0	0	0	0	0	0	0
12	50-54	55	52.5	NA	0.4244118	0.02794118	53.12	33.5	42.72	0	0	0	0	0	0	0	0
13	55-59	60	57.5	0.45	0.4523529	0.02794118	77.97	48.7	77.94	0	0	0	0	0	0	0	0
14	60-64	65	62.5	NA	0.4802941	0.02794118	118.62	75.7	125.98	0	0	0	0	0	0	0	0
15	65-69	70	67.5	0.58	0.5082353	0.02794118	179.57	120.4	184.04	0	0	0	0	0	0	0	0
16	70-74	75	72.5	NA	0.5361765	0.02794118	237.26	165.6	250.96	0	0	0	0	0	0	0	0
17	75-79	80	77.5	NA	0.5641176	0.02794118	304.43	215.9	319.14	0	0	0	0	0	0	0	0
18	80-84	85	82.5	NA	0.5920588	0.02794118	363.45	267.4	387.22	0	0	0	0	0	0	0	0
19	85+	90	87.5	NA	0.6200000	0.02794118	391.68	288.5	NA	0	0	0	0	0	0	0	0

Figure A.2: Initial CRC Dataframe

A.1.2 *Setting Initial Parameters*

The following code contains the parameter values used in all models. Parameters ending with a "1" (or no number) refer to values obtained from Calabrese and Shibata [17] , while those ending in "2" refer to those obtained elsewhere. k refers to the number of barriers to cancer; u refers to the mutation rate; Nm refers to the

number of colon stem cells; Ns refers to the number of colon stem cells infected by JCV.

```
#Constant Parameters
k1<-6 #number of barriers to cancer (from Hanahan)
k2<-k1-3 #estimated number of barriers left after JCV infection.
u1<-3*(10^-9)*(1000) # "If three genes are at risk in a pathway, then the
    probability of mutation (u) of any one of the three genes in a single division
    is 3 * 10-6 instead of 1* 10-6 with a single gene target[of 1000bp] " Calabrese
    2010
u2<-3*(10^-8)
Nm<-8*15000000 # Number of stem cells in the colon. 8 stem cells per crypt,
    15,000,000 crypts in the colon (From Calabrese)
Nm2<-5*15000000
Ns=Nm*0.95 #Estimated number of actively infected colon crypt cells
Ns2<-Nm2*0.95
```

A.1.3 *Functions Used to Calculate Probability of Cancer*

The following code was used to create functions that calculate the probability of colon cancer for any given age. For the mutation model, the inputs are the current age, number of stem cells in the colon (N), and the mutation rate (μ). The infection model includes all of the above, but with the addition of age of infection, so that the total amount of time an individual has been infected by JCV can be calculated.

```
        #Probability Models


#Mutation model
page1<-function(age,N,u){
        d1=age*365*0.25 #Estimated From Calabrese. Number of stem cell divisions.
        p0<-1-((1-(1-(1-u)^d1)^k1)^N) #Calabrese
        p0      }
```

```
#Infection Model
page2<-function(agenow,ageinfection,N,u){
        d2=(agenow-ageinfection)*365*0.25 #Number of infected cell divisions=Number
            of years infected with JCV * Number of days in a year * 1 division every
             4 days
        p1=(1-(1-(1-(1-u)^d2)^k2)^(N))
        p1      }
```

A.1.4 *Mutation Model*

As the mutation model argues that the oncogenic process begins at birth, calculating the cumulative probability of cancer in the mutation model is simply a matter of using the above function (**page1**) on each age group. The vector CRC["ProbMutation"] collects the results when using the parameter values found in Calabrese and Shibata [17], while CRC["ProbMutationNew"] collects the probabilities when the new parameters are used.

```
CRC["ProbMutation"]<-page1(CalabreseAgeRange,Nm,u1)
CRC["ProbMutationNew"]<-page1(CalabreseAgeRange,Nm2,u2)
```

A.1.5 *Infection Model*

Calculating the cumulative probability of cancer with infection is a bit more difficult than when dealing with mutation alone. The reason is because one must take into account how long an individual has been infected, which is the difference in current age and age of infection. Thus, one must create a matrix that has each possible current age as one row, and each possible age of infection as one column. Such a matrix was created using the following code:

```
agenow<-CalabreseAgeRange
ageinfect<-CalabreseAgeRange
agenowVageinfect<-matrix(nrow=length(agenow),ncol=length(ageinfect),dimnames=list(c(
    "AgeNow0","AgeNow5","AgeNow10","AgeNow15","AgeNow20","AgeNow25","AgeNow30","
    AgeNow35","AgeNow40","AgeNow45","AgeNow50","AgeNow55","AgeNow60","AgeNow65","
    AgeNow70","AgeNow75","AgeNow80","AgeNow85","AgeNow90"),c("AgeInfect0","
    AgeInfect5","AgeInfect10","AgeInfect15","AgeInfect20","AgeInfect25","AgeInfect30
    ","AgeInfect35","AgeInfect40","AgeInfect45","AgeInfect50","AgeInfect55","
    AgeInfect60","AgeInfect65","AgeInfect70","AgeInfect75","AgeInfect80","
    AgeInfect85","AgeInfect90")))
```

The empty matrix that is created can be found in Figure A.3.

	AgeInfect0	AgeInfect5	AgeInfect10	AgeInfect15	AgeInfect20	AgeInfect25	AgeInfect30	AgeInfect35	AgeInfect40	AgeInfect45	AgeInfect50	AgeInfect55	AgeInfect60	AgeInfect65	AgeInfect70
AgeNow0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow5	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow15	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow20	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow25	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow30	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow35	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow40	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow45	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow50	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow55	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow60	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow65	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow70	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow75	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow80	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow85	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
AgeNow90	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

	AgeInfect75	AgeInfect80	AgeInfect85	AgeInfect90
AgeNow0	NA	NA	NA	NA
AgeNow5	NA	NA	NA	NA
AgeNow10	NA	NA	NA	NA
AgeNow15	NA	NA	NA	NA
AgeNow20	NA	NA	NA	NA
AgeNow25	NA	NA	NA	NA
AgeNow30	NA	NA	NA	NA
AgeNow35	NA	NA	NA	NA
AgeNow40	NA	NA	NA	NA
AgeNow45	NA	NA	NA	NA
AgeNow50	NA	NA	NA	NA
AgeNow55	NA	NA	NA	NA
AgeNow60	NA	NA	NA	NA
AgeNow65	NA	NA	NA	NA
AgeNow70	NA	NA	NA	NA
AgeNow75	NA	NA	NA	NA
AgeNow80	NA	NA	NA	NA
AgeNow85	NA	NA	NA	NA
AgeNow90	NA	NA	NA	NA

Figure A.3

A.1.5.1 *Parameters from Calabrese and Shibata [17]*

The empty matrix from above can be filled in with the probability of cancer for a given period of infection by using the **page2** function, using current age and age of infection as parameters. Note that N_s is used in these calculations (as opposed to N), as it reflects the number of infected colon stem cells (95% of colon stem cells). Only half of the matrix was filled in, since an individual cannot be infected before they were born (i.e. agenow>ageinfect). The code used to fill in the matrix is as follows:

```
for(i in 1:length(agenow)){
        for(k in 1:length(ageinfect)){
                ifelse(agenow[i]>=ageinfect[k],
                agenowVageinfect[i,k]<-page2(agenow[i],ageinfect[k],Ns2,u2),
                agenowVageinfect[i,k]<-0 ) } }
```

The filled in matrix can be found in Figure A.4

```
> agenowVageinfect
       AgeInfect0 AgeInfect5 AgeInfect10 AgeInfect15 AgeInfect20 AgeInfect25 AgeInfect30 AgeInfect35 AgeInfect40 AgeInfect45 AgeInfect50 AgeInfect55 AgeInfect60 AgeInfect65 AgeInfect70
AgeNow0   0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow5   0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow10  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow15  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow20  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow25  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow30  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow35  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow40  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow45  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow50  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow55  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000  0.0000000
AgeNow60  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000  0.0000000
AgeNow65  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000  0.0000000
AgeNow70  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331  0.0000000
AgeNow75  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157  0.2530331
AgeNow80  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081  0.9026157
AgeNow85  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  0.9996081
AgeNow90  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000  1.0000000
       AgeInfect75 AgeInfect80 AgeInfect85 AgeInfect90
AgeNow0   0.0000000  0.0000000  0.0000000          0
AgeNow5   0.0000000  0.0000000  0.0000000          0
AgeNow10  0.0000000  0.0000000  0.0000000          0
AgeNow15  0.0000000  0.0000000  0.0000000          0
AgeNow20  0.0000000  0.0000000  0.0000000          0
AgeNow25  0.0000000  0.0000000  0.0000000          0
AgeNow30  0.0000000  0.0000000  0.0000000          0
AgeNow35  0.0000000  0.0000000  0.0000000          0
AgeNow40  0.0000000  0.0000000  0.0000000          0
AgeNow45  0.0000000  0.0000000  0.0000000          0
AgeNow50  0.0000000  0.0000000  0.0000000          0
AgeNow55  0.0000000  0.0000000  0.0000000          0
AgeNow60  0.0000000  0.0000000  0.0000000          0
AgeNow65  0.0000000  0.0000000  0.0000000          0
AgeNow70  0.0000000  0.0000000  0.0000000          0
AgeNow75  0.0000000  0.0000000  0.0000000          0
AgeNow80  0.2530331  0.0000000  0.0000000          0
AgeNow85  0.9026157  0.2530331  0.0000000          0
AgeNow90  0.9996081  0.9026157  0.2530331          0
```

Figure A.4

Each element of the matrix above gives the probability of cancer given a particular length of infection (i.e. current age i - age of infection k), and thus provides the probability of colon cancer given each combination of current age and age of infection. However, one must also take into account the probability that the individual was actually infected by JCV at age k, which can be accomplished by multiplying each element by the average incidence of JCV. Afterwards, one can sum across each row to determine the cumulative probability of colon cancer given infection by JCV for each current age, yielding the predicted *prevalence* of colon cancer for each age. The logic behind this is that prevalence is equal to the total number of cases, which is the sum of all new cases (incidence), past and present, for each current age. For example, the prevalence of CRC at age 60 includes all individuals that developed CRC at age $30, 35, 40 \ldots 60$, which is equivalent to summing across each row of the above matrix. These values were then

stored in CRC["ProbInfection"]. The code used to accomplish this is as follows (Note: if the cumulative probability of colon cancer exceeded 1, it was replaced with a value of 1):

```
agenowVageinfect[which(is.na(agenowVageinfect==TRUE))]=0 #replace NA values with 0
for(i in 1:length(CalabreseAgeRange)){
        ifelse(sum(agenowVageinfect[i,1:length(ageinfect)])<1,
        CRC["ProbInfection"][i,1]<-sum(agenowVageinfect[i,1:length(ageinfect)])*
            AvgJCVIncidence,
        CRC["ProbInfection"][i,1]<-1) }
```

A.1.5.2 *New Parameters*

The same method as described in A.1.5.1 was used to calculate the expected prevalence of colon cancer with infection, but using the new parameter values. The "AgeNowVsAgeInfect" matrix can be found in Figure A.5.After all calculations were completed, they were stored in CRC["ProbInfectionNew"].

```
agenow<-CalabreseAgeRange
ageinfect<-CalabreseAgeRange
agenowVageinfect<-matrix(nrow=length(agenow),ncol=length(ageinfect),dimnames=list(c(
    "AgeNow0","AgeNow5","AgeNow10","AgeNow15","AgeNow20","AgeNow25","AgeNow30","
    AgeNow35","AgeNow40","AgeNow45","AgeNow50","AgeNow55","AgeNow60","AgeNow65","
    AgeNow70","AgeNow75","AgeNow80","AgeNow85","AgeNow90"),c("AgeInfect0","
    AgeInfect5","AgeInfect10","AgeInfect15","AgeInfect20","AgeInfect25","AgeInfect30
    ","AgeInfect35","AgeInfect40","AgeInfect45","AgeInfect50","AgeInfect55","
    AgeInfect60","AgeInfect65","AgeInfect70","AgeInfect75","AgeInfect80","
    AgeInfect85","AgeInfect90")))

for(i in 1:length(agenow)){
        for(k in 1:length(ageinfect)){
                ifelse(agenow[i]>=ageinfect[k],
                agenowVageinfect[i,k]<-page2(agenow[i],ageinfect[k],Ns2,u2),
                agenowVageinfect[i,k]<-0 ) } }
```

```
agenowVageinfect[which(is.na(agenowVageinfect==TRUE))]=0 #replace NA values with 0
for(i in 1:length(CalabreseAgeRange)){
        ifelse(sum(agenowVageinfect[i,1:length(ageinfect)])<1,
        CRC["ProbInfectionNew"][i,1]<-sum(agenowVageinfect[i,1:length(ageinfect)])*
            AvgJCVIncidence,
        CRC["ProbInfectionNew"][i,1]<-1) }
```

```
> agenowVageinfect
          AgeInfect0  AgeInfect5  AgeInfect10  AgeInfect15  AgeInfect20  AgeInfect25  AgeInfect30  AgeInfect35  AgeInfect40  AgeInfect45  AgeInfect50  AgeInfect55  AgeInfect60  AgeInfect65
AgeNow0  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow5  1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow10 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow15 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow20 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow25 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow30 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow35 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow40 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow45 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow50 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow55 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow60 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00
AgeNow65 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00
AgeNow70 5.010814e-04 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07
AgeNow75 6.162628e-04 5.010814e-04 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06 1.463412e-06
AgeNow80 7.478484e-04 6.162628e-04 5.010814e-04 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05 4.936039e-06
AgeNow85 8.969306e-04 7.478484e-04 6.162628e-04 5.010814e-04 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05 1.169141e-05
AgeNow90 1.064592e-03 8.969306e-04 7.478484e-04 6.162628e-04 5.010814e-04 4.012189e-04 3.155885e-04 2.431026e-04 1.826568e-04 1.331617e-04 9.352747e-05 6.265583e-05 3.945599e-05 2.283689e-05
          AgeInfect70  AgeInfect75  AgeInfect80  AgeInfect85 AgeInfect90
AgeNow0  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow5  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow10 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow15 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow20 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow25 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow30 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow35 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow40 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow45 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow50 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow55 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow60 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow65 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow70 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow75 1.819378e-07 0.000000e+00 0.000000e+00 0.000000e+00           0
AgeNow80 1.463412e-06 1.819378e-07 0.000000e+00 0.000000e+00           0
AgeNow85 4.936039e-06 1.463412e-06 1.819378e-07 0.000000e+00           0
AgeNow90 1.169141e-05 4.936039e-06 1.463412e-06 1.819378e-07           0
```

Figure A.5

A.1.6 *Converting Data from Prevalence to Incidence*

The probability values generated reflect the cumulative probability (i.e. prevalence) of colon cancer for each age group. However, the data from SEE [2] are in the form of incidence per age group, per 100,000 individuals. As such, the prevalence data were converted to incidence by taking the difference in probability of colon cancer between each age group. Afterwards, the incidence values were multiplied by 100,000 so as to provide the expected incidence of CRC per 100,000 individuals. This was accomplished by using the following code:

```
        #Calculate Incidence of Colon Cancer, per 100,000


#Mutation
```

```
for(i in 2:length(CRC["ProbMutation"][,1])){
        CRC["Incidence_Mutation"][i,]<-(CRC["ProbMutation"][i,]-CRC["ProbMutation"
            ][(i-1),])*100000
                }

for(i in 2:length(CRC["ProbMutationNew"][,1])){
                CRC["Incidence_MutationNew"][i,]<-(CRC["ProbMutationNew"][i,]-CRC["
                    ProbMutationNew"][(i-1),])*100000
                }

#Infection

for(i in 2:length(CRC["ProbInfection"][,1])){
                CRC["Incidence_Infection"][i,]<-(CRC["ProbInfection"][i,]-CRC["
                    ProbInfection"][(i-1),])*100000
                }

for(i in 2:length(CRC["ProbInfectionNew"][,1])){
                CRC["Incidence_InfectionNew"][i,]<-(CRC["ProbInfectionNew"][i,]-CRC[
                    "ProbInfectionNew"][(i-1),])*100000
                }
```

A.1.7 *Results*

A.1.7.1 *Dataframe*

After being filled in by the above code, the CRC dataframe looks as follows:

> CRC

	SeerAges	CalabreseAgeRange	AverageAgeCancer	JCVPrevalence	JCVEstimatedPrev	JCVIncidence	CRC_2000_2006	CRC_2003_2007	CalabreseData	ProbMutation	ProbMutationNew	Incidence_Mutation
1	<1	0	0.0	NA	0.1310294	NA	0.00	0.0	0.00	0.000000e+00	0	0.00000000
2	1-4	5	2.5	0.11	0.1450000	0.01397059	0.00	0.0	0.00	0.000000e+00	0	0.00000000
3	5-9	10	7.5	0.14	0.1729412	0.02794118	0.00	0.0	0.00	5.329070e-08	0	0.00532907
4	10-14	15	12.5	0.24	0.2008824	0.02794118	0.06	0.0	0.07	5.728749e-07	0	0.05195842
5	15-19	20	17.5	0.22	0.2288235	0.02794118	0.26	0.2	0.18	3.184115e-06	0	0.26112396
6	20-24	25	22.5	NA	0.2567647	0.02794118	0.81	0.6	0.47	1.208359e-05	0	0.88994798
7	25-29	30	27.5	0.34	0.2847059	0.02794118	1.86	1.2	1.46	3.591729e-05	0	2.38336959
8	30-34	35	32.5	NA	0.3126471	0.02794118	4.06	2.7	2.82	9.020377e-05	0	5.42864823
9	35-39	40	37.5	0.39	0.3405882	0.02794118	7.74	4.6	5.59	2.001532e-04	0	10.99494054
10	40-44	45	42.5	NA	0.3685294	0.02794118	14.67	9.2	11.14	4.040750e-04	0	20.39218698
11	45-49	50	47.5	0.34	0.3964706	0.02794118	27.65	16.9	21.59	7.570941e-04	0	35.38190365
12	50-54	55	52.5	NA	0.4244118	0.02794118	53.12	33.5	42.72	1.335372e-03	0	57.82779444
13	55-59	60	57.5	0.45	0.4523529	0.02794118	77.97	48.7	77.94	2.240572e-03	0	98.51998517
14	60-64	65	62.5	NA	0.4802941	0.02794118	118.62	75.7	125.98	3.604599e-03	0	136.40273079
15	65-69	70	67.5	0.50	0.5082353	0.02794118	179.57	120.4	184.04	5.594405e-03	0	198.98058930
16	70-74	75	72.5	NA	0.5361765	0.02794118	237.26	165.6	250.96	8.416646e-03	0	282.22405037
17	75-79	80	77.5	NA	0.5641176	0.02794118	304.43	215.9	319.14	1.232194e-02	0	390.52921474
18	80-84	85	82.5	NA	0.5920588	0.02794118	363.45	267.4	387.22	1.760815e-02	0	528.62107868
19	85+	90	87.5	NA	0.6200000	0.02794118	391.68	288.5	NA	2.462230e-02	0	701.41507843

