Somatic Microsatellite Variability in Cancer and DNA Repair Disorders

Zalman Vaksman

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics and Computational Biology

Chair: Harold R Garner
Carla V Finkielstein
Stanley A Hefta
Jason A Holiday

December 11th 2014
Blacksburg Virginia

Keywords: Microsatellite, Somatic Variability, Liver Cancer, Colorectal Cancer, DNA repair Disorders, Werner syndrome, Cockayne's syndrome, Xeroderma pigmentosum, Rothmond-Thomson syndrome, RecQ, ERCC8, XPA

Somatic Microsatellite Variability in Cancer and DNA Repair Disorders

Zalman Vaksman

ABSTRACT

Microsatellites (MSTs) are short tandem repeats of motifs, 1 – 6 nucleotide in length, and are considered mutational "hot-spots" and show a greater degree of somatic variability and population polymorphisms than surrounding DNA sequences. MSTs provide for a unique computational alignment problem for many commonly used algorithms, and therefore required software tools to be developed to specifically address these issues. For this work we developed a novel approach to extract MSTs from next-gen sequencing data that can robustly detect signatures of MST mutation bias and somatic variation occurring in next-gen data including a high frequency of in-phase indels. Somatic variability, novel genomic polymorphisms that arise within a cell population not found in the progenitors, plays a critical role in cellular reprogramming leading to the development and progression of cancer. MST mutation rates are between 10 and 1000 time higher than that of surrounding DNA. MSTs are found ubiquitously throughout the genome including in nearly all transcribed regions and 10-20% of coding genomic regions. Currently the only established DNA repair defect that that has been directly linked to MST instability is replication coupled mismatch repair (MMR). An initial analysis of the utility of the software was conducted with DNA repair impaired cell lines. The results demonstrated the utility in identifying the consequences of DNA repair impairments on genomic stability. There were major objectives of the finding including 1) complimenting genomics of matched DNA samples with in-sample quantification of variation and 2) demonstrating that DNA repair proficient cells and those with different defects in DNA repair can have different somatic MST variability (SMV) profiles that may be potential markers for these defects.

DNA repair disorders are debilitating conditions that result in physical and neurological abnormalities robbing the individual of a normal quality of life and life span. The various conditions that fall into this class are known as progeroid disorders and they provide a very important glimpse into the aging process on a genomic level. The conditions for four cohorts analyzed here were; Cockayne's syndrome, caused by the loss of the ERCC8 gene, also known as CSA; xeroderma pigmentosum, caused by the loss of the XPA or XPB genes; Werner syndrome, caused by the loss of the RecQL2 gene; and Rothmond-Thomson syndrome, caused by the loss of the RecQL4 gene. The goal of this project was to determine if impaired excision repair genes CSA or global XPA and B or excision repair supporting helicases BLM or RecQL4 leads to MST destabilization. Comparing cohorts from excision repair disorders with a co-sequenced normal cohort we found that CSA both RecQ helicases had an effect on the exome somatic variability of MSTs. On the other hand the effects of XPA/B were inconclusive.

MST instability (MSI), defined as acquired/lost primary alleles in tumors for a small set of microsatellite loci, has been implicated and is a clinically relevant marker for colorectal cancer. Conversely, no clinically actionable genetic markers have been found for liver cancer, a cancer with a very high mortality rate. Here we explore the use SMV defined as the presence of minor alleles at MST loci, as a complementary measure of MSI as a genetic marker for colorectal and liver cancer by analyzing Illumina sequenced genomes from The Cancer Genome Atlas. Our data shows that SMV may distinguish a subpopulation of African American patients with colorectal cancer, ~33% of the population in this study. Further, for liver cancer, a higher rate of SMV may be indicative of earlier age of onset. In conclusion, the work presented here suggests that MSI should be expanded to include SMV, not only instability.

DEDICATION:

To my wife Natalie, who never gave up on me and kept battling Eva and Raanan to let me work on my dissertation. My partner in life and work.

To Eva who's favorite quotes for the last two months were "No, no daddy work! Daddy home!", "Daddy, brush teeth and Wowo game" and "Daddy, stay, sleep with Eva". To Raanan who always had a smile and a diaper for me between 3 and 5 am morning, usually as soon as I got to bed. To all three (Natalie, Eva and Raanan) of you who never let me forget to come home at night and gave me plenty of reasons to keep going.

To my sister and Ilya who truly understood why I gave up a great job and career to do this.

To my parents, this really needs no explanation, I think everyone understands the role they played in this process. To that I say THANK YOU!

To Dr. Jeffry Levi (Seton Hall University, Dept of Psychology), Dr. Laksmi Putcha (NASA, Johnson Space Center) and Dr. Heidi Kaplan (University of Texas, Houston Medical School, Microbiology) my mentors in life. You kept me from straying to far off the path.

To Skip, my mentor, who kept me focused when there were too many distractions. The only person who understood that programming was not a difficult skill to learn and only gave credit for the science.

And finally to my committee, you really came through for me at the last moment. Dr.s Carla Finkielstein, Stan Hefta and Jason Holiday thank you. You guys really are what a PhD committee should be, especially at the end when its most difficult.

ACKNOWLEDGEMENTS

This work was funded by the VBI MIS Division director's funds, VBI Genomics Research Lab Small Grant (CLF-1172), high performance computing was supported by a grant from the NSF (OCI-1124123) and NSF S-STEM grant (DUE-0850198). I would like to thank the system administrators in the VBI computational core for technical support. I would like to thank members of the VBI Genomics Research Lab and Beckman-Coulter Next-gen sequencing group. All the sequencing was split between these two centers; both essential to the project's success.

Table of Contents

LIST OF FIGURES:

Chapter 1:

Chapter 2:

LIST OF TABLES:

Chapter 1: Introduction and overview: In search of somatic variation in microsatellites using Next-Gen sequencing: mechanism of mutation and detection method development.

OVERVIEW

Microsatellites (MSTs) are short tandem repeats of motifs, 1 – 6 nucleotide in length, and are considered mutational "hot-spots" and show a greater degree of somatic variability and population polymorphisms than surrounding DNA sequences. Multiple studies have shown that MST mutation rates can exceed the rates of non-repetitive DNA sequences by 10 – 1000 [1-4]. The high mutation rate is a result of the unique physical features associated with MSTs such as complex DNA structures from AT or GC rich sequences or the formation of DNA loops that make the locus vulnerable to DNA breaks or polymerase slippage [1,5-7]. The consequence of these features is a shift in the mutational spectrum from SNPs in non-repetitive DNA to indels. However, although the mutational rate of MSTs is distinctly higher than other DNA sequences, most MSTs, including single nucleotide repeats, are generally stable with only a handful of motifs and loci showing population variability [8]. In the past 20 years a form of somatic variability, MST instability (MSI), has gained great interest because of its link to the prognosis and treatment of colorectal, endometrial and other cancers. For example, MSI-high colorectal and ovarian tumors respond better to chemotherapy, thereby reducing the need for surgical intervention and improving prognosis [9,10]. Conversely, breast cancer patients with MSI-high tumors show an increased likelihood of relapse after successful initial treatment [11].

Novel genomic polymorphisms that arise within a cell population play a critical role in cellular reprogramming, leading to the development of cancer and other aging disorders [12]. Somatic mutations, as they are known, result from stress or chemically induced

DNA damage, transcription or inappropriate nucleotide insertion during replication. Suppression of somatic mutations is essential for genomic stability, therefore cells have evolved multiple repair mechanisms to compensate for DNA damage [13-15]. Congenital DNA Repair Deficiency Disorders (DRDDs) is a conglomerate of inherited conditions, whereby affected individuals are unable to appropriately repair either single or double strand DNA damage. These often lead to genomic instability, cell cycle arrest, cell death and a high rate of neuronal cell pruning [16-19]. These disorders are characterized by neurological deficiencies, poor skin pigmentation, short stature, photosensitive and a predisposition to cancer at various stages of life. Although genomic instability is a hallmark of DRDDs it is often measured with regards to gain or loss of heterozygosity (GOH or LOH), or chromosomal structural variation rather than a direct measurement of the increase in somatic variability. As a result, little is known about the prevalence of somatic mutation or changes in somatic variability in these patients, including those that have developed cancer. With the only major exception being Lynch syndrome, a mismatch repair deficiency that leads to MSI colorectal or endometrial cancer [9,10,20,21].

***MST definition and sequencing alignment**:*

Microsatellites (MSTs) are defined as short tandem repeats of 1 – 6 nucleotide motifs, such as AAAA or ACACAC, which can recur in as few as 3 cycles to spanning thousands of nucleotides. There are several ambiguities as to what actually qualifies as a MST, purity of the sequence and minimal length are the most variable aspects described in the literature[1,2,4]. MSTs as short as 5 nucleotides have been described in the

literature, these usually only apply to single nucleotide runs. The minimal number of full cycles has also varied from study to study, with no clear resolution. However, the general consensus is that 3 full cycles (other than single nucleotide repeats) are sufficient to call these loci MSTs, with no real maximum. The other issue is sequence purity. Sequence purity plays a key role in mutagenesis (discussed later) and therefore can sway experimental results. Therefore, the generally accepted rule is a minimum of 85% purity is acceptable; meaning a 20 nucleotide MST sequence can have 2-3 non-motif nucleotides or 1-2 indels. For the purpose of the work described here the operational definition of a MST is a repetitive sequence spanning 8 or more nucleotides with 3 or more cycles and 85% purity, allowing for 2 frameshifting indels or 3 SNPs for every 20 nucleotides.

Although it is an exaggeration to state that researchers have generally ignored MSTs, research has generally been very focused on a small number of disorders that are caused by tri-nucleotide repeat expansions. The predominant reason for this is that MSTs are more difficult to work with than non-repetitive regions. MSTs are susceptible to polymerase error, sequencing error and computational alignment difficulties [1,3]. MSTs are often AT or GC rich leading to a higher rate of Polymerase Chain Reaction (PCR) errors than non-repetitive DNA. Further, polymerase slippage and processivity are major issues for shorter motif MSTs such as single – tri-nucleotide repeats [22-24]. However, with the advancement of genomic sequencing technology and computational genome sequencing analysis another issue has come to the forefront, sequencing alignment.

MSTs provide for a unique computational alignment problem for many commonly used algorithms, and therefore required software tools to be developed to specifically address these issues. The two major issues are SNP/indel calling and genotyping [25-28]. Since MSTs are repetitive elements, the genotype is not assigned based on a nucleotide at a chromosomal position but is assigned based on the repeat length. For example, a genotype is not a T/T at chr1:892376 but $(CTT)_5$ at chr1:892376-892391 or the actual sequence as a whole. This creates a problem for mapping and genotyping software, for it requires the read to contain the full MST with flanking sequences, otherwise the mapper software can't correctly align the read to the MST and therefore the allele genotyper can't correctly identify the allele. The reason for this is explained in figure 1-1. Until recently, the highly accurate methods (Illumina and SOLiD) for high-throughput sequencing would only produce 35-50 nucleotide read lengths, the significance of which is that only MSTs smaller then 20-25 bases could fit within the read at sufficient coverage to genotype [3,4]. To resolve these issues, all current mapping software masks or disables alignments at repeated DNA sequences, focusing only on the non-repeat regions (see instruction manuals for BWA [29] and Novoalign (Novocraft, Selangor Malaysia) .

The issues with mapping MSTs meant that until recently only a small number of very short MSTs, and mainly single or di- nucleotide repeats could be reliably studied using next-gen genome sequencing. However, with the recent increases in read lengths, base calling accuracy and coverage, as well as the significant decrease in cost, new analysis methods for MST genotyping have also been developed. Over the past 3 years multiple groups, including the Garner lab, have developed highly accurate MST genotypers.

RepeatSeq, lobSTR and our in-lab software take advantage of the aligners flanker mapping and provide genotyping based on in-read MST allele length, but removes the ability to call SNPs [25-27,30,31]. In this dissertation I describe a modification to one of our in-lab developed programs, GenoTan [27,31], to enable us to study somatic variability in microsatellites (SMV) using Next Generation Sequencing (NGS) data (described in chapter 2). Our new MST multi-allele caller has been designed to report read coverage and sequence variation from the predominant allele for all captured alleles, including embedded SNPs.

MST polymorphism trends:

MST polymorphism rates and modes can vary greatly based on motif size and nucleotide composition as well repeat cycles of the MST motif. Polymerase error rates have been recorded for various MST unit lengths and sequences both *in-vitro* and *in-vivo*. As previously stated, error rates for MSTs have been found to be over 1000 times greater than non-repetitive DNA. Single-nucleotide repeats have been shown to be most vulnerable, with an error rates for $[A/T]_{8+}$ repeats shown to be in excess of 3.1×10^{-2} per nucleotide for non-processive polymerases [22-24,32,33]. Similar rates, between 9×10^{-4} and 1.7×10^{-2} per nucleotide, have been found for di, tri and tetra-nucleotide $[NN/N/N]_{7-19}$ repeats *in-vitro*. *In-vivo* mutation rates for single – tetra-nucleotide repeats in human cell lines lacking mismatch repair, which is required to repair polymerase slippage, have found to be as high as 5×10^{-5} per nucleotide for single-nucleotide runs. Error rates for di –tetra-nucleotide repeats range between 0.2 and 5×10^{-5} per nucleotide depending on the motif and cycle number [4,24]. Significantly higher rates have also been shown in

various studies, however, most are for highly A/T or GC enriched sequences such as GGC, AT, AAAC/AAAT repeats.

Analysis of genomes found in the 1000 Genomes Project (1KG), a publicly available compendium of sequenced genomes from healthy volunteers from various backgrounds, has enabled the measurement of MST polymorphism rates in the general human population. Results from various analyses have demonstrated that the preferred mode of MST mutations is expansions and contractions (indels) rather than SNPs, as found for non-repetitive DNA. A recent genomic population survey of 178 individuals (phase 1) from the 1KG project, published by Montgomery et al. [34], has shown that approximately 40% of all population polymorphisms containing indel variants reside within MSTs. Considering that MSTs make up only 2-3% of the genome it is clear that MSTs are enriched for indels. The Makova group, which has done the most extensive analysis of this data for MST variations, determined polymorphism rates in the sequenced population for single – tetra-nucleotide repeats in Asian, European and African individuals [4,34]. Results from their studies show that rates are a function of both motif cycles and total length. Single nucleotide repeats show a log-linear increase in polymorphism incidence rate from ~0.05% for 2 motif copies to ~ 6 – 7% for 10 motif units, independent of race. On the other hand di – tetra nucleotide repeats show a bi-phasic increase in polymorphism incidence rate with an exponential rate increase from ~ 0.01 - 0.05% incidence for 2 motif copies until crossing the 1% mark, at which point they begin an asymptotic phase terminating at 10% incidence rates [4]. Interestingly, the 1% threshold crossing point for the mono – tri-nucleotides MST motifs is at 8.3, 4.67 and 3.4

cycles respectively, or ~ 8, 9 and 9 nucleotides in length. These results confirm that the total MST sequence length, not only motif length and nucleotide sequence, play a role in MST stability [3,4]. Again, microsatellites are currently defined as tandemly repeated motif without a set minimal number cycles, however based on work by the Makova group and Dechering et al. and Lai et al [3,35,36] a definition to include a minimal threshold should be used to remove noise from the analysis. For single nucleotide MSTs a 9 nucleotide "run" appears to be optimal [3,36]. Similarly, unpublished modeling data from our work, which was prepared for chapter 2, suggest a minimal threshold of 8 nucleotides for single nucleotide MSTs. A minimal threshold for di and tri-nucleotide is less clear however, all the reports, including our unpublished data suggest 3 – 5 repeat units show a significant increase in mutational rates per nucleotide as compared to either non-repetitive DNA or MSTs with less cycles [3,36-38]. For the work presented in this dissertation, we ultimately based our definition of MSTs to tandem repeat containing loci that meet the following criteria: 1) minimal length of 8 nucleotides, 2) minimum of 3 complete cycles and 3) 85% purity. Although purity is known to play an important role in mutational rates of MSTs, little work has been done on the topic.

Brief overview of polymerase slippage and repair:

MSTs are significantly more polymorphic in the general population than non-repetitive regions. The predominant model of mutation for microsatellites is polymerase slippage during replication or repair due to the misalignment of the DNA after denaturation, causing an expansion or deletion of the MST. In this model, the replication DNA helicase, bound to the leading strand polymerases (Pol ε) processes the DNA at similar

rates to the leading strand polymerase. Once a repeat is encountered, the helicase opens the DNA at the same rate, however the polymerase is significantly hampered and one of two things occur, leading to the same result [39-41]. The slowdown either allows the repeat to self-anneal, as in the case of tri-nucleotide repeats, or to maintain coupling with the helicase the polymerase bypasses the sequence [42,43]. In either case, the newly formed daughter strand is shortened and the DNA strands re-anneal to accommodate the shorted sequence leavening the skipped portion of the DNA to be recognized by MisMatch Repair (MMR) proteins. On the other hand, if the loop is formed on the daughter strand, as the DNA self-anneals and renatures after replication, this causes insertions leading to lengthening of the repeat sequence. This second scenario is most common for hairpin loop forming sequences such as tri and tetra-nucleotide repeats as well as some of di-nucleotide repeat motifs [44]. Slippages occur in both leading and lagging strands as well as hairpins and loops can occur in both strands simultaneously, leading to a "4-way junction" which is more difficult for a cell to repair correctly, and can require double strand break genes for repair [43].

Replication stalling and restarting also has a significant effect on MST expansions and contractions. Polymerases at MST loci in the replication forks, especially long-track MSTs, show difficulty in traversing the region and can completely stall in the process. In such cases it is believed that the polymerase will back up and utilize the daughter strand from the opposite leading or lagging strand to generate the sequence and restart replication [41,43,45]. This creates a 4-way fork for which resolution is not completely clear [46]. But this is consistent with large scale expansions found that obey the threshold

hypothesis of MST expansions described for disorders such as ataxia, fragile X and other long-track tri-nucleotide repeat expansions [43,46]. A more in-depth review of the various DNA loops created by MSTs after DNA denaturation is found in Mirkin 2007 as well as McMurray 2008 [40,47].

Mismatch repair and MST instability:

During DNA synthesis insertion of erroneous nucleotides occurs with various frequencies, based on the DNA region replicated and recessive polymerase. Studies in cell lines by the Eckert lab estimate the error rate for replicative pol δ and ε or repair (pol κ) polymerases to exceed 10^{-3} for single, 10^{-4} for di-nucleotide and tetra-nucleotide repeats, depending on length and motif sequence [32]. Unrepaired, the errors can integrate into the genome and cause genomic instability leading to cellular dysfunction and potentially cancer (figure 1 – 2). These nucleotide insertion errors are not corrected by the polymerase but are repaired by the trailing proofreading mechanism. The mismatched nucleotides produce looping bulky adducts that are recognized by 1 of 2 MMR complexes, analogs of the prokaryotic MutS [39,48,49]. However, unlike prokaryotes, eukaryotic cells do not have MutH strand recognition to determine the maternal strand. That is bypassed by the fact that eukaryotic cells do not have a nicase and therefore have to use the existing nicks from the incomplete replication [50].

Replication mismatches, either SNPs or indels, are recognized by one of two DNA binding complexes that closely trail the replication fork. Mismatch recognition is based on error size; short 1-2 nucleotide mismatches are detected by MSH2/MSH6 (MutSα)

heterodimer complex while indels and longer mismatches are detected by MSH2/MSH3 (MutSβ) complex MutS [48,49]. Repair of erroneous inserts on the leading strand is initiated once the MutS complex forces a dissociation of the polymerase from the DNA without resolution. The unresolved DNA is used for rescission. As for the lagging strand, the complex stalls polymerase processing and rescission is conducted with the use of the regions between Okazaki fragments. Mismatch recognition enables the MutSα or β complex to associate with the eukaryotic MutL analog composed of MLH1 and either PMS1 or PMS2 [51-53]. The various MutS-MutL combinations differ in their activity at DNA mismatches enabling the recognition complex to associate with proteins required for repair [52,53]. Association of the MutS-MutL with PCNA, a helicase, dissociates the polymerase from the DNA and stalls replication. Rescission, by the exonuclease EXO1, removes the erroneous and surrounding nucleotides from the daughter strand. This step is followed by re-association of the polymerase with the DNA to complete the repair and reinitiate replication at the stalled fork [39]. A simplified visual model of this process is found in figure 1 – 3.

The function of MMR is to remove the unmatched base insertions and deletions that emerge as a result of DNA polymerase errors during DNA synthesis. MSTs are especially susceptible to MMR protein dysfunction. Familial impairment of MMR is known as Lynch-syndrome, which represents 7-10% of all cases of colorectal cancer [20,54,55]. Lynch-syndrome is an autosomal-dominant DNA repair deficiency disorder that causes an augmented progression of carcinogenesis due to mutations in a small number of MMR genes. MLH1 and MSH2 contribute approximately 70%+ of Lynch syndrome patients

while MSH6 and PMS2 make up to ~27% of the rest [50,56,57]. The remaining 3-5% of the patients have an EPCAM mutation, which is upstream of MSH2. Lynch syndrome can be classified into two categories; hereditary site- specific nonpolyposis colonic cancer (HSSNC) predominantly leading to early onset of proximal colon nonpolyposis colorectal carcinoma and cancer family syndrome (also known as Lynch syndrome II), which is more associated with endometrial cancer rather then colonic [32,56,57].