	Incidence_MutationNew	ProbInfection	ProbInfectionNew	Incidence_Infection	Incidence_InfectionNew
1	0	0.000000000	0.000000e+00	0.0000	0.0000000000
2	0	0.006873652	4.942346e-09	687.3652	0.0004942346
3	0	1.000000000	4.469597e-08	99312.6348	0.0039753625
4	0	1.000000000	1.787836e-07	0.0000	0.0134087671
5	0	1.000000000	4.963813e-07	0.0000	0.0317597610
6	0	1.000000000	1.116746e-06	0.0000	0.0620364811
7	0	1.000000000	2.188569e-06	0.0000	0.1071823333
8	0	1.000000000	3.890617e-06	0.0000	0.1702047742
9	0	1.000000000	6.431294e-06	0.0000	0.2540676914
10	0	1.000000000	1.004864e-05	0.0000	0.3617341691
11	0	1.000000000	1.501051e-05	0.0000	0.4961877287
12	0	1.000000000	2.161440e-05	0.0000	0.6603890882
13	0	1.000000000	3.018738e-05	0.0000	0.8572973009
14	0	1.000000000	4.108650e-05	0.0000	1.0899124545
15	0	1.000000000	5.469839e-05	0.0000	1.3611892860
16	0	1.000000000	7.143919e-05	0.0000	1.6740797859
17	0	1.000000000	9.175452e-05	0.0000	2.0315327526
18	0	1.000000000	1.161197e-04	0.0000	2.4365147928
19	0	1.000000000	1.458393e-04	0.0000	2.8919668708

Figure A.6:

A.1.7.2 *Plots*

For the purposes of plotting, the incidence data were converted to log-scale, so as to emphasize the different predictions between the mutation models and the infection models. If the incidence was 0, then $\log(0)$ yields $-\infty$, which obviously cannot fit onto the plot. Therefore, values that did yield $-\infty$ were replaced with -25.22388, which reflects the smallest predicted probability. The code used to accomplish this is as follows:

```
        #Adjust data so the log can be taken: if prob=0, then log(0)=-Inf, which won
            't fit on a graph. Set min to -25.22388, the smallest probability
            produced in Stepwise Model that was not -Inf
logIMOld<-log(CRC[,"Incidence_Mutation"])
logIMNew<-log(CRC[,"Incidence_MutationNew"])
logIIOld<-log(CRC[,"Incidence_Infection"])
logIINew<-log(CRC[,"Incidence_InfectionNew"])


#Replace -Inf with -25


#Mutation, Old Parameters
for(i in 1:length(logIMOld)){
        ifelse(logIMOld[i]==-Inf,
```

```
                logIMOld[i]<--25.22388,
                logIMOld[i]<-logIMOld[i]) }

#Mutation, New Parameters
for(i in 1:length(logIMNew)){
        ifelse(logIMNew[i]==-Inf,
                logIMNew[i]<--25.22388,
                logIMNew[i]<-logIMNew[i])       }

#Infection, Old Parameters
for(i in 1:length(logIIOld)){
        ifelse(logIIOld[i]==-Inf,
                logIIOld[i]<--25.22388,
                logIIOld[i]<-logIIOld[i]) }

#Infection, New Parameters
for(i in 1:length(logIINew)){
        ifelse(logIINew[i]==-Inf,
                logIINew[i]<--25.22388,
                logIINew[i]<-logIINew[i]) }
```

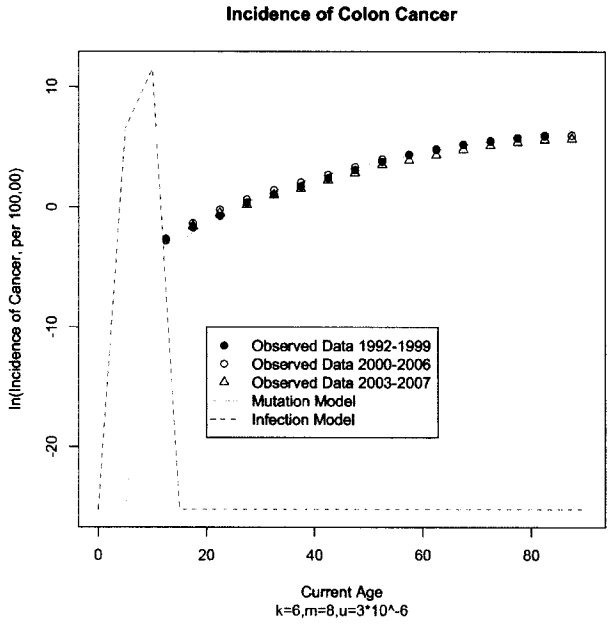
After transforming the data, the following code was used to generate plots comparing the observed data to each model's predictions, given the parameters used in Calabrese and Shibata [17] (the resulting plot can be found in FigureA.7a :

```
#Log Plot of Predicted Cancer Incidence Probability,As a Function of Age, With Real
    Data.Infection and Mutation Models, Old Parameters   xrange<-CalabreseAgeRange
ymax<-max(CRC[,"Incidence_Infection"]) lymax<-log(ymax)
lymin<--25.22388 #Based off results from stepwise model. Otherwise,log(CRC[,"
    Incidence_MutationNew"]) yields -Inf for all values because (CRC[,"Incidence_
    MutationNew"]) is 0
yrange<-seq(lymin,lymax,(lymax-lymin)/(length(xrange)-1))
```

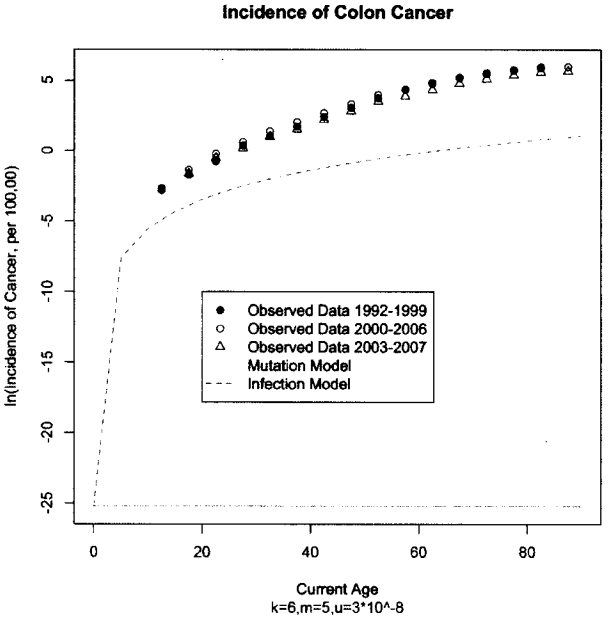
```
plot(xrange, yrange, type="n", xlab="Current Age",ylab="ln(Incidence of Cancer, per
    100,00)",main="Incidence of Colon Cancer",sub="k=6,N=8,u=3*10^-6") points(
    AverageAgeCancer,log(CalabreseData),pch=19,lty=1)
points(AverageAgeCancer,log(CRC_2000_2006),pch=1,lty=1)
points(AverageAgeCancer,log(CRC_2003_2007),pch=2,lty=1)
lines(CalabreseAgeRange,logIMOld,col="red")
lines(CalabreseAgeRange,logIIOld,col="blue",lty=2)
legend(20,-10,c("Observed Data 1992-1999","Observed Data 2000-2006","Observed Data
    2003-2007","Mutation Model","Infection Model"),col=c(1,1,1,"red","blue"),lty=c(
    NA,NA,NA,1,2),pch=c(19,1,2,rep(NA,2)))
```

A similar set of code was used to plot the results of each model's predictions, given the new parameters (the resulting plot can be found in Figure A.7b):

```
#Log Plot of Predicted Cancer Incidence Probability,As a Function of Age, With Real
    Data. Both Models, New Parameters
xrange<-CalabreseAgeRange
ymax<-max(CRC[,"CRC_2000_2006"])
lymax<-log(ymax)
lymin<--25.22388 #Based off results from stepwise model. Otherwise,log(CRC[,"
    Incidence_MutationNew"]) yields -Inf for all values because (CRC[,"Incidence_
    MutationNew"]) is 0
yrange<-seq(lymin,lymax,(lymax-lymin)/(length(xrange)-1))
plot(xrange, yrange, type="n", xlab="Current Age",ylab="ln(Incidence of Cancer, per
    100,00)",main="Incidence of Colon Cancer",sub="k=6,N=5,u=3*10^-8")
points(AverageAgeCancer,log(CalabreseData),pch=19,lty=1)
points(AverageAgeCancer,log(CRC_2000_2006),pch=1,lty=1)
points(AverageAgeCancer,log(CRC_2003_2007),pch=2,lty=1)
lines(CalabreseAgeRange,logIMNew,col="red")
lines(CalabreseAgeRange,logIINew,col="blue",lty=2)
legend(20,-10,c("Observed Data 1992-1999","Observed Data 2000-2006","Observed Data
    2003-2007","Mutation Model","Infection Model"),col=c(1,1,1,"red","blue"),lty=c(
    NA,NA,NA,1,2),pch=c(19,1,2,rep(NA,2)))
```



(a) Old Parameters



(b) New Parameters

Figure A.7:

A.2 ESTIMATING INCIDENCE OF COLON CANCER: GENOMIC INSTABILITY MODEL

A problem of the static model described above is that it assumes all cells have an equal probability of acquiring enough mutations to knock out each barrier. However, it seems more realistic that a subset of cells will acquire the first mutation, and that this subset will have a head start in the race towards cancer. Thus, the following model tracks how many cells have acquired x mutations. This task is accomplished by determining the probability that any one cell has acquired a mutation. That probability is then multiplied by the current number of colon stem cells, yielding the expected number of stem cells that have acquired x mutations. This process is repeated until all barriers to cancer have been removed, and as such the number of cells carrying x mutations decreases over time. Furthermore, it has been observed that the mutation rate increases throughout the development of cancer cells, a process known as genomic instability. This model tries to capture the impact of genomic instability by increasing the mutation rate by 150% every time a barrier is removed. Finally, this model uses the "new parameters" described above (i.e. $\mu_0 = 3 \times 10^{-8}$ and $m = 5$ stem cells per crypt).

A.2.1 *Importing the Observed data*

Just as in the static model, the first step of this code is to import the observed data on CRC incidence and JCV prevalence:

```
Seer0307<-read.csv("/Users/chandlergatenbee/Documents/UofL/CancerProbPaper/Observed
    Data/CRC03-07.csv") #Colon cancer data from SEER 2003-2007
Seer0006<-read.csv("/Users/chandlergatenbee/Documents/UofL/CancerProbPaper/Observed
    Data/CRC00-06.csv") #Colon cancer data from SEER 2000-2006
CalabreseAgeRange<-seq(0,90,5) #Age Values used by Calabrese to calculate
    probability of cancer SeerAges<-Seer0307[,1] #Age categories used by SEER
CRC_2000_2006<-Seer0006[,2] #Getting actual values
```

```
CRC_2003_2007<-Seer0307[,2] #Geetting actual values
JCVPrevalence<-c(NA,0.11,0.14, 0.24, 0.22, NA, 0.34, NA, 0.39, NA, 0.34, NA, 0.45,
    NA, 0.5, NA, NA, NA, NA) #JCV Prevalence by age From Knowles 2003
AverageAgeCancer<-rep(0,length(CalabreseAgeRange))
        for(i in 2:length(CalabreseAgeRange)){
                AverageAgeCancer[i]<-mean(CalabreseAgeRange[(i-1):i])
                        }
JCVAvgPrevalence<-rep(NA,length(JCVPrevalence))
        for(i in 2:length(JCVPrevalence)){
                JCVAvgPrevalence[i]<-mean(JCVPrevalence[(i-1):i])
                        }
CalabreseData<-c(0, 0, 0, 0.07, 0.18, 0.47, 1.46, 2.82, 5.59, 11.14, 21.59, 42.72,
    77.94, 125.98, 184.04, 250.96, 319.14, 387.22, NA) #Actual data used in
    Calabrese; from SEER 1992-1999
```

A.2.2 *Creating Empty Vectors*

Two vectors were created to collect the results of this model: "PrevInfectInit" collects the prevalence data generated by the Infection Initiation Model (i.e. Infection Model); "PrevMutation" collects the prevalence data generated by the Mutation model:

```
PrevInfectInit<-rep(0,length(CalabreseAgeRange)) Incidence_InfectInit<-rep(0,length(
    CalabreseAgeRange))

PrevMutation<-rep(0,length(CalabreseAgeRange)) Incidence_Mutation<-rep(0,length(
    CalabreseAgeRange))
```

A.2.3 *Estimating JCV Incidence*

As in the static model, JCV incidence was estimated by linearly interpolating missing values of the JCV prevalence regression line, and then taking the difference between the prevalence values for each age group. The same code used above was used here as well:

```
#Force Regression Line to go through origin: no evidence that JCV is vertically
    transmitted
RegPrev<-lm(JCVPrevalence~AverageAgeCancer)
JCVEstimatedPrev<-coef(RegPrev)[2]*AverageAgeCancer+coef(RegPrev)[1]

#Plot to Verify Predicted Values fit Regression Line
xrangelm<-AverageAgeCancer
yrangelm<-seq(0,0.5,0.5/(length(xrangelm)-1))
plot(xrangelm,yrangelm,xlab="Age of Infection", ylab="Prevelance of JCV",main="
    Estimated Prevalence of JCV Infection", type="n")
points(AverageAgeCancer,JCVPrevalence,pch=1,col=1) #Observed Prevalence
points(AverageAgeCancer,JCVEstimatedPrev,pch=2,col=2) #Estimated Prevalence
abline(RegPrev)
legend(0, 0.5, c("Actual Prevalence of JCV", "Estimated Prevelance of JCV", "
    Estimated Prevelance for Each Age"), lty=c(NA,1,NA), pch=c(1,NA,2), col=c(1,1,2)
    )

#Calculate JCV incidence
JCVIncidence<-rep(NA,length(AverageAgeCancer))
        for(i in 2:length(JCVEstimatedPrev)){
                JCVIncidence[i]<-JCVEstimatedPrev[i]-JCVEstimatedPrev[(i-1)]
                        }
AvgJCVIncidence<-mean(JCVIncidence[-1])
```

A.2.4 *Setting Initial Parameters*

As noted above, the "new" parameter values were used in this simulation, as their description is more recent and consistent with stem cell theory. The only addition to these parameters was "mutincrease", which describes the degree of genomic instability, here estimated as a 1.5x increase in the mutation rate every time a barrier is removed.

```
#Constant Parameters
u<-3 #Total number of genes that can knock possibly out each barrier.~4 per barrier.
     Note, on COSMIC (Catolog of Somatic Mutations In Cancer) there are only 7 genes
     known to have mutations related to cancer of the GI tract (http://www.sanger.ac
    .uk/ perl/ genetics/ CGP/ cosmic?action=byhist&ss=NS&ss=lymph_node&sn=
    gastrointestinal_tract_%28site_indeterminate%29&s=3&hn=carcinoid-endocrine_
    tumour&hn=other&hn=carcinoma).
k1<-6 #number of barriers to cancer (from Hanahan)
k2<-3 #estimated number of barriers remaining after JCV infection
u1<-u*(10^-8) #Muations for stem cell lineages. Mutation rate from Frank 2003, and
    is per gene per divsion (they suggest the division rate for stem cells is
    actually between 10^-7 and 10^-10, so 10^-8 is between, although it is still on
    the higher side). u1=#genes that can knockout pathway * stem cell mutation rate
    * Number of genes active in a cell (Frank 2004)
Nm<-5*(15000000) # Number of stem cells in the colon. 5 stem cells per crypt,
    15,000,000 crypts in the colon (From Calabrese). Assuming these cells are CD133+
Ns=Nm*0.95 #Estimated number of stem cells infected by JCV
mutincrease<-1.5  #Amount mutation rate is increased by after each barrier is
    knocked down: Genomic Instability
```

A.2.5 *Creating the Dataframe*

After all vectors were created, they were collected in a dataframe called "CRC" (the resulting dataframe can be found in Figure A.8):

```
        #Main Data frame
CRC<-data.frame(SeerAges, CalabreseAgeRange, AverageAgeCancer, JCVPrevalence,
    JCVEstimatedPrev, JCVIncidence, CRC_2000_2006, CRC_2003_2007, CalabreseData,
    PrevMutation, Incidence_Mutation, PrevInfectInit, Incidence_InfectInit)
```

```
> CRC
   SeerAges CalabreseAgeRange AverageAgeCancer JCVPrevalence JCVEstimatedPrev JCVIncidence CRC_2000_2006 CRC_2003_2007 CalabreseData PrevMutation Incidence_Mutation PrevInfectInit
1        <1                0              0.0            NA        0.1310294          NA         0.00          0.0          0.00            0                 0             0
2       1-4                5              2.5          0.11        0.1450000  0.01397059         0.00          0.0          0.00            0                 0             0
3       5-9               10              7.5          0.14        0.1729412  0.02794118         0.00          0.0          0.00            0                 0             0
4     10-14               15             12.5          0.24        0.2008824  0.02794118         0.06          0.0          0.07            0                 0             0
5     15-19               20             17.5          0.22        0.2288235  0.02794118         0.26          0.2          0.18            0                 0             0
6     20-24               25             22.5            NA        0.2567647  0.02794118         0.81          0.6          0.47            0                 0             0
7     25-29               30             27.5          0.34        0.2847059  0.02794118         1.86          1.2          1.46            0                 0             0
8     30-34               35             32.5            NA        0.3126471  0.02794118         4.06          2.7          2.82            0                 0             0
9     35-39               40             37.5          0.39        0.3405882  0.02794118         7.74          4.6          5.59            0                 0             0
10    40-44               45             42.5            NA        0.3685294  0.02794118        14.67          9.2         11.14            0                 0             0
11    45-49               50             47.5          0.34        0.3964706  0.02794118        27.65         16.9         21.59            0                 0             0
12    50-54               55             52.5            NA        0.4244118  0.02794118        53.12         33.5         42.72            0                 0             0
13    55-59               60             57.5          0.45        0.4523529  0.02794118        77.97         48.7         77.94            0                 0             0
14    60-64               65             62.5            NA        0.4802941  0.02794118       118.62         75.7        125.98            0                 0             0
15    65-69               70             67.5          0.50        0.5082353  0.02794118       179.57        120.4        184.04            0                 0             0
16    70-74               75             72.5            NA        0.5361765  0.02794118       237.26        165.6        250.96            0                 0             0
17    75-79               80             77.5            NA        0.5641176  0.02794118       304.43        215.9        319.14            0                 0             0
18    80-84               85             82.5            NA        0.5920588  0.02794118       363.45        267.4        387.22            0                 0             0
19      85+               90             87.5            NA        0.6200000  0.02794118       391.68        288.5            NA            0                 0             0
   Incidence_InfectInit
1                     0
2                     0
3                     0
4                     0
5                     0
6                     0
7                     0
8                     0
9                     0
10                    0
11                    0
12                    0
13                    0
14                    0
15                    0
16                    0
17                    0
18                    0
19                    0
```

Figure A.8

A.2.6 *Creating Genomic Instability*

A vector containing the increase in mutation rate was created using the code below. A vector was created so that it's elements could be accessed later during the modeling process. The resulting vector can be found in Figure A.9.

```
#Genomic Instability


GI<-c(1,rep(NA,k1-1))
        for(i in 2:length(GI)){
                GI[i]<-GI[i-1]*mutincrease }
```

```
> GI
[1] 1.00000 1.50000 2.25000 3.37500 5.06250 7.59375
```

Figure A.9

A.2.7 *Functions Used to Calculate Probability of Cancer*

Functions similar to those used in the static model were used here to calculate each model's probability of cancer. The primary difference is that an individual's total number of divisions is divided equally among each barrier (i.e. "(age/k1)". For example, in the mutation model an individual that is 30 years old has 5 years worth of divisions to knock out the first barrier, 5 years worth of divisions to knock out the second barrier, and so on. In the infection model the number of divisions is divided among the years an individual has been infected: If now 60 and infected at 30, then the individual has 5 years worth of divisions to knock out the fourth barrier (JCV knocked out the first 3) , 5 years worth of divisions to knock out the 4th barrier, and 5 years worth of divisions to knock out the final 6th barrier:

```
#Probability Models


#Mutation
page1<-function(age,N,u){
        d1=(age/k1)*365*0.25 #Estimated From Calabrese. Number of stem cell
            divisions. Number of divisions divided equally between barriers
        p1<-(1-((1-(1-(1-u)^d1)^1)^N)) #Probability of 1 cell knocking out 1 barrier
             after d1/k1 divisions
        p1      }


#Infection
page2<-function(agenow,ageinfection,N,u){
        d2=((agenow-ageinfection)/k2)*365*0.25 #Number of infected cell divisions=
            Number of years infected with JCV * Number of days in a year * 1
            division every 4 days. Number of divisions divided equally between
            barriers
        p2=(1-(1-(1-(1-u)^d2)^1)^(N)) #Probability of 1 cell knocking out 1 barrier
            after d2/k2 divisions
        p2      }
```

A.2.8 *Genomic Instability Mutation Model*

A.2.8.1 *Calculating Probability of Oncogenesis with Decreasing Cell Population and Increasing Mutation Rate*

The vector "mprev" was created to store the predicted prevalence of colon cancer for each age group, per 100,000 individuals. A *for-loop* was created to calculate the probability of oncogenesis for each age group. The first step of this loop was to reset all parameters back to their original values every time the probability of cancer was being calculated for a new age group: "mutrate" is a vector of the mutation rate at each step; "Nm" is the initial number of stem cells; "mN" is a vector that stores the number of stem cells that have knocked down k barriers (the first value is Nm); and "mprob" is a vector of the probability that a cell has knocked down the kth barrier. A sample of these vectors and values can be found in Figure A.10(they are from the last age group, 90 years old).

The first step of calculating the probability of cancer was to calculate the probability that 1 cell would knock down the first barrier, given the individual's age and the initial mutation rate. This value was stored as "mprob[1]". After this initial probability was determined, another loop was initiated. This loop first calculates how many cells are expected to have knocked down the first barrier. This is accomplished by multiplying "mprob[1]" by the initial number of stem cells ("mN[(i-1)]"; i starts at 2). After the number of stem cells that have removed the first barrier has been calculated, the probability of one cell knocking down the second barrier is calculated, using the next mutation rate ("mutrate[i]"). This loop thus tracks how many cells are expected to have removed a barrier, as well as the probability of 1 cell knocking down the next barrier given the increased mutation rate.

The above loop is repeated until the 5th barrier has been removed. At this point, "mprob" has 5 probabilities, each reflecting the probability that 1 cell has knocked out a barrier, up to and including the 5th barrier. "mN" also contains 5 values, each reflecting the number of cells that are expected to have been able to knock out each barrier, up to, but not including, the 5th barrier. Using these values, "mN[k1]" (i.e. mN[6]) calculates the number of cells expected to have knocked out out the 5th barrier. This is the sub-population of cells that can knockout the final barrier. The probability of this is calculated using **page1**, that sub-population of cells ("mN[k1]"; k=6), and the final (highest) mutation rate ("mutrate[k1]"). This final value, "mprob[k1]", thus reflects the probability that the sub-population of cells that already knocked down 5 barriers knocked down the final, 6th, barrier. After this final probability had been calculated, it is multiplied by 100,000 and stored in "mprev", which keeps track of the prevalence of colon cancer per 100,000 individuals, for each age group.

The above loop is then repeated for each age group (*j*). After the final prevalence value was collected, all prevalence values for each age group were moved to CRC[,"PrevMutation"].

```
mprev<-rep(NA,length(CalabreseAgeRange)) #Predicted prevalence of colon cancer for
    each age group, assuming the mutation model

for(j in 1:length(CalabreseAgeRange)){

        #Reset Values for each Age Group
        mutrate<-GI*u1 #Vector of increase in mutation rate
        Nm<-5*(15000000) #Initial number of colon stem cells
        mN<-c(Nm,rep(NA,k1-1)) #Vector storing number of cells remaining after each
            barrier is knocked down, up to 5th barrier
        mprob<-rep(NA,k1)       #Vector storing probabilities of knocking down each
            barrier

        #Calculate Probablity
```

```
        mprob[1]<-page1(CalabreseAgeRange[j],1,mutrate[1]) #Probability of 1 cell
            knocking down barrier 1

        for(i in 2:(k1-1)){
                mN[i]<-mprob[(i-1)]*mN[(i-1)]   #Number of cells that would have
                    knocked out barrier
                mprob[i]<-page1(CalabreseAgeRange[j],1,mutrate[i]) #Having knocked
                    down previous barriers, Probability of 1 cell knocking down
                    barrier
                                                }

        mN[k1]<-mprob[k1-1]*mN[k1-1]    #Number of cells that would have knocked out
             barrier 5
        mprob[k1]<-page1(CalabreseAgeRange[j],mN[k1],mutrate[k1]) #Probability of
            knocking down all 6 barriers

        mprev[j]<-mprob[k1]*100000 #Predicted prevalence of CRC, per 100,000
            individuals

        }

CRC[,"PrevMutation"]<-mprev
```

```
> mutrate
[1] 3.000000e-08 4.500000e-08 6.750000e-08 1.012500e-07 1.518750e-07 2.278125e-07
> Nm
[1] 7.5e+07
> mN
[1] 7.500000e+07 3.079624e+03 1.896798e-01 1.752382e-05 2.428387e-09 5.047581e-13
> mprob
[1] 4.106166e-05 6.159185e-05 9.238636e-05 1.385763e-04 2.078573e-04 1.110223e-16
```

Figure A.10

A.2.8.2 *Calculating Incidence of Colon Cancer with Mutation Model*

As the above results are in the form of prevalence data, they must be transformed into incidence data so that they are comparable to the observed data. This was ac-

complished calculating the difference in prevalence between any two age groups, and then stored in CRC["Incidence_Mutation"] and "Mutation":

```
#Calculate Incidence


for(i in 2:length(mprev)){
        CRC[i,"Incidence_Mutation"]<-CRC[i,"PrevMutation"]-CRC[(i-1),"PrevMutation"]
             }


Mutation<-CRC[,"Incidence_Mutation"]
```

A.2.9 *Genomic Instability Infection Model*

A.2.9.1 *Calculating Probability of Cancer Given Length of Infection*

As in the stepwise mutation model, a for loop was used to calculate the probability of cancer for each age group. After the probability was calculated, the parameters were reset to their original values: "mutrate" is a vector of the mutation rates; "Ns" is the initial number of infected stem cells; "iiN" is a vector containing the number of infected stem cells that have knocked out a barrier (starting at 0 barriers); "iiprob" is a vector containing the probability that the kth barrier has been removed. An example set of these values can be found in Figure A.11 (these values are for 90 year old individuals that were infected at 90 years old, which explains why iiN and iiprob are zero).

The logic underlying the stepwise infection model is the same as that underlying the mutation model. However, in the case of the infection model, JCV has already knocked out three barriers, and so only 3 barriers remain. Thus, a complicated for loop seemed unnecessary and the cell population sizes and probabilities were calculated in a series. Of note is that the mutation rate starts at the level expected if three barriers have already been removed, while the cell population size is remains at the initial size (see Figure ??). Finally, as in the static model,

a matrix was created so as to capture the probability of cancer given the length of infection (current age - age of infection). This matrix was filled in with the probability of cancer for each combination of current age and age of infection, and can be found in Figure A.12.

```
#Create Matrix
agenow<-CalabreseAgeRange
ageinfect<-CalabreseAgeRange
agenowVageinfect<-matrix(nrow=length(agenow),ncol=length(ageinfect),dimnames=list(c(
    "AgeNow0", "AgeNow5", "AgeNow10", "AgeNow15", "AgeNow20", "AgeNow25", "AgeNow30"
    , "AgeNow35", "AgeNow40", "AgeNow45", "AgeNow50", "AgeNow55", "AgeNow60", "
    AgeNow65", "AgeNow70", "AgeNow75", "AgeNow80", "AgeNow85", "AgeNow90"), c("
    AgeInfect0", "AgeInfect5", "AgeInfect10", "AgeInfect15", "AgeInfect20", "
    AgeInfect25", "AgeInfect30", "AgeInfect35", "AgeInfect40", "AgeInfect45", "
    AgeInfect50", "AgeInfect55", "AgeInfect60", "AgeInfect65", "AgeInfect70", "
    AgeInfect75", "AgeInfect80", "AgeInfect85", "AgeInfect90")))

#Calculate Probability of Cancer for each current age Vs. age of infection

for(i in 1:length(agenow)){
        for(j in 1:length(ageinfect)){
                #Reset Parameters for each Age Group
                mutrate<-GI*u1 #mutation rates
                Ns<-5*(15000000)*0.95 #initial cell population size
                iiN<-c(Ns,rep(NA,k2-1)) #collects number of cells that have knocked
                    out each barrier
                iiprob<-rep(NA,k2) #collects probability of 1 cell knocking out a
                    barrier

        #Calculate Probability of Knocking out Each Barrier, given N stem cells and
            a u mutation rate. Infection Already knocked out 3 barriers

                #Probability of knocking-out Barrier 4
```

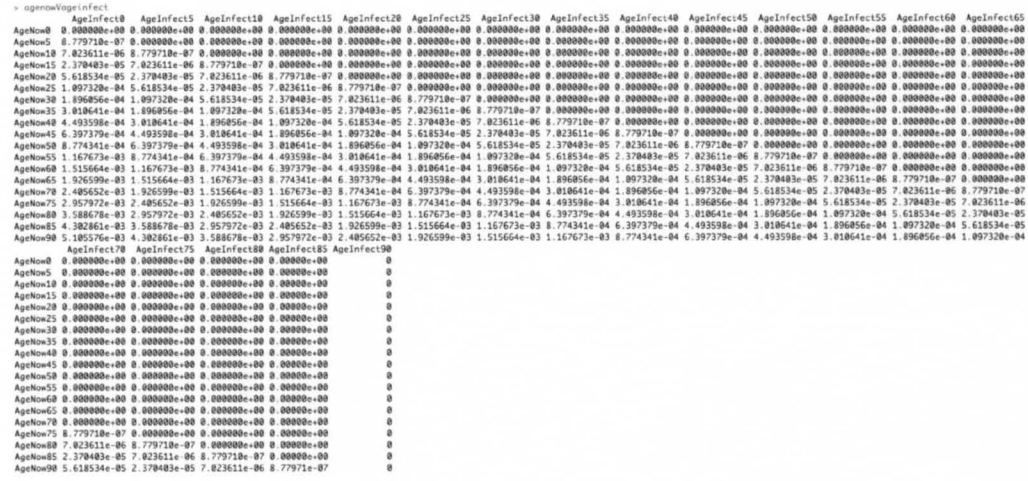
```
            iiprob[1]<-page2(agenow[i],ageinfect[j],1,mutrate[4]) #Probability
                of 1 cell knocking down barrier 4

            # Probability of knocking-out Barrier 5
            iiN[2]<-iiprob[1]*Ns    #Number of cells that would have knocked out
                 barrier 4
            iiprob[2]<-page2(agenow[i],ageinfect[j],1,mutrate[5]) #Having
                knocked down previous barriers, this is the probability of 1
                cell knocking down barrier 5

            # Probability of knocking-out Barrier 6
            iiN[3]<-iiprob[2]*iiN[2]        #Having knocked down previous
                barriers,Number of cells that would have knocked out barrier 5
            iiprob[3]<-page2(agenow[i],ageinfect[j],iiN[3],mutrate[6]) #
                Probability of knocking down all 6 barriers, given the number of
                 cells that removed 5 barriers

            ifelse(iiprob[3]>0,
                    agenowVageinfect[i,j]<-iiprob[3],
                    agenowVageinfect[i,j]<-0) #Makes sure all probabilites are
                        positive
            }
        }
```

```
> mutrate
[1] 3.000000e-08 4.500000e-08 6.750000e-08 1.012500e-07 1.518750e-07 2.278125e-07
> Ns
[1] 71250000
> iiN
[1] 71250000        0        0
> iiprob
[1] 0 0 0
```

Figure A.11

A.2.9.2 *Calculating Cumulative Probability of Cancer for Each Age Group*

As in the static model, the cumulative probability of colon cancer given JCV infection is found by summing the probabilities of developing cancer for each



```
> agenowVageinfect
          AgeInfect0   AgeInfect5  AgeInfect10  AgeInfect15  AgeInfect20  AgeInfect25  AgeInfect30  AgeInfect35  AgeInfect40  AgeInfect45  AgeInfect50  AgeInfect55  AgeInfect60  AgeInfect65
AgeNow0  0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow5  8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow10 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow15 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow20 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow25 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow30 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow35 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow40 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow45 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow50 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow55 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00 0.000000e+00
AgeNow60 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00 0.000000e+00
AgeNow65 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07 0.000000e+00
AgeNow70 2.405652e-03 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06 8.779710e-07
AgeNow75 2.957972e-03 2.405652e-03 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05 7.023611e-06
AgeNow80 3.588678e-03 2.957972e-03 2.405652e-03 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05 2.370403e-05
AgeNow85 4.302861e-03 3.588678e-03 2.957972e-03 2.405652e-03 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04 5.618534e-05
AgeNow90 5.105576e-03 4.302861e-03 3.588678e-03 2.957972e-03 2.405652e-03 1.926599e-03 1.515664e-03 1.167673e-03 8.774341e-04 6.397379e-04 4.493598e-04 3.010641e-04 1.896056e-04 1.097320e-04
          AgeInfect70  AgeInfect75  AgeInfect80 AgeInfect85 AgeInfect90
AgeNow0  0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow5  0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow10 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow15 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow20 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow25 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow30 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow35 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow40 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow45 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow50 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow55 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow60 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow65 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow70 0.000000e+00 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow75 8.779710e-07 0.000000e+00 0.000000e+00 0.00000e+00          0
AgeNow80 7.023611e-06 8.779710e-07 0.000000e+00 0.00000e+00          0
AgeNow85 2.370403e-05 7.023611e-06 8.779710e-07 0.00000e+00          0
AgeNow90 5.618534e-05 2.370403e-05 7.023611e-06 8.77971e-07          0
```

Figure A.12

age group, given all possible ages of infection (i.e. summing across the "Current Age" rows in the above matrix). This cumulative probability is first multiplied by the average incidence of JCV infection and then multiplied by 100,000, yielding the prevalence of JCV induced colon cancer per 100,000 individuals. These prevalence values are then stored in CRC[,"PrevInfectInit"]:

```
iiprev<-rep(NA,length(agenow))

for(i in 1:length(iiprev)){

        iiprev[i]<-sum(agenowVageinfect[i,])*100000*AvgJCVIncidence

        }


CRC[,"PrevInfectInit"]<-iiprev
```

A.2.9.3 *Calculating Incidence of JCV Induced Colon Cancer for Each Age Group*

Again, the prevalence data need to converted into incidence data, which was then stored in CRC[i,"Incidence_InfectInit"] and the vector "Infection_Initation":

```
#Calculate Incidence


for(i in 2:length(iiprev)){

        CRC[i,"Incidence_InfectInit"]<-CRC[i,"PrevInfectInit"]-CRC[(i-1),"
            PrevInfectInit"] }
```

```
Infection_Initation<-CRC[,"Incidence_InfectInit"]
```

A.2.10 *Results*

A.2.10.1 *Dataframe*

The CRC dataframe created after running the above code can be found in Figure A.13

> CRC

	SeerAges	CalabreseAgeRange	AverageAgeCancer	JCVPrevalence	JCVEstimatedPrev	JCVIncidence	CRC_2000_2006	CRC_2003_2007	CalabreseData	PrevMutation	Incidence_Mutation	PrevInfectInit
1	<1	0	0.0	NA	0.1318294	NA	0.00	0.0	0.00	0.000000e+00	0.000000e+00	0.000000000
2	1-4	5	2.5	0.11	0.1450000	0.01397059	0.00	0.0	0.00	0.000000e+00	0.000000e+00	0.002385011
3	5-9	10	7.5	0.14	0.1729412	0.02794118	0.00	0.0	0.00	0.000000e+00	0.000000e+00	0.021464674
4	10-14	15	12.5	0.24	0.2088824	0.02794118	0.06	0.0	0.07	0.000000e+00	0.000000e+00	0.085856758
5	15-19	20	17.5	0.22	0.2288235	0.02794118	0.26	0.2	0.18	0.000000e+00	0.000000e+00	0.238484431
6	20-24	25	22.5	NA	0.2567647	0.02794118	0.81	0.6	0.47	0.000000e+00	0.000000e+00	0.536571635
7	25-29	30	27.5	0.34	0.2847059	0.02794118	1.86	1.2	1.46	0.000000e+00	0.000000e+00	1.051635837
8	30-34	35	32.5	NA	0.3126471	0.02794118	4.06	2.7	2.82	0.000000e+00	0.000000e+00	1.869477382
9	35-39	40	37.5	0.39	0.3405882	0.02794118	7.74	4.6	5.59	0.000000e+00	0.000000e+00	3.090164715
10	40-44	45	42.5	NA	0.3685294	0.02794118	14.67	9.2	11.14	0.000000e+00	0.000000e+00	4.828014712
11	45-49	50	47.5	0.34	0.3964706	0.02794118	27.65	16.9	21.59	0.000000e+00	0.000000e+00	7.211567378
12	50-54	55	52.5	NA	0.4244118	0.02794118	53.12	33.5	42.72	0.000000e+00	0.000000e+00	10.383554172
13	55-59	60	57.5	0.45	0.4523529	0.02794118	77.97	48.7	77.94	0.000000e+00	0.000000e+00	14.580559216
14	60-64	65	62.5	NA	0.4802941	0.02794118	118.62	75.7	125.98	0.000000e+00	0.000000e+00	19.734472660
15	65-69	70	67.5	0.50	0.5082353	0.02794118	179.57	120.4	184.04	0.000000e+00	0.000000e+00	26.269435469
16	70-74	75	72.5	NA	0.5361765	0.02794118	237.26	165.6	250.96	0.000000e+00	0.000000e+00	34.384774923
17	75-79	80	77.5	NA	0.5641176	0.02794118	304.43	215.9	319.14	1.110223e-11	1.110223e-11	44.053430107
18	80-84	85	82.5	NA	0.5920588	0.02794118	363.45	267.4	387.22	1.110223e-11	0.000000e+00	55.742166703
19	85+	90	87.5	NA	0.6200000	0.02794118	391.68	288.5	NA	1.110223e-11	0.000000e+00	69.611480389

	Incidence_InfectInit
1	0.000000000
2	0.002385011
3	0.019079663
4	0.064392085
5	0.152627673
6	0.298087204
7	0.515064202
8	0.817841545
9	1.220687333
10	1.737849997
11	2.383552666
12	3.171986794
13	4.117305044
14	5.233613444
15	6.534962889
16	8.035339454
17	9.748655184
18	11.688736596
19	13.869313686