MSTs and DNA repair – implications for disorders:

Again, MSTs are considered mutational "hot-spots", but unlike non-repetitive DNA, MSTs have a distinct bias for indels, rather than SNPs. For many years it was believed that MSTs are prone to expansion, and this was based on data from CAG and CGG tri-nucleotide repeats associated with specific disorders. However, recent evidence suggests that dynamic changes are very complicated and are dependent on multiple variables including MST repeat units, nucleotide length, and motif. Evidence suggests that MSTs most prone to polymorphism are single nucleotide runs, specifically A and T nucleotide runs. Difference in AT and GC content also plays a key role in the rate of mutation [4,8,21,30,58,59]. Many MSTs, especially those in promoter and exonic regions, are under significant selective pressure and therefore MST genomic localization is also important [59-62]. These SMV trends or biases are significantly changed in cells with impaired MMR. Cells with impaired MutS or MutL complexes (Lynch or Muir-Torre syndrome), two of the three essential complexes required for removal and replacement of incorrect nucleotides, show a significant increase in SMV regardless of genomic localization [20,21]. For these disorders, although the predominant mutated motif is

mono-nucleotide, other motifs including di- and tetra-nucleotide MSTs, also show an increase in somatic mutations [20,21,31]. Further, analysis of sequenced exomes of cells lacking a functional BRCA2 gene or those with an impaired Fanconi anemia pathway revealed that unlike the LOH found in these cell lines, the MutL/S (-) colorectal tumor cell line DLD-1 displayed a significant increase in the number of heterozygotic loci [31]. These results suggest a difference between single strand break repair (SSBR) and double strand break repair (DSBR).

Aging, progeroid disorders and DNA repair:

Aging is a progressive deterioration of tissue and function that is poorly understood from a molecular biology perspective. On a genetic/genomic level, aging is associated with an accumulation of somatic mutations in dividing and terminally differentiated cells [63-65]. The gradual increase in mutational load can lead to either an induction of apoptosis or a general reprogramming of cells. The latter is thought to be a major cause of increased cancer susceptibility during aging. Mutations can be induced through a variety of methods leading to various effects on the genome. Single strand damage can include nucleotide mismatch during replication, nucleotide cross-linking or chemical changes such as amination/deamination and oxidation, among many. These types of damage lead to SNPs or short indels. On the other hand double strand damage such as chemically or radiation induced inter-strand crosslinking or double strand breaks, can lead to the loss of large genomic regions, in excess of 3kb. Mutations in any of the single or double stand DNA repair genes can have dire consequences at both the cellular and organism level. Disorders associated with loss of DNA repair show various common characteristics that

include stunted physical growth and development and a condition known as progeria (accelerated aging) [16,66]. A list short list of DNA repair disorders, those currently of interest to our lab, are listed in table 1 - 1.

To maintain genomic stability, cells have evolved multiple mechanisms for repair including multiple single and double strand DNA repair processes. To date only MMR is known to affect MSTs but recent evidence suggests that nucleotide excision repair (NER), a single strand DNA photoproduct and cross-links repair complex, is also responsible for trinucleotide instability [67-69]. NER consists of two distinct pathways a general global NER (GG-NER) and transcription coupled NER (TC-NER) [39,70,71].

Impairment of either GG or TC-NER pathways results in a rapid accumulation of mutations and a significantly shortened life span [72-75]. Cockayne's syndrome is caused by mutations in either ERCC8 or ERCC6 (the genes are also known as CSA and CSB respectively) leading to deficiency in TC-NER [74,76,77]. This disorder is considered one of the most severe of the progeroid conditions with death usually resulting before age 20. Individuals with Cockayne's syndrome usually experience a retardation in neuronal, skeletal/muscular and organ development. The disorder is also characterized by extreme photosensitivity. Surprisingly, unlike other progeria-like disorders Cockayne's syndrome patients do not have a predisposition to cancer, the reason for which is unknown. Xeroderma Pigmantosum, a progeroid disorder caused by impairments in GG-NER, is less severe, with a significantly longer life span [78-80]. Like with Cockayne's syndrome, patients also exhibit photo-sensitivity but also a high cancer rate [81,82].

Xeroderma pigmantosum (XP) associated genes, 12 known to date, include most of the ERCC gene family as well as several other genes, including polymerase and rescission genes [78,83].

Neither TC or GG-NER have yet been implicated in MSI cancers, however recent work has shown that CSA and XP genes do play a role in MST expansion [84-86]. The McMurrey group as well as Zhoa and coworkers have shown in several papers that in mice and human cell cultures, ERCC6, ERCC1 and ERCC2 play a role in CTG expansions by inhibiting MMR during crosslink repair, that can lead to fragile-X syndrome and other diseases [66,84-87]. On the other hand, a mouse cell line knockdown of CSB or XPG can lead to transcription induced CAG contractions by interaction with TFII [68,87,88]. In humans, the effects of impaired NER are currently unknown since no MSI studies have been conducted on any of the NER associated disorders.

The ATP dependent RecQ helicase family of proteins are DNA binding proteins that unwind and denature DNA strands during DNA or RNA synthesis and therefore play a key role in DNA replication, transcription and repair. The human RecQ family consists of 5 proteins RecQL1 – 5 of which 3 are associated with well-established etiologies [89-91]. Mutations leading to loss of or impairment of RecQL2 (WRN) causes a condition known as Werner syndrome, RecQL3, also known as BLM, is associated with Bloom syndrome while functionally impaired RecQL4 (RTS) is the cause of Rothmond-Thomson syndrome. All three syndromes are considered progeroid disorders and are characterized by small stature, premature aging, diabetes, retarded neuronal development

and cancer predisposition, including various rare cancers [90-98] with differences in life-span, phenotype severity and facial features among the disorders. WRN, BLM and RTS interact directly and indirectly via both single and double strand DNA repair pathways. RTS and BLM are both directly associated with base excision repair, while WRN and BLM have both direct and indirect roles in NER and homologous recombination [39,99,100]. However, like with Cockayne's syndrome and XP patients, the effect of RecQ homolog inactivation on MSTs is unknown in these patients.

Hypothesis and experimental approach:

MSTs are complex DNA motifs that due to their physiochemical structure are mutational hotspots. These motifs show a significant increase in polymorphism within a cell line and across same species populations relative non-repetitive DNA [1]. Due to their mutation rate, certain MSTs have been used as forensic "fingerprint" markers for personal identification [101]. MSTs mutations, especially tri-nucleotide repeats, are also associated with various disorders (~40) such as Huntington's Coria and fragile X [1,41,102]. MST instability, defined as an increased global MST mutation rate due to MMR dysfunction, is associated with the prognosis and treatment options for colorectal and endometrial cancers [20,103,104]. Although current tests for MSI have been based on a small number (5 loci) of genetic markers known as the Bethesda markers [32,104,105], the use of Next-Gen sequencing enables researchers and healthcare professionals to probe more loci and expand the usefulness of MSI to other disease types [20,32].

This thesis is a compilation of work conducted on MST stability and somatic variability in select congenital DNA repair disorders as well as liver and colorectal cancer. **My overarching hypothesis is that MST somatic variability can be used to distinguish between cancer subpopulations, and may represent a new actionable clinical diagnostic approach. Further, the patterns of SMV are distinguishable based on specific DNA pathways that are defective in the various DNA repair disorders.** The hypothesis is divided into three experimental components; the first is the development of the method by which we analyze SMV in cell or patient samples using next-gen sequencing data. The second is a proof of concept using cell lines that model various disorders. The third is the use of the method to analyze existing cancer datasets, colorectal cancer and liver cancer, as demonstration of the potential utility.

Approach – SMV analysis: The introduction of NGS enabled detailed research on global genomic scale, however, due to their sequencing error rates and alignment difficulties MSTs have largely been understudied. With the recent increases in read lengths, base calling accuracy and coverage, as well as the significant decrease in cost, have led to the desire to measure MSTs accurately to determine their utility to the further understanding of biology and potential for medical impact. This incentivized us to develop new analysis methods for somatic variation [106] and MSTs genotyping. Over the past 3 years have also led multiple groups to also develop accurate (debate-able) MST genotypers [25-28]. In the second chapter we describe a modification to one of our in-lab developed programs, GenoTan [27], to

enable us to study SMV using NGS data (described in detail in the approach sections of the proposal). The new MST multi-allele caller was developed to overcome issues with sequence analysis of repetitive DNA sequences. To determine alleles, the software uses the whole targeted sequence as a unit and alleles are variant sequences that are substantiated by a defined number of reads at a given locus. Unlike other MST callers previously published [25,27] our software outputs all the alleles that have passed the various filters as well as outputting the differences between the genotype allele and the non-genotype alleles (minor alleles or variant alleles).

TABLES:

Table 1-1: A list of commonly studied DRDDs, the genes affected, cancer predisposition and age of onset for cancer. HR – Homologous Recombination, CLR – Crosslink Repair, NHEJ – Nonhomologous End Joining, MMR – Mismatch Repair, BER – Base Excision Repair, NER – Nucleotide Excision Repair.

Syndrome Name	Gene/s	Pathway	Cancer predisoposition	Onset
Predominantly double strand break repair				
Bloom Syndrome	BLM	HR	leukemia, lymphoma, colon, breast and more	Early - before 20
Rothmund-Thomson	RecQL4	HR	basal cell carcinoma, squamous cell carcinoma, and Bowen's diseas	By 30 a rate of 90%
Fanconi Anemia	FANC(A-R)	HR and CLR	leukemia, liver tumor and solid tumors	Early onset but depends on the gene
Werner Syndrome	WRN	NHEJ and BER	colorectal, skin, thyroid, and pancreatic and soft tissue	Most get cancer by 25
BRCA1/2 mutation	BRCA1/2	HR	breast, overian, lymphoma and others	Mid to late-life
Predominantly single strand repair				
Lynch Syndrome	MLH1	MMR	colorectal, endometrial and ovarian cancer	Early to mid life, depends on severity
Cockayne Syndrome	ERCC6-8	BER and NER	No or Unknown	
Xeroderma Pigmentosum	XP (A-G)	NER	skin cancer (melanoma and non-melanoma) and central nervous system	Onset before 15, 40% survival by age of 20
Trichothiodystropia	ERCC2/3	NER	skin cancer (melanoma and non-melanoma) and central nervous system	Early - mid life

FIGURES:

Figure 1-1: This figure highlights the various problems associated with aligning sequencing reads to MSTs. Here the reference is represented by read R. Aligning a read 1 with one flanker sequence that terminates in the middle of the MST can't be used to determine the length nor the position of the nucleotides. Even if the read contains a MST that is the same length as the reference sequence, as seen with reads 2 and 3 but without flanking nucleotides it is still impossible to predict the actual MST length. In this case, the MST can be longer than the reference but there is no way of determining this. The importance in maintaining flanker sequence length is depicted in read type 4, where the flanker sequence is below the set threshold thereby making it impossible to determine if it is part of the flanker sequence or an indel. This factor is even more apparent with read type 5, here a mutation in the flanker sequence makes it difficult to distinguish between the flanking region and the MST. In cases such as read 5 the determination is made by the alignment software (such as BWA or GATK), with confirmation by the MST caller. Read type 6 is the most commonly used read type for alignment of MSTs, and is the most reliable for accurately calling the microsatellite genotype. This read has the minimum number of flanking sequences and is aligned to the reference. Although in this example the MST sequence is the same as the reference, an expansion or a contraction will also be easily aligned to the reference using the flanking regions.

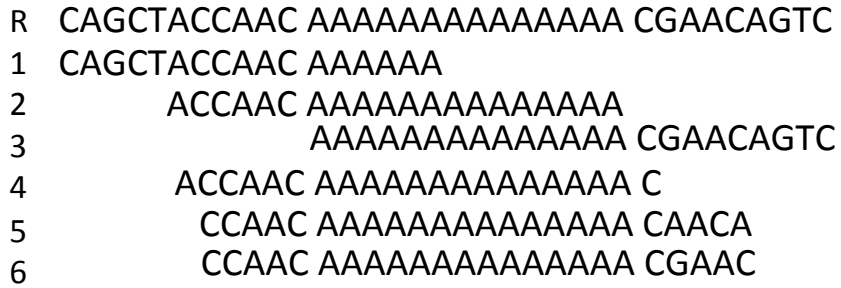


Figure 1-2: A model depicting a comparison of the different malformed DNA products that are created during replication. The first is a DNA mismatch due to inappropriately inserted nucleotide. Incomplete or complete palindromic repeats such as tri-nucletide repeats, self-anneal forming a hairpin. This forms a large malformation for which MutSβ but not MutSα has a high affinity. Single nucleotide repeats on the other hand form B loops with as few as a single nucleotide or more. For a loop with 1 or 2 nucleotides MutSα will initiate the resolution process however a bigger loop may require more complex repair and is the domain of MutSβ. Once a bulky product is detected by a functional mismatch repair complex the resolution maintains the mother strands' periodicity. However in the absence of a functional mismatch repair the slippage is incorporated in the next round of division or unrelated damage repair as an indel.

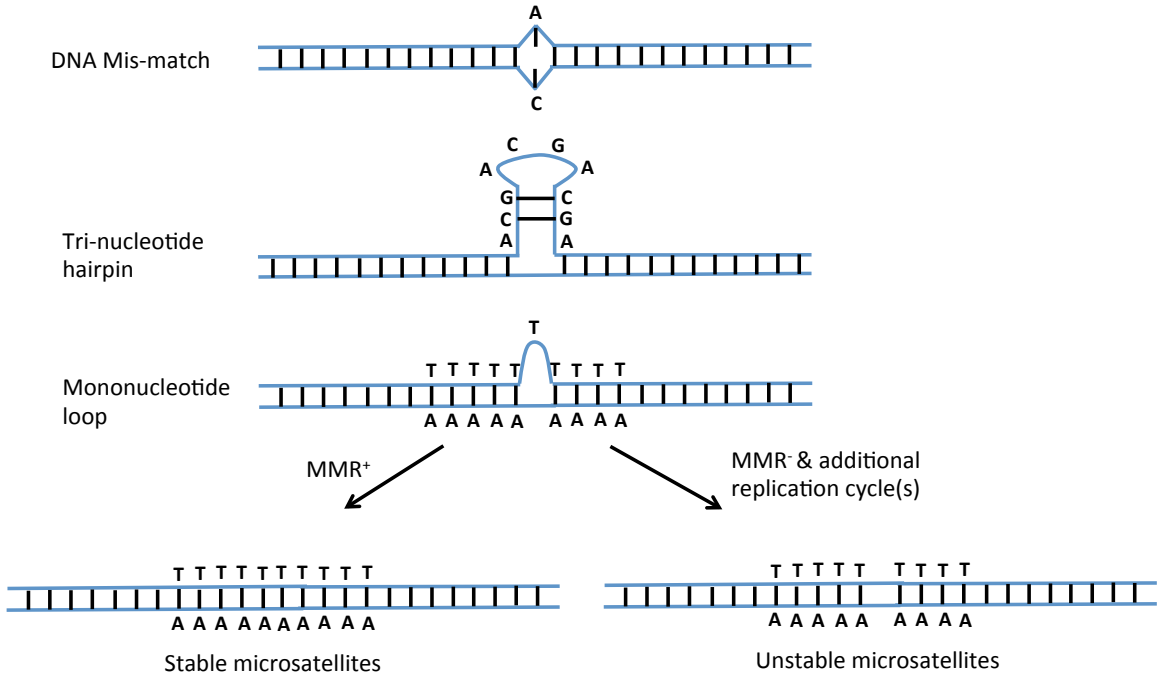


Figure 1-3: Polymerase slippage and replication coupled Mismatch repair of microsatellites. A) Slippage by a replication polymerase at a microsatellite causes a structural DNA adduct. B) A replication coupled mismatch repair complex heterodimer MutSα displays a high affinity for the bulky DNA product. C) Once bound to the malformed DNA strand; through conformational changes MutSα displays a high affinity to the heterodimeric complex MutL. Once coupled the replication fork polymerase stalls or depending on which MutS and MutL complexes formed the polymerase is ejected from the replication fork. D) Coupling of MutL/S supercomplex recruits various DNA protecting and rescission proteins, e.g. EXO1. It is believed the complex does not recruit a nicases to initiate a strand break instead the exonulease initiates rescission only on the daughter strand at the nearest unprocessed open flap or in between Okazaki fragments. Once rescission is complete E) the polymerase restarts DNA synthesis.

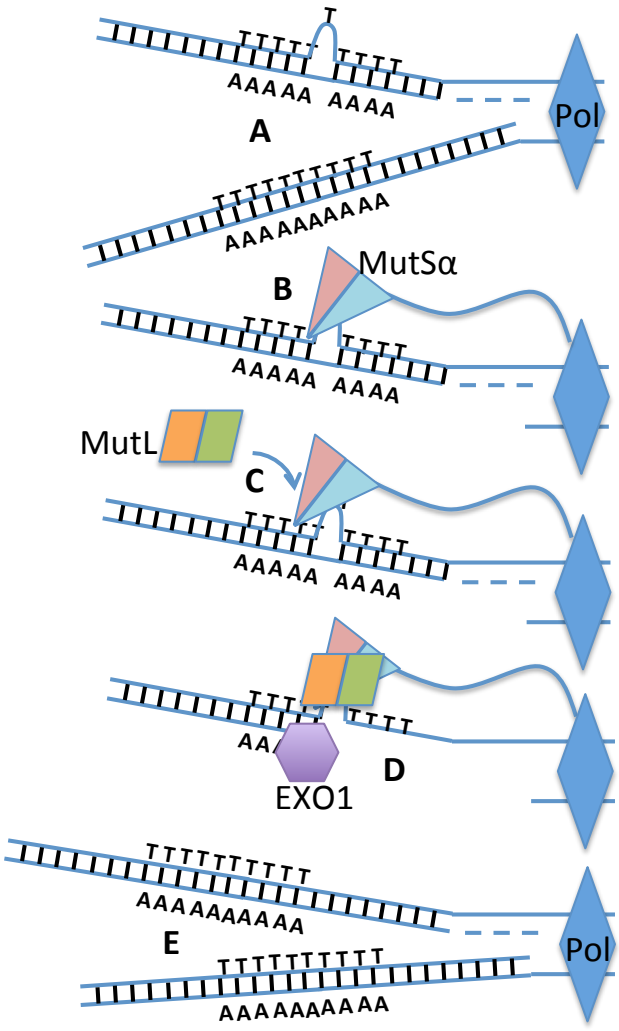
A) Mononucleotide loop during DNA synthesis (dotted line represents Okazaki fragments)

B) Looped DNA is recognized by MutSα MMR proteins.

C) Polymerase stalls and MutL MMR proteins are recruited.

D) Resection of the nascent lagging strand.

E) Re-synthesis across the region.



REFERENCES:

1. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-477.
2. Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, et al. (2013) Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda) 3: 451-463.
3. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, et al. (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol Evol 2: 620-635.
4. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, et al. (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 5: 606-620.
5. Volker J, Gindikin V, Klump HH, Plum GE, Breslauer KJ (2012) Energy landscapes of dynamic ensembles of rolling triplet repeat bulge loops: implications for DNA expansion associated with disease states. J Am Chem Soc 134: 6033-6044.
6. Barros P, Boan F, Blanco MG, Gomez-Marquez J (2009) Effect of monovalent cations and G-quadruplex structures on the outcome of intramolecular homologous recombination. FEBS J 276: 2983-2993.
7. Grzeskowiak K, Ohishi H, Ivanov V (2005) Circular dichroism spectra of d(CGCGCGCGCGCG): evidence for intermediate models in the B-to-Z transition. Nucleic Acids Symp Ser (Oxf): 249-250.
8. Natalie C Fonville LJM, Zalman Vaksman, Harold R Garner (Submitted) Microsatellites in the exome are predominantly single-allelic and invariant. . Genome Biology.
9. Caliman LP, Tavares RL, Piedade JB, AC DEA, K DEJDDC, et al. (2012) Evaluation of microsatellite instability in women with epithelial ovarian cancer. Oncol Lett 4: 556-560.
10. Adem C, Soderberg CL, Cunningham JM, Reynolds C, Sebo TJ, et al. (2003) Microsatellite instability in hereditary and sporadic breast cancers. Int J Cancer 107: 580-582.
11. Regitnig P, Moser R, Thalhammer M, Luschin-Ebengreuth G, Ploner F, et al. (2002) Microsatellite analysis of breast carcinoma and corresponding local recurrences. J Pathol 198: 190-197.
12. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.
13. Harris RS, Kong Q, Maizels N (1999) Somatic hypermutation and the three R's: repair, replication and recombination. Mutat Res 436: 157-178.
14. Stratton MR (2011) Exploring the genomes of cancer cells: progress and promise. Science 331: 1553-1558.
15. Kunz C, Saito Y, Schar P (2009) DNA Repair in mammalian cells: Mismatched repair: variations on a theme. Cell Mol Life Sci 66: 1021-1038.

16. Knoch J, Kamenisch Y, Kubisch C, Berneburg M (2012) Rare hereditary diseases with defects in DNA-repair. Eur J Dermatol 22: 443-455.
17. Lin Y, Wilson JH (2012) Nucleotide excision repair, mismatch repair, and R-loops modulate convergent transcription-induced cell death and repeat instability. PLoS One 7: e46807.
18. Menck CF, Munford V (2014) DNA repair diseases: What do they tell us about cancer and aging? Genet Mol Biol 37: 220-233.
19. Heijink AM, Krajewska M, van Vugt MA (2013) The DNA damage response during mitosis. Mutat Res 750: 45-55.
20. Kim TM, Laird PW, Park PJ (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 155: 858-868.
21. Yoon K, Lee S, Han TS, Moon SY, Yun SM, et al. (2013) Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. Genome Res 23: 1109-1117.
22. Hile SE, Yan G, Eckert KA (2000) Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res 60: 1698-1703.
23. Bagshaw AT, Pitt JP, Gemmell NJ (2008) High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots. BMC Genomics 9: 49.
24. Abdulovic AL, Hile SE, Kunkel TA, Eckert KA (2011) The in vitro fidelity of yeast DNA polymerase delta and polymerase epsilon holoenzymes during dinucleotide microsatellite DNA synthesis. DNA Repair (Amst) 10: 497-505.
25. Highnam G, Franck C, Martin A, Stephens C, Puthige A, et al. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 41: e32.
26. Gymrek M, Golan D, Rosset S, Erlich Y (2012) lobSTR: A short tandem repeat profiler for personal genomes. Genome Res 22: 1154-1162.
27. Tae H, Kim DY, McCormick J, Settlage RE, Garner HR (2013) Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. Bioinformatics.
28. McIver LJ, McCormick JF, Martin A, Fondon JW, 3rd, Garner HR (2013) Population-scale analysis of human microsatellites reveals novel sources of exonic variation. Gene 516: 328-334.
29. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
30. Lauren J McIver NCF, Enusha Karunasena, Harold R Garner (Submitted) Microsatellite genotyping reveals a signature in breast cancer exomes. Breast Cancer Research and Treatment.
31..
32. Hile SE, Shabashev S, Eckert KA (2013) Tumor-specific microsatellite instability: do distinct mechanisms underlie the MSI-L and EMAST phenotypes? Mutat Res 743-744: 67-77.
33. Hite JM, Eckert KA, Cheng KC (1996) Factors affecting fidelity of DNA synthesis during PCR amplification of d(C-A)n.d(G-T)n microsatellite repeats. Nucleic Acids Res 24: 2429-2434.

34. Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, et al. (2013) The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. Genome Res 23: 749-761.
35. Dechering KJ, Cuelenaere K, Konings RN, Leunissen JA (1998) Distinct frequency-distributions of homopolymeric DNA tracts in different genomes. Nucleic Acids Res 26: 4056-4062.
36. Lai Y, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol 20: 2123-2131.
37. Dieringer D, Schlotterer C (2003) Two distinct modes of microsatellite mutation processes: evidence from the complete genomic sequences of nine species. Genome Res 13: 2242-2251.
38. Brandstrom M, Ellegren H (2008) Genome-wide analysis of microsatellite polymorphism in chicken circumventing the ascertainment bias. Genome Res 18: 881-887.
39. McMurray CT (2010) Mechanisms of trinucleotide repeat instability during human development. Nat Rev Genet 11: 786-799.
40. Mirkin EV, Mirkin SM (2007) Replication fork stalling at natural impediments. Microbiol Mol Biol Rev 71: 13-35.
41. Samadashwily GM, Raca G, Mirkin SM (1997) Trinucleotide repeats affect DNA replication in vivo. Nat Genet 17: 298-304.
42. Viguera E, Canceill D, Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. EMBO J 20: 2587-2595.
43. Shishkin AA, Voineagu I, Matera R, Cherng N, Chernet BT, et al. (2009) Large-scale expansions of Friedreich's ataxia GAA repeats in yeast. Mol Cell 35: 82-92.
44. Hile SE, Eckert KA (2008) DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellite sequences. Nucleic Acids Res 36: 688-696.
45. Fouche N, Ozgur S, Roy D, Griffith JD (2006) Replication fork regression in repetitive DNAs. Nucleic Acids Res 34: 6044-6050.
46. Yang Z, Lau R, Marcadier JL, Chitayat D, Pearson CE (2003) Replication inhibitors modulate instability of an expanded trinucleotide repeat at the myotonic dystrophy type 1 disease locus in human cells. Am J Hum Genet 73: 1092-1105.
47. Delagoutte E, Goellner GM, Guo J, Baldacci G, McMurray CT (2008) Single-stranded DNA-binding protein in vitro eliminates the orientation-dependent impediment to polymerase passage on CAG/CTG repeats. J Biol Chem 283: 13341-13356.
48. Iaccarino I, Marra G, Palombo F, Jiricny J (1998) hMSH2 and hMSH6 play distinct roles in mismatch binding and contribute differently to the ATPase activity of hMutSalpha. EMBO J 17: 2677-2686.
49. Huang J, Kuismanen SA, Liu T, Chadwick RB, Johnson CK, et al. (2001) MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. Cancer Res 61: 1619-1623.

50. Hewish M, Lord CJ, Martin SA, Cunningham D, Ashworth A (2010) Mismatch repair deficient colorectal cancer in the era of personalized treatment. Nat Rev Clin Oncol 7: 197-208.
51. Manley K, Shirley TL, Flaherty L, Messer A (1999) Msh2 deficiency prevents in vivo somatic instability of the CAG repeat in Huntington disease transgenic mice. Nat Genet 23: 471-473.
52. Owen BA, W HL, McMurray CT (2009) The nucleotide binding dynamics of human MSH2-MSH3 are lesion dependent. Nat Struct Mol Biol 16: 550-557.
53. Tian L, Gu L, Li GM (2009) Distinct nucleotide binding/hydrolysis properties and molar ratio of MutSalpha and MutSbeta determine their differential mismatch binding activities. J Biol Chem 284: 11557-11562.
54. Siah SP, Quinn DM, Bennett GD, Casey G, Flower RL, et al. (2000) Microsatellite instability markers in breast cancer: a review and study showing MSI was not detected at 'BAT 25' and 'BAT 26' microsatellite markers in early-onset breast cancer. Breast Cancer Res Treat 60: 135-142.
55. Steinke V, Holzapfel S, Loeffler M, Holinski-Feder E, Morak M, et al. (2014) Evaluating the performance of clinical criteria for predicting mismatch repair gene mutations in Lynch syndrome: a comprehensive analysis of 3,671 families. Int J Cancer 135: 69-77.
56. Stoffel EM, Kastrinos F (2013) Familial Colorectal Cancer, Beyond Lynch Syndrome. Clin Gastroenterol Hepatol.
57. Cleary SP, Cotterchio M, Jenkins MA, Kim H, Bristow R, et al. (2009) Germline MutY human homologue mutations and colorectal cancer: a multisite case-control study. Gastroenterology 136: 1251-1260.
58. Siddle KJ, Goodship JA, Keavney B, Santibanez-Koref MF (2011) Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics 27: 895-898.
59. Denver DR, Morris K, Kewalramani A, Harris KE, Chow A, et al. (2004) Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of Caenorhabditis elegans. J Mol Evol 58: 584-595.
60. Mestrovic N, Castagnone-Sereno P, Plohl M (2006) Interplay of selective pressure and stochastic events directs evolution of the MEL172 satellite DNA library in root-knot nematodes. Mol Biol Evol 23: 2316-2325.
61. Williams LE, Wernegreen JJ (2013) Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont. Genome Biol Evol 5: 599-605.
62. Payseur BA, Jing P, Haasl RJ (2011) A genomic portrait of human microsatellite variation. Mol Biol Evol 28: 303-312.
63. Bavarva JH, Tae H, McIver L, Garner HR (2014) Nicotine and oxidative stress induced exomic variations are concordant and overrepresented in cancer-associated genes. Oncotarget 5: 4788-4798.
64. Freitas AA, de Magalhaes JP (2011) A review and appraisal of the DNA damage theory of ageing. Mutat Res 728: 12-22.
65. Hasty P, Campisi J, Hoeijmakers J, van Steeg H, Vijg J (2003) Aging and genome maintenance: lessons from the mouse? Science 299: 1355-1359.

66. Budworth H, McMurray CT (2013) Bidirectional transcription of trinucleotide repeats: roles for excision repair. DNA Repair (Amst) 12: 672-684.
67. Hubert L, Jr., Lin Y, Dion V, Wilson JH (2011) Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. Hum Mol Genet 20: 4822-4830.
68. Lin Y, Hubert L, Jr., Wilson JH (2009) Transcription destabilizes triplet repeats. Mol Carcinog 48: 350-361.
69. Concannon C, Lahue RS (2014) Nucleotide excision repair and the 26S proteasome function together to promote trinucleotide repeat expansions. DNA Repair (Amst) 13: 42-49.
70. Parsons JL, Dianov GL (2013) Co-ordination of base excision repair and genome stability. DNA Repair (Amst) 12: 326-333.
71. Liu Y, Wilson SH (2012) DNA base excision repair: a mechanism of trinucleotide repeat expansion. Trends Biochem Sci 37: 162-172.
72. Andrade LN, Nathanson JL, Yeo GW, Menck CF, Muotri AR (2012) Evidence for premature aging due to oxidative stress in iPSCs from Cockayne syndrome. Hum Mol Genet 21: 3825-3834.
73. Kozmin SG, Jinks-Robertson S (2013) The mechanism of nucleotide excision repair-mediated UV-induced mutagenesis in nonproliferating cells. Genetics 193: 803-817.
74. Nardo T, Oneda R, Spivak G, Vaz B, Mortier L, et al. (2009) A UV-sensitive syndrome patient with a specific CSA mutation reveals separable roles for CSA in response to UV and oxidative DNA damage. Proc Natl Acad Sci U S A 106: 6209-6214.
75. Andressoo JO, Hoeijmakers JH, Mitchell JR (2006) Nucleotide excision repair disorders and the balance between cancer and aging. Cell Cycle 5: 2886-2888.
76. Marteijn JA, Lans H, Vermeulen W, Hoeijmakers JH (2014) Understanding nucleotide excision repair and its roles in cancer and ageing. Nat Rev Mol Cell Biol 15: 465-481.
77. Berquist BR, Wilson DM, 3rd (2012) Pathways for repairing and tolerating the spectrum of oxidative DNA lesions. Cancer Lett 327: 61-72.
78. Lehmann AR, McGibbon D, Stefanini M (2011) Xeroderma pigmentosum. Orphanet J Rare Dis 6: 70.
79. Hayashi M (2008) Roles of oxidative stress in xeroderma pigmentosum. Adv Exp Med Biol 637: 120-127.
80. Fan W, Luo J (2008) RecQ4 facilitates UV light-induced DNA damage repair through interaction with nucleotide excision repair factor xeroderma pigmentosum group A (XPA). J Biol Chem 283: 29037-29044.
81. Jiang J, Zhang X, Yang H, Wang W (2009) Polymorphisms of DNA repair genes: ADPRT, XRCC1, and XPD and cancer risk in genetic epidemiology. Methods Mol Biol 471: 305-333.
82. Sturgis EM, Zheng R, Li L, Castillo EJ, Eicher SA, et al. (2000) XPD/ERCC2 polymorphisms and risk of head and neck cancer: a case-control analysis. Carcinogenesis 21: 2219-2223.

83. Butt FM, Moshi JR, Owibingire S, Chindia ML (2010) Xeroderma pigmentosum: a review and case series. J Craniomaxillofac Surg 38: 534-537.
84. Martins S, Pearson CE, Coutinho P, Provost S, Amorim A, et al. (2014) Modifiers of (CAG)(n) instability in Machado-Joseph disease (MJD/SCA3) transmissions: an association study with DNA replication, repair and recombination genes. Hum Genet 133: 1311-1318.
85. Zhao XN, Usdin K (2014) Gender and cell-type-specific effects of the transcription-coupled repair protein, ERCC6/CSB, on repeat expansion in a mouse model of the fragile X-related disorders. Hum Mutat 35: 341-349.
86. Kovtun IV, Johnson KO, McMurray CT (2011) Cockayne syndrome B protein antagonizes OGG1 in modulating CAG repeat length in vivo. Aging (Albany NY) 3: 509-514.
87. Lin Y, Wilson JH (2009) Diverse effects of individual mismatch repair components on transcription-induced CAG repeat instability in human cells. DNA Repair (Amst) 8: 878-885.
88. Hubert L, Jr., Lin Y, Dion V, Wilson JH (2011) Topoisomerase 1 and single-strand break repair modulate transcription-induced CAG repeat contraction in human cells. Mol Cell Biol 31: 3105-3112.
89. Kitano K (2014) Structural mechanisms of human RecQ helicases WRN and BLM. Front Genet 5: 366.
90. Lowy AM, Kordich JJ, Gismondi V, Varesco L, Blough RI, et al. (2001) Numerous colonic adenomas in an individual with Bloom's syndrome. Gastroenterology 121: 435-439.
91. Diaz A, Vogiatzi MG, Sanz MM, German J (2006) Evaluation of short stature, carbohydrate metabolism and other endocrinopathies in Bloom's syndrome. Horm Res 66: 111-117.
92. Chang C, Shiah HS, Hsu NY, Huang HY, Chu JS, et al. (2014) Jejunal Cancer with WRN Mutation Identified from Next-Generation Sequencing: A Case Study and Minireview. Case Rep Surg 2014: 126924.
93. Vidal V, Bay JO, Champomier F, Grancho M, Beauville L, et al. (1998) The 1396del A mutation and a missense mutation or a rare polymorphism of the WRN gene detected in a French Werner family with a severe phenotype and a case of an unusual vulvar cancer. Mutations in brief no. 136. Online. Hum Mutat 11: 413-414.
94. Goto M, Miller RW, Ishikawa Y, Sugano H (1996) Excess of rare cancers in Werner syndrome (adult progeria). Cancer Epidemiol Biomarkers Prev 5: 239-246.
95. Cai MY, Liang H, Li M, Bi Y, Chen X, et al. (2010) Lamin C protein deficiency in the primary fibroblasts from a new laminopathy case with ovarian cystadenoma. Chin Med J (Engl) 123: 2237-2243.
96. Carlson AM, Lindor NM, Litzow MR (2011) Therapy-related myelodysplasia in a patient with Rothmund-Thomson syndrome. Eur J Haematol 86: 536-540.
97. Simon T, Kohlhase J, Wilhelm C, Kochanek M, De Carolis B, et al. (2010) Multiple malignant diseases in a patient with Rothmund-Thomson syndrome with RECQL4 mutations: Case report and literature review. Am J Med Genet A 152A: 1575-1579.

98. Manthei KA, Keck JL (2013) The BLM dissolvasome in DNA replication and repair. Cell Mol Life Sci 70: 4067-4084.
99. Trego KS, Chernikova SB, Davalos AR, Perry JJ, Finger LD, et al. (2011) The DNA repair endonuclease XPG interacts directly and functionally with the WRN helicase defective in Werner syndrome. Cell Cycle 10: 1998-2007.
100. Bouwman P, Jonkers J (2012) The effects of deregulated DNA damage signalling on cancer chemotherapy response and resistance. Nat Rev Cancer 12: 587-598.
101. Butler JM (2005) Forensic DNA typing : biology, technology, and genetics of STR markers. Amsterdam ; Boston: Elsevier Academic Press. xvii, 660 p. p.
102. Russo MT, Blasi MF, Chiera F, Fortini P, Degan P, et al. (2004) The oxidized deoxynucleoside triphosphate pool is a significant contributor to genetic instability in mismatch repair-deficient cells. Mol Cell Biol 24: 465-474.
103. Xiao H, Yoon YS, Hong SM, Roh SA, Cho DH, et al. (2013) Poorly differentiated colorectal cancers: correlation of microsatellite instability with clinicopathologic features and survival. Am J Clin Pathol 140: 341-347.
104. Lynch HT, de la Chapelle A (1999) Genetic susceptibility to non-polyposis colorectal cancer. J Med Genet 36: 801-818.
105. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, et al. (2010) Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One 5: e15661.
106. Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, et al. (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res 22: 568-576.

Chapter 2: Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies

Vaksman Z, Fonville NC, Tae H, Garner HR (2014) Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies. PLoS One 9: e110263

ABSTRACT

Microsatellites (MST), tandem repeats of 1-6 nucleotide motifs, are mutational hot-spots with a bias for insertions and deletions (INDELs) rather than single nucleotide polymorphisms (SNPs). The majority of MST instability studies are limited to a small number of loci, the Bethesda markers, which are only informative for a subset of colorectal cancers. In this paper we evaluate non-haplotype alleles present within next-gen sequencing data to evaluate somatic MST variation (SMV) within DNA repair proficient and DNA repair defective cell lines. We confirm that alleles present within next-gen data that do not contribute to the haplotype can be reliably quantified and utilized to evaluate the SMV without requiring comparisons of matched samples. We observed that SMV patterns found in DNA repair proficient cell lines without DNA repair defects, MCF10A, HEK293 and PD20 RV:D2, had consistent patterns among samples. Further, we were able to confirm that changes in SMV patterns in cell lines lacking functional BRCA2, FANCD2 and mismatch repair were consistent with the different pathways perturbed. Using this new exome sequencing analysis approach we show that DNA instability can be identified in a sample and that patterns of instability vary depending on the impaired DNA repair mechanism, and that genes harboring minor alleles are strongly associated with cancer pathways. The MST Minor Allele Caller used for this study is available at https://github.com/zalmanv/MST_minor_allele_caller.

INTRODUCTION

Microsatellites (MSTs) are regions of repetitive DNA at which 1-6 nucleotides are tandemly repeated; and are present ubiquitously throughout the genome, both in gene and intergenic regions. Observations of somatic variation in MSTs have demonstrated that MST mutation rates are between 10 and 1000 time higher than that of surrounding DNA [1,2], rendering microsatellites mutational "hot-spots" [3,4]. The increased mutational rate of MSTs is thought to be primarily due DNA polymerase slippage and mis-alignment of the slipped structure due to local homology [5-7]. This difference in primary mutational mechanism suggests that, unlike non-repetitive DNA whose mutational spectrum is primarily SNPs, microsatellites are more prone to INDELs [4,7,8]. Specifically MSTs are prone to INDELs that are 'in-phase' or result in expansion or contraction by complete repeat units. For example, a dimer microsatellite will typically expand or contract by 2N nucleotides while a trimer will expand or contract by 3N [1].

MSTs are found in and around a significant number of coding and promoter regions and specific microsatellite variations have been linked to over 40 disorders, such as the CAG microsatellite whose expansion is associated with Huntington's disease and the CGG repeat whose expansion is associated with Fragile X [1,9]. In addition, a more general increase in MST instability has been associated with colon cancer, which, if detected, results in better prognosis and can influence treatment [10,11]. Currently, MST instability is clinically defined based on the results of a kit that tests somatic variation of 18 - 21 "susceptible" loci (PowerPlex®21, Promega). Although the test has been shown to be effective for identifying MST unstable colon cancer [12], it is significantly less effective

for most other disorders including other cancers [13-15]. The ability to capture and discern variation patterns exome-wide would provide a more accurate and useful clinical data for a broader range of disorders. In recent reports next-gen sequencing has been used to uncover MST instability in intestinal and endometrial cancers by observing genotype changes in MSTs between tumor and healthy tissue [14,15].

The goal of this research was to identify patterns of somatic variation in MSTs as a possible marker for genomic instability. We hypothesize that the variable nature of MSTs and the quantification of minor allele content makes them ideal candidates for in-depth next-gen analysis and that somatic variation of microsatellite loci can be quantified using high-depth sequencing. A broadening of the definition of MST instability to include changes in somatic variability and using an exome/genome-wide approach may enable a more accurate diagnosis of patients then what is currently provided by PowerPlex®21.

Somatic variability, novel genomic polymorphisms that arise within a cell population not found in the progenitors, plays a critical role in cellular reprogramming leading to the development and progression of cancer [16]. Suppression of mutations is essential for genomic stability, therefore cells have evolved multiple mechanisms to repair damaged or unpaired nucleotides [17,18]. Currently the only established DNA repair defect that that has been directly linked to MST instability is mismatch repair (MMR). MMR impairments have been shown to increase somatic variation at MSTs in both cell lines and tumors [19-21]. Although the role other DNA repair mechanisms such as inter-strand crosslink repair (as seen in Fanconi anemia genes) and homologous recombination (HR)

play in MST instability is less clear, both are important for genomic and chromosomal stability (reviewed by [22,23]).

In this study we first show that we can robustly detect signatures of MST mutation bias and somatic variation occurring in cell lines in next-gen data including a high frequency of in-phase INDELs. We are then able to construct a pattern of somatic MST variation (SMV) by using DNA repair proficient cell lines. Our results indicate that ~5% of microsatellite loci show somatic variation, i.e. have at least one additional non-haplotype allele present. Finally, we are able to differentiate between cell lines with known defects in various DNA repair mechanisms (mismatch repair, DNA crosslink repair, homologous recombination), which correlate with an altered distribution of loci with non-haplotype alleles. These findings suggest that signatures that distinctly define specific defective DNA repair mechanisms can be gleaned from next-gen sequencing data and that this information has the potential to be utilized for detection of individuals with altered levels of somatic variation that are at increased risk of disease or the evaluation of patient's tumor that may yield clinically actionable information.

METHODS

Cells, DNA prep and sequencing: HEK (human embryonic kidney) and MCF10A (immortalized breast epithelial) and HEK293 (human embryonic kidney) cells were obtained from ATCC. PD20 and PD20 RV:D2 (FANCD2 and FANCD2 retrovirally corrected) cell lines were obtained from the Fanconi Anemia Foundation (Eugene OR). Sequencing data for Capan-1 cells was previously published by Barber and coworkers [12].

PD20, PD20 RV:D2 and HEK293 cells were grown at 37^{o}C with 5% CO_2, in DMEM supplemented with 10% FBS (Invitrogen) and 1X pen/strep (Invitrogen) to 80% confluence. MCF10A cells were grown to confluence in DMEM/F12 medium (Invitrogen, Carlsbad, CA), supplemented with 5% horse serum (Invitrogen), antibiotics-1X Pen/Strep (Invitrogen), 20 ng/mL EGF (Peprotech, Rocky Hill, NJ), 0.5 mg/mL hydrocortisone (Sigma), 100 ng/mL cholera toxin (Sigma), and 10 µg/mL insulin (Sigma) at 37^{o}C with 5% CO_2. All cell lines were collected by trypsinazation and prepared for DNA extraction. DNA was extracted using the Qiagen DNAeasy kit (Qiagen) as per manufacturers instructions.

Since PD20 RV:D2 were derived from PD20 cells by retroviral insertion of the corrected FANCD2 gene we confirmed the maintenance of the corrected version using the sequencing data. Further, a comparison of growth-curves showed an order of magnitude more cells 48 hours after exposure to the DNA interstrand cross-linker Cisplatin, confirming a partial rescue phenotype.

Sequencing and analysis pipeline: Exome paired-end libraries were prepared using the Agilent (Chicago, IL) SureSelectXT Human All Exon V4 capture library. 2x100 bp reads were obtained using an Illumina (San Diego, CA) HiSeq 2500 instrument in Rapid Run mode on a HiSeq Rapid v1 flowcell. Indexed reads were de-multiplexed with CASAVA v1.8.2.

Paired-end sequencing reads were trimmed using fastX_Toolkit and aligned to HG19/GRCh37 human reference genome (http://www.genome.ucsc.edu) using BWA-mem. The output was then sorted, indexed and PCR duplicates were removed using SAMTOOLS [24]. Bam files were then locally realigned and target loci marked using GATK IndelRealigner and TargetIntervals. MST alleles were retrieved and analyzed using software described in the next section.