Figure A.13

A.2.10.2 *Plots*

As in the static model, the incidence values from the stepwise models were transformed to the log scale so as the emphasize the difference between the Mutation and Infection models' predictions. Again, any -Inf values were replaced with -25, for the reasons stated above. These transformations were accomplished using the following code:

```
#Adjust data so the log can be taken: if prob=0, then log(0)=-Inf, which won't fit
    on a graph. Set min to -25
```

```
logMutation<-log(Models[,"Mutation"])
logInfection<-log(Models[,"Infection_Initation"])

#Replace -Inf with -25.22388

        #Mutation
                for(i in 1:length(logMutation)){
                        ifelse(logMutation[i]==-Inf,
                        logMutation[i]<--25.22388,
                        logMutation[i]<-logMutation[i])
                        }

        #Infection
                for(i in 1:length(logInfection)){
                        ifelse(logInfection[i]==-Inf,
                        logInfection[i]<--25.22388,
                        logInfection[i]<-logInfection[i])
                        }
```

Once log-transformed, the data was plotted against the observed data using the code below. The resulting plot can be found in Figure A.14.

```
#Log Plot
xrange<-CalabreseAgeRange
ymax<-max(CRC[,"CRC_2000_2006"])
lymax<-log(ymax)
lymin<-max(log(Models[,"Mutation"]))
yrange<-seq(lymin,lymax,(lymax-lymin)/(length(xrange)-1))
plot(xrange, yrange, type="n", xlab="Current Age",ylab="ln(Incidence of Cancer, per
    100,000)",main="Predicted and Observed Incidence of Colon Cancer",sub="N=5 u=3*
    10^-8,GI=1.5")
points(AverageAgeCancer,log(CalabreseData),pch=19,lty=1)
points(AverageAgeCancer,log(CRC_2000_2006),pch=1,lty=1)
points(AverageAgeCancer,log(CRC_2003_2007),pch=2,lty=1)
lines(CalabreseAgeRange,logMutation,col="red",lty=1)
```

```
lines(CalabreseAgeRange,logInfection,col="blue",lty=2)
legend(20,-10,c("Observed Data 1992–1999","Observed Data 2000–2006","Observed Data
    2003–2007","Mutation Model,k=6","Infection Initation Model,k=3"),col=c(1,1,1,"
    red","blue"),lty=c(NA,NA,NA,1,2),pch=c(19,1,2,rep(NA,2)))
```

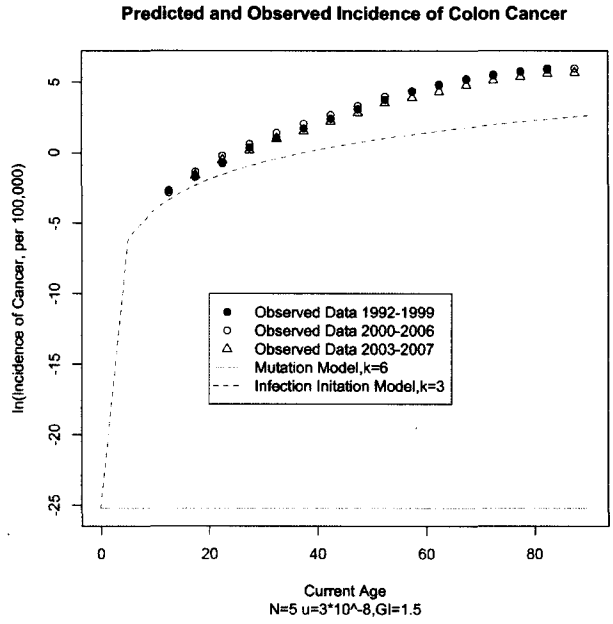


Figure A.14

Finally, infection's impact on the probability of developing cancer was calculated by dividing the probability of cancer with infection by the probability of cancer by mutation alone. A value of 1 had to be added to each probability as many of the mutation probabilities were equal to 0 (see FigureA.13). The resulting plot can be found in Figure A.15.

```
      #Plot of the Increase in Probabilty
RatioIncreaseProb<-(iiprev+1)/(mprev+1)

plot(RatioIncreaseProb~CalabreseAgeRange, main="Impact of Infection On Probability
    of Cancer", ylab= "Probability Ratio", xlab="Age", ylim=c(0,max(
    RatioIncreaseProb)))
```

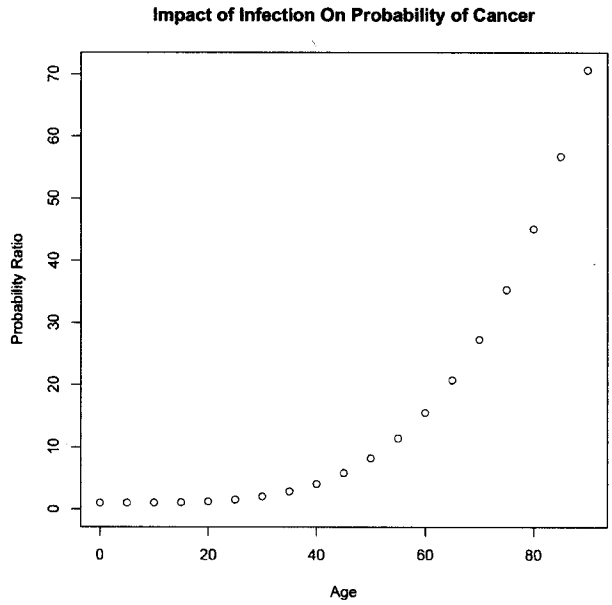

Impact of Infection On Probability of Cancer
Probability Ratio
0
10
20
30
40
50
60
70
0
20
40
60
80
Age

Figure A.15

APPENDIX B R CODE FOR GEOMETRIC MODEL

There are ten models total, each which has a different combination of genes per barrier and the stem cell mutation rate. The two stem cell mutation rates are $\mu = 10^{-10}$ and $\mu = 10^{-11}$, while there can be 3,6,9,12 or 16 genes per barrier. In this example code, there are 3 genes per barrier, and the stem cell mutation rate is $\mu = 10^{-10}$. In order to provide examples, some of the parameters are changed, primarily the number of stem cells in the colon and the number of repetitions. When conducting runs to generate the data presented in Chapter 6, $Nm = 5 \times 15000000$ and $reps = 10000$.

B.1 SETUP

The first few lines setup the parameters used in the model:

```
> genes.per.barrier <- 3
> stem.cell.mutation.rate <- 1000 * 10^-10
> genomic.instability <- 2
> u<-genes.per.barrier
> #Number of barriers in Mutation Model
> k1<-6
> #Number of barriers in Infection Model
> k2<-3
> #initial stem cell mutation rate
> u1 <- u*stem.cell.mutation.rate
> #umber of stem cells in the colon.
```

```
> Nm<- 10# Set to 5*(15000000) in model
> #5 stem cells per crypt, 15,000,000 crypts (Calabrese).
> #Estimated number of stem cells infected by JCV
> Ns=Nm*0.95
> #Redundent, but used in model
> mutincrease<-genomic.instability
> # number of times to run model. Set to 10 here for example
> #purposes, but #is set to 1000 when collecting all results
> reps<-10
> # prints when every 10th iteration is completed
> printscale<-seq(0,reps,10)
```

The next chunk of code creates a vector what the mutation rate will be after genomic instability increases the mutation rate when a barrier is removed:

```
>          GI<-c(1,rep(NA,k1-1))
> for(i in 2:length(GI)){
+ GI[i]<-GI[i-1]*mutincrease
+ }
> mutrate<-GI*u1
> mutrate

[1] 3.0e-07 6.0e-07 1.2e-06 2.4e-06 4.8e-06 9.6e-06
```

B.2 MUTATION MODEL

This part of the code creates 6 different vectors that determine how many trials will occur before the first success (days,mutation) in each of Nm cell lines. This process is repeated for all six barriers. Note that the mutation increases as each barrier is removed, thus simulating genomic instbaility .After each vector is cre-

ated, they can be added together to find the total number of divisions it takes to remove all six barriers.

After the total number division have been calculated, the mean, max, min, and standared deviation are calculated. Afterwards, the number of divisions are converted to years, assuming that stem cells divide once every four days, i.e. $\frac{\text{N divisions}}{1} \times \frac{4\text{days}}{1\text{division}} \times \frac{1\text{year}}{365\text{days}} = \text{age in years}$. The results are exported to a .csv.Note that in these examples the number of years is high. However, when there are the vector is 7.5×10^7 elements long, some values will be lower.

This procedure is repeated for each value of rep, which in this case is 10 individuals. As the data are generated they are added to a data frame.

```
> #
> #rgeom(n=Nm, prob=mutrate[i])
> #
> for(i in 1:reps){
+ #number of divisions required before first mutation.
+ N.div.for.mutaiton.r1<-rgeom(Nm,mutrate[1])
+ N.div.for.mutaiton.r1
+
+ #number of divisions required before second mutation.
+ N.div.for.mutaiton.r2<-rgeom(Nm,mutrate[2])
+ N.div.for.mutaiton.r2
+ # number of divisions required before third mutation.
+ N.div.for.mutaiton.r3<-rgeom(Nm,mutrate[3])
+ N.div.for.mutaiton.r3
+ # number of divisions required before fourth mutation.
+ N.div.for.mutaiton.r4<-rgeom(Nm,mutrate[4])
+ N.div.for.mutaiton.r4
+ # number of divisions required before fifth mutation.
+ N.div.for.mutaiton.r5<-rgeom(Nm,mutrate[5])
```

```
+ N.div.for.mutaiton.r5
+ #number of divisions required before sixth mutation.
+ N.div.for.mutaiton.r6<-rgeom(Nm,mutrate[6])
+ N.div.for.mutaiton.r6
+ #Total Number of Divisions
+ Mutation.Total.Divisions <- N.div.for.mutaiton.r1 +
+ N.div.for.mutaiton.r2 +
+ N.div.for.mutaiton.r3 + N.div.for.mutaiton.r4 +
+ N.div.for.mutaiton.r5 + N.div.for.mutaiton.r6
+ #Calculate mean,standard deviaition, min, and max
+ mean.years<-mean(Mutation.Total.Divisions)*4*(1/365)
+ stdev.years <- sd(Mutation.Total.Divisions)*4*(1/365)
+ min.years <- min(Mutation.Total.Divisions)*4*(1/365)
+ max.years <- max(Mutation.Total.Divisions)*4*(1/365)
+ #
+ Model <- "Mutation"
+ iteration <- i
+ #
+ #Build data frame
+ if (i==1){
+ #Initiate Data frame
+         Mutation.Data <- data.frame(Model,mean.years,
+                 stdev.years,min.years,max.years)
+         }
+ else{
+ #Add to data frame
+ New.Data <-  data.frame(Model,mean.years,
+                 stdev.years,min.years,max.years)
+ Mutation.Data <- rbind(Mutation.Data,New.Data)
```

```
+         }
+ }
> Mutation.Data
```

	Model	mean.years	stdev.years	min.years	max.years
1	Mutation	58943.96	28289.22	33575.781	120091.5
2	Mutation	98563.89	61002.97	23388.011	208677.3
3	Mutation	54854.77	37885.94	16042.586	154127.4
4	Mutation	61352.80	34357.45	7745.896	136633.1
5	Mutation	73895.62	36610.36	12873.534	119736.3
6	Mutation	86052.36	35054.35	42257.238	145139.3
7	Mutation	63963.59	28471.03	32581.655	119885.1
8	Mutation	66089.51	36094.76	25705.140	131563.8
9	Mutation	67877.19	40096.74	14720.285	126313.1
10	Mutation	83613.25	39647.85	13310.718	140384.5

B.3 INFECTION MODEL

The infection model is implemented in much the same way as the mutation mode. One difference is that the population of cells is 95% of the cell population in the mutation model. A second difference is the the model starts using the fourth mutation rate because the three previous barriers have all ready been removed. Finally, the model

```
> for(i in 1:reps){
+ # number of divisions required before first mutation in 4 barrier
+ Infection.N.div.for.mutaiton.r4<-rgeom(Ns,mutrate[4])
+ # number of divisions required before first mutation in 5th barrier
+ Infection.N.div.for.mutaiton.r5<-rgeom(Ns,mutrate[5])
+ # number of divisions required before first mutation in 5th barrier
```

```
+ Infection.N.div.for.mutaiton.r6<-rgeom(Ns,mutrate[6])
+ #Total Number of Divisions
+ Infection.Total.Divisions  <- Infection.N.div.for.mutaiton.r4 +
+ Infection.N.div.for.mutaiton.r5 + Infection.N.div.for.mutaiton.r6
+ Model <- "Infection"
+ iteration <- i
+ mean.years<-mean(Infection.Total.Divisions)*4*(1/365)
+ stdev.years <- sd(Infection.Total.Divisions)*4*(1/365)
+ min.years <- min(Infection.Total.Divisions)*4*(1/365)
+ max.years <- max(Infection.Total.Divisions)*4*(1/365)
+ if (i==1){
+         #Initiate Data frame
+         Infection.Data <- data.frame(Model,mean.years,
+                 stdev.years,min.years,max.years)
+ }
+ else{
+         #Add to data frame
+         New.Data <-  data.frame(Model,mean.years,
+                 stdev.years,min.years,max.years)
+ Infection.Data <- rbind(Infection.Data,New.Data)
+ }
+ }
> Infection.Data
```

	Model	mean.years	stdev.years	min.years	max.years
1	Infection	8547.683	3373.122	4925.808	14949.26
2	Infection	7196.665	4658.234	1640.921	14864.99
3	Infection	10975.610	7244.343	3424.252	23861.61
4	Infection	7297.763	4685.403	3631.321	19027.34
5	Infection	7867.534	5550.063	2564.679	17520.11

```
6  Infection   6748.620    3537.967  1665.644  12705.71
7  Infection   8132.100    6050.912  1092.427  18647.27
8  Infection   5970.298    4333.681  1669.677  13285.05
9  Infection   9644.497    7154.300  1686.542  24798.86
10 Infection   6233.749    3317.813  2677.786  12019.61
```

Like the Mutation model, the mean, standard deviation, min, and max are calculated and converted to age in years

B.4 BINNING DATA

The observed data are in groups of 5 years, so the results of the Geometric model also need to binned in to 5 year age groups. This is done by creating a function that bins the data and converts to incidence per 100,000 individuals.

```
> bin.data <- function(summary.data,age.of.metastatic.tumors){
+ indv <-  length(summary.data[,1])
+ ages1<- (length (subset(age.of.metastatic.tumors,
+         age.of.metastatic.tumors < 1 &
+         age.of.metastatic.tumors  > 0)))/indv*100000
+ ages4<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 5 &
+         age.of.metastatic.tumors  >= 1)))/indv*100000
+ ages9<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 10 &
+         age.of.metastatic.tumors  >= 5)) )/indv*100000
+ ages14<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 15
+         & age.of.metastatic.tumors  >= 10)))/indv*100000
+ ages19<- (length (subset(age.of.metastatic.tumors ,
```

```
+         age.of.metastatic.tumors  < 20 &
+         age.of.metastatic.tumors  >= 15)))/indv*100000
+ ages24<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 25 &
+         age.of.metastatic.tumors  >= 20)))/indv*100000
+ ages29<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 30 &
+         age.of.metastatic.tumors  >= 25)))/indv*100000
+ ages34<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 35 &
+         age.of.metastatic.tumors  >= 30)))/indv*100000
+ ages39<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 40 &
+         age.of.metastatic.tumors  >= 35)))/indv*100000
+ ages44<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 45 &
+         age.of.metastatic.tumors  >= 40)))/indv*100000
+ ages49<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 50 &
+         age.of.metastatic.tumors  >= 45)))/indv*100000
+ ages54<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 55 &
+         age.of.metastatic.tumors  >= 50)))/indv*100000
+ ages59<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 60 &
+         age.of.metastatic.tumors  >= 55)))/indv*100000
+ ages64<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 65 &
+         age.of.metastatic.tumors  >= 60)))/indv*100000
```

```
+ ages69<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 70 &
+         age.of.metastatic.tumors  >= 65)))/indv*100000
+ ages74<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 75 &
+         age.of.metastatic.tumors  >= 70)))/indv*100000
+ ages79<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 80 &
+         age.of.metastatic.tumors  >= 75)))/indv*100000
+ ages84<- (length (subset(age.of.metastatic.tumors ,
+         age.of.metastatic.tumors  < 85 &
+         age.of.metastatic.tumors  >= 80)))/indv*100000
+ ages85up<- (length (subset(age.of.metastatic.tumors ,
+         #age.of.metastatic.tumors  <=100 &
+         age.of.metastatic.tumors  >= 85)))/indv*100000
+ #
+ d.f <- c(ages1, ages4, ages9, ages14,
+ ages19, ages24, ages29, ages34, ages39,
+ ages44, ages49, ages54, ages59, ages64,
+ ages69, ages74, ages79, ages84,ages85up)
+ #
+ return(d.f)
+ }
```

The bin.data function is applied to both the infection results to put them into an age group.

```
> options(width=60)
> Mutation.Min.Results <- Mutation.Data$min.years
> Mutation.Binned <- bin.data(Mutation.Data,Mutation.Min.Results)
> Mutation.Binned
```

AGE	MORTALITY
< 20	0%
$20 - 34$	0.6%
$35 - 44$	2.5%
$45 - 54$	8.6%
$55 - 64$	16.5%
$65 - 74$	22%
$75 - 84$	29%
85+	20.8%

Table B.1: Mortality From Colorectal Cancer, $2005 - 2009$ [2]

```
 [1] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
[10] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
[19] 1e+05

> #
> Infection.Min.Results <- Infection.Data$min.years
> Infection.Binned<- bin.data(Infection.Data,Infection.Min.Results)
> Infection.Binned

 [1] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
[10] 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00 0e+00
[19] 1e+05
```

B.5 CONVERTING INCIDENCE TO PREVALENCE

Since each repetition is one individual, the results are incidence of colorectal cancer. The prevalence of in each age can be determined by summing up the incidence values of earlier ages. However, some people die before they move to the next age group, so mortality from colorectal cancer is worked into the calculation of prevalence by multiplying the prevalence value by the number of people that did survive colorectal cancer. The mortality rates used in this procedure are found in Table B.1.

```
> Calculate.Prevalence.With.Mortality <- function(binned.data){
+ ages1<- binned.data[1]  #ages < 1
+ total <- ages1
+ ages4<- binned.data[2]+total #ages 1-4
+ total <- ages4
+ ages9 <- binned.data[3]+total #ages 5-9
+ total <- ages9
+ ages14 <- binned.data[4]+total #ages 10-14
+ total <- ages14
+ ages19 <- binned.data[5]+total #ages 15-19
+ total <- ages19
+ #
+ #Mortality in 20-34=0.6%
+ ages24 <- binned.data[6]*(1-0.006)+total #ages 20-24
+ total <- ages24
+ ages29 <- binned.data[7]*(1-0.006)+total #ages 25-29
+ total <- ages29
+ ages34 <- binned.data[8]*(1-0.006)+total #ages 30-34
+ total <- ages34
+ #
+      #Mortality in 35-44=0 2.5%
+ ages39 <- binned.data[9]*(1-0.025)+total #ages 35-39
+ total <- ages39
+ ages44 <- binned.data[10]*(1-0.025)+total
+ #ages 40-44
+ total <- ages44
+ #
+ #Mortality in 45-54=0 8.6%
+ ages49 <- binned.data[11]*(1-0.086)+total #ages 45-49
```

```
+ total <- ages49
+ ages54 <- binned.data[12]*(1-0.086)+total #ages 50-54
+ total <- ages54
+ #
+ #Mortality in 55-64=16.5%
+ ages59 <- binned.data[13]*(1-0.165)+total #ages 55-59
+ total <- ages59
+ ages64 <- binned.data[14]*(1-0.165)+total #ages 60-64
+ total <- ages64
+ #
+ #Mortality in 65-74=22%
+ ages69 <- binned.data[15]*(1-0.22)+total #ages 65-69
+ total <- ages69
+ ages74 <- binned.data[16]*(1-0.22)+total #ages 70-74
+ total <- ages74
+ #
+ #Mortality in 75-84=29%
+ ages79 <- binned.data[17]*(1-0.29)+total #ages 75-80
+ total <- ages79
+ ages84 <- binned.data[18]*(1-0.29)+total #ages 80-85
+ total <- ages84
+ #
+ #Mortality in 85+ =20.8%
+ ages85up <- binned.data[19]*(1-0.208)+total #ages 85+
+ #
+   d.f <- c(ages1, ages4, ages9, ages14, ages19,
+ ages24, ages29, ages34, ages39, ages44, ages49,
+ ages54, ages59, ages64, ages69, ages74, ages79,
+ ages84,ages85up)
```

```
+
+ return(d.f) }
```