Microsatellite minor-allele software: A catalogue of MST loci was generated from the HG19/GRCh37 reference genome using Tandem Repeats Finder [25] (with the following parameters: 2.7.7.80.10.18.6). The list was filtered to remove any loci that were shorter than 8 nucleotides, had less than 3 copies of a given motif unit or were below 85% sequence purity. Duplicated loci were identified based on sequence purity and sequence length and were removed.

MSTs were analyzed using a custom MST minor-allele caller based on GenoTan and ReviSTER software [26,27], which were developed by this group to improve MST

haplotype predictions (https://github.com/zalmanv/MST_minor_allele_caller). The minor-allele caller extracts marked MSTs from bam files using SAMTOOLs. MST loci are called based on predicted alignments and an adjustable length flanking sequence (this study used either 5 or 7 nucleotide sequence). Reads with low base call scores (below a base score of 28) for nucleotides within the repeats and those with mapping quality score below 10% were eliminated. Alleles are initially called only when two or more reads, verified in both directions of a paired-end run, have the same sequence. All alleles for a given locus are binned with the number of supporting paired-end reads. The final number of alleles is computed based on a user specified minimal requirement of substantiating reads (for this study the minimum number of substantiating reads is either 2 or 3 reads per allele). If more than one allele per locus was found, zygosity and the sequence length difference from the most common allele were recorded. Heterozygotic loci were called using the following criteria as described and confirmed in the GenoTan and ReviSTER manuscripts [26,27]: 1) it is the second most common allele, 2) The number of confirming reads is greater than 25% of the total reads for the locus or greater than 50% of the depth for the most common allele, if the total is below 25% of the total depth.

In addition to MST loci, we also generated a somatic variability profile for non-MST loci. To make the data comparable we randomly selected 3 million loci, each consisting of 15 nucleotides segments, from the HG19 genome. We then filtered out any loci that intersected with our MST and were left with over 2 million loci. The same pipeline as for MSTs was used to generate the data for non-MST loci. This data yielded information on the number of loci with minor alleles and type of mutation (SNPs and INDELs).

Sequence validation and allele calls validated by independent Sanger sequencing method. The MST minor-allele caller we use in this paper is a modified version of a published and experimentally verified code, however to further validate the multi-allele capability of the modified code 30 loci, including 17 showing multiple alleles, were verified using Sanger sequencing. Figure S2-1A shows the data from the minor-allele caller output at one of these loci, chr10:72639137-72639161, at which we would predict at least 3 alleles to be present in this sample (MCF10A) with lengths of 21, 23, and 25 nucleotides. Sanger sequencing confirmed that multiple alleles were present, with the alleles being greater than 21 nucleotides long (figure S2-1B). Of the 30 loci 28 loci verified the genotype and 14 of 17 loci with minor alleles also had visible minor alleles by Sanger sequencing.

Modeling error rates to establish rules that differentiate errors from high confidence minor alleles: Two methods were used to generate models of NGS runs for chromosomes 17 and 21; 1) Wgsim (https://github.com/lh3/wgsim) a commonly used paired-end read generator and 2) in-house designed generator. Both methods were set to have a per nucleotide error rate between 0.5% and 5%. The major difference between the two methods was that wgsim was used to obtain modeling data with fairly similar coverage (read depth) across the reference chromosome while the lab-designed algorithm allowed for a more variable coverage as is observed in a typical next-gen sequencing run. The generated fastq files were run through the same pipeline as actual real sequencing data. The accuracy of the pipeline was analyzed by the verification of the predicted alignment.

Predicted error rates ranged between 1.3% and 1.9%, with the majority of errors due to misalignments.

RESULTS

We modified a previously published and verified MST genotyper [26] to enumerate all possible alleles present within next-gen data, as opposed to only capturing the most common (haplotype) alleles. We first characterized the error which may cause false positive allele calls via a parametric sensitivity study conducted on *in-silico* generated data, and showed that our measure can then be used to accurately quantify minor alleles and thus be used to distinguish between mutational mechanisms that are exhibited in different cell lines. To accomplish this, we establish a baseline SMV profile from DNA repair proficient cell lines, and compared this to what is seen in cell lines with various DNA repair defects.

Characterizing the effect of sequencing error on minority allele calling: This analysis evaluates each MST locus to establish the one or two alleles that define the genotype, then it robustly calls additional non-haplotype or 'minor' alleles that are present at lower frequency within next-gen data. However, the accuracy of such minority allele calls can be significantly affected by sequencing errors found within the raw reads that map to each locus. To minimize the number of false positive 'alleles', we first established the minimal number of reads necessary for confirming an allele in the presence of typical next-gen errors. It has been established by a number of studies that 3 reads mapped to a loci is sufficient to properly call major alleles [28-30]. To corroborate this, we created an *in-silico* sequencing data set for chromosomes 21 and 17, with randomly generated errors ranging from 0.5% to 5% which mimicked next-gen sequencing data in both the error types that were created and read coverage per locus (results depicted in figure 2-1).

We first determined the parameters required to optimize the measurement of the fraction of loci *without* minor alleles in sequencing data with the above-mentioned error rates. Alignment and zygosity calling accuracy is displayed in table S2-1. The sequencing data generator produced between 8 and 10.5 million reads that contained over 58,000 targeted MSTs. Over 98.5% of the reads mapped correctly with an accuracy of over 99.8% in coding regions (regions captured by exome sequencing). The accuracy of zygosity calls was over 99.98% for all error rates. Next we varied the minimum number of reads covering a locus required to call an allele. Changing the threshold from 2 confirming reads (figure 2-1A) to 3 confirming reads (figure 2-1B) statistically and significantly decreased the fraction of loci with more alleles than the haplotype number (1 if homozygotic or 2 if heterozygotic). Using a threshold of 2 confirming reads per allele, the fraction of loci without minor alleles identified (due to sequencing errors being interpreted as alleles) was 19 - 62% for simulated data sets with error rates ranging between 5% - 0.5% respectively (figure 2-1A), indicating that requiring only 2 reads to identify an allele leads to a high level of false alleles. By increasing the threshold to 3 confirming reads the percent of loci without minor alleles increases to 73 – 99% for the same data set (figure 2-1B). By increasing to 4 confirming reads per allele we further increase the number of loci without minor alleles 87% - 99% (figure S2-2A). However, at error rates close to the actual HiSeq rates (of ~1%), we only saw a modest increase in the number of loci without minor alleles, a change from 97% (3 reads per allele) to 99% (4 reads per allele). This is in contrast to an increase from 61% with 2 reads per allele to 97% with 3 confirming reads per allele.

We next examined how sequencing error might affect the number of alleles present in our data. To do this we used modeling data with error rates similar to the actual HiSeq error rate (1%) and 2.5% error (figure 2-2), and determined the average read depth per locus with increasing alleles. For the *in-silico* generated data, we found a linear increase in the total read depth as the number of alleles increased (using 2 - 4 confirming reads per allele) up to 8 alleles (figures 2-2 and S2-2). A comparison of these results to actual sequencing data from our cell lines (discussed in more detail later) shows that when 3 or more reads are required to confirm an allele, the number of alleles called for a given read depth is greater than what would be expected from error, even at a rate of 2.5% which is substantially more than the observed next-gen error rate of 1% (figure 2-2B and S2-2B), i.e. more alleles are called at a lower read depth in the actual data than would be present due to error. Based on these results, requiring a minimum of 3 reads covering a locus to confirm an allele minimizes the number of 'false' alleles being identified due to sequencing error.

Polymerase slippage vs. nucleotide misincorperation: Another potential source of error in calling alleles from sequencing data is amplification errors induced during the library preparation process [31]. These errors would likely be present at higher frequency than errors generated during sequencing [31,32]; therefore cannot be minimized by solely increasing the minimum read coverage (as above). Somatic mutation of MSTs is primarily associated with polymerase slippage [33,34], which is thought to cause the characteristic INDEL bias [31,35,36]. In contrast, nucleotide mis-incorperation errors

during *in-vitro* amplification would be predicted to lead primarily to SNPs in sequencing data [37]. Both of the mentioned DNA synthesis methods would lead to an increase in the number of loci with non-haplotype alleles, however with a predicted variation pattern that is distinctly different. To differentiate between the two predicted SMV patterns including minority alleles, and to assess the influence of nucleotide mis-incorporation/amplification error on our results, we compared a standard exome sequence from cells which are proficient for DNA repair (described later) that did not undergo whole genome amplification (WGA) with data from the sequencing of a single cell [38] which would be expected to have no somatic variation within the sample, but has necessarily undergone WGA to generate the quantity of DNA necessary for sequencing. Therefore, for the WGA sample, presumably all non-haplotype alleles present are due to amplification error. As expected, genome amplification increases the number of loci with non-haplotype alleles (figure 2-1) to 11.3% and 7% of the total with a threshold of 2 and 3 reads, respectively. The DNA repair proficient cells, which did not undergo extensive amplification, were only decreased by 1.7%, from 7% to 5.3%, by altering the minimum read cutoff. From this it can be concluded that neither errors during library prep nor during the sequencing run account for more than 4 percent of the total non-haplotype alleles detected.

Approximately 85% of mutations found within microsatellite loci in the WGA single-cell data were SNPs, which is expected as a consequence of polymerase errors during amplification. These results were comparable to those predicted by our model, which showed that ~88% of the total minor alleles were composed of alleles carrying SNPs

rather than INDELs (Figure 2-3). In contrast, SNPs account for only 36% (±3.4%) of the total minor alleles in DNA repair proficient cell lines. In addition, although for all the DNA repair proficient cell lines the most common MST motifs with minor alleles observed were mono-nucleotide repeats found within 56% - 66% of loci, loci containing tri-nucleotide motifs accounted for over 55% of the total loci with minor alleles in the WGA data (table S2-2). These results further support the hypothesis that this approach can differentiate between distinct MST mutational profiles: INDELs, particularly at mono-nucleotide runs predominantly reflect DNA repair proficient biological SMV whereas SNPs in MSTs, particularly at tri-nucleotide motif containing loci are predominantly amplification-induced errors or potentially due to altered DNA maintenance capacity. This is further supported by a similar study that has found that the majority of MSTs that are variable within the normal population (individuals sequenced as part of the 1,000 Genomes Project) are predominantly INDELs at mono-nucleotide runs [30].

MST vs non-MST regions: MSTs are considered to be more susceptible to mutations than the surrounding non-repetitive DNA regions [3,14,39]. Because of this, one could expect that non-MST regions would have less somatic variability (non-MST equivalent of SMV) than MST regions. In order to perform a fair comparison with the MST data, 2 million segments consisting of 15 nucleotides each were randomly selected throughout the genome. The same analysis as was performed on loci containing MSTs was also applied to these non-MST regions. It was found that for these non-MST loci the average fraction of loci that were homozygotic was 98.9% with a standard deviation of 0.2, while only

96.7% of the MST containing loci was homozygotic. Even more significant, only 2% (standard deviation of 0.2) of the non-MST loci (homozygotic and heterozygotic) had minor alleles, while 5.1 % of the MST loci harbored minor alleles (table 2-1). Further, a comparison of SNP and INDEL distributions indicated that, unlike MST regions where INDEL variations prevail (64%), SNPs account for the majority (96.9%) of the differences in minor alleles at non-MST loci (table 2-2). Taken together, these results confirm that, consistent with the literature, MSTs are more susceptible to mutation [2-4,34].

Reproducibility within a cell line: The objective of this study is to characterize the pattern of SMV from DNA repair proficient cells and then compare to cell populations in which DNA repair is compromised. SMV changes associated with disease will likely be subtle and require highly reproducible control data. To test the reproducibility of SMV measurements within a cell line, two biological replicate cultures of PD20 RV:D2 (PD20 RV:D2-1 and PD20 RV:D2-2) cells were grown separately and sequenced. PD20 RV:D2 are fibroblasts derived from an individual with Fanconi Anemia subgroup D2, retroviraly complimented with a functional copy of FANCD2 [40]. Using a minimum read depth cutoff of 15 to genotype a given loci, we successfully called over 280K and 250K loci (at an average depth of 52 and 45 reads per locus) for PD20 RV:D2-1 and 2 respectively. Both samples showed a similar SNP to INDEL ratio, with INDELs making up over ~67% of the minor alleles (table 2-2). A genotype analysis showed that approximately 96.8% of called loci were homozygous while heterozygosity was observed in ~3.2% of the loci called (table 2-1). Comparison of those loci that were called in both samples shows that

haplotype discordance (i.e. homo- or heterozygotic using standard genotyping) was 1.1% (table 2-3), of which 92% were due the fraction of reads supporting a second allele being below the haplotype threshold (see method) and was therefore counted as a minor allele instead of a second haplotype allele, as is the convention in established genotype callers. Only 173 discordant loci were due to sequence differences between the two samples.

For the purpose of this study SMV is defined by the presence of variant MST alleles that are supported by a minimum of 3 confirming reads but do not contribute to haplotype. An analysis of variant MST alleles found a total of 5.4% and 5.3% of MST loci in the PD20 RV:D2-1 and 2 samples, respectively, had 1 or more minor alleles (table 2-1). The concordance of loci without minor alleles in either sample is 93.9% while 3.4% of loci have at least one minor allele in both samples. By concordance we mean a locus has minor alleles or the same haplotype in multiple samples. Conversely, discordance, where a locus in only one of the compared samples had minor alleles, was 2.7% (table 2-3). To confirm the significance of these values, we calculated the probabilities of concordance and discordance based on a cohort of randomly selected loci (5.4% and 5.3% of a total samples), which was < 0.25% concordant, and compared with our results. Using a Pearson's goodness of fit X^2, we verified that the concordant loci are not randomly distributed ($p < 0.0001$). To determine within cell line reproducibility we compared the percent of loci having minor alleles by chromosome as a whole and binned into a million base regions. A linear regression model comparing the percent of loci with minor alleles for each chromosome (as depicted in figure S2-3) shows a significant correlation ($R^2 = 0.85$ and $p < 0.001$) between two independently cultured samples (Figure 2-4A).

Similarly, a comparison of the binned chromosome also shows a significant correlation ($R^2 = 0.60$ and $p < 0.001$, figure 2-4B). Visualization of the distribution of fraction of MST loci showing somatic variation in a representative chromosome (chr1), depicted in figure 2-5, indicates specific chromosomal regions that may harbor SMV "hot-spots". An evaluation of MST loci in translated (exon) regions found over 820 genes containing MSTs with a minimum of 2 minor alleles in both PD20 RV:D2 samples, with some of genes found within segments of chromosome 1 with increased SMV depicted in figure 2-5 (a complete list of exonal MSTs with the minor alleles called, for all cell lines discussed in this paper are available in supplemental spreadsheet 1).

Taken together these results support our hypothesis that this method truly reflects SMV rather than error generated during sequencing and that the results are highly reproducible. The data further suggests that within an individual or cell line, specific genomic regions may contain MSTs that are more susceptible to somatic variability.

Reproducibility between cell lines: To begin to establish a SMV baseline for DNA repair proficient cells, we compared the haplotype, minor allele and SNP/INDEL distributions for two DNA repair proficient cell lines and the PD20 RV:D2 cells discussed above. MCF10A cells are immortalized breast epithelial cells derived from a healthy human female and HEK293 cells are a human embryonic kidney cell line derived from a healthy male fetus. Sequencing produced over 45 million reads with over 170K microsatellite loci called at an average depth of 42 reads per locus for HEK293 cells and over 190K microsatellite loci called at an average depth of 39 reads per locus for MCF10A cells.

Considering major alleles only, 96.4% and 97.0% of all MST loci, respectively, are homozygotic (table 2-1). The average fraction of loci with minor alleles for all three cell lines was 5.1% with a standard deviation of 0.4%. Although MCF10A cells had fewer loci with minor alleles than the PD20 RV:D2 and HEK293 cells (4.5% compared with 5.3% and 5.4% respectively, table 2-1), and showed a difference in the fraction of secondary alleles with SNPs compare to INDELS (table 2-2), MCF10A was not considered an outlier (using Grubb's test for outliers). When we compared the haplotype and minor allele concordance between two non-related cell lines, MCF10A and PD20 RV:D2, we found that 3.8% of loci have different genotypes with only 60% due to haplotype differences. For those loci with minor alleles, discordance is 4.0% and concordance is only 2.0%, the result is significantly above what would be anticipated by chance with Pearson's X^2 (i.e. < 0.3%,). Interestingly, a full factorial comparison of the fraction of loci with minor alleles for each chromosome (as depicted in figure S2-4), using a linear regression model, found a non-significant correlation ($R^2 = 0.061$ and $p < 0.23$, figure 2-4C). However, a correlation using the 1 million base bins is significant with an R^2 value of 0.33 and a $p < 0.0001$ (figure 2-4D), supporting the concept that certain regions contain minor allele susceptibility hot spots. These results demonstrate substantial reproducibility between unrelated independently grown DNA repair proficient cell lines even when the samples are derived from different tissues of origin. These results also suggest that a baseline profile of SMV can be established for DNA repair proficient cells to compare to cell lines with DNA repair defects.

SMV in cells with compromised DNA repair capacity: Thus far we have established that (1) three DNA repair proficient cell lines show similar SMV with low variability both within and between cell lines and that (2) we can differentiate between different SMV trends based on the ratio of INDELs to SNPs. However, the larger goal of this study is to compare SMV patterns between cell lines representative of healthy individuals and those that may have altered DNA repair capacity. To test this, we evaluated 3 cell lines commonly used to study DNA repair and stability. DLD-1 cells are MST instability (MSI) high colon cancer cell line, impaired in Mismatch repair (MMR), selected as positive controls for this study [41]. Capan-1 cells were sequenced previously [12] and are a BRCA2- cell line that can propagate in culture. PD20 cells are from a FANCD2(-) cell line from which the PD20 RV:D2 cells were derived [40]. Both the Capan-1 cells and the PD20 cells have mutations in genes that are involved in normal DNA repair (homologous recombination and interstrand crosslink repair, respectively).

For DLD-1 and PD20 cells, the number of loci that passed filters ranged between 185K and 260K with an average depth of between of 56 and 62 reads per locus respectively. Only 124K loci were called for Capan-1 cells, with an average depth of 71 reads per locus. To capture MST differences between the DNA repair proficient and DNA repair defective cell lines we first evaluated haplotypes and the presence of minor alleles for each cell line. Both DLD-1 and Capan-1 cells significantly differ with respect to haplotype distribution from DNA repair proficient cells (table 2-4). Capan-1 cells showed a significant decrease in heterozygotic loci, 2.1% compare to 3.3% for DNA repair proficient, which was anticipated due to the known trend for loss of heterozygosity in

these cells as reported in the literature due to gene conversion in the absence of BRCA2 [42,43]. In contrast, there was an increase (5.5%) in hetereozygotic loci in DLD-1 cells, which can potentially be attributed to increased mutation due to the MMR defects responsible for the MSI in DLD-1 cells. Surprisingly, haplotype distribution analysis at non-MST loci shows that DLD-1 cells, but not Capan-1 differ significantly from DNA repair proficient (1.8% compared to 1.2% for DLD-1 and Capan-1 respectively). This was unexpected because neither mutation mechanism (homologous recombination nor MMR) would necessarily be restricted to MST vs non-MST regions. A comparison of SNPs and INDELs in the DNA repair impaired cell lines showed Capan-1 cells significantly differed from the DNA repair proficient mean in the fraction of SNPs, with 47% and 91% for MST and non-MST loci respectively (table 2-5). Conversely, DLD-1 and PD20 cells were not found to be different from DNA repair proficient cell lines. For the DNA repair proficient cells the mean fraction of loci with minor alleles was 5.1% with a SD of 0.4%. Capan-1 cells showed again, a greater susceptibility to mutation with a significant increase (6.2%) in the number of loci with minor alleles (table 2-4). In contrast, PD20 and DLD-1 cells both show a significant decrease in loci with minor alleles, 3.1% and 3.2% respectively. This was surprising, particularly because the PD20 cells showed a decrease with respect to their corrected cell line PD20 RV:D2. Concordance of loci with minor alleles between the two related cell lines, PD20 and PD20 RV:D2, was 2.5% while discordance was 3.1%, which was significantly above chance (Pearson's X^2). However, it was greater than the concordance between PD20 RV:D2 and MCF10A, which is to be expected since PD20 and PD20 RV:D2 are related strains (Table 2-3).

Because Capan-1 cells displayed the highest disparity in mutation rate from DNA repair proficient cell lines, including changes in SNP:INDEL ratios, we decided to check the concordance of genotype and minor allele containing loci between them and PD20 RV:D2s (table 2-3). Genotype concordance for the loci that were found in both samples, was over 97.3%, even higher than when we compared PD20 RV:D2 with MCF10As. When comparing the loci with minor alleles ~2% of the total had minor alleles in both samples (were concordant) however 12% were found to have minor alleles in only one samples, meaning discordance (table 2-3). Although this is strikingly different, for the PD20 RV:D2 cells to MCF10A comparison, the concordance rate is still significantly greater than expected by chance. Very similar results were obtained when Capan-1 cells were compared to MCF10A cells. These results offer additional support the hypothesis that some MST loci are more susceptible to mutations than others.

For DLD-1 cells, the increase in heterozygotic loci coupled with the significant reduction in the number of minor alleles is counterintuitive. This suggests the possibility of a proliferation of a small number of subpopulations. If our hypothesis is correct we would anticipate two things to occur: 1) an increase the average depth of reads that define the second allele and 2) an increase in the read depth supporting minor alleles without an increase in the number. To test our hypothesis we first compared the fraction of total reads covering the second allele regardless of haplotype and reads covering only minor alleles. As depicted in figure 2-6, DLD-1 cells show greater than a 4% increase with respect to the DNA repair proficient average in the fractional coverage of the second

allele and more than 8% increase (figure 2-6A and B) for the percent coverage supporting minor alleles. Both were statistically significant. Neither Capan-1 nor PD20 were found to be different from the DNA repair proficient group for either of these parameters. These results suggest a population bottleneck where only a small number of distinct subpopulations are the predominant contributors of the reads captured by the sequencer.

SMV in exons: MSTs are present ubiquitously throughout the genome and are found in over 16% of exons[1]. Although MST expansions or contractions in promoter and interexonal regions can affect transcription, mutations in exons are the most frequently implicated in downstream effects, consistent with exons being under significant selective pressure. An analysis of heterozygotic loci found that exons had significantly less heterozygotic loci, a reduction of over 1.2% compared to untranslated regions (2.4% and 3.8% respectively, figure 2-7A). However the difference in the fraction of loci with minor alleles in exons and untranslated regions was not significant (5.1% and 5.6%, figure 2-7B). In the previous sections we showed that DLD-1 cells, a strain defective in MMR, was found, unexpectedly, to have a significant reduction in the number of MST loci with minor alleles and an increase in heterozygotic loci. Based on this comparison it appears that the results are due to the increased difference between translated and untranslated regions. As shown in figure 2-8A, the fraction of MST loci with minor alleles in exons is 1.1% (compared to 4.7% in untranslated regions) while the fraction of loci that are heterozygotic is 1.7%, compared to 7.9% in untranslated regions (figure 2-8B). These results further support hypothesis that DLD-1 cells have undergone a population bottleneck.

To determine the potential genetic implications of minor allele hot spots, we focused on the analysis of genes affected, specifically we inspected genes containing MST loci found in exons that with 2 or more alleles that did not contribute to the haplotype (minor alleles). This data is provided in a spreadsheet as supplemental materials. The spreadsheet lists the MST loci (based on the HG19 genome), gene name, cell genotype, total number of alleles, variants called and other pertinent information. Of the 2603 genes whose exons harbor minor allele containing loci found in at least one of the 4 DNA repair proficient samples sequenced 47% were found to have 2 or more minor alleles in more then one sample and 9.5% were found in all 4 samples (figure 2-7A). A Genome Ontology (GO) analysis of the 247 genes harboring MSTs with multiple minor alleles in all 4 samples found only a borderline ($p < 0.01$, we use a lower p then 0.05 to compensate for the number of comparisons) significant enrichment of GOTERM categories that included transcription factors, regulators, repressors and DNA binding genes. In addition, there was no significant enrichment for any KEGG pathway categories or cataloged disorders. Conversely, of the ~1100 minor allele harboring genes found in the DNA repair impaired cell lines, only 3 (0.27%) were found in all three cell lines while 95% are in only 1 of the three cell lines (figure 2-7B), which suggests this concordance pattern was primarily random. Further, no genes with multiallelic MSTs were found in all of the sequenced samples and only 18 were found in 6 of the 7 cell line samples. A KEGG pathway enrichment analysis of the minor allele harboring genes found in the DNA repair impaired cell lines suggests a pattern associated with various cancer pathways. Significant KEGG terms enriched were general cancer, colorectal cancer,

myeloma, cervical cancer and cell adhesion (with $p < 0.001$). Together, these results support the hypothesis that specific MST loci in repair proficient cells are more susceptible to somatic mutations but the genes associated with them are not associated with any specific categorized pathway. In contrast, for cells that have impairments in DNA repair pathways, somatic mutations in MSTs appear in higher frequency in loci that are specific to the DNA repair deficiency, and these mutations are implicated in disease, specifically cancer.

DISCUSSION

Somatic mutation can lead to subpopulations of cells carrying mutated alleles. These are examined in cancers, as tumors can be considered to contain subpopulations of cells, i.e. the tissues are not gnomically homogenous [44,45]. Tumors usually carry an allele or set of alleles that confirm their abnormal growth. These alleles, when detected in the tumor but not parent cells, can be the basis for important clinical treatment decisions [11,38,45]. In cell populations with increased somatic mutation rates, like those with altered DNA repair capacity, there may be a concordant increase in subpopulation diversity. As a subpopulation propagates the mutations become more abundant, which becomes detectable in next-gen sequencing data [31,32]. A major assumption of our analysis is that an increase in the number of alleles detected in next-gen sequencing data is reflective of an increase in cell subpopulations or somatic mutation present in the sequenced sample. In this paper we evaluate allele frequencies at MSTs in various cell populations as a quantifiable indicator of variation.

The data presented here evaluate both the standard genotype and minor alleles that are present in next-gen data to establish a baseline for SMV in DNA repair proficient cells and compare this to cells with altered DNA repair capacity. The focus on cell lines with known etiologies is to establish the viability and robustness of our approach. The results show the utility in identifying the consequences of DNA repair impairments on genomic stability. There are several major objectives/findings from this analysis including (1) complimenting genomic analysis away of matched DNA samples with in-sample quantification of variation, (2) demonstrating that DNA repair proficient cells and those

with different defects in DNA repair can have different SMV profiles that may be potential markers for these defects and (3) a quantitative measure of the fraction of loci that exhibit minor alleles may be reflective of subpopulations of cells with different genomic content, potentially those cells that may contribute to tumor formation. MST instability is important in the prognosis and selection of treatment for various cancers, and better, more accurate identification methods are always being sought [10,11].

These data demonstrate that the SNP:INDEL ratio at MSTs can be used to distinguish between different *in-vivo* mutational mechanisms and PCR amplified genomes. Both the WGA single cell sample and the Capan-1 cell line showed an increase in SNPs compared to INDELs at MST loci, however the fractions differed greatly. This is consistent with what was expected from both nucleotide mis-incorporation errors by polymerases (WGA single cell sample) and defects in DNA repair (Capan-1). Neither DLD-1 nor PD20 cells, which are defective in MMR and interstrand cross-link repair, respectively, had a significant alteration of the ratio of SNPs:INDELs at MST loci.

Capan-1 cells displayed a reduction of heterozygotic loci as compared to DNA repair proficient cell lines. This was expected since Capan-1 cells are a BRCA2- cells (impaired in homologous recombination) and have been shown to exhibit a loss of heterozygocity [43]. However, our analysis also indicates a significant increase in the fraction of loci with minor alleles. This could be due to two reasons: 1) Capan-1 cells are a hypotriploid with over 35 structural rearrangements (www.path.cam.ac.uk/%7epawefish/index.html) and with multiple chromosomal regions having more than three copies [46,47]. The

minor alleles in Capan-1 cells can therefore be part of the genotype rather than somatic variation. Conversely, 2) Capan-1 cells have been reported to have an extremely high rate of INDELs and SNPs, significantly higher than expected from the hyperploidy [12]. The results shown here could be due to increased mutation rate shown with this cell line [12] and further support general genomic instability in Capan-1 cells.

Unexpectedly, although DLD-1 cells are a MST unstable cell line, they did not display either of our predicted markers for increase in MST mutation rate: 1) an increase in the number of minor alleles, as was seen with Capan-1 cells, or 2) a decrease in the number heterozygotic loci and the number of minor alleles, as we found in Capan-1 and PD20 cells (table 2-5). Conversely, DLD-1 cells showed both a significant increase in the number of heterozygotic loci and a reduction in the fraction of loci with more than two alleles. Further, they displayed a great reduction in both the fraction of loci with minor alleles and heterozygotic loci in exons (conserved chromosomal regions). We hypothesize that this is the result of defective MMR leading to an increase in mutations that have become fixed in the population. Alternatively, this may have resulted from a bottleneck in the growth of the cell population. If this was the case, the increase in heterozygotic loci allele may be a product of a limited set of surviving cell subpopulations. If a subpopulation with an un-repaired mutation, reached a sufficient proportion of the population due to the bottleneck it would generate sufficient reads for the locus to be mistakenly called heterozygotic. This point is reinforced by the significant increase in the portion of the total number of reads covering the second allele while the fraction of loci with minor alleles and the number of minor alleles per locus are

decreased. This is important to note because it suggests that we can not only distinguish between different mutational mechanisms using the minor alleles in next-gen sequencing, but may also be able to identify cells that have experienced a growth-limiting condition as we expand this work in the future.

The work presented here is a proof-of concept of an approach to assess somatic variation in MSTs using next-gen sequencing. Using this analysis we were able to establish a SMV profile in DNA repair proficient cell lines which we can use to compare to cells with potential or known alterations in DNA repair capacity to begin to evaluate exome or whole genome sequenced samples without requiring a matched genomic sample as baseline. Based on the results presented here this approach can be used to ascertain both scientifically and clinically relevant information. Scientifically, even with known mutations the consequences on the genome as a whole is still relatively unknown. Clinically, somatic variation is a measure of genomic stability and this approach might be used as an addition to current MST instability criteria.

DATA AVAILABILITY: We affirm that we adhere to all PLoS One policies on data sharing of materials associated with this paper. All software developed for this project is available online (https://github.com/zalmanv/MST_minor_allele_caller). All sequencing data generated for this study can be downloaded from the NCBI SRA using accession number SAMN02615820, SAMN02615821, SAMN02615822, SAMN02615823, SAMN02615824. Any other data or programs mentioned here are available upon request.

FIGURES

Figure 2-1: Effects of sequencing error and the minimum number of reads required to call an allele on the number of alleles called in sequencing data. Modeling data with different error frequencies (0.5% - 5%) showed an increase in loci with multiple alleles as error increased when both 2 (A) and 3 (B) reads were minimally required to call an allele. In contrast, standard exome sequencing data from DNA repair proficient cells (PD20 RV:D2 cells) and exome sequencing after whole genome amplification from a single cell were insensitive to the cut-off used.

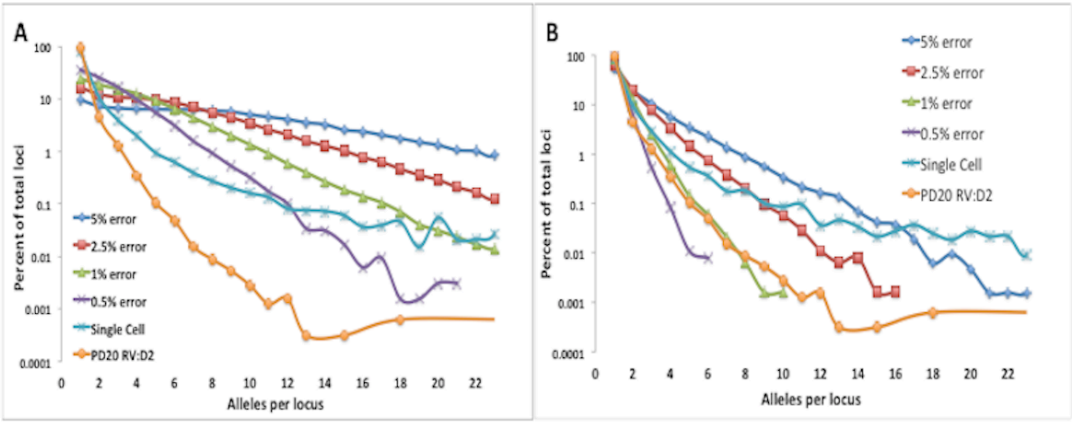


Figure 2-2: Variation in average depth per locus cannot explain the number of loci with minor alleles. The average read depth at loci with increasing numbers of alleles using A) 2 and B) 3 confirming reads per allele for in-silico generated data using 1% and 2.5% induced error rate for 4 different cell lines.

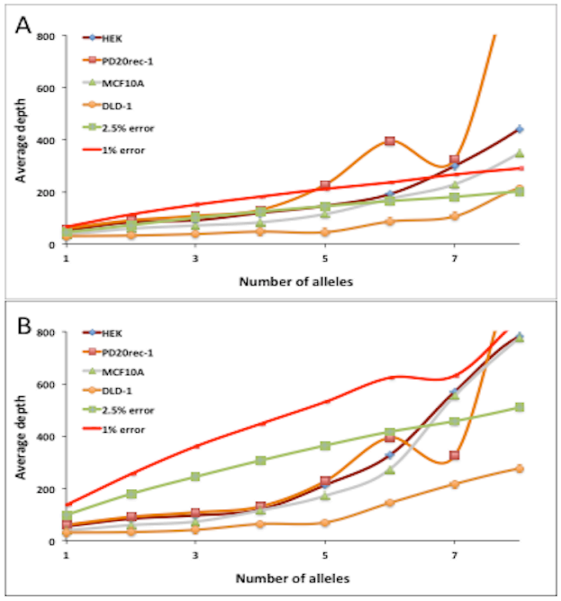
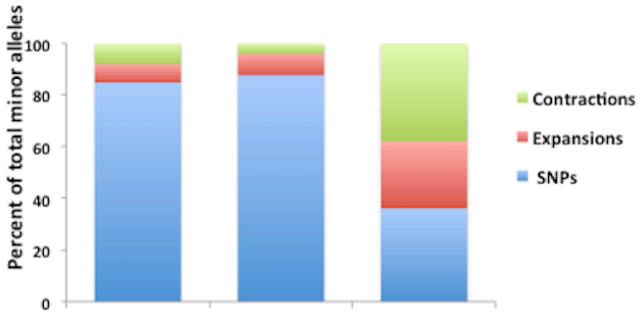


Figure 2-3: DNA repair proficient cells vary significantly from the *in-silico* modeling and single cell sequencing analysis with respect to SNPs and INDELs. The percent of SNPs, expansion and contractions for single cell sequencing and the *in-silico* model as well as the mean and standard deviation for the control cell lines. * significant difference $p < 0.01$.



Percent (%)	Single Cell	Model (1% error)	Repair Proficient Mean	SD
SNPs	84.9 *	87.8 *	36.2	3.4
Expansions	7.3 *	8.3 *	26.0	0.7
Contractions	7.7 *	3.9 *	37.9	2.9

Figure 2-4: A regression analysis indicates a significant within and between cell line correlation in the fraction of loci with one or more minor alleles. Full factorial plots of the fraction of loci with minor alleles by chromosome, regression line and correlation coefficient for A) PD20 RV:D2-1 and 2 C) PD20 RV:D2-1, 2, MCF10A and HEK293. Also full factorial plots of the fraction of loci with minor alleles for the corresponding 1 million base segments of all the chromosomes, a regression line and the correlation coefficient for B) PD20 RV:D2-1 and 2 D) PD20 RV:D2-1, 2, MCF10A and HEK293.

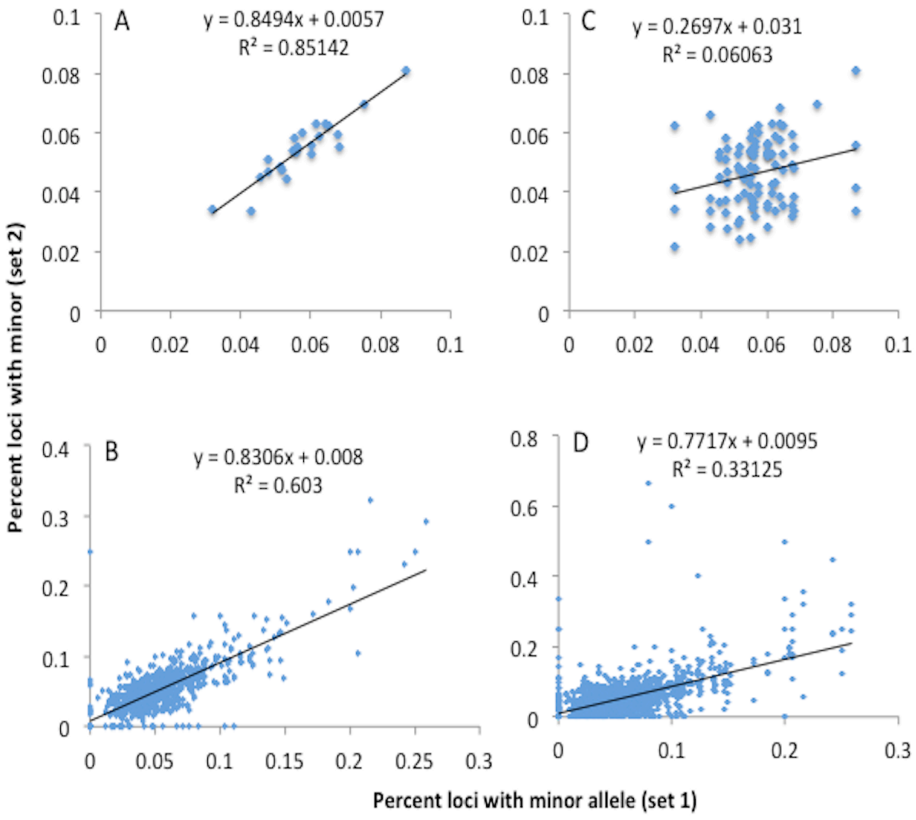


Figure 2-5: The distribution of MST loci showing somatic variability for chromosome 1 binned into 1 million base regions in PD20 and the derived PD20 RV:D2 cell line. The horizontal line demarcates outlier segments, based on a X^2 distribution. All genes shown were found to contain exonal MSTs that with at least 2 minor alleles in both PD20 RV:D2 samples and were found in regions that exceeded the demarcated level. Genes shown in red were found to contain exonal MSTs with at least 2 minor alleles in all 4 DNA repair proficient cell line samples and those shown in blue were found in 3 of the 4 samples. The chromosome image shown at the bottom was obtained from http://en.wikipedia.org/wiki/Chromosome_1 _(human).

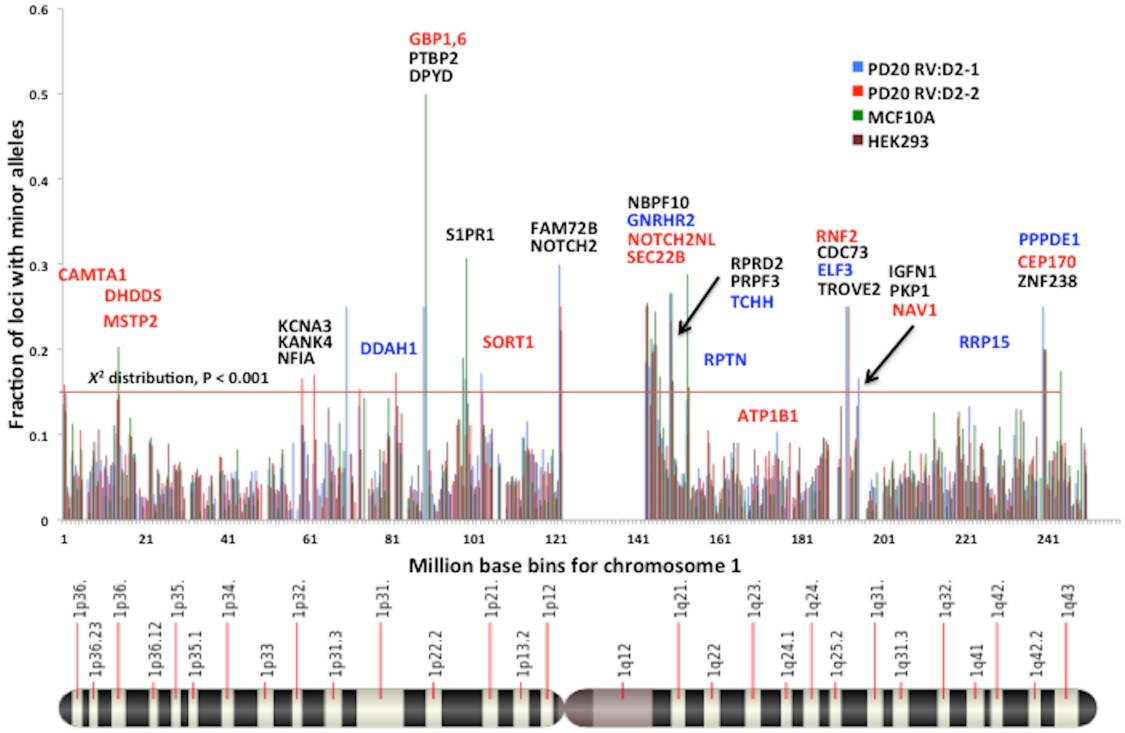


Figure 2-6: An increase in the fraction of reads substantiating the second alleles if present, and all minor alleles. The average fraction of reads representing A) all minor alleles (only for loci with minor alleles) and B) the second allele in both heterozygotic and homozygotic loci that have at least one minor allele, for DLD-1, PD20 and Capan-1 cells were compared to the average of the DNA repair proficient cell lines. The (+) denotes a significant difference from DNA repair proficient ($p < 0.01$) with z-test.

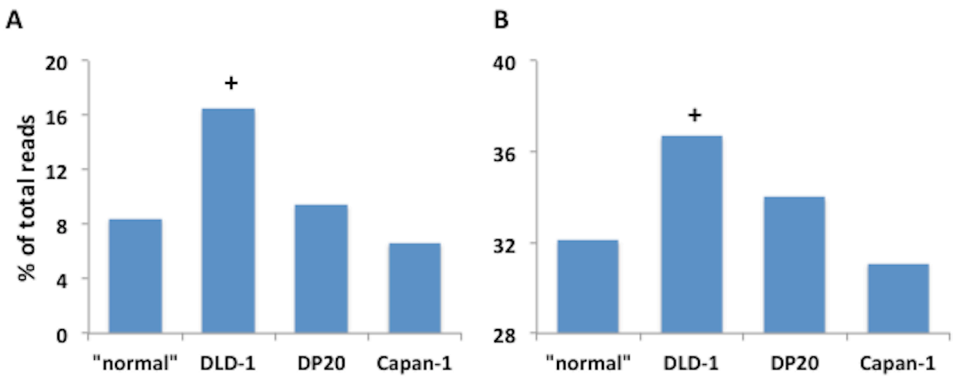


Figure 2-7: A comparison of the percent of heterozygotic loci and loci exhibiting SMV in exons and untranslated genomic regions in DNA repair proficient and impaired cell lines. A) The percent of MST loci that for which minor alleles were found and B) percent of heterozygotic MST loci, in exons and untranslated regions. Depicted in both figures are the means for the DNA repair proficient cell lines and the individual percentage for PD20, DLD-1 and capan-1 cell lines. (+) $p < 0.05$ as compared to DNA proficient cells and (*) $p < 0.001$ as compared to DNA proficient cells in measurement of the difference between exons and untranslated regions.