This prevalence function can then be applied to the results generated by the Geometric model

```
> options(width=60)
> Mutation.Prevalence <- Calculate.Prevalence.With.Mortality(Mutation.Binned)
> Mutation.Prevalence

 [1]     0     0     0     0     0     0     0     0     0
[10]     0     0     0     0     0     0     0     0     0
[19] 79200

> Infection.Prevalence <- Calculate.Prevalence.With.Mortality(Infection.Binned)
> Infection.Prevalence

 [1]     0     0     0     0     0     0     0     0     0
[10]     0     0     0     0     0     0     0     0     0
[19] 79200
```

B.6 BUILD FINAL DATA FRAME

The prevalence values calculated above can be added to a summary data frame that includes both observed values and modeled values.

```
> Seer.Age.Group <- c("<1","1-4","5-9","10-14","15-19","20-24","25-29",
+ "30-34","35-39","40-44","45-49","50-54","55-59","60-64","65-69",
+ "70-74","75-79","80-84","85+")
> Age.Group<-seq(0,90,5)
> #
> #Observed Data From SEER 1992-1999
```

```
> #Build Data Frame
> Observed.Data<-c(0,0,0,0.07,0.18,0.47,1.46,2.82,5.59,11.14,21.59,
+ 42.72,77.94,125.98,184.04,250.96,319.14,387.22,NA)
> Model <- rep("Observed",length(Observed.Data))
> Prevalence <-Calculate.Prevalence.With.Mortality(Observed.Data)
> Obs.Data<-data.frame(Model,Seer.Age.Group,Age.Group,Prevalence)
> Data<- Obs.Data
> #
> #Mutation Data
> Model <- rep("Mutation",length(Observed.Data))
> Prevalence <-Mutation.Prevalence
> Mutation.Data<- data.frame(Model,Seer.Age.Group,Age.Group,Prevalence)
> Data<-rbind(Data,Mutation.Data)
> #
> #Infection Data
> Model <- rep("Infection",length(Observed.Data))
> Prevalence <-Infection.Prevalence
> Infection.Data<- data.frame(Model,Seer.Age.Group,Age.Group,Prevalence)
> Data<-rbind(Data,Infection.Data)
> Data
```

	Model	Seer.Age.Group	Age.Group	Prevalence
1	Observed	<1	0	0.00000
2	Observed	1-4	5	0.00000
3	Observed	5-9	10	0.00000
4	Observed	10-14	15	0.07000
5	Observed	15-19	20	0.25000
6	Observed	20-24	25	0.71718
7	Observed	25-29	30	2.16842
8	Observed	30-34	35	4.97150

9	Observed	35-39	40	10.42175
10	Observed	40-44	45	21.28325
11	Observed	45-49	50	41.01651
12	Observed	50-54	55	80.06259
13	Observed	55-59	60	145.14249
14	Observed	60-64	65	250.33579
15	Observed	65-69	70	393.88699
16	Observed	70-74	75	589.63579
17	Observed	75-79	80	816.22519
18	Observed	80-84	85	1091.15139
19	Observed	85+	90	NA
20	Mutation	<1	0	0.00000
21	Mutation	1-4	5	0.00000
22	Mutation	5-9	10	0.00000
23	Mutation	10-14	15	0.00000
24	Mutation	15-19	20	0.00000
25	Mutation	20-24	25	0.00000
26	Mutation	25-29	30	0.00000
27	Mutation	30-34	35	0.00000
28	Mutation	35-39	40	0.00000
29	Mutation	40-44	45	0.00000
30	Mutation	45-49	50	0.00000
31	Mutation	50-54	55	0.00000
32	Mutation	55-59	60	0.00000
33	Mutation	60-64	65	0.00000
34	Mutation	65-69	70	0.00000
35	Mutation	70-74	75	0.00000
36	Mutation	75-79	80	0.00000
37	Mutation	80-84	85	0.00000

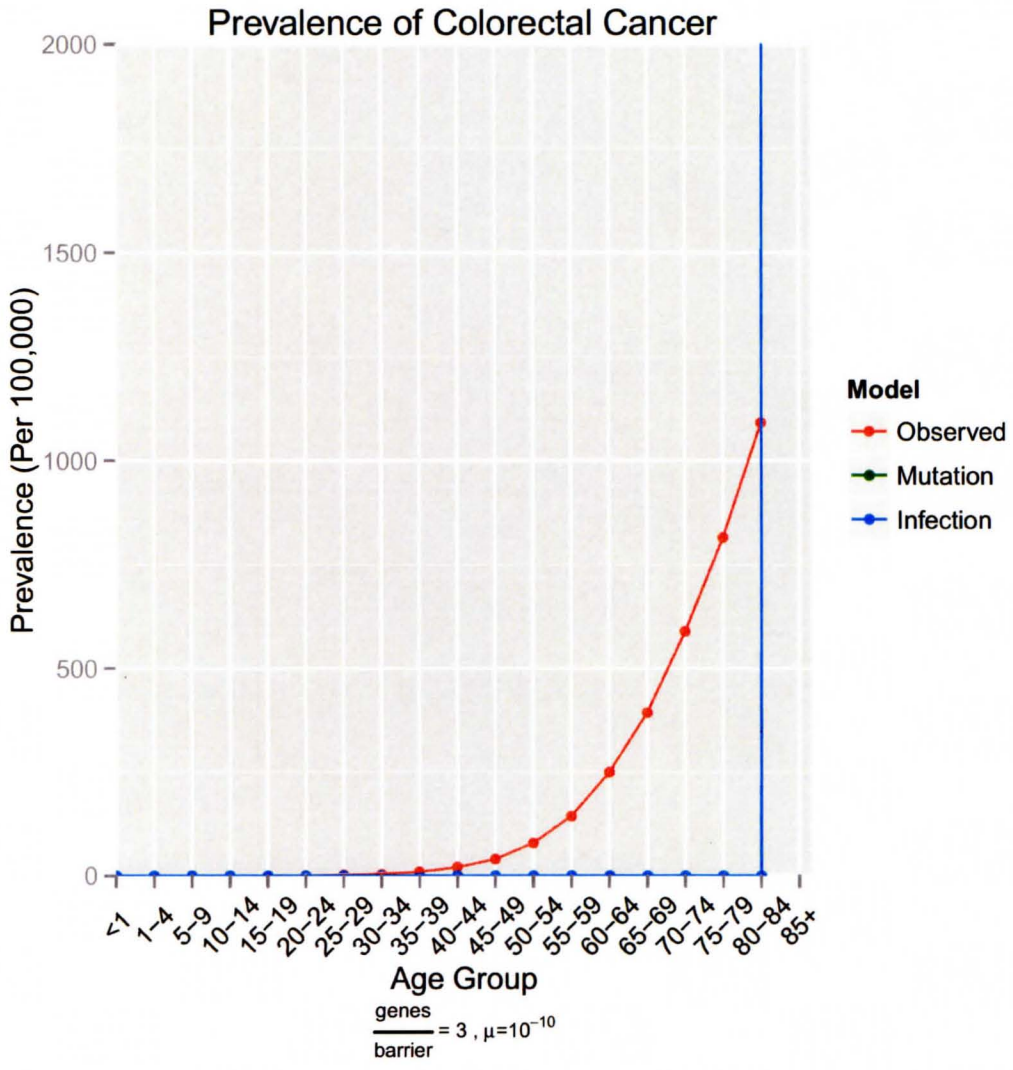
```
38  Mutation          85+       90 79200.00000
39 Infection          <1         0     0.00000
40 Infection          1-4        5     0.00000
41 Infection          5-9       10     0.00000
42 Infection        10-14       15     0.00000
43 Infection        15-19       20     0.00000
44 Infection        20-24       25     0.00000
45 Infection        25-29       30     0.00000
46 Infection        30-34       35     0.00000
47 Infection        35-39       40     0.00000
48 Infection        40-44       45     0.00000
49 Infection        45-49       50     0.00000
50 Infection        50-54       55     0.00000
51 Infection        55-59       60     0.00000
52 Infection        60-64       65     0.00000
53 Infection        65-69       70     0.00000
54 Infection        70-74       75     0.00000
55 Infection        75-79       80     0.00000
56 Infection        80-84       85     0.00000
57 Infection          85+       90 79200.00000
```

B.7 PREVALENCE PLOT

The modeled and observed prevalence values can be plotted using ggplot2.

```
> library(ggplot2)
> library(gridExtra)
> library(scales)
```

```
> Prevalence.Plot <- ggplot(data=Data,aes(x=Age.Group,
+         y=Prevalence,group=Model,colour=Model)) +
+ geom_point() +
+ geom_line(aes(x=Age.Group,y=Prevalence)) +
+ coord_cartesian(ylim=c(0, 2000))+
+ scale_x_discrete(breaks=Age.Group,
+         labels= Seer.Age.Group,name="Age Group") +
+ scale_y_continuous(name="Prevalence (Per 100,000)") +
+ opts(title="Prevalence of Colorectal Cancer",
+         axis.text.x = theme_text(angle = 45))
> grid.arrange(Prevalence.Plot, sub = textGrob(
+         expression(paste(frac(genes,barrier)," = 3 , ",mu, "=",10^{-10})),
+         hjust = 0.87,vjust=0.3, gp = gpar(cex = 0.7))
+         )
> Prevalence.Plot
```



B.8 EUCLIDIAN DISTANCE AND PLOT

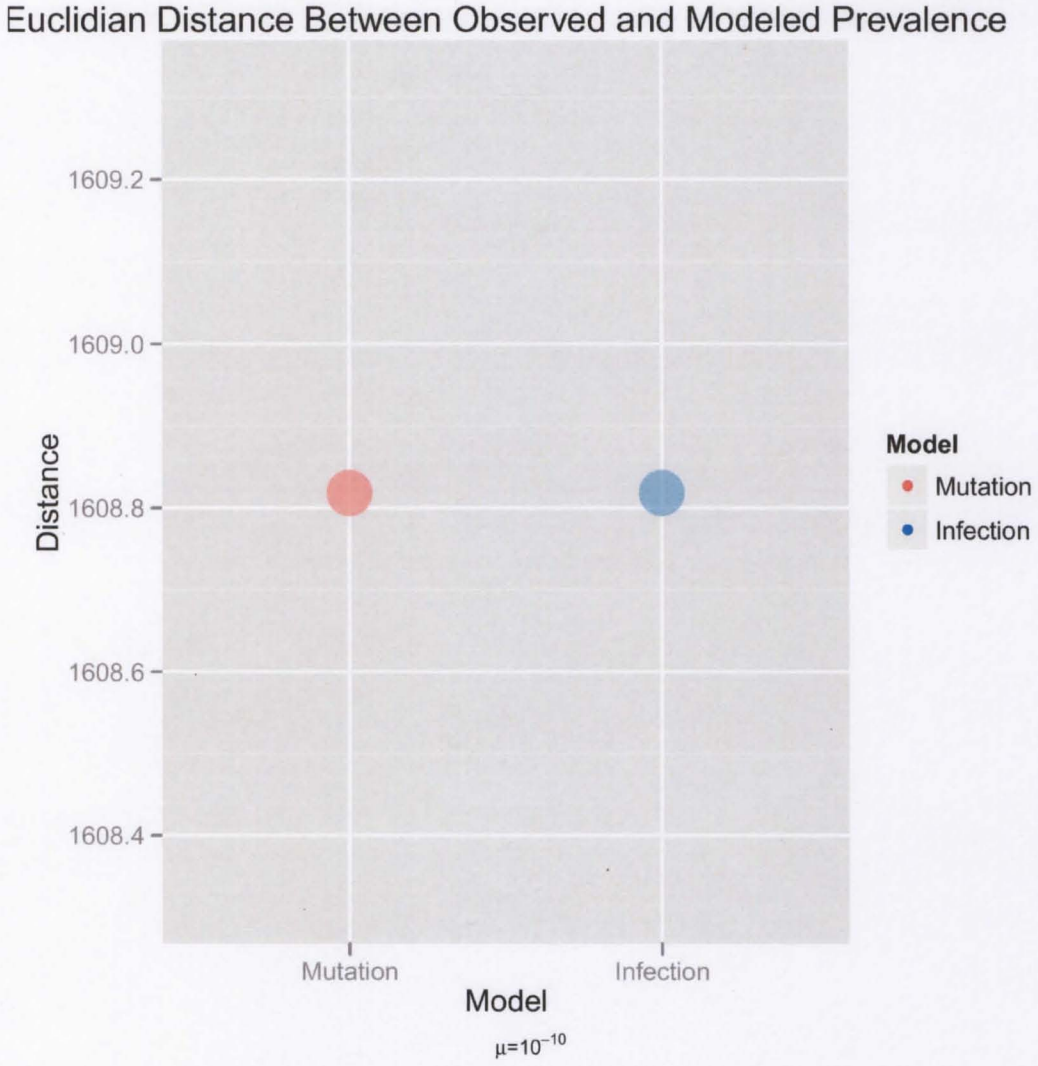
The distance between each modeled prevalence point and the observed prevalence point can be calculated using **R** *dist* function. However, to make sure the distance is point to point, the data must be compared as a row

```
> Mut.V.Obs<-rbind(Mutation.Data$Prevalence,Obs.Data$Prevalence)
> Mut.V.Obs
```

```
     [,1] [,2] [,3] [,4] [,5]    [,6]    [,7]   [,8]
[1,]    0    0    0 0.00 0.00 0.00000 0.00000 0.0000
[2,]    0    0    0 0.07 0.25 0.71718 2.16842 4.9715
         [,9]    [,10]    [,11]    [,12]    [,13]    [,14]
[1,]  0.00000  0.00000  0.00000  0.00000   0.0000   0.0000
[2,] 10.42175 21.28325 41.01651 80.06259 145.1425 250.3358
       [,15]    [,16]    [,17]    [,18] [,19]
[1,]   0.000   0.0000   0.0000    0.000 79200
[2,] 393.887 589.6358 816.2252 1091.151    NA
> Mut.Dist<-dist(Mut.V.Obs)
> Mut.Dist
         1
2 1608.818
> EDistance <- Mut.Dist[1]
> EDistance
[1] 1608.818
> Model<-"Mutation"
> Dist.Data<-data.frame(Model,EDistance)
> #Infection Data
> Inf.V.Obs<-rbind(Infection.Data$Prevalence,Obs.Data$Prevalence)
> EDistance<-dist(Inf.V.Obs)[1]
> Model<-"Infection"
> New.Dist.Data<-data.frame(Model,EDistance)
> Dist.Data<-rbind(Dist.Data,New.Dist.Data)
> Dist.Data
      Model EDistance
1  Mutation  1608.818
2 Infection  1608.818
```

Once the distance has been calculated it can be plotted to get a sense of how well each model replicates the observed data.

```
> Dist.Plot <- ggplot(Dist.Data,
+         aes(x=Model,y=EDistance,color=Model,size=100,alpha=0.5))+
+ geom_point()+
+ scale_size(range = c(2, 15))+
+ guides(size=FALSE,alpha=FALSE)+
+ coord_trans(y="log2")+
+ opts(title="Euclidian Distance Between Observed and Modeled Prevalence")+
+ ylab("Distance")
> grid.arrange(Dist.Plot,
+         sub=textGrob(expression(paste(mu, "=",10^{-10})),
+         hjust = 0.87,vjust=0.3, gp = gpar(cex = 0.7)))
> Dist.Plot
```


Euclidian Distance Between Observed and Modeled Prevalence
1609.2
1609.0
1608.8
1608.6
1608.4
Distance
Mutation
Infection
Model
μ=10^−10
Model
Mutation
Infection

APPENDIX C ABM ODD

C.1 OVERVIEW

C.1.1 *Purpose*

This model was designed to better understand the roles of mutation and infection in the development of colon cancer. Is mutation alone sufficient to generate the emergence of metastatic tumors, or is infection needed to "kick-start" the process of oncogenesis? How does the order of mutations affect the probability of developing a metastatic tumor? Does this order change if infection is present? If infection does turn out to play an important role, how much does the age of infection by JCV affect the probability of developing a metastatic tumor?

C.1.2 *Entities, State Variables, and Scales*

The model has six entities: patches, stem cells, metastatic stem cells, transit cells, VEGF molecules, and vessels.

There are 3721 patches laid out in a grid, creating a non-wrapping square world that is 61x61 patches, centered around patch $(0, 0)$. Each patch is meant to represent the cells underlying stem and transit cells of the colon crypt. This world is subdivided into a colon crypt and metastatic tissue. The colon crypt is centered around patch $(-17, -17)$, and is 25x25 patches, creating a total of 625 patches. The crypt is further subdivided into the inner and outer crypt. The inner crypt, also centered around patch $(-17, -17)$, is 15x15 patches, and is colored yellow. The

outer crypt is composed of the remaining crypt patches and is colored pink. The metastatic tissue is centered around patch (13, 13), and is 30x30 patches, for a total of 900 patches. All remaining patches are colored black. The state variables of each patch can be found in Table C.1 on page 204.

PATCH STATE VARIABLES
oxygen
infected

Table C.1: Patch State Variables

Stem cells are located in a circular layout around the center of the colon crypt, and face outwards. Stem cells are responsible for producing the initial transit cells in the model. Each stem cell has several state variables, which are listed in Table C.3 on page 205

STEM CELL STATE VARIABLES	
count-stem-cell-divisions	CIF-II-removed?
stem-cell-mutation-rate	count-barriers-removed
count-neutral-mutations	order-of-barrier-removal (list)
count-deleterious-mutations	age-of-barrier-removal (list)
count-beneficial-mutations	infected?
count-pro-growth-mutations	stem-cell-age-of-infection
count-anti-growth-mutations	transformed?
count-anti-apoptosis-mutations	completely-transformed?
count-metastasis-mutations	age-of-transformation
count-CIF-II-mutations	successful-invasion?

count-telomerase-mutations	parent
order-of-mutations (list)	most-recent-event
age-of-mutation (list)	tissue
pro-growth-removed?	count-JCV-NCRR-mutations
anti-growth-removed?	JCV-Mad-1-d98-survival
anti-apoptosis-removed?	JCV-Mad-1-d98-evolved?
metastasis-removed?	JCV-Mad-1-d98-days-active
telomerase-removed?	JCV-Mad-1-d98-deactivated?

Table C.3: Stem-Cell-State-Variables

Metastatic stem cells have the same state variables as regular stem cells (see Table C.3 on page 205); they simply have the ability to survive in the metastatic tissue if they successfully invade.

Transit cells have many of the same state variables as stem cells, except that track telomere length, telomerase mutations, oxygen levels, and hypoxic state. See Table C.5 on page 206 for a complete list of the transit cell state variables.

TRANSIT CELL STATE VARIABLES	
count-transit-cell-divisions	count-barriers-removed
transit-cell-mutation-rate	age-of-barrier-removal (list)
count-neutral-mutations	infected?

count-deleterious-mutations	transit-cell-age-of-infection
count-beneficial-mutations	transformed?
count-pro-growth-mutations	completely-transformed?
count-anti-growth-mutations	age-of-transformation
count-anti-apoptosis-mutations	successful-invasion?
count-metastasis-mutations	oxygen
count-telomerase-mutations	metastatic?
count-CIF-II-mutations	telomere-length
order-of-mutations (list)	parent
age-of-mutation (list)	most-recent-event
pro-growth-removed?	tissue
anti-growth-removed?	count-JCV-NCRR-mutations
anti-apoptosis-removed?	JCV-Mad-1-d98-survival
metastasis-removed?	JCV-Mad-1-d98-evolved
telomerase-removed?	JCV-Mad-1-d98-days-active
CIF-II-removed?	JCV-Mad-1-d98-deactivated

Table C.5: Transit Cell State Variables

VEGF molecules do not own any state variables, but the vessels they are attracted to do, and can be found in Table C.6 on page 207.

VESSEL STATE VARIABLES
lifespan
time-alive

Table C.6: Vessel State Variables

Finally, there are many global variables, which can be found in Table C.8 on page 219.

In this model, one time step is equal to one day.

C.1.3 *Process Overview and Scheduling*

The schedule of this ABM is as follows:

1. determine-if-simulation-should-end (global procedure)
2. tick (global procedure)
3. determine-if-mutant-phenotype-on
4. determine-location
5. infection-modeled? = true [infection] (occurs only if infection is being modeled)
6. hit-and-run? (if true, the "Hit and Run" infection model is run)
7. stem-cell-determine-division-type (symmetric or asymmetric; a global procedure)
8. stem-cell-replace-metastatic-stem-cell
9. metastatic-stem-cell-fill-tissue
10. metastatic-stem-cell-determine-division-type (symmetric or asymmetric; a global procedure)

11. transit-cell-consume-oxygen
12. transit-cell-division-mutation-and-movement (transit cell procedure)
13. angiogenesis (turtle procedure)
14. oxygen-replenish (patch procedure)
15. oxygen-diffusion (global procedure)
16. oxygen-recolor-patches (patch procedure)
17. transit-cell-death (turtle procedure)
18. maintain-population-cap (turtle procedure)
19. evaluate-state-and-record-data (global procedure)

This schedule is repeated until one of two events occur: 1) the crypt reaches 100 years, or 2) a metastatic tumor forms and at least one cell in that tumor has completely removed all of the cancer barriers.

The details of each sub-model can be found in Section C.3.3

C.2 DESIGN CONCEPTS

C.2.1 *Basic Principles*

The basic principle of this model is the formation of metastatic tumors, and the roles that mutation and infection play in this process. It is commonly accepted that stem cell mutations play a critical role in the development of colon cancer Michor et al. [85], Boman and Huang [11], but there is also experimental evidence indicating that infection by JC Virus may also play a role in tumor formation [35, 69, 82, 89, 24]. Thus, this model aims to determine the roles that mutation and infection play in the development of colon cancer. It is hoped this model can answer this question by determining how many crypts develop metastatic

tumors by mutation alone, and how many develop metastatic tumors with infection and mutation.

It has also been argued that a particular order of mutations is most likely to cause colon cancer [85, 39], although this is not universally accepted [132]. This model will also look for a relationship between the order of mutations and timing of tumor development and tumor size. Furthermore, as infection is being considered, it seems worthwhile to determine if there is any relationship between the age of infection and the age of tumor formation, and the strength of any such relationship.

C.2.2 *Emergence*

The accumulation of mutations and presence of infection can lead to changes in cellular behavior that result in the emergence of metastatic tumors.

C.2.3 *Adaptation*

In this model, cellular behavior changes to match patterns observed in cancers (i.e. the agents exhibit indirect-objective-seeking behavior). Behaviors change whenever there is a mutation in a beneficial gene belonging to one of six categories: pro-growth genes; anti-growth genes; anti-apoptosis genes; metastasis genes; telomerase genes; CIF-II pathway genes.

The effect of disrupting each pathway was hypothesized using the descriptions provided in [50, 153, 67].

Mutations in pro-growth genes allow cells to divide every time step, so long as they are in the inner-crypt and their telomeres have not deteriorated.

Mutations in anti-growth genes allow transit cells to divide in the outer crypt.

Mutations in anti-apoptosis genes allow cells to avoid death by accumulation of deleterious mutations. For stem cells, this means that these cells are never lost

during symmetric division. For transit cells, this means that these cells are never killed when the population reaches its maximum size (*normal-max-pop-size*). For both cell types, any anti-apoptotic mutations also increases the mutation rate via genomic instability.

A single mutation in a metastasis gene gives stem and transit cells the ability to detect the neighbor patch with the most oxygen, and then move to that patch . If angiogenesis has occurred, and a cell has two metastatic mutations and is close to a blood vessel (produced by angiogenesis) the cell will try to invade the metastatic tissue. However, successful invasion is not certain, and is controlled by the parameter *probability-of-successful-invasion*.

Mutations in the telomerase gene prevent the degradation of telomeres, essentially giving transit cells the ability to divide without limit.

If a cell is infected by JCV, and has acquired mutations in CIF-II, then JCV will remove the pro-growth, anti-growth, and apoptosis barriers. In the "Full Transformation" version of the model, JCV completely removes each of those barriers, while in the "Partial Transformation" version JCV only partially removes each of those barriers. Thus, mutations in the CIF-II pathway are required for transformation by JCV.

C.2.4 *Objectives*

As the cells use indirect-objective seeking behavior, there are no specific agent objectives.

C.2.5 *Learning*

No agents in this model learn from past experiences.

C.2.6 *Prediction*

No agents in this model make predictions.

C.2.7 *Sensing*

Un-mutated cells must decide if they can divide. They do so by determining if there is an unoccupied patch within a cone with an angle of 180° and radius of 1.5 pathes. If there is such an empty patch, the cell will divide and move to the empty patch; if not, the cell will remain where it is and not divide.

During angiogenesis, VEGF molecules sense the closest vessels and move towards them.

C.2.8 *Interaction*

During asymmetric division, stem cells produce daughter transit cells, who inherit their parent's mutations. Likewise, daughter transit cells inherit their parent transit cell's mutations.

During angiogenesis, VEGF molecules are produced by hypoxic patches, and those VEGF molecules are attracted to the closest vessel.

C.2.9 *Stochasticity*

The order in which cells are chosen to move/divide is decided randomly.

The number of mutations per division is modeled by choosing a Poisson random number, with an appropriate mean (see C.3.3.6 for a detailed description)

Determining whether stem cell division will be symmetric or asymmetric is decided randomly (see C.3.3.6 for details)

The direction that stem cells face each run is chosen randomly, so that there is no bias on a particular stem cell orientation (i.e. those that directly face a corner could divide more frequently).

C.2.10 *Collectives*

There are three primary collectives in this model: stem cells, metastatic stem cells, and transit cells. Stem cells usually produce transit cells (during asymmetric division), but occasionally produce a stem cell (during symmetric division). Metastatic stem cells also arise from stem cells that accumulate sufficient metastatic mutations. Finally, transit cells produce other transit cells.

C.2.11 *Observation*

This model tracks the population size of the colon crypt and metastatic tissue so as to determine the presence and size of polyps and tumors, and the age at which they occur. This model also tracks the order and timing of each mutation in each cell, as well as the order and timing of when cancer barriers are removed in each cell. From this one can determine if a certain order of mutations leads to more rapid tumor development. Finally, the age of infection, whether the infection has become chronic, and time to transformation are tracked so as to help determine the role of infection.

The results of experiments will be recorded in .csv files. At the end of each run, the following data will be recorded in a row in the experiment's output file:

- date-and-time
- run-number
- infection-modeled?
- mutations-added-by-transformation

- age
- age-when-infection-became-chronic
- metastatic-tumor-formed?
- colon-tumor-formed?
- count-crypt-transit-cells
- mean-transit-cells
- mean-stem-cell-div-rate
- count-transformed-cells
- age-of-colon-tumor-formation
- age-of-metastasis
- count-metastatic-cells
- pro-growth-spread?
- anti-growth-spread?
- anti-apoptosis-spread?
- telomerase-spread?
- metastasis-spread?
- cif-ii-spread?
- m.tumor.cell.type
- m.tumor.cell.number
- m.tumor.cell.parent
- m.tumor.cell.mutations

- m.tumor.cell.mutations.timing
- m.tumor.cell.barriers.removed
- m.tumor.cell.barriers.timing
- stem-cell-with-most-mutations
- count-mutations-of-most-fit-stem-cell
- mutations-of-most-fit-stem-cell
- timing-mutations-of-most-fit-stem-cell
- count-barriers-of-most-fit-stem-cell
- barriers-of-most-fit-stem-cell
- timing-barriers-of-most-fit-stem-cell
- metastatic-stem-cell-with-most-mutations
- count-mutations-of-most-fit-metastatic-stem-cell
- mutations-of-most-fit-metastatic-stem-cell
- timing-mutations-of-most-fit-metastatic-stem-cell
- count-barriers-of-most-fit-metastatic-stem-cell
- barriers-of-most-fit-metastatic-stem-cell
- timing-barriers-of-most-fit-metastatic-stem-cell

In addition to this data, more detailed data on each run will be also be collected, in the event that more resolution is needed (see C.3.3.18 for details.)

C.3 DETAILS

C.3.1 *Initialization*

The model is initialized by creating the world that is 61x61 patches, and centered around patch $(0, 0)$. Each patch is meant to represent the cells underlying stem and transit cells of body. This world is subdivided into a colon crypt and metastatic tissue. The colon crypt is centered around patch $(-17, -17)$, and is 25x25 patches, creating a total of 625 patches. The crypt is further subdivided into the inner and outer crypt. The inner crypt, also centered around patch $(-17, -17)$, is 15x15 patches, and is colored yellow. The outer crypt is composed of the remaining crypt patches and is colored pink. The metastatic tissue is centered around patch $(13, 13)$, and is 30x30 patches, for a total of 900 patches. All remaining patches are colored black.

Five stem cells are created around the the center of the colon crypt, and laid out in a circle. Each each stem cell faces outward, but their position in the circle is chosen randomly during the setup procedure. Below is a table of the initial global parameter values

MONITORS

PARAMETER	VALUE	SOURCE
mean-transit-cells	0	
mean-oxygen-per-patch-in-crypt	0	
mean-stem-cell-div-rate	0	
count-metastatic-cells	0	
age-of-metastasis	0	
metastatic-tumor-formed?	false	
colon-tumor-formed?	false	

count-transformed-cells	0
count-metastatic-cells	0
age-of-colon-tumor-formation	0
pro-growth-spread?	false
anti-growth-spread?	false
anti-apoptosis-spread?	false
telomerase-spread?	false
metastasis-spread?	false
cif-ii-spread?	false

CRYPT

PARAMETER	VALUE	SOURCE
crypt-center-x	-17	
crypt-center-y	-17	
outer-crypt-width	25	
inner-crypt-width	15	
number-of-stem-cells	5	[12]
number-of-metastatic-stem-cells	0	
number-of-transit-cells	0	
normal-max-pop-size	300	[12]
max-pop-size	5000	
probability-of-asymmetric-stem-cell-division	0.95	[79]
metastatic-cell-type		
initial-telomere-length	9	

MUTATION

PARAMETER	VALUE	SOURCE
initial-stem-cell-mutation-rate	10^{-9}	
initial-transit-cell-mutation-rate	10^{-8}	
ratio-transit-mutation-to-stem-mutation	10	[19, 43, 16]
candidate-genes-per-barrier	variable	
mutations-needed-to-remove-each-barrier	2	[143]

GENOME

PARAMETER	VALUE	SOURCE
genomic-instability	2	[143]
genome-length	7×10^9bp	[92]
genes-per-genome	70,000	[92]
average-gene-length	1000bp	[43, 17]
count-netural-genes-bp	varies	see C.3.3.6
count-genomic-instability-bp	varies	see C.3.3.6
count-beneficial-genes-bp	varies	see C.3.3.6
count-deleterious-genes-bp	varies	

METASTATIC TISSUE

PARAMETER	VALUE	SOURCE
metastatic-tissue-center-x	13	
metastatic-tissue-center-y	13	
metastatic-tissue-width	30	
probability-of-successful-invasion	$\frac{1}{1000}$	[67]

successful-metastasis?	false
metastatic-vessel-detection-radius	5

INFECTION

PARAMETER	VALUE	SOURCE
probability-of-JCV-spreading-to-neighbor	0.02	calibrated
frequency-of-infection-years	1	calibrated
chronic-infection?	false	
cell-type-transformed	""	
transformation-location	""	
count-transformed-cells	0	

JCV

PARAMETER	VALUE	SOURCE
JCV-NCRR-region-length	430	[47]
JCV-mutation-rate-per-day	1.074	[31]
JCV-mutation-in-NCRR	4.6182×10^{-4}	
JCV-genomic-instability	7.5	[81]

OXYGEN

PARAMETER	VALUE	SOURCE
amount-of-oxygen-per-patch	1	
oxygen-in-non-tissue	0.5	calibrated
oxygen-replaced-by-patch-each-time-step	0.25	calibrated

oxygen-needed-to-move	0.1	calibrated
oxygen-consumed-by-cell	1	calibrated
oxygen-metabolized-during-division	1	calibrated
oxygen-diffused-by-cell	1	calibrated
oxygen-of-hypoxic-cells	6×10^{-4}	calibrated

ANGIOGENESIS

PARAMETER	VALUE	SOURCE
vessel-detection-radius	5	
vessel-forward-movement	1	
oxygen-added-by-vessel	0.5	calibrated

Table C.8: Initial Parameters

C.3.2 *Input Data*

There is no input data in this model.

C.3.3 *Submodels*

C.3.3.1 *Determine If Simulation Should End*

This sub-model determines whether or not the simulation should continue. The simulation will end if either of the following conditions are true: a metastatic tumor has formed; the crypt has reach 100 years of age.

C.3.3.2 *Determine if Mutant Phenotype is On*

The model as a whole assumes that it takes more than one mutation to remove each barrier, but that any single mutation would still have some effect on the cell's behavoir. Thus, a sub-model is needed to determine if a mutant phenotype is expressed, based on the number of mutations in each barrier. This is accomplished by using a Bernoulli reporter to determine if the phenotype is expressed, where the probability that the mutant phenotype is expressed is equal to number of mutations in that barrier divided by how mutations it takes to remove each barrier. For example, if a cell has one pro-growth mutation (i.e. *count-pro-growth-mutations* = 1) and it takes two mutations to remove each barrier (i.e. *mutations-needed-to-remove-each-barrier* = 2) , then the probability that the pro-growth phenotype is expressed is equal to $\frac{1}{2} = 0.5$. Thus, that particular cell will exhibit the pro-growth phenotype 50% of the time.

If a stem cell has one metastatic mutation it will find the neighbor with the most oxygen, and then move to that neighbor (even if there is already another cell there). If the stem cell has two metastatic mutations, and is there is a blood vessel the radius of 1 patch, then the stem cell will attempt to invade the metastatic tissue. The probability that invasion will be successful is determined using the Bernoulli reporter *report-successful-invasion?*, where the probability of success is *probability-of-successful-invasion*. If invasion is successful, the stem cell moves to the center of the metastatic tissue, hatches a metastatic stem cell, and dies. This

leaves only the metastatic stem cell, but it inherits all of the parent stem cells variables (i.e. mutations, barriers removed, mutation rate, ect...).

Similarly, if a transit cell has one metastatic mutation it will move to the neighbor with the most oxygen, and if it has two mutations it will try to invade the metastatic tissue. See C.3.3.11 for a full description of the metastatic mutations interact with the other mutations.

VERIFICATION

This sub-model was verified by creating a world with 100 stem cells, with the option to manually add a mutation to any barrier in all of the stem cells. Next to the "world view" are histograms of how many stem cells are expressing the mutant phenotype during that tick. If one adds a single mutation, then the histogram hovers around 50 (i.e. ~50 of the 100 stem cells are expressing the mutant phenotype); if an additional mutation is added to that barrier, all of the stem cells will express the mutant phenotype.

C.3.3.3 *Determine Location*

This sub-model simply reports which type of tissue (colon crypt or metastatic tissue) each cell is in when a critical event occurs. Such critical events include mutations, barrier removal, transformation, etc...See C.3.3.18 for a complete list of events recorded during each run.

C.3.3.4 *Infection*

JCV is a common virus, and most individuals tend to be infected by adolescence Viscidi et al. [144], Knowles et al. [68]. Thus, if *infection?* is set to "true", then infection is modeled by giving JCV the opportunity to infect each crypt every *frequency-of-infection-year* years. During each infection attempt, a random transit cell is chosen to be infected. This infected transit cell can then infect one of its randomly chosen neighbors, with the probability of successful infection being equal to *probability-of-JCV-spreading-to-neighbor*. If all of the cells in the crypt become

infected, then the cell is considered to have become chronic, and the global parameter *chronic-infection?* is set to true, and the age at which the infection became chronic is also recored.

Initially, an infected cell's phenotype remains the same as other wild-type cells of the same breed (i.e. transit cell or stem cell). This is because the cell's CIF-II pathway is preventing the expression of viral oncogenes [153]. However, if the CIF-II pathway is "removed" by the accumulation of "beneficial" mutations, then viral oncoproteins are expressed. JCV produces several oncoproteins that, in animal models, strongly interfere with apoptosis (p53), and cell division (pRb,β-catenin, MAPK (via PP2A)), resulting in multi-nucleation, increased doubling time, growth in anchorage dependent conditions, uncontrolled cell growth, and chromosomal instability (reviewed in [82]). In the "Full Transformation" model, these processes can be modeled by completely removing the apoptosis barrier (chromosomal instability and uncontrolled cell growth), the pro-growth barrier (growth in increased doubling time and anchorage dependent conditions; i.e. no contact inhibition), and the anti-growth barrier (uncontrolled cell growth). In the "Partial Transformation" model, each of these barriers are only partially removed; however, if there is already a mutation in one of those barriers, then JCV will serve to completely remove that barrier. In the "No Infection" model, mutations in if the CIF-II pathway do not have any effect.

Note that transformation is modeled by adding mutations to the pro-growth, anti-growth, and anti-apoptosis barriers whenever *CIF-II-removed?* is true. However, it is not hypothesized that removing the CIF-II pathway literally adds mutations to each of these pathways; this method was chosen simply because it allows one to easily change the phenotype of the cell using the existing code.

VERIFICATION

The infection model was visually verified by creating a model in which cells do not move, and then randomly selecting a cell to infect, and then changing the shape and size of that infected cell. One could then watch the

infection spread among neighboring cells, each changing to the infected size and shape.

Varying the *probability-of-JCV-spreading-to-neighbor* parameter between 0 and 1 also ensured that the infection sub-model is behaving as expected. Low values of *probability-of-JCV-spreading-to-neighbor* resulted in a large number of time steps until the entire crypt is infected, while large values of *probability-of-JCV-spreading-to-neighbor* resulted in few time steps until every cell is infected.