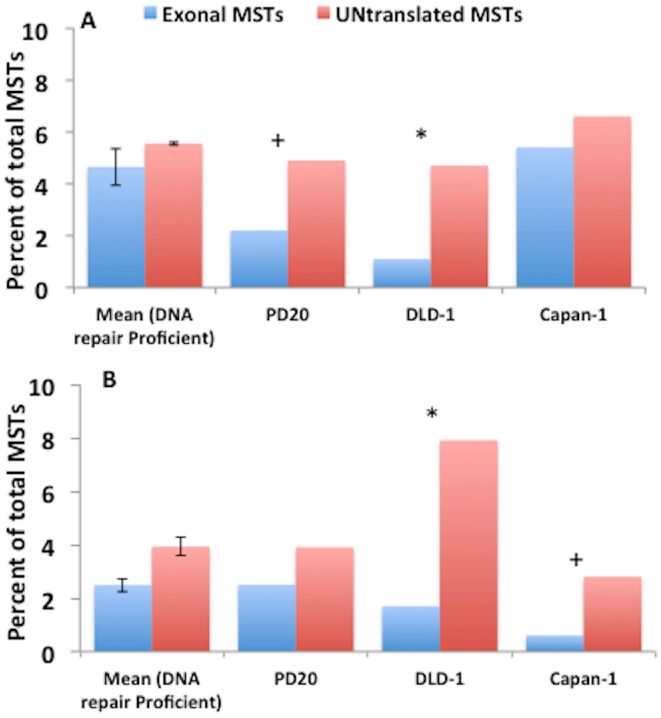


Figure 2-8: The distribution of genes that show SMV in DNA repair deficient cell lines appears random while those in the DNA repair proficient cell lines show significant similarity. The percent of genes with MSTs that with MSTs that have a minimum of 2 minor alleles in A) DNA repair proficient cell lines and B) DNA repair deficient cell lines that are found in all the or some of the sequenced samples. In figure B) the genes that are present in all three DNA repair deficient cell lines is 0.3% and the slice of the pie chart is not visible due to the small percentage.

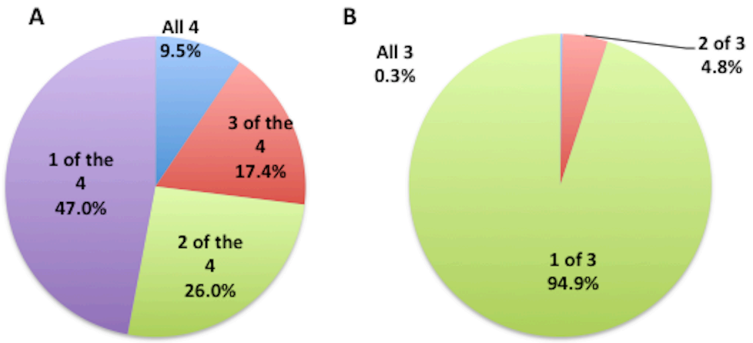


Table 2-1: Exome sequencing data indicates that MST and non-MST haplotype and somatic polymorphism are reproducible in DNA repair proficient cell lines.

	Microsatellite loci				Repair Proficient		Non-Microsatellite loci				Repair Proficient	
Percent (%)	PD20 RV:D2-1	PD20 RV:D2-2	MCF10A	HEK293	Mean	SD	PD20 RV:D2-1	PD20 RV:D2-2	MCF10A	HEK293	Mean	SD
Homo-zyg	96.8	96.8	96.4	97.0	96.7	0.3	99.0	99.0	98.6	99.1	98.9	0.2
Hetero-zyg	3.2	3.2	3.6	3.0	3.3	0.3	1.0	1.0	1.4	0.9	1.1	0.2
Multi-alleles	5.4	5.3	4.5	5.3	5.1	0.4	1.7	2.0	2.1	2.1	2.0	0.2

Table 2-2: MST and non-MST containing loci from exome sequencing of DNA repair proficient cells, but not from sequencing of a single cell after whole genome amplification, show the expected high ratio of INDELs (expansions and contractions) to SNPs.

	Microsatellite loci				Repair Proficient		Non-microsatellite loci				Repair Proficient	
Percent (%)	PD20 RV:D2-1	PD20 RV:D2-2	MCF10 A	HEK293	Mean	SD	PD20 RV:D2-1	PD20 RV:D2-2	MCF10A	HEK293	Mean	SD
SNPs	33.6	32.7	41.4	36.9	36.2	3.4	96.9	96.6	96.8	97.2	96.9	0.2
Expansions	26.2	27.0	25.3	25.5	26.0	0.7	1.3	1.6	1.6	1.4	1.5	0.1
Contractions	40.3	40.3	33.3	37.5	37.9	2.9	1.8	1.8	1.6	1.3	1.6	0.2

Table 2-3: Percent concordance/discordance of haplotype and loci with minor alleles for cell lines.

	Genotype	More then haplotype alleles		Haplotype Allele number
	Discordance	Concordance	Discordance	Concordance
PD20 RV:D2-1 & -2	1.06	3.43	2.69	93.88
PD20rec-1 & PD20	1.15	2.50	3.07	94.43
PD20rec-1 & MCF10A	3.79	1.99	3.95	94.10
PD20rec-1 & Capan-1	2.68	1.92	12.68	85.40
MCF10A & Capan-1	2.19	1.24	13.62	85.10

Table 2-4: Haplotype distribution and somatic polymorphism rate differ in DNA repair defective cell lines compared to DNA repair proficient cell lines.

	Microsatellite loci					Non-microsatellite loci				
	Repair Proficient		Repair impaired cell lines			Repair Proficient		Repair impaired cell lines		
Percent (%)	Mean	SD	PD20	DLD-1	Capan-1	Mean	SD	PD20	DLD-1	Capan-1
Homo-zyg	96.7	0.3	97.2 #	94.5 #	97.9 #	98.9	0.2	98.8	98.2 #	99.2
Hetero-zyg	3.3	0.3	2.8	5.5 #	2.1 #	1.1	0.2	1.2	1.8 #	0.8
Multi-alleles	5.1	0.4	3.1 #	3.2 #	6.2 #	2.0	0.2	1.2 #	1.2 #	3.7 #

significantly different $p<0.01$ - z-test

Table 2-5: SNP and INDEL fractions differ in DNA repair defective cell lines compared to DNA repair proficient cells.

	Repair Proficient		Microsatellite loci Repair impaired cell lines			Repair Proficient		Non-microsatellite loci repair impaired cell lines		
Percent (%)	Mean	SD	DP20	DLD-1	Capan-1	Mean	SD	PD20	DLD-1	Capan-1
SNPs	36.2	3.4	35.7	36.9	47.6 #	96.9	0.2	95.4 #	94.9 #	90.8 #
Expansions	26.0	0.7	26.3	29.7	21.2 #	1.5	0.1	2.1 #	2.2 #	2.8 #
Contractions	37.9	2.9	38.0	33.3	31.2	1.6	0.2	2.5 #	2.9 #	6.4 #

significantly different $p<0.01$ - z-test

REFERENCES

1. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-477.
2. Fonville NC, Ward RM, Mittelman D (2011) Stress-induced modulators of repeat instability and genome evolution. J Mol Microbiol Biotechnol 21: 36-44.
3. Bagshaw AT, Pitt JP, Gemmell NJ (2008) High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots. BMC Genomics 9: 49.
4. Payseur BA, Jing P, Haasl RJ (2011) A genomic portrait of human microsatellite variation. Mol Biol Evol 28: 303-312.
5. Delagoutte E, Goellner GM, Guo J, Baldacci G, McMurray CT (2008) Single-stranded DNA-binding protein in vitro eliminates the orientation-dependent impediment to polymerase passage on CAG/CTG repeats. J Biol Chem 283: 13341-13356.
6. Hile SE, Eckert KA (2008) DNA polymerase kappa produces interrupted mutations and displays polar pausing within mononucleotide microsatellite sequences. Nucleic Acids Res 36: 688-696.
7. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, et al. (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 5: 606-620.
8. Leclercq S, Rivals E, Jarne P (2010) DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. Genome Biol Evol 2: 325-335.
9. Budworth H, McMurray CT (2013) Bidirectional transcription of trinucleotide repeats: roles for excision repair. DNA Repair (Amst) 12: 672-684.
10. Xiao H, Yoon YS, Hong SM, Roh SA, Cho DH, et al. (2013) Poorly differentiated colorectal cancers: correlation of microsatellite instability with clinicopathologic features and survival. Am J Clin Pathol 140: 341-347.
11. Hong SP, Min BS, Kim TI, Cheon JH, Kim NK, et al. (2012) The differential impact of microsatellite instability as a marker of prognosis and tumour response between colon cancer and rectal cancer. Eur J Cancer 48: 1235-1243.
12. Barber LJ, Rosa Rosa JM, Kozarewa I, Fenwick K, Assiotis I, et al. (2011) Comprehensive genomic analysis of a BRCA2 deficient human pancreatic cancer. PLoS One 6: e21639.
13. Lacroix-Triki M, Lambros MB, Geyer FC, Suarez PH, Reis-Filho JS, et al. (2010) Absence of microsatellite instability in mucinous carcinomas of the breast. Int J Clin Exp Pathol 4: 22-31.
14. Yoon K, Lee S, Han TS, Moon SY, Yun SM, et al. (2013) Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. Genome Res 23: 1109-1117.
15. Kim TM, Laird PW, Park PJ (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 155: 858-868.
16. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.
17. Harris RS, Kong Q, Maizels N (1999) Somatic hypermutation and the three R's: repair, replication and recombination. Mutat Res 436: 157-178.

18. Kunz C, Saito Y, Schar P (2009) DNA Repair in mammalian cells: Mismatched repair: variations on a theme. Cell Mol Life Sci 66: 1021-1038.
19. Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, et al. (2013) Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda) 3: 451-463.
20. Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70: 431-435.
21. Eckert KA, Mowery A, Hile SE (2002) Misalignment-mediated DNA polymerase beta mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry 41: 10490-10498.
22. Roy R, Chun J, Powell SN (2012) BRCA1 and BRCA2: different roles in a common pathway of genome protection. Nat Rev Cancer 12: 68-78.
23. Kottemann MC, Smogorzewska A (2013) Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature 493: 356-363.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
25. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
26. Tae H, Kim DY, McCormick J, Settlage RE, Garner HR (2013) Discretized Gaussian mixture for genotyping of microsatellite loci containing homopolymer runs. Bioinformatics.
27. Tae H, McMahon KW, Settlage RE, Bavarva JH, Garner HR (2013) ReviSTER: an automated pipeline to revise misaligned reads to simple tandem repeats. Bioinformatics 29: 1734-1741.
28. McIver LJ, McCormick JF, Martin A, Fondon JW, 3rd, Garner HR (2013) Population-scale analysis of human microsatellites reveals novel sources of exonic variation. Gene 516: 328-334.
29. Lauren J McIver NCF, Enusha Karunasena, Harold R Garner (Submitted) Microsatellite genotyping reveals a signature in breast cancer exomes. Breast Cancer Research and Treatment.
30. Natalie C Fonville LJM, Zalman Vaksman, Harold R Garner (Submitted) Microsatellites in the exome are predominantly single-allelic and invariant. . Genome Biology.
31. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, et al. (2012) Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A 109: 14508-14513.
32. Gundry M, Vijg J (2012) Direct mutation analysis by high-throughput sequencing: from germline to low-abundant, somatic variants. Mutat Res 729: 1-15.
33. Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci U S A 95: 10774-10778.
34. Jarne P, Lagoda PJ (1996) Microsatellites, from molecules to populations and back. Trends Ecol Evol 11: 424-429.
35. Kanagawa T (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). J Biosci Bioeng 96: 317-323.

36. Meyerhans A, Vartanian JP, Wain-Hobson S (1990) DNA recombination during PCR. Nucleic Acids Res 18: 1687-1691.
37. Brodin J, Mild M, Hedskog C, Sherwood E, Leitner T, et al. (2013) PCR-induced transitions are the major source of error in cleaned ultra-deep pyrosequencing data. PLoS One 8: e70388.
38. Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148: 873-885.
39. Mestrovic N, Castagnone-Sereno P, Plohl M (2006) Interplay of selective pressure and stochastic events directs evolution of the MEL172 satellite DNA library in root-knot nematodes. Mol Biol Evol 23: 2316-2325.
40. Ohashi A, Zdzienicka MZ, Chen J, Couch FJ (2005) Fanconi anemia complementation group D2 (FANCD2) functions independently of BRCA2- and RAD51-associated homologous recombination in response to DNA damage. J Biol Chem 280: 14877-14883.
41. Chen TR, Hay RJ, Macy ML (1983) Intercellular karyotypic similarity in near-diploid cell lines of human tumor origins. Cancer Genet Cytogenet 10: 351-362.
42. Holt JT, Toole WP, Patel VR, Hwang H, Brown ET (2008) Restoration of CAPAN-1 cells with functional BRCA2 provides insight into the DNA repair activity of individuals who are heterozygous for BRCA2 mutations. Cancer Genet Cytogenet 186: 85-94.
43. Butz J, Wickstrom E, Edwards J (2003) Characterization of mutations and loss of heterozygosity of p53 and K-ras2 in pancreatic cancer cell lines by immobilized polymerase chain reaction. BMC Biotechnol 3: 11.
44. Tang DG (2012) Understanding cancer stem cell heterogeneity and plasticity. Cell Res 22: 457-472.
45. Schor SL (1995) Fibroblast subpopulations as accelerators of tumor progression: the role of migration stimulating factor. EXS 74: 273-296.
46. Sirivatanauksorn V, Sirivatanauksorn Y, Gorman PA, Davidson JM, Sheer D, et al. (2001) Non-random chromosomal rearrangements in pancreatic cancer cell lines identified by spectral karyotyping. Int J Cancer 91: 350-358.
47. Grigorova M, Staines JM, Ozdag H, Caldas C, Edwards PA (2004) Possible causes of chromosome instability: comparison of chromosomal abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. Cytogenet Genome Res 104: 333-340.

SUPPLIMENTERY material for chapter 2

Table S2-1: In-silico model mapping and genotyping accuracy.

	In-silico error rate			
	5.0%	2.5%	1.0%	0.5%
Total mapped reads with MSTs	8202071	8863891	9915873	10481522
MST call accuracy rate	98.57%	98.53%	98.52%	98.33%
MST accuracy in coding regions or introns	99.23%	99.63%	99.62%	99.81%
% correct zygosity calls	99.96%	99.99%	99.98%	100%
# incorrect zygosity calls (total)	21 (58520)	3 (57204)	15 (58863)	0 (58558)

Table S2-2: The total minor alleles sorted by MST motif length indicate that single cell exome amplification alters the distributions observed in DNA repair proficient cell lines.

MST motif Length	PD20 RV:D2-1	PD20 RV:D2-2	MCF10A	HEK293	Single Cell
1-nt	65.2	66.2	55.7	62.3	12.0 #
2-nt	10.4	9.9	12.6	10.7	13.1
3-nt	8.0	7.8	11.3	8.8	55.2 #
4-nt	3.1	3.1	3.5	3.5	2.8
5-nt	7.0	7.0	8.0	7.4	13.8 #
6-nt	6.3	6.1	8.9	7.3	3.1

Table S2-3: Percent concordance/discordance of haplotype and loci with minor alleles for cell lines.

	Genotype	More then haplotype alleles		Haplotype Allele number
	Discordance	Concordance	Discordance	Concordance
PD20 RV:D2-1 & -2	1.06	3.43	2.69	93.88
PD20rec-1 & PD20	1.15	2.50	3.07	94.43
PD20rec-1 & MCF10A	3.79	1.99	3.95	94.10
PD20rec-1 & Capan-1	2.68	1.92	12.68	85.40
MCF10A & Capan-1	2.19	1.24	13.62	85.10

Figure S2-1: Effects of sequencing error and the minimum number of reads required to call an allele on the number of alleles called in sequencing data. Modeling data with different error frequencies (0.5% - 5%) showed an increase in loci with multiple alleles as error increased when both 2 (A) and 3 (B) reads were minimally required to call an allele. In contrast, standard exome sequencing data from DNA repair proficient cells (PD20 RV:D2 cells) and exome sequencing after whole genome amplification from a single cell were insensitive to the cut-off used.

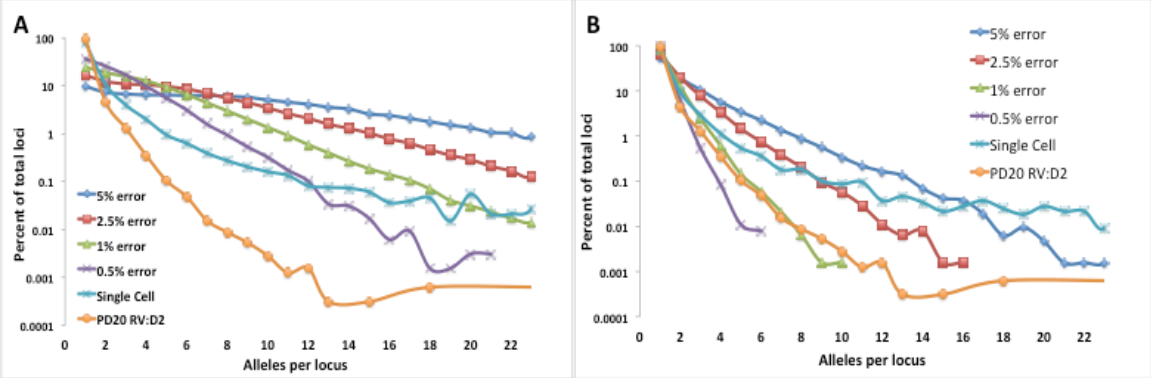


Figure S2-2: Variation in average depth per locus cannot explain the number of loci with minor alleles. The average read depth at loci with increasing numbers of alleles using A) 2 and B) 3 confirming reads per allele for in-silico generated data using 1% and 2.5% induced error rate for 4 different cell lines.

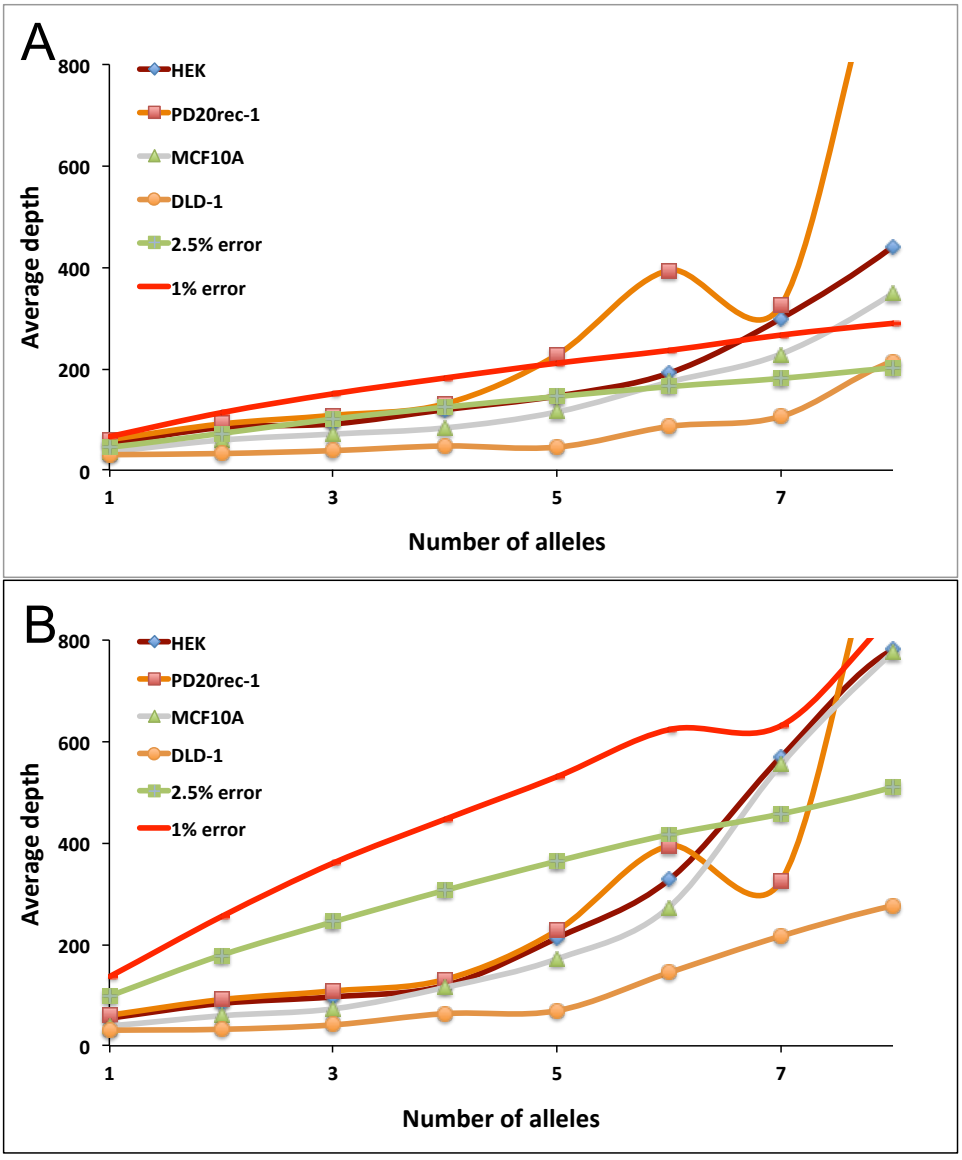


Figure S2-3: Effects of sequencing error and the minimum number of reads required to call an allele on of the number of alleles called in sequencing data. (A) Modeling data with different error frequencies (0.5% - 5%) showed an increase in loci with multiple alleles as error increased when 4 reads were minimally required to call an allele. (B) The average read depth at loci with increasing numbers of alleles using 4 confirming reads per allele for in-silico generated data using 1% and 2.5% error rate and 4 different cell lines.

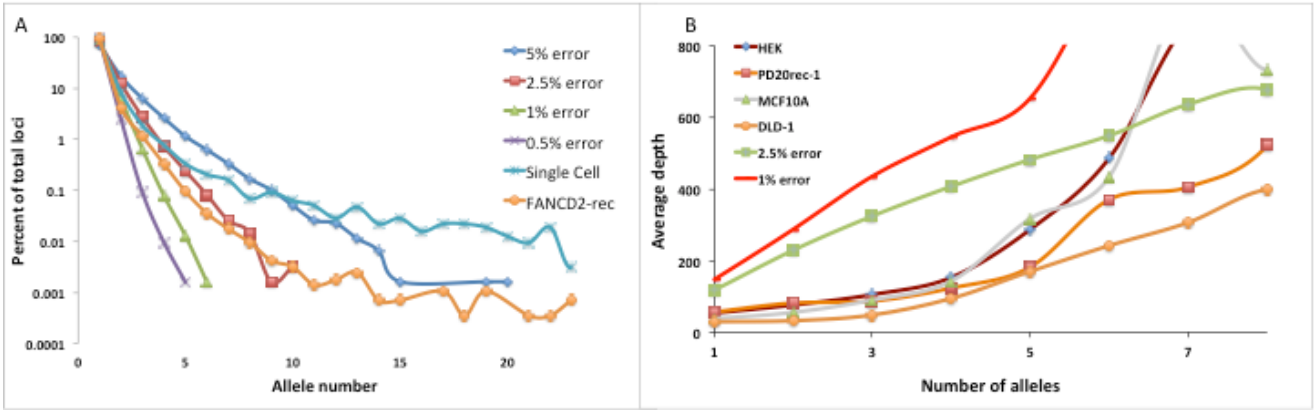


Figure S2-4: The fraction of loci with minor alleles per chromosome for both PD20 RV:D2 samples analyzed.

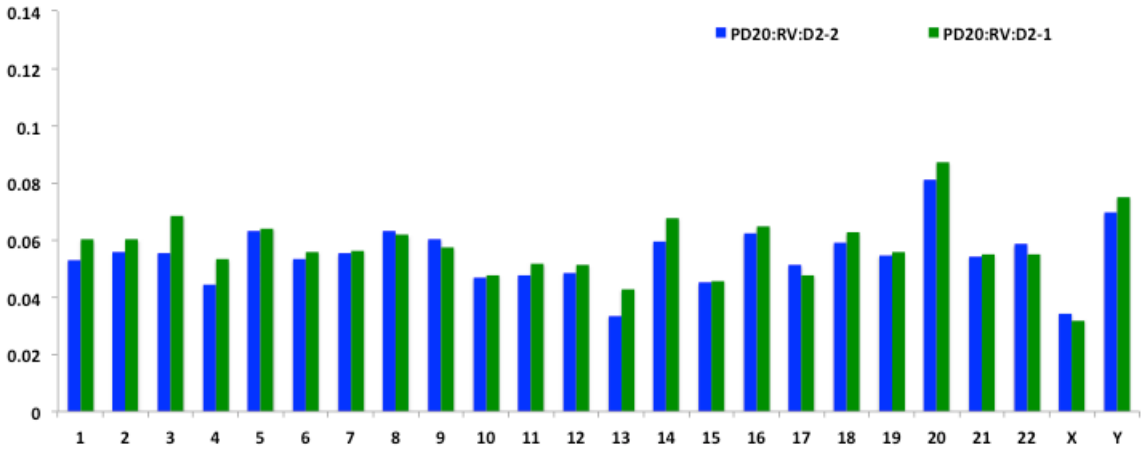


Figure S2-5: A regression analysis indicates a significant within and between cell line correlation in the fraction of loci with one or more minor alleles. Full factorial plots of the fraction of loci with minor alleles by chromosome, regression line and correlation coefficient for A) PD20 RV:D2-1 and 2 C) PD20 RV:D2-1, 2, MCF10A and HEK293. Also full factorial plots of the fraction of loci with minor alleles for the corresponding 1 million base segments of all the chromosomes, a regression line and the correlation coefficient for B) PD20 RV:D2-1 and 2 D) PD20 RV:D2-1, 2, MCF10A and HEK293.

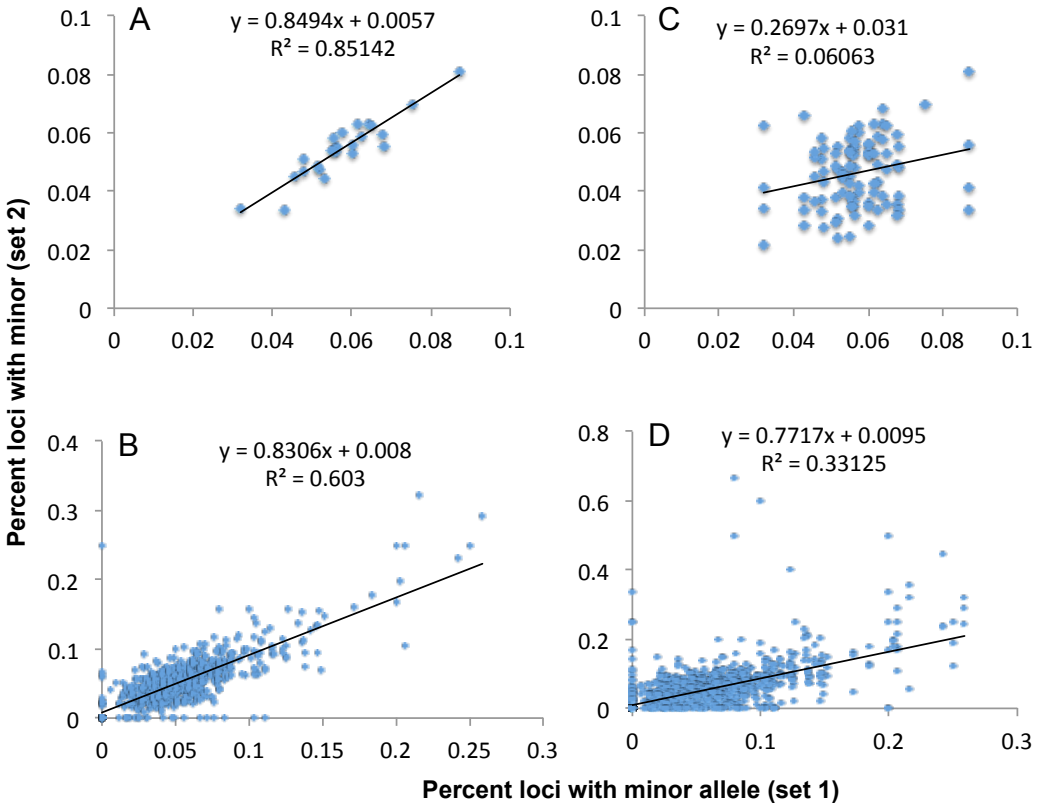
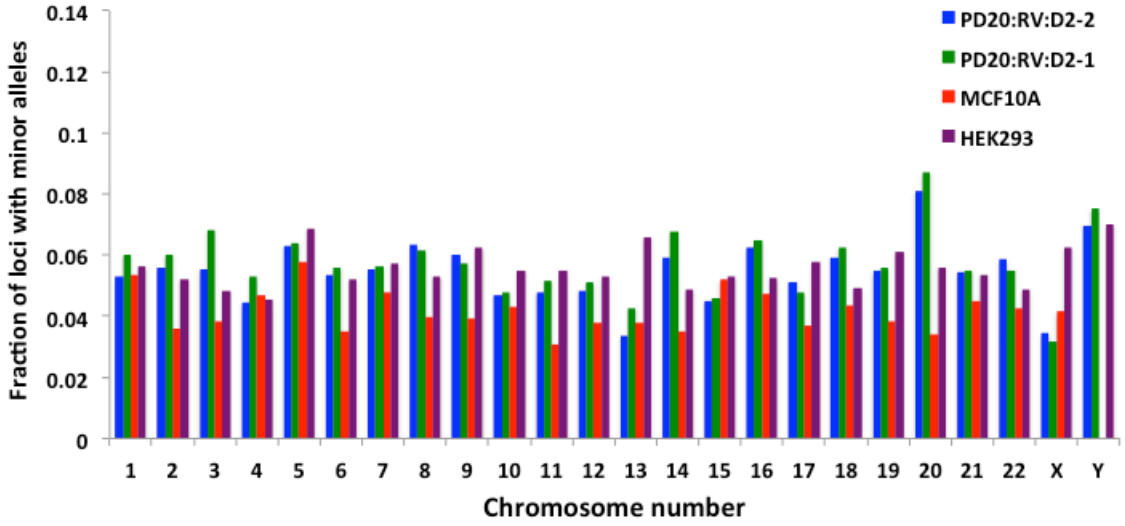


Figure S2-6: The fraction of loci with minor alleles per chromosome for all the DNA repair proficient cell line samples analyzed.



Chapter 3: Effects of impaired of nucleotide excision repair and WRN and RecQL4 on microsatellite somatic variation on a genome wide scale.

ABSTRACT

In a recent report we described a novel approach to characterize somatic microsatellite (MST) variation (SMV) by capturing haplotype and low frequency alleles from HiSeq data. Based on sequencing data from established cell lines we were able to establish a SMV baseline and differentiate between various DNA repair defective cell lines including the mismatch repair defective DLD-1. MSTs are present in or near nearly every gene and since MSTs are fragile mutational "hotspots" with a more than a 3:1 indels to SNP ratio, maintenance of MST integrity is essential. Little is known about transcription-coupled (TC) MST resolution, but evidence suggests that nucleotide excision repair (NER) is involved. The goal of this project was to determine if impaired TC-NER (CSA) or global (GG-NER) (XPA and B) or excision repair supporting helicases BLM or RecQL4 leads to MST destabilization. Comparing exome data from co-sequenced controls, the CSA patient cohort had a higher general and exonal SMV rate. The effect of XPA/B on its patient cohort was less conclusive. Cohorts with impaired BLM or RecQL4 showed similar a lower SMV rate, however, based on the increase in the fraction of reads supporting low frequency alleles, we would argue that the likely reason for the apparent MST stability is a population bottleneck. In conclusion, this work demonstrates that CSA and the RecQ helicases are important to genomic stability, and the effects are seen on a global scale. Further, this work supports the hypothesis that TC-NER is important for the stability of MSTs in transcribed regions. And finally, this is the first publication that uses next-gen sequencing to study genome stability on these sets of disorders.

INTRODUCTION

Somatic variability (SV) describes genetic polymorphisms that occur within a single cell population. This is different than a similar term, somatic variation, which describes differences between germline and somatic populations [1,2]. Although both are a measure of genomic stability, SV is the more difficult parameter to explore as it requires an accounting of the germline genotype as well as low frequency alleles that arise within a single population that do not contribute to genotype. This distinction plays an important role in aging and carcinogenesis since both are associated with the accumulation of mutations over time in a germline population [1,3,4]. Studies in individuals with impairments in DNA repair such as Bloom (BlS), Cockayne's (CS), Lynch, Werner, syndromes (WS), Xeroderma Pigmentosum (XP) and others show that these individuals undergo a process akin to accelerated aging commensurate with mutation rates [4-8]. And other than CS, all the other disorders just mentioned are associated with cancer rates similar to that of aged individuals.

An understanding of genomic, chromosomal, and more specifically microsatellite (MST) stability is essential to discern cancer predisposition, even in healthy individuals [9,10]. The degree with which MSTs are polymorphic makes them ideal forensic markers, and their ubiquitous presence throughout the genome, in regulatory, coding and intragenic regions, denotes their potential clinical significance [11-13]. Studies on replicative and repair polymerase error in MSTs show that MSTs are 10 – 1000 time more vulnerable to mutation than non-repetitive DNA sequences. Rate of polymerase error correction is highly dependent on MST motif nucleotide number, sequence and the number of cycles

that the motif is repeated at the locus, [14-16] however, the high fidelity of mismatch repair (MMR) coupled replication and repair-based DNA synthesis is responsible for the removal of the vast majority of mutations for most MST motifs [17]. Impairment of MMR MutL or MutS complex function leads to what is thought to be a genome-wide MST instability phenomenon associated with colorectal cancer [18-20].

In a previous chapter (chapter 2) we evaluated a novel approach to assess MST stability and polymorphism rates obtained from high-coverage (millions of nucleotides) low-depth tens to low hundreds reads per locus) sequencing. The study results confirmed the value in identifying the impact of DNA repair impairments on genomic stability. The focus on cell lines with characterized etiologies was to establish the viability and robustness of our novel approach. The major findings from that analysis included: 1) demonstrating low variation in controlled, matched samples; 2) establishing a baseline SMV rate for DNA repair proficient cells; and 3) demonstrating that SMV profile changes in DNA repair defective cells are reflective of the impairment and can be used as a marker. Another significant finding was that the fraction of loci with low frequency alleles may be reflective of distinct subpopulations that may contribute to tumor formation or cellular reprogramming leading to various abnormalities.

MSTs form bulky adducts during DNA synthesis that are recognized by replication coupled MMR complexes. Similar loops are formed during translation, but the MMR complex is displaced and the method by which the MST structure is resolved with high fidelity is not well understood [21-26]. Wilson and colleagues have recently argued that a likely candidate is transcription coupled nucleotide excision repair (TC-NER) alone or in

conjunction with MMR MSH2 and 3 complexes. NER is also involved in repair of single strand chromosomal aberrations, malformed DNA structures and adducts such as single strand nucleotide crosslinks (T=T). NER consists of two distinct pathways, one is transcription coupled, TC-NER; while the other is a non-transcription, global pathway (GG-NER). Impairment of TC-NER ERCC8/6 genes (CSA and CSB) cause Cockayne's syndrome, while Xeroderma pigmentosum (XP) is caused by GG-NER impairment to a host of XP genes, both known and unknown [5,8,27,28]. Excision repair for non-bulky nucleotide abnormalities such as deamination and other chemical modifications is maintained by base excision repair (BER) [29]. Base excision repair is essential, and therefore, there is no disorder caused by inhibition of primary BER gene, for defects are lethal. However, some of the human RecQ family of helicases interact with both BER and NER [6,7,30]. The RecQ super family consists of 5 members known as RecQL1 – 5. Deficiency in RecQL2 causes Werner's syndrome (WRN) while mutations in RecQL3causes Blooms syndrome (BLM). Loss of function mutations in RecQL4 causes Rothmond-Thomson syndrome (RTS). Although the interaction of the RecQ family of genes with both sets of excision repair genes has been characterized, the role they play in the repair process and the outcome, in terms of genomic stability, is very poorly understood.

The work described in this paper is an initial step in characterizing the role that the various excision repair pathways have on MST stability. MST stability is assessed by capturing both the contributing haplotype and the low frequency alleles (minor alleles) from exome enriched next-gen sequencing data. The captured alleles are binned and the

polymorphism rate, which we term SMV, is determined. The samples sequenced for this study were from donors with impaired NER or RecQ function, including CSA, XPA and B, RecQL2 and RecQL4. The results here implicate both NER pathways in MST stability, however many of the effects were in opposite directions. The differences between CSA and XPA/B cohorts were statistically significant, however, in most cases, either one or both were the same as the normal subject cohort. In summary, the data presented here suggests that GG-NER and TC-NER may have differing effects on MST stability, while SMV trends in the absence of either RecQL2 or L4 suggest that they may be involved in MST destabilization.

METHODS

DNA prep and sequencing: For this study we obtained DNA from patients with impairments in the following genes; Werner syndrome – WRN (RecQL2), Bloom syndrome – BLM (RecQL3), Cockayne's syndrome – CSA (ERCC8) and CSB (ERCC6), and Xeroderma Pigmentosum – XPA (ERCC1). The samples for this study were purchased from the Coriell cell repository catalog (Coriell Institute, Camden, NJ). The DNA samples, which were purchased pre-prepared from the Coriell Cell Repository catalog (Coriell Institute, Camden, NJ), were isolated by proliferating patient fibroblasts. For the normal controls, DNA was also obtained from Coriell using proliferated fibroblasts. The data from the aging part of this study was obtained from a previous study by our lab [31].

Sample prep and Sequencing: Allocated DNA samples were prepared for Illumina sequencing using the Agilent (Chicago, IL) SureSelectXT Human All Exon V4 capture library. The prepared exomes, paired-end libraries with inserts sizes of 2x100 bp, were sequenced using the Illumina (San Diego, CA) HiSeq 2500 with the HiSeq Rapid v1 flowcell in rapid run mode. Reads obtained were indexed and de-multiplexed with CASAVA v1.8.2 software provided by Illumina. All the sequencing data from these patient DNA samples is freely available upon request or at NCBI. Data for the aging subjects are available at NCBI Short Read Archive and all the appropriate information to obtain those samples is available in Bavarva et al. [31].

Analysis pipeline: Fastq files obtained from the sequencing runs were all run using the same pipeline with the same parameters. Initially, reads in Fastq files were trimmed for regions with low/poor base read scores using the fastX_Toolkit using a minimum score of 25. All trimmed fastq files were then aligned to the HG19/GRCh37 (http://www.genome.ucsc.edu) reference genome using the BWA-mem function [32]. The output sam files were indexed and converted to bam files using SAMTOOLS [33]. GATK/Picard [34] was then used to remove PCR duplicates and mark MST and non-repetitive targeted regions, which were locally realigned using IndelRealigner and TargetIntervals commands. The data for non-MST loci from the multi-allele caller described in [35] and the next section, was confirmed using the output from SAMTOOLs pile-up function (mpileup). The mpileup function is used to obtain a nucleotide-by-nucleotide (only for non-masked regions) variation data that is used by variant callers for genotyping. Mpileup results were analyzed using a costume PERL code. MSTs were analyzed using our previously described multi-allele caller [35]; the method and parameters used are described in the next section.

Microsatellite minor-allele software: MSTs were cataloged using Tandem Repeats Finder [36] (with the following parameters: 2.7.7.80.10.18.6). The list generated was filtered to remove MST loci candidates that were less than 8 nucleotides long and greater than 3 complete motif cycles. Sequence purity selected for this set is 85% for SNPs and 80% if indels are present. Also, removed were sequences that corresponded to inhibitory RNAs and known LINE/SINE elements. In addition to the MST catalog, a second list of 2 million non-MST loci was generated from the HG19 reference. The loci were 15

nucleotides long and at least 50 nucleotides away from any MST. Both loci lists were run using our in-lab developed minor-allele caller.

The minor-allele caller [35] is a tool specifically developed to identify the genotype alleles including alleles that do not contribute to the genotype; here they are termed minor alleles. This program is not designed for individual SNP or indel identification but assesses somatic variability for a specific sequence for which a list haad been developed. A more detailed description and the validation of the procedure is described by Vaksman et al. [35] (chapter 2) with a short description here of the parameters set for this study. The MST alleles were called based on alignments of 7 nucleotide flanking sequences on either side of the MST in a read. Reads with alignment scores below 10% were eliminated. Only loci with a total read depth of 15 or more reads were used for the analysis and an allele was called only when 3 or more substantiating reads were identified. Genotyping (of the two primary alleles) is part of the caller's function and the method is described in Vaksman et al. [35].

RESULTS

Next-gen sequencing enables high-resolution analysis of DNA mutations on a global genomic scale for most DNA sequences, with MSTs being one of the few exceptions. Although many MSTs, such as those in telomeres, are still too long to sequence, advances in sequencing and computational technologies have increased our ability to study them. Recently we published a new method for extracting Hiseq (Illumina, San Diego, CA) reads containing MSTs. The reads were parsed and used to genotype and quantify the distribution of non-haplotype alleles. We used data from DNA repair proficient cell lines to determine a "normal" baseline SMV and compare the results to DNA repair deficient cell lines as a proof-of-concept analysis. To expand on that work, here we analyze sequencing data from patients with loss of function mutations in the TC-NER CSA gene and the GG-NER XPA/B genes as well as the multi-pathway associated helicases RecQL2 and RecQL4.

Sample characterization: For this study, DNA samples from expanded patient fibroblasts were purchased from the Coriell Institute (Camden, NJ). The samples included 6 normal healthy adults, 6 from adult Werner syndrome patients, 5 from young Cockayne's syndrome patients, 5 from XPA/B patients and 5 from RTS patients. A listing of the samples by Coriell ID along with gender, race, causative mutation and cell line immortalization status can be found in table 3-1. The DNA was purified using the SureSelect exome enrichment kit (see methods) and resulting sequencing confirmed an on-target enrichment rate of over 81%. Due to the unique method for extracting MSTs from sequencing reads, coverage for these loci includes only reads that contain the MST

with 7 nucleotide flanking sequences (full MST + 14 nucleotides) on the 5' and 3' ends. The average number of reads fully spanning MSTs for each group is listed in table 3-2. For the DNA repair proficient "normal" sample depth was 47 reads per MST locus. Similarly, the mean depth for CSA, XPA/B, WRN and RTS samples was 53, 38, 35 and 39 reads read per locus (SE 7, 2, 0.4 and 6), respectively (Table 3-2). Although there is a strong correlation between total read count and total loci that passed minimum cutoff (R^2 = 0.94, figure 3-1A) we found no correlation between total read count and SMV rate (figure 3-1B).

SMV in healthy individuals: In the previous report we established a baseline SMV rate using DNA repair proficient cell lines [35]. We found that the SMV rate variance for unrelated, optimally growing, cell lines was sufficiently less than 20% of the SMV rate, which was sufficient to enable us to measure significant increases and decreases in SMV rates in DNA deficient cell lines. Because most of the cells used in that study were from old highly passaged lines with abnormal karyotypes [37,38] and grown under controlled conditions, we cannot create an optimal matched baseline from healthy subjects.