Transformation was verified by manually adding a single pro-growth, anti-growth, anti-apoptosis, and CIF-II mutation to each stem cell. One could then step through the model and inspect the stem-cell's *pro-growth-removed?*, *anti-growth-removed?*, *apoptosis-removed?*, and *CIF-II-removed?* parameters. Whenever *CIF-II-removed?* was true, each of the other three barriers had a mutation added to it. However, when *CIF-II-removed?* was false, the mutations returned to its previous value. This verifies that the transformation procedure only adds mutations when CIF-II is removed.

CALIBRATION

The parameters *frequency-of-infection-year* years and *probability-of-JCV-spreading-to-neighbor* were calibrated so that the modeled incidence of infection had a pattern similar to that observed in Viscidi et al. [144], Knowles et al. [68]. After several parameter sweeps, *frequency-of-infection-year* was set to 1, and *probability-of-JCV-spreading-to-neighbor* was set to 0.02. In other words, each crypt is exposed to infection once a year, and there is only a 2% chance that each infected cell would be able to infect its neighbor.

C.3.3.5 *Hit and Run Model of Infection*

An alternative infection model is built upon the research of [113], who argue that JCV may increase the risk of colorectal by some sort of "hit and run" mecha-

nism. Under this hypothesis, JCV first infects the kidneys, mutates into the Mad-1 strain, giving it the ability to infect colorectal cells. This mutation event is then followed by a deletion in the non-coding control region (NCRR), which may allow JCV to express its oncoproteins in the colon cells, as well as increasing CIN, a process that may result in the formation of a colon cancer stem cell. Such transformation may occur for 14-21 days, which how long LT can be detected after transfection [111].

The infection component of the "Hit and Run" model is conducted in the same manner as described in C.3.3.4. However, in this case mutations in the CIF pathway are not required for transformation. Instead, mutations in the NCRR region allow for transformation. In this case, JCV mutations events in JCV's NCRR region are determined by the reporter *report-JCV-NCRR-mutation?*, which uses a Bernoulli random distribution, where the probability of success is *JCV-mutation-in-NCRR*, which is calculated as *JCV-NCRR-region-length* $\times$ *JCV-mutation-per-day.* If a cell is dividing and has *infected?=true* and *report-JCV-NCRR-mutation?=true*, the cell acquires one NCRR mutation. Ensuring that NCRR mutations only occur when the host cell divides was accomplished by inserting the *JCV-mutation* within each cell's division procedure. In this Hit and Run model, a single NCRR mutation simulates cell's acquired ability to infect the colon, but does not otherwise affect the cell's phenotype. If a second mutation occurs, and *infected?=true*, the cell is expresses its oncoproteins, removing the pro-growth, anti-growth, and apoptosis barriers. Additionally, the expression of these oncoproteins induces CIN, increasing the cell's current mutation rate 7.5 fold [81].

Once the JCV starts expressing the oncoproteins it is assigned a random lifespan between 14-21 days, which is recorded under the *JCV-Mad-1-d98-survival* parameter of the cell. Similarly, once the cell accumulates two NCRR mutations it has its *JCV-Mad-1-d98-evolved?* set to *true*, and *JCV-Mad-1-d98-days-active* set to 0. Every time the cell divides it increases *JCV-Mad-1-d98-days-active* by 1, and when *JCV-Mad-1-d98-days-active* = *JCV-Mad-1-d98-survival* the cell returns to

its previous mutation rate, removes the effects of JCV's oncoproteins on the barriers, resets *JCV-Mad-1-d98-days-active* back to zero, and sets *JCV-Mad-1-d98-evolved?* back to false. Returning to the previous mutation rate is accomplished by dividing the infected cell's mutation rate by *JCV-genomic-instability*, thus removing JCV induced genomic instability. Similarly, returning to the previous number of barriers removed is accomplished by subtracting *mutations-added-by-transformation* from *count-anti-growth-mutations*, *count-pro-growth-mutations*, and *count-anti-apoptosis-mutations*. By doing so, any additional mutations accumulated during the genomically unstable transformation period will remain present. FInally, *JCV-Mad-1-d98-days-active* is reset to zero because it is assumed that JCV, and not the cell, become inactive. This means that the cell can be re-infected. If this occurs, the cell will acquire a new value for JCV-Mad-1-d98-survival, JCV-Mad-1-d98-evolved? set to true, and the process will repeat.

VERIFICATION

The Hit and Run model was verified by creating a world in which all cells were infected and NCRR mutations could be added manually, one by one. It was verified that when a cell has only one NCRR mutation there is no phenotypic change, but when there are two NCRR mutations *pro-growth-removed?*, *anti-growth-removed?*, and *apoptosis-removed?* were all set to true. Similarly, the mutation rate was increased 7.5 fold. By inspecting each transformed cell, one is also able to confirm that *JCV-Mad-1-d98-days-active* increases by one every day, and that the cell does die when *JCV-Mad-1-d98-days-active* = *JCV-Mad-1-d98-survival*. By creating links between parent and daughter cells, one can also verify that daughter cells inherit the parental values *of JCV-Mad-1-d98-days-active* and *JCV-Mad-1-d98-survival*. By following these same cells, one can also verify that the mutation rate returns to its previous state, that the effect of JCV's oncoproteins on the barriers is removed, and that *JCV-Mad-1-d98-days-active* returns to zero and *JCV-Mad-1-d98-evolved?* returns to false when *JCV-Mad-1-d98-days-active* = *JCV-Mad-1-d98-survival*.

Furthermore, by manually adding NCRR mutations to those same cells after *JCV-Mad-1-d98-evolved?* is reset to false, one can verify that the cell is assigned a new value for *JCV-Mad-1-d98-survival*.

That NCRR mutations only occur when the cell divides was verified by setting *JCV-mutation-in-NCRR* to 1, and filling the crypt with cells. In this scenario the stem cells cannot divide, and so no NCRR mutations occur. However, when the crypt is not filled, the cells are able to divide and if they are infected they can acquire NCRR mutations. This was further verified by having each infected cell print a message that it divided and had a mutation in the NCRR.

C.3.3.6 *Stem Cell Determine Division Type*

Stem cells can either divide asymmetrically or symmetrically [12]. During asymmetric division, each stem cell produces one daughter transit cell and one daughter stem cell. During symmetric division, the stem cell produces either 2 stem cells or 2 transit cells. It has been suggested that symmetric division occurs 5% of the time [79].

To model this process, one can use the reporter *asymmetric-division?*, which uses a Bernoulli random distribution, where the probability of success is *probability-of-asymmetric-stem-cell division*. If "true" is returned, asymmetric division occurs; if "false" is returned, symmetric division occurs.

ASYMMETRIC DIVISION

If asymmetric division occurs, then any stem cells that have *pro-growth-removed?* and *anti-growth-removed?* as false will call the *stem-cell-normal-division* procedure. In this procedure, each the stem cell will undergo mutation (see the C.3.3.6 for details of the mutation procedure) and hatch a transit cell if there is an empty patch within a cone having an angle of 120° and radius of 1.5 patches. If there is no such empty patch, the stem cell will not undergo mutation or division. If this

stem cell is mobile (because it has one metastatic mutation), it will only be able to divide in the inner crypt.

Any stem cells with *pro-growth-removed?* as true but *anti-growth-removed?* as false, call the *stem-cell-pro-growth-division* procedure. When this procedure is called, the stem cell will always undergo mutation and division, producing a transit cell that moves to a randomly chosen patch with a cone with an angle of 120° and radius of 1.5 patches. If this stem cell is mobile (because it has one metastatic mutation), it will only be able to divide in the inner crypt.

Any stem cells with *anti-growth-removed?* as true but *pro-growth-removed?* as false, call the *stem-cell-anti-growth-division* procedure. When this procedure is called, and if the cell is in the inner-crypt, the cell will divide as in the *stem-cell-normal-division* procedure . However, if this stem cell is mobile (because it has one metastatic mutation), it gains the ability to divide in both the inner crypt and the outer crypt.

Any stem cells with *anti-growth-removed?* as true and *pro-growth-removed?* as true, call the *stem-cell-anti-and-pro-growth-division* procedure. When this procedure is called, the stem cell will always undergo mutation and division, producing a transit cell that moves to a randomly chosen patch with a cone with an angle of 120° and radius of 1.5 patches . Furthermore, if this stem cell is mobile (because it has one metastatic mutation), it also gains the ability to divide in both the inner crypt and the outer crypt.

SYMMETRIC DIVISION

If *asymmetric-division?* returns false, then symmetric division will occur. During symmetric division, the stem cell with the most deleterious mutations is killed and replaced by a daughter from the stem cell with the fewest deleterious mutations. However, it is important to note that any stem cells with *anti-growth-removed?* as true will be exempt from symmetric division. It is thus possible that the true least fit stem cell could survive symmetric division (assuming it has *anti-growth-removed?* as true). All of the other stem

cells (i.e. not the least or most fit stem cells) will follow the asymmetric division procedure rules. Symmetric division is modeled this way to capture the idea that the stem cell crypt has evolved to minimize the risk of disease by periodically killing off the least fit (most deleterious) cells [12].

STEM CELL MUTATION

The number of each type of mutation added during each division is determined by drawing a random number from a Poisson distribution, with an appropriate mean (λ). Mutations of each type accumulate over the life of the cell. Finally, the mutation rate will increase proportionately with *count-anti-apoptosis-mutations*$\times$*genomic-instability*.

NEUTRAL MUTATIONS

The human genome is ~7×10^9bp, and there are 70,000 genes, each of which has an average length of 1000bp (reviewed in [92]). Assuming that mutations anywhere in any of these 70,000 genes will either result in a deleterious of beneficial phenotype, then mutations elsewhere must be neutral. Thus, one can assume that $7 \times 10^9 - 70000\,(1000) = 6.93 \times 10^9$bp of the genome are neutral. If the stem cell mutation rate is 1×10^{-9}, then one should expect $10^{-9} \times 6.93 \times 10^9 = 6.93$ neutral mutations per stem cell division. Thus, to determine how many neutral mutations will occur during each division, one can draw a random number from a Poisson distribution with $\lambda = 6.93$ mutations per division.

BENEFICIAL MUTATIONS

Assuming there are 6 barriers to cancer, 3 candidate genes that can inhibit each one of those barriers, each with an average length of 1000bp, then there are $6 \times 3 \times 1000 = 18,000$bp, that if mutated will remove a barrier to cancer. As these mutations increase the cell's ability to survive and replicate, they can be considered beneficial mutations. Thus, one

can expect that there are $1 \times 10^{-9} \times 18000 = 1.8 \times 10^{-5}$ beneficial mutations during each stem cell division. Thus, one can determine how many beneficial mutations will occur by drawing a random number from a Poisson distribution with $\lambda = 1.8 \times 10^{-5}$ mutations per division.

If a beneficial mutation does occur, then a random number X is drawn from a Uniform distribution that has a range from 1 to 6. This number then determines which kind of beneficial mutation will occur. If $X = 1$ there is a pro-growth mutation; if $X = 2$ there is an anti-growth mutation; if $X = 3$ there is an anti-apoptosis mutation and the mutation rate is increased by multiplying it by *genomic-instability*; if $X = 4$ there is a telomerase mutation; if $X = 5$ there is a metastatic mutation; if $X = 6$ there is a CIF-II mutation. A Uniform distribution is used because it is assumed that the mutation is equally likely to "land" in any one of the beneficial genes, since they all have equal lengths.

GENOMIC INSTABILITY MUTATIONS

Assuming there are 7 genomic instability genes, there are total of $7 \times 1000 = 7000$bp, that if mutated will increase the mutation rate. Therefore, the expected number of genomic instability mutations per division is $\lambda = 7000 \times 10^{-9} = 7 \times 10^{-6}$. Each mutation that lands in a genomic instability gene linearly increases the mutation rate by a factor of two. For example, one mutation doubles the mutation rate, 3 mutations increase the mutation rate by a factor of 6, and so on until the mutation rate is increased 14 fold.

DELETERIOUS MUTATIONS

Assuming that there are 70,000 genes, each with an average gene length of 1000bp, then there are $70000\,(1000) = 7 \times 10^7$bp that are not neutral when mutated. If there are 6 barriers to cancer, 3 candidate genes that can inhibit each one of those barriers, then there are $6 \times 3 \times 1000 =$

18,000bp, that if mutated will remove a barrier to cancer. These mutations can be considered beneficial. Finally, there are 7000bp that induced genomic instability when mutated. Thus, the remaining $7 \times 10^7 - 18,000 - 7000 = 69975000$bp, if mutated, will be deleterious. One should then expect $1 \times 10^{-9} \times 69982000 = 0.069982$ deleterious mutations per each stem cell division. Thus, to determine how many deleterious mutations there will be during each stem cell division, one can draw a random number from a Poisson distribution with $\lambda = 0.069982$ mutations per division.

PARTIAL BARRIER REMOVAL

Whether or not a mutation will be expressed is determined using the reporter *express-mutation?*. This reporter will randomly report a true or false value, using a Bernoulli distribution where the probability of success is $p = \frac{\textit{count-mutations-in-barrier}}{\textit{mutations-needed-to-remove-each-barrier}}$. For example, if *mutations-needed-to-remove-each-barrier* = 2, and *count-pro-growth-mutations* = 1, then $p = \frac{1}{2} = 0.5$. If "true" is returned, the mutation will have its effect; if "false" is returned the mutation will not have its effect. NetLogo does not include a built in Bernoulli distribution, but Grimm and Railsback do provide the code for how to create such a distribution [108].

Partial barrier removal is included in the model to capture the idea that it likely takes more than a single mutation to completely remove a barrier to cancer, but disruptions in the pathways(s) by single mutations still have some affect on the phenotype.

VERIFICATION

The rate of symmetric division was verified simply by adding a monitor that recorded how frequently symmetric division occurred. Once also had the option to change the rate of asymmetric division, which did result in a

change in the monitored rate of symmetric division (i.e. increasing the probability of asymmetric division decreased the rate of symmetric division).

The code used to identify the least fit and most fit stem cells during symmetric division was verified by outputting each stem-cell's number of deleterious mutations, and which cells were being identified as the most fit, least fit, and other stem cells. The code was verified because the cell with the most deleterious mutations was successfully identified, colored green, and then killed off in the next round of division. Similarly, the most fit stem cell was successfully identified, changing its shape, and producing a daughter cell (with the same shape) during the next round of division.

The code for partial barrier removal was verified by creating a world with 100 stem cells. Next to the world were histograms of how many cells had the pro-growth phenotype, anti-growth phenotype, etc...One could then manually add mutations to all of the stem cells. The parameter *mutations-needed-to-remove-each-barrier* was set to two, so when one mutation was added to all of the stem cells, the histogram showed that ~50% of the cells exhibited the mutant phenotype, verifying the the mutated cells expressed the mutant phenotype ~50% of the time. When a second mutation was added, all of the stem cells expressed the mutant phenotype.

Normal stem cell division (i.e. only divide when there is an empty patch ahead) was verified by creating a world with only two stem cells, each of a different color. One can place a stationary transit cell in front of the stem cell, and then ask the stem cell to identify all patches that it could have its daughter transit cell move to. This verifies that the code works because the stem cell will identify all patches within their cone of vision, except for that with a transit cell already on it.

The pro-growth phenotype was verified by filling the crypt with stationary transit cells and then setting *pro-growth-removed?* to true for one stem cell, and then adding a single metstatic mutation. When the simulation is run,

the stem cells produce transit cells, even though there are already transit cells on every patch. However, it is only able to divide in the outer crypt

The anti-growth phenotype was verified by giving one of the stem cells one metastatic mutation (making it mobile), and then setting *anti-growth-removed?* to true. This mobile stem cell will then divide in divide in both the inner and outer crypt.

The CIF-II deficient phenotype was verified by infecting one stem cell, adding two CIF-II mutations, and adding one metastatic mutation. Stationary transit cells were then added to every patch in the crypt. When the simulation was run, the mutated and transformed stem cell then had *cif-ii-removed?*, *pro-growth-removed?*, *anti-growth-removed?*, and *anti-apoptosis-removed?* as true, verifying that transformation removed these other barriers. Furthermore, the mutated and transformed cell was able to move around the crypt (because of the metastatic mutation) and divide everywhere, even when *proability-of-asymmetric-division* is set to 0 (verifying the anti-apoptosis phenotype).

It was also verified that stem cells with *apoptosis-removed?* as true were excluded from being identified as the least fit cell. This was accomplished by choosing one stem cell and manually adding 100 deleterious mutations, setting *apoptosis-removed?* to true, and changing its color to black. Throughout the simulation, this stem cell was never killed off, even though it had the most deleterious mutations.

See C.3.3.7 for a description of how the behavior of stem cells with metastatic mutations was verified.

C.3.3.7 *Stem Cell Replace Metastatic Stem Cell*

This model assumes that the crypt can determine if it has too few stem cells, and will respond by having the most-fit stem cell (i.e. that stem cell with the fewest deleterious mutations) hatch one daughter stem cell. The only time the crypt will have too few stem cells will be if one stem cell acquired a metastatic mutation,

giving it the ability to roam around the crypt. Again, if this occurs, the most-fit stem cell hatches a daughter stem cell, returning the number of non-metastatic stem cells to its normal amount.

VERIFICATION

The creation of metastatic stem cells from stem cells was verified by creating a world where the *probability-of-successful-invasion* was set to 1, vessels could be added manually, and metastatic mutations could also manually be added. After adding vessels and one metastatic mutation, the mutated stem cell moves toward the neighbors with the most oxygen, which in this case are the neighbors with vessels. The mutated stem cell will keep moving around the vessels as long as it only has one metastatic mutation. One can then add a second metastatic mutation to this stem cell, after which the stem cell moves the metastatic tissue, hatches 1 metastatic stem cell, which then hatches four more metastatic stem cells.

C.3.3.8 *Metastatic Stem Cell Fill Tissue*

After a stem cell invades the metastatic tissue, it hatches *number-of-stem-cells* - 1 metastatic stem cells. After there are *number-of-stem-cells* metastatic stem cells, the metastatic stem cells are laid out in a circle.

C.3.3.9 *Metastatic Stem Cell Determine Division Type (symmetric or asymmetric; a global procedure)*

This procedure is identical to the stem cell division type procedure, except that it applies only to metastatic stem cells in the metastatic tissue.

C.3.3.10 *Transit Cell Consume Oxygen*

Each transit cell will consume a certain amount of oxygen every time step. If the underlying patch has enough oxygen, then the transit cell will consume *oxygen-consumed-by-cell* units of oxygen. If there is less than *oxygen-consumed-by-cell* units

of oxygen in the patch, then the transit cell will only consume half of the available oxygen.

For a description of the calibration process for how much oxygen each cell consumes, see C.3.3.14.

C.3.3.11 *Transit Cell Division, Mutation, and Movement (turtle procedure)*

Every time a transit cell divides it undergoes mutation (see C.3.3.11 for details), uses up some oxygen (determined by the parameter *oxygen-metabolized-during-division*; if there is not enough oxygen the cell will not divide), divides its oxygen equally between itself and it's daughter, and decreases it's telomere length by one unit (so long as there aren't any telomerase mutations). After hatching the daughter cell, the parent then moves to a different patch, using some more oxygen in the process (determined by the parameter *oxygen-needed-to-move*). Where exactly the parent cell can divide and move to is determined by which (if any) mutant phenotypes it has.

If the transit cell has no pro-growth, anti-growth, or metastatic mutations, it will call the *transit-cell-normal-division* procedure. In this procedure, the transit cell will only divide if it is in the inner crypt and there is an empty patch with its cone of vision, which has an angle of 180° and radius of 1.5 patch units. If it can divide, the parent then randomly choses and moves to one of the empty patches in it's cone of vision. If the cell is in the outer crypt, it will use the same rules to decide to move (i.e. there must be an empty patch)

If the transit cell has no pro-growth, anti-growth, but one metastatic mutation, it will call the *transit-cell-normal-metastasis-division* procedure. In this procedure, the transit cell will only divide if there is an empty patch with its cone of vision, which has an angle of 180° and radius of 1.5 patch units. If the parent can divide, it moves to the empty neighbor with the greatest amount of oxygen.

If the transit cell has *pro-growth-removed?* as true (see C.3.3.2 for details) , but not anti-growth, or metastatic mutations, it will call the *transit-cell-pro-growth-*

division procedure. In this procedure, the transit cell will only divide if it is in the inner crypt, but it does not require that there be an empty patch within it's cone of vision. If it can divide, the parent then randomly choses and moves to one of the patches in it's cone of vision. If the cell is in the outer crypt, it will use the same rules to decide to move (i.e. it can move to any patch within its cone of vision).

If the transit cell has *pro-growth-removed?* as true, one metastatic mutation, but no anti-growth mutations, it will call the *transit-cell-pro-growth-metastasis-division* procedure. This procedure is very similar to the *transit-cell-pro-growth-division* procedure, but randomly choosing a patch in the cone of vision, the parent moves to the neighboring patch with the most oxygen, regardless of whether or not that patch is already occupied. Again, these transit cells can only divide if they are in the inner crypt.

If the transit cell has *anti-growth-removed?* as true (see C.3.3.2 for details) , but no pro-growth or metastatic mutations, it will call the *transit-cell-anti-growth-division* procedure. In this procedure, division requires that there be an empty patch within the transit cell's cone of vision, but these transit cells can divide in both the inner and outer crypt. If it can divide, the parent then randomly choses and moves to one of the empty patches in it's cone of vision.

If the transit cell has *anti-growth-removed?* as true, one metastatic mutation, but no pro-growth mutations, it will call the *transit-cell-anti-growth-metastasis-division* procedure. This procedure is very similar to the *transit-cell-anti-growth-division* procedure, but instead of randomly choosing an empty patch in the cone of vision, the parent moves to the empty neighbor patch with the most oxygen.

If the transit cell has both *anti-growth-removed?* and *pro-growth-removes?* as true (see C.3.3.2 for details), but metastatic mutations, it will call the *transit-cell-anti-and-pro-growth-division* procedure. In this procedure, the transit cell does not require that there be an empty patch within it's cone of vision, and it can divide in both the inner and outer crypt. When these cells divide, they randomly chose

and move to one of the patches in it's cone of vision, regardless of whether or not it is already occupied.

If the transit cell has both *anti-growth-removed?* and *pro-growth-removes?* as true and one metastatic mutation, it will call the *transit-cell-anti-and-pro-growth-metastasis-division* procedure. This procedure is very similar to the *transit-cell-anti-and-pro-growth-division* procedure, but instead of randomly choosing a patch in the cone of vision, the parent moves to the neighbor patch with the most oxygen, even if it is already occupied.

If any transit cell has *telomerase-removed?* as true, the telomeres do not decrease in length after division.

For a full description of the apoptosis mutations, see C.3.3.17. For a description of how CIF-II mutations affect the transit cell's phenotype, see C.3.3.4.

MUTATION

The mutation procedure for transit cells is nearly identical to that of the stem cells, except that the mutation rate is ten times higher, increasing the expected number of mutations per division accordingly.

Transit cells will also exhibit partial barrier removal, using the same process as described in the stem cell mutation section.

VERIFICATION

The code used for mutation is exactly the same as that used for stem cells (see C.3.3.6), except that the transit cells use the *transit-cell-mutation-rate* instead of *stem-cell-mutation-rate*, and so was already verified.

Normal transit cell movement was verified by randomly placing stationary transit cells around the crypt, and then following mobile transit cells. The mobile transit cells identify all potential target patches by changing the patch color, and then moves to one of them. The movement code was verified because patches with stationary transit cells on them were not identified as target patches, and the mobile transit cells did not move to them.

Furthermore, it was verified that the parent moves to the target patch by changing the color of the parent cell.

The pro-growth mutation was verified by removing the pro-growth barrier in one stem cell (i.e. setting *count-pro-growth-mutations* to two), and changing the color of all of its daughter cells to green (instead of red). One can then step through each tick, verifying that, when in the inner crypt, the mutated cells will move to occupied patches, and divide (visualized by creating links between the parent and daughter cells). One can also observe that the mutated transit cells cannot divide in the outer crypt, which can be verified by inspecting such a cell and making sure that *count-transit-cell-divisions* does not increase.

The anti-growth code was verified by removing the anti-growth barrier in one of the stem cells (i.e. setting *count-anti-growth-mutations* to two), and stepping through the model until some of the mutated cells (colored brown instead of red) reached the outer crypt. Once the mutated cell reaches the outer crypt, one can inspect it and continue to step through the model, allowing the mutated cell to move and divide in the outer crypt. *Count-transit-cell-divisions* continues to increase (so long as telomeres remain and there are empty patches in the cone of vision), verifying that the anti-growth code.

It was also verified that the pro-growth and anti-growth phenotypes worked together; that is, transit cells with both of these barriers removed can divide anywhere in the crypt, even if there are not any empty patches within their cone of vision. This was verified by removing both barriers in one stem cell (i.e. setting *count-anti-growth-mutations* to two and setting *count-pro-growth-mutations* to two), and filling the outer crypt with stationary transit cells. The mutated cells have the ability to move to and divide on occupied patches in both the inner and outer crypt, verifying that code for the two mutant phenotypes work together.

The anti-apoptosis phenotype was verified by randomly hatching 300 transit cells, each with a random number of deleterious mutations, ranging between 1 and 100. Afterwards, 20 of those transit cells had the apoptosis barrier removed, their number of deleterious mutations set to 200, and their shape changed to a square (instead of a circle). Next, the population cap procedure was executed, but all of the cells with apoptosis removed survived the cap (see C.3.3.17). This verifies that anti-apoptotic cells a exempt for the fitness search conducted during the population cap procedure, giving them the ability to survive even though they have the most deleterious mutations.

The metastasis phenotype was verified in the same manner as described in C.3.3.7, except that if the transit cell successfully invades the metastatic tissue it still behaves as if it were in the crypt (i.e. it still uses the *transit-cell-division-mutation-and-movement* procedure).

The telomerase mutation was simply verified by removing the telomerase barrier in one of the transit cells, and then inspecting it to make sure that *telomere-length* did not decrease even when the cell underwent division.

C.3.3.12 *Angiogenesis (turtle procedure)*

Angiogenesis is the production of new blood vessels during hypoxic stress. Such stress can occur if there are too many cells and not enough oxygen; in this situation, the body responds by producing new blood vessels to supply oxygen to the extra cells. This process is modeled by asking any patches that oxygen levels below *oxygen-of-hypoxic-cells* and no other vessels within *vessel-detection-radius* to sprout a new blood vessel. If a patch is hypoxic (i.e. oxygen levels below *oxygen-of-hypoxic-cells*), but there is a blood vessel within *vessel-detection-radius*, it will sprout a VEGF molecule. The VEGF molecule then detects the closest vessel and moves towards it one patch unit for each tick. Once the VEGF molecule is within 0.5 patch units of the nearest vessel, it stimulates the vessel to produce one more

vessel, which moves forward *vessel-forward-movement* patch units. After such vessel growth, the stimulating VEGF molecule dies.

Each blood vessel is randomly assigned a lifespan, ranging from 1-250 days Chen et al. [21].

During each tick, the vessels add *oxygen-added-by-vessel* units of oxygen to the underlying patch, thus increasing the amount of oxygen available to the crypt.

VERIFICATION

Angiogenesis should only occur when there is not enough oxygen in the crypt for the number of cells present, which would occur if the population increased beyond its normal size. Such a population increase would only occur if mutations drive the cells to divide more frequently than normal, as occurs when many of the cancer barriers are removed. This was verified by running the model and manually adding mutations to the stem cells. When there are no mutations, angiogenesis does not occur. However, after several barriers are removed, angiogenesis begins to occur, increasing the amount of oxygen available, and allowing the population to increase from ~250 cells to 5000 cells (see C.3.3.17 for a description of why the population is capped at 5000 cells).

C.3.3.13 *Oxygen Replenish (patch procedure)*

Each tick, *oxygen-replaced-by-patch-each-time-step* is added to each patch in the colon crypt. This process is meant to simulate the process of the underlying bed of blood vessels supplying oxygen to support the cells in the crypt. Until a cell successfully invades the metastatic tissue, all patches in the metastatic tissue have their oxygen kept at *amount-of-oxygen-per-patch*, so as to simulate homeostasis prior to invasion.

C.3.3.14 *Oxygen diffusion (global procedure)*

Oxygen diffusion is modeled using NetLogo's built in primitive, *diffuse*. Thus, *oxygen-diffused-by-cell* percent of the patches oxygen is divided equally among the patch's neighbors. Also, all patches that do not represent tissue (i.e. the black patches) have their oxygen levels set to *oxygen-in-non-tissue*. These two processes are repeated 20 times each tick, so as to stimulate constant oxygen diffusion throughout the day Chen et al. [21].

CALIBRATION OF OXYGEN AND ANGIOGENESIS PARAMETERS

It has been observed that colon stem cells divide approximately once every four to five days, there are ~250 cells in each colon crypt, and angiogenesis does not occur under normal growth conditions Potten et al. [103], Booth and Potten [12], Kleinsmith [67]. As oxygen levels determine whether or not a transit cell can divide or move (making room for other cells), a series of seven parameter sweeps were conducted on all oxygen parameters to narrow down a final set of values to sweep. The values tested in this final sweep were : *oxygen-added-by-vessel* ranged from $0.1 - 0.5$ in 0.1 increments; *oxygen-replaced-by-patch-each-time-step* ranged from $0.25 - 1$ in 0.25 increments; *oxygen-metabolized-during-division* ranged from $0.25, 0.5, 1.0$; *oxygen-of-hypoxic-cells* ranged from $3 \times 10^{-4} - 8 \times 10^{-4}$ in increments of 1×10^{-4}; *oxygen-consumed-by-cell* was tested at $0.25, 0.5, 1.0$; *oxygen-diffused-by-cell* was tested at $0.5, 0.75, 1.0$; *oxygen-in-non-tissue* was tested at $0.5, 0.75, 1.0$. All values were tested in combination using NetLogo's "Behavior Space", resulting in a total of 4860 runs, each of which lasted 500 ticks. The final set of parameter values, which can be found in C.8, resulted in a average stem cell division rate of 0.2212 divisions per day (or one division every 4.5 days), an average of 255.766 transit cells in the crypt, and 0 vessels.

Of note is that mutations were turned off during the parameter sweeps mutation. This was done because mutations affect the cell's phenotype, chang-

ing the rules about when and where division can occur, and thus the stem cell's division rates and total population size (which could induce angiogenesis). By turning off mutation, one can thus get a better idea of the stem cell's division rate and population size under normal conditions.

C.3.3.15 *Oxygen Recolor Patches (patch procedure)*

The shade of each patch can be changed to reflect the amount of oxygen that it has. This can easily be accomplished using NetLogo's *scale-color* primitive, using oxygen as the number input.

C.3.3.16 *Transit Cell Death (turtle procedure)*

In a normal crypt, transit cells die in one of two ways: they move outside the crypt and are shed, or they completely lose their telomeres.

C.3.3.17 *Maintain Population Cap (turtle procedure)*

In this model there are actually two maximum population sizes, *normal-max-pop-size* and *max-pop-size*. If the population grows larger than *normal-max-pop-size*, the number of *excess cells* is determined by subtracting the current population size from *normal-max-pop-size*. Next, a sub-population of the least fit cells (i.e. most deleterious mutations) of size 1.5 $\times$ number of excess cells is found. The oxygen levels of each of these "least fit" cells is then determined. In the end, *excess-cells* least fit cells with the lowest oxygen levels are then killed off, bringing the population size back to *normal-max-pop-size*. The idea behind this is that when resources become scarce, the least fit cells with the fewest resources (in this case oxygen) would be the most susceptible to death.

It is important to note that cells with *anti-apoptosis-removed?* as true are excluded from the search for the least fit cells. This is because the apoptosis mutations prevent cell death, even in the presence of deleterious mutations and limited resources. This means that if enough cells have *anti-apoptosis-removed?* as

true, the population grow much larger than *normal-max-pop-size*. Due to limited computing power, a second population size limit had to be created, so as to prevent the computer from freezing while it attempts to track tens to hunderds of thousands of cells. This absolute limit is set by *max-pop-size*, and when it is reached the number of excess cells is calculated (as above) and *excess-cells* are randomly chosen from all cells to be killed off, returning the population size back to *max-pop-size*.

VERIFICATION

This procedure was verified by filling the crypt with 300 transit cells, each with a random number of deleterious mutations, ranging between 1 and 100, and a random oxygen level, ranging between 0 and 1. The color of each cell was also scaled according to how many deleterious mutations they had; the lighter the color, the more deleterious mutations. One could then identify the 1.5 $\times$ number of excess cells with the most deleterious mutations, and change their shape to a small circle. The average number of deleterious mutations in the ID'd cells was higher than the average number of deleterious mutations in the cells not identified as being the most deleterious, verifying that the search procedure was working correctly. Next, of those least fit cells, *excess-cells* were identified that had the lowest oxygen levels. Again, the average oxygen levels of these cells was lower than the average oxygen levels of the cells not identified as being "least fit" or most hypoxic. Together, this verifies that the population cap procedure successfully identifies the least fit cells, and then finds the most hypoxic of those cells. Finally, one can kill of the least fit and most hypoxic cells, returning the population level back to *max-pop-size*.

C.3.3.18 *Evaluate state and Record Data (global procedure)*

In addition to the summary data (described in C.2.11), major events of each run are also recored, in the event that more details are needed about what led to tumor formation. Such major events include:

- Stem cell mutation
- Metastatic stem cell mutation
- Stem cell barrier removal
- Metastatic stem cell barrier removal
- Stem cell transformation
- Metastatic stem cell transformation
- Stem cell symmetric division
- If a mutation has spread to all stem cells, either by symmetric division or mutation
- Formation of colon tumor (i.e. cell in crypt has pro-growth, anti-growth, apoptosis, and telomerase barriers completely removed)
- Successful invasion into metastatic tissue
- Formation of metastatic tumor (i.e. cell invaded metastatic tissue and has pro-growth, anti-growth, apoptosis, metastasis, and telomerase barriers completely removed)

Every time a major event occurs in a run, the following data is recorded by the cell experiencing the major event:

- Date and time
- Infection.Modeled (true/false)

- Mutations.Added.By.Transformation (0,1,2)
- Age
- Tissue
- Event
- Cell.Type (i.e. transit cell, stem cell, metastatic stem cell)
- Cell.Number
- Parent
- N.Beneficial.Mutations
- Beneficial.Mutations (list of mutations)
- Timing.of.Beneficial.Mutations (list)
- N.Barriers.Removed
- Barriers.Removed (list of barriers removed)
- Timing.of.Barrier.Removal (list)
- Cell.Infected (true/false)
- Cell.Transformed (true/false)
- Cell.Invasion (true/false)
- N.Cells.in.Crypt
- N.Metastatic.Cells
- Mean.Stem.Cell.Div.Rate
- Metastasis.Occurred (true/false)
- Age.of.Metastasis

- Colon.Tumor.Formed (true/false)
- Metastatic.Tumor.Formed (true/false)
- N.Transformed.Cells
- Age.When.Infection.Became.Chronic

APPENDIX D ACRONYMS

CDK Cyclin Dependent Kinase

EC Endothelial cell

VEGF Vascular Endothelial Growth Factor

CAM Cell-Cell Adhesion Molecules

MSI Microsatellite Instability

CIN Chromosomal Instability

LOH Loss of Heterozygosity

CIF Cellular Interfering Factor

FAP Family Adenomatous Polyposis

HNPCC Hereditary Nonpolyposis Colorectal Cancer

APC Adenomatous Polyposis Coli

KRAS Kristen Rat Sarcoma Virus

MMR Mismatch Repair Enzymes

ACF Aberrant Crypt Foci

GSK3 Glycogen Synthase Kinase 3

JCV JC Virus

PML Progressive Multifocal Leukoencephalopathy

NCRR Non-Coding Regulatory Region

T-ag Large T Antigen

t-ag Small T Antigen

COP Calabrese model with Original Parameters

CNP Calabrese model with New Parameters

GII Genomic Instability Infection Model

GIM Genomic Instability Mutation Model

ABM Agent Based Model

CURRICULUM VITAE

PERSONAL INFORMATION

Name	Chandler D. Gatenbee
Address	Department of Biology University of Louisville Louisville, KY 40292
E-mail	cdgateo2@louisville.edu
DOB	Washington D.C. - September 3, 1982

EDUCATION

2011-Present	Ph.D. Candidate in Biology Focus in Ecology, Evolution, and Behavioral Biology University of Louisville 2011-Present
2006-2008	M.S. Biological Anthropology Focus in Human Population Genetics University of Utah
2000-2005	B.S. Biological Anthropology, Minor in Biology Focus in the Natrual Sciences University of Louisville

AWARDS

2012	Doctoral Dissertation Completion Award, School of Interdisciplinary and Graduate Studies, University of Louisville
2005	Anthropology Merit Award, Department of Anthropology, University of Louisville

2004 Anthropology Student Travel Award, Department of Anthropology, University of Louisville

PUBLICATIONS

2012 C.D Huff, D.J. Witherspoon, Y. Zhang, C. Gatenbee, L.A. Denson, S. Kugathasan, H. Hakonarson, A. Whiting,C.T. Davis, W. Wu, J. Xing, W.S. Watkins, M.J. Bamshad, J.P. Bradfield, K.Bulayeva, T. Simonson, L.B. Jorde, and S.L Guthery. Crohn's Disease and Genetic Hitchhiking at IBD5, *Molecular Biology and Evolution* (2012) 29(1): 101-111

2011 R. Fernández-Botrán, Z. Ahmed, F.A. Crespo, C. Gatenbee, J. Gonzalez, D.W. Dickson, I. Litvan. Cytokine Expression and Microglial Activation in Progressive Supranuclear Palsy, Parkinsonism Related Disorders, (2011) 17(9):683-8.

2009 R. Pennington, C. Gatenbee, B. Kennedy, H. Harpending, and G. Cochran. Group Differences in Proneness to Inflammation, Infection, Genetics and Evolution (2009) 9(6):1371-80.

2008 S. Guthery, C. Gatenbee, Y. Zhang, M. Bamshad, and L. Jorde. M2058 The Distribution of the IBD5 Haplotype Among Worldwide Human Populations. Gastroenterology (2008) 134(4):A-460)

2007 C.R. Tillquist and C.D. Gatenbee. Icelandic Genetics Database. Encyclopedia of Epidemiology. SAGE Reference Project

2007 C.R. Tillquist, B. Schwallie, C. Gatenbee, F.A. Crespo, and P. Killoran. Report of ancient DNA analyses of samples from the Old Frankfort Cemetery. Submitted to Kentucky Archaeological Survey.

REFEREED JOURNALS

2010-Present Reviewer for *Infection, Genetics and Evolution*