To establish a baseline for normal human subjects, we obtained 4 samples from healthy individuals of whom 2 were related to a child with Fanconi anemia, another related to a CSA patient and 1 related an individual with an unclassified disorder. Data for two more subjects were obtained from a previous publication [31] by our lab discussing genomic changes in aging. The subjects' DNA used was derived from fibroblasts when one individual was 42 and the other was 45 years of age. To establish the baseline, we

computed the sample haplotype distribution and the fraction of loci that have a minimum of one minor allele. This data was analyzed for both MST and non-MST loci. A mean of 3.57% (SE 0.43) of the loci were found to be heterozygotic for MST loci and 1.19% (SE 0.078) were found to be heterozygotic for non-repetitive loci (table 3-2, see "Proficient"). The fraction of MST loci with minor alleles was 3.39% for MSTs and 1.34% for non-MST loci (SE 0.83 and 0.36 respectively). For both measures, the difference between MST loci and non-MST loci was statistically significant ($p<0.001$). These results support our previous finding that a baseline heterozygosity and SMV rate has a low variance and therefore can be established using cells obtained from an unrelated healthy population.

SMV in DNA repair impaired cells: In the previous section we established a baseline rate for SMV and SV for non-MST loci using normal unrelated healthy adults. As with the Vaksman et al. study, the baseline variability for SMV is low, with a standard error of 0.43% and a range between 4.3% and 2.7%, to enable us to statistically compare the rates of individuals with impaired DNA repair, without requiring a large cohort of subjects, to obtain statistical significance. A power analysis based on the results obtained from the DLD-1, PD20 and repair proficient cell lines in chapter 2 indicated that a cohort of 5 - 6 subjects per disorder is sufficient to statistically differentiate between normal and impaired subjects. Since the power analysis was based on a DLD-1, a MMR deficient cell line [37,39], the focus here was to study SMV patterns in other disease based single strand DNA repair pathways.

Although MSTs form bulky DNA products, it is unknown if these structures are either recognized or repaired by excision repair. To study the role of transcription coupled or global nucleotide excision repair on MST stability, we compared sequencing data obtained from expanded patient cell lines (see methods) donated by individuals with non-functional CSA, XPA or XPB genes. For all studies described in this paper, we grouped the XPA and XPB data. Since there are no disorders caused by mutations in the primary base excision pathway, we will compensate by using samples from RTS and WRN patients, since the helicases are known to interact with break repair pathways. The results of the haplotype analysis for the four DNA repair impaired cohorts displayed in table 3-3 show that the fraction of heterozygotic loci for XPA/B (2.5% SE 0.13) and WRN (2.4% SE 0.09) is significantly lower than the normal cohort (3.0% SE 0.43). No haplotype differences were found between any of the cohorts and controls for non-MST loci. However, a within subject haplotype comparison shows MST loci have 2 – 4 fold greater heterozygotic then non-MST loci.

SMV dynamics are measured by comparing the fraction of MST loci that have one or more low frequency alleles (minor alleles). A linear regression (ANOVA) comparison of control and cohort groups indicates that individuals with a defective CSA gene display a significantly higher SMV rate (4.84% SE 0.6) while the group with impaired WRN gene have a significantly lower number of minor alleles (2.6% SE 0.1) than the control group (3.39% SE 0.8, table 3-3). Interestingly, the SMV rate difference between TC-NER and GG-NER deficient patients, CSA and XPA/B, is significant and appear to diverge in opposite directions compare to the control group.

MST motifs: The rates of MST polymorphism are known to be dependent on motif sequence, length, purity and cycle number [40-42]. Single nucleotide repeats are especially vulnerable to mutations, and they account for more than 50% of polymorphic MST loci in a given dataset (see chapter 4). The significance of the homopolymeric runs will be discussed in greater detail in the next chapter. For this dataset, the overall average of loci that are polymorphic for single nucleotide repeats was only 12%, significantly lower than the 22% for the data in chapter 2 and over 40% for the data in chapter 4. However, the chapter 4 data cannot be compared to these samples since the sequenced DNA samples were extracted directly from the tumor and not expanded prior to DNA collection. The fraction of heterozygotic single nucleotide loci and those that have one or more minor alleles for each cohort are shown in figure 3-2. Our analysis showed that XPA/B and WRN cohorts had on average ~25% fewer single nucleotide heterozygotic loci as well as lower (35% less) SMV rate. Although the data are not shown here, similar results were obtained when single nucleotides repeats were removed from the analysis.

Exome SMV: Exons are highly conserved genomic sequences with a low tolerance for expansions or contractions. Both frameshift and inframe indel mutations can have catastrophic effects on cells and therefore heterozygosity and mutation rates in exones are significantly lower than in other genomic regions [21,35]. Data from the previous chapter demonstrated a lower exonal SMV and SV (for non-MST loci) rates in DNA repair proficient and deficient cells (figurer 2-7). Similar results were found when analyzing data from the current patient-based study. The overall fraction of heterozygotic MSTs

was between 2 – 4X greater than exonic MST loci, a statistically significant difference for every cohort, while the difference in exonic non-MST loci was ~ 30% of overall SMV rate (figure 3-3-4A and table 3-4). Conversely, the overall SMV rate was ~ 2X greater, a statistically significant difference ($p<0.05$, t-test) than exonic SMV for all cohorts except XPA/B, for whom no difference was found. For non-MST loci, although the difference was not significant for any of the individual cohorts, the overall trend was a slightly higher SV rate for loci found in exons (figure 3-4B). Further, SMV rate in CSA and XPA/B was found to be 39% and 45% higher (figure 3-4B), respectively, than the repair proficient cohort, a significant difference ($p<0.05$ ANOVA, Dunnett's). The difference between the repair proficient cohort controls and both RecQL2 and L4 deficient cohorts was -29% and -32%, respectively, a significant reduction in rate ($p<0.05$ ANOVA, Dunnett's). The differences in heterozygosity in both MST and non-MST loci were anticipated. Although an increase in SMV rate in exons in the CSA patient cohort was expected, it indicates that it has a potential role in stabilizing MSTs during transcription. A similar increase in XPA/B patients was not anticipated. However, since NER genes have multiple roles within the different sub-pathways, it does suggest that it too has a role in stabilizing MSTs in transcribed regions.

Increase in low frequency events: To detect low frequency events using a low depth approach, variant frequencies must exceed detection probabilities so as to be able to be captured and sequenced consistently. A set of highly polymorphic MST loci have been identified in chapter 2 (supplementary spreadsheet) and also with samples tested in chapter 4 (data not shown). Here, we hypothesize that increased SMV rates are directly

related to increase in mutational load, a factor that should be detected by: 1) an increase in the number of captured variants; or 2) an increase in the fraction of reads that support lower frequency alleles. Up to this point we focused on fraction of loci with sequence variants. In this section we discuss other parameters and measures of SMV.

Mutation rates were probed by calculating the fraction of reads contributing to lower frequency alleles and novel allele rates. In a previous section, we established SMV rates, defined as the fraction of loci with one or more minor alleles, per cohort (figure 3-3 and table 3-3). To determine increases in mutability of moderately susceptible MSTs, we calculated the fraction of MST loci with 2 or more minor alleles loci (figure 3-3), as well as a per locus average number of minor alleles (table 3-5). The increase in the fraction of MST loci with 2+ minor alleles found in CSA patients (1.52% SE 1.8), compared to control (0.82% SE 0.22), was significant, as were the reductions seen in XPA and WRN patient cohorts (0.48% for both, SE 0.04 and 0.06, respectively). The average number of reads covering each minor allele was significantly higher in WRN and RTS subjects (12.07 and 13.07 respectively), compared to 9.98 for controls, (table 3-5). Although the average number of minor alleles per locus in WRN and RTS cohorts was higher, it was just below the cutoff for significance. These results are intricate, but it is clear that impairments in CSA function does increase SMV. Based on all the previous results, it appears that dysfunctional WRN and RTS helps stabilize MSTs, thus reducing SMV. Together, these results suggest that the opposite should be true. However, as with the DLD-1 cells discussed in the previous chapter; reduced SMV rate along with an increase

in minor allele coverage, would suggest a population bottleneck and a proliferation of a small number of subpopulations.

DISCUSSION

DNA repair disorders are debilitating conditions that result in physical and neurological abnormalities robbing the individual of a normal quality of life and life span. Affected individuals experience varying severities of extreme photosensitivity, small stature, infertility and rapid aging, a process characterized by graying or loss of hair at childhood, wrinkling of the skin, gradual loss of eye sight and in many cases an early loss of life [5,7,8,43,44]. The various conditions that fall into this class are known as progeroid disorders and they provide a very important glimpse into the aging process on a genomic level [45,46]. These disorders are caused by the disruption of two cellular processes involving genomic stability, i.e. DNA repair and replication. The conditions for four cohorts analyzed here were; Cockayne's syndrome, caused by the loss of the ERCC8 gene, also known as CSA; xeroderma pigmentosum, caused by the loss of the XPA or XPB genes; Werner syndrome, caused by the loss of the RecQL2 gene; and Rothmond-Thomson syndrome, caused by the loss of the RecQL4 gene [7,8,47].

We selected these disorders for analysis because of their underlying etiologies, which include CSA/ERCC8 loss of transcription coupled nucleotide excision repair, general global nucleotide excision repair through the loss of function of XPA or XPB and multi-pathway helicase dysfunction in the RecQ subfamily of genes involved in excision repair, double strand repair and replication. Although these disorders vary in many regards, even

within disorder subtypes, the commonality is the lack of long term DNA stability in these patients. This aspect of the disorder specific symptoms opens the door to very intriguing scientific queries which after further investigation may help discover new approaches for treatments for those patients and other populations whose genomes are undergoing DNA instability [7,24,48,49].

The primary question of interest is sequence stability at well-defined mutational "hotspots", MSTs [50,51]. The most severe and most applicable disorder for this study is Cockayne's syndrome. Recent publication by the Wilson group at NIH [23,48,52] implicate TC-NER and MST stability/instability during RNA production. Since sequencing was done on exome enriched DNA, the potential bias is obvious. Our results reflected this assessment,giving the clearest difference with respect to the normal unaffected population. First, the CSA impaired group was the only group to have a higher SMV rate than controls, WRN and RTS also were statistically different, but their mean SMV rate was ~40% less than the controls while the CSA group's mean was ~30% greater than the control cohort. The CSA cohort was also found to have a high percentage of loci with 2 or more minor alleles and a higher average of heterozygotic MST loci, although not statistically significant. Interestingly, this trend was only associated with MST loci and not with the non-MST loci. The reason for this is not clear, but we suspect that the issue is not that CSA is only involved in MST stability, but the fact that mutation rate in non-MST loci is sufficiently low as to make it difficult to accurately capture it. This supports the implication that CSA is either directly or indirectly associated with MST fidelity.

Genes involved in WRN and RTS are helicases involved in both single and double strand repair. It is well established that genomic instability is a hallmark of both disorders, however the results here at first glance seem to indicate the absence of defects in either of these two helicases increases MST stability. Cohorts of both syndromes displayed a reduced SMV rate and a statistically significant reduction in the fraction of heterozygotic loci. Differences in heterozygosity can easily be attributed to the loss of heterozygosity (LOH) which for both disorders are known to cause reduction in SMV, including for WRN a reduction in the fraction of loci with more than one minor allele. The interesting and surprising aspect was the significant increase in the fraction of reads that support minor alleles, 12+% for WRN and 13+% for RTS, compared with controls. In this case, either the CSA or the XPA/B cohorts were significantly or substantially different from the controls. This observation created an interesting paradigm, these individuals show a reduction in mutation rates but an increase in their carriers. In evolutionary theory, this is a classic example of a population bottleneck, a loss of genetic plurality and a move towards homogeneity. In our previous work, a similar response was found in DLD-1 colon cancer cells [35].

The within-sample overall genetic variation in non-repetitive sequences is significantly lower than what is observed in MSTs; at nearly a 3:1 ratio for SMV and heterozygosity rates. This can be anticipated since MST mutation rate is several orders of magnitude greater than non-MST loci. However, it is interesting that in the coding regions, the MST polymorphism rate is nearly identical or even lower than what is found in non-MST loci.

The likely reason for this may be that since MST mutation is predominantly through indels, the resulting mutation is more likely to be catastrophic, resulting in cell death and terminating propagation. That is, for indels in exons the selection pressure is very high. Even a low-level increase in MST instability will therefore lead to massive cell death in cultures. This would explain the distinctly lower rate of polymorphisms observed in cell cultures compared to samples collected directly from the patients. This explanation is also relevant to the discussion in the previous paragraph on population bottlenecks, loss of MST stability in or near coding regions may explain the bottleneck.

In conclusion, the work described here describes changes in somatic variability in MSTs and non-repetitive regions. This work is the first to conducted using next-gen sequence analysis on individuals diagnosed with Cockaynes, Werner and Rothmond-Thomson syndromes. Also this is one of the first reports to conduct a comparative genomics analysis of progeriod disorders. Because of the rapid accumulation of nearly every kind of imaginable mutation or chromosomal aberration, in some cases with a strong bias such as Fanconi anemia or Lynch syndrome, these genomes have the potential to explain the deterioration and malignancies associated with aging, radiation or toxin exposure.

FIGURES:

Figure 3-1: A) A strong correlation was found when comparing the number of on-target reads and total number of loci with coverage over 15 reads B) but there is no relationship between the total on-target reads and SMV rates.

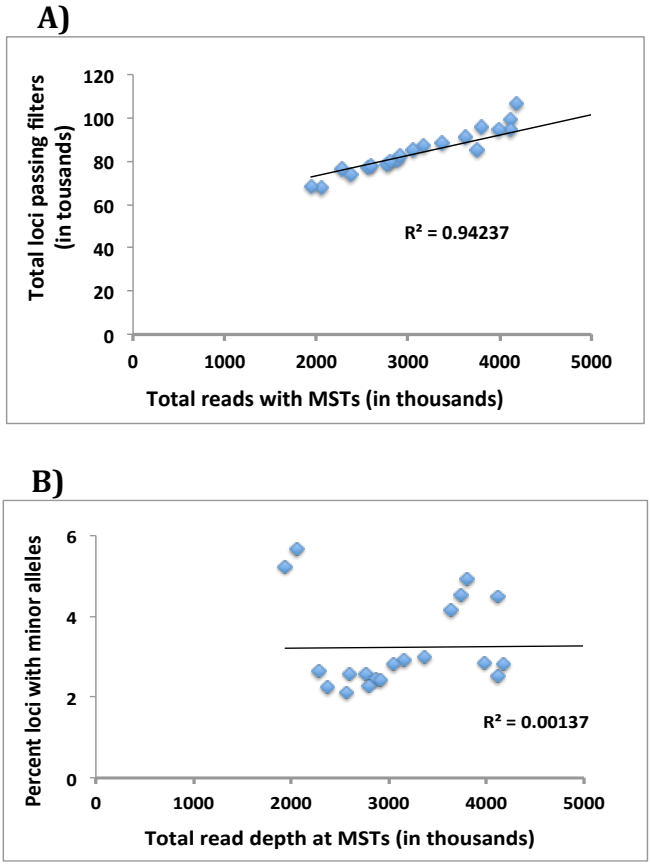
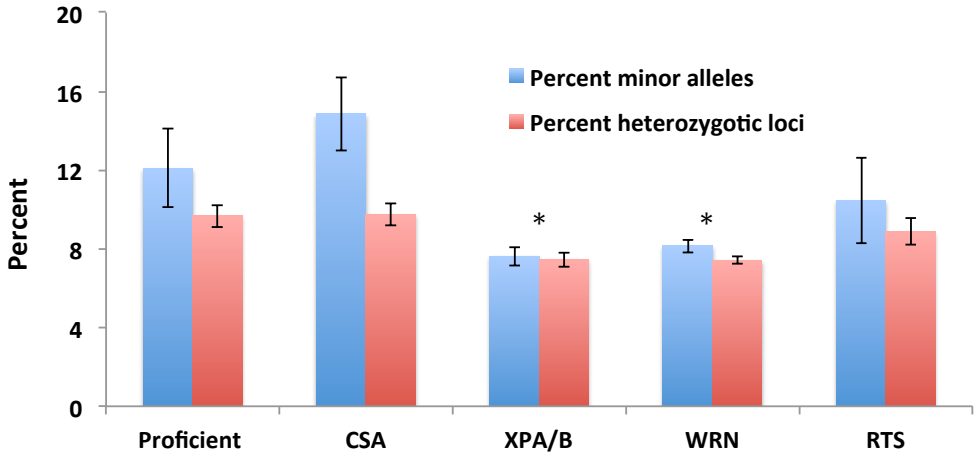
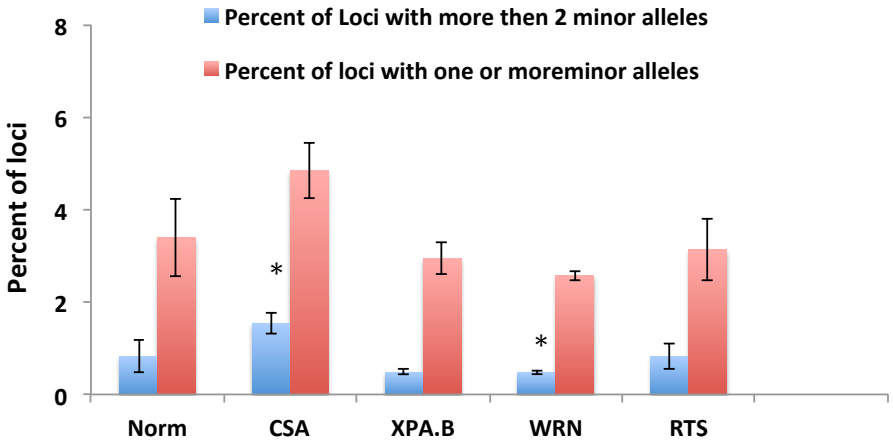


Figure 3-2: The fraction of heterozygotic (orange) single nucleotide MST loci and those with one or more minor alleles (blue) are significantly lower for the XPA/B and WRN cohorts.



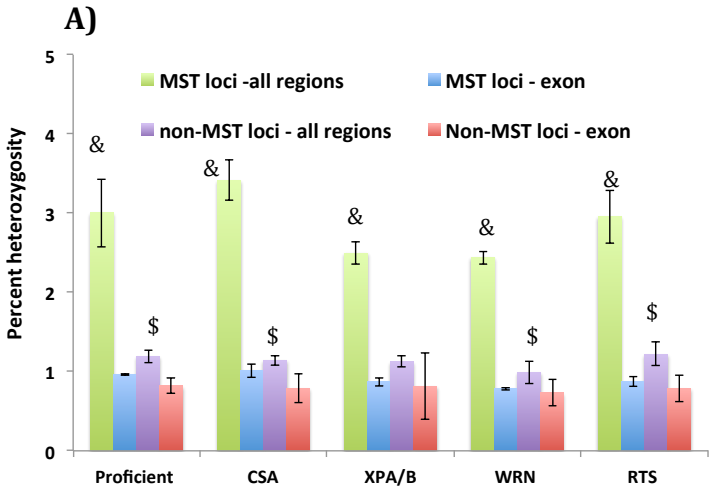
* - Significantly different from Proficient cells $p < 0.01$ ANOVA, Fisher's PLSD

Figure 3-3: Dysfunction of CSA leads to reduced MST stability while loss of WRN function causes MST stabilization, i.e. a reduction in polymorphic loci. In orange, the fraction of loci with at least 1 identified minor allele for both impaired NER and RecQ patient cohorts. In blue are the same cohorts with the fraction of loci that had 2 or more minor alleles.

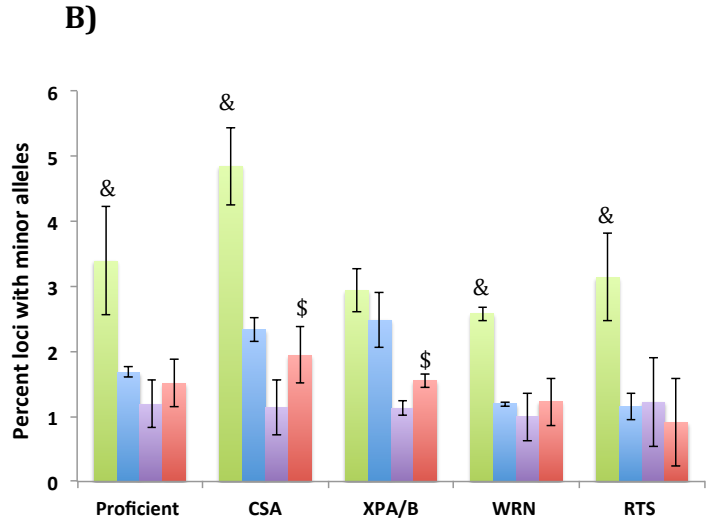


* - Significantly different from Proficient cells $p < 0.01$ ANOVA, Fisher's PLSD

Figure 3-4: SMV rates in exons are significantly greater in CSA and XPA/B patients. A) The fraction of heterozygotic MST loci is significantly greater than non-MST loci in different for any of the disorders tested. B) The fraction of MST and non-MST loci with minor alleles are statistically different in both NER deficient cohorts. SMV rates in both cohorts were significantly greater than controls.



& - Significantly different from exonic MSTs loci in the same cohort $p < 0.01$ t-test
$ - Significantly different from exonic non-MSTs loci in the same cohort $p < 0.05$ t-test



& - Significantly different from exonic MSTs loci in the same cohort $p < 0.01$ t-test
$ - Significantly different from exonic non-MSTs loci in the same cohort $p < 0.05$ t-test

Table 3-1: The available meta-data for the subjects used in this study. For the normal controls we listed the 4 subjects that were sequenced specifically for this paper, the other 2 subjects were described in [53].

Sample # Coriell	Disorder	Gender	Age	Race	Mutation only if classified		Transformed line
CorSample1	Normal	Male	45	Caucasian	NA	NA	Untransformed
CorSample2	Normal	Male	44	Caucasian	NA	NA	Untransformed
NA19272	Normal	Female	23	Caucasian	NA	NA	Untransformed
NA20731	Normal	Female	34	Caucasian	NA	NA	Epstein-Barr
NA01857	CSA	Male	13	Caucasian	CSA- not stated		
NA12496	CSA	Male	3	Caucasian	CSA - unstated		Epstein-Barr
NG06244	CSA	Female	4	Caucasian	CSA - unstated		Untransformed
NG07075	CSA	Female	11	Caucasian	73G>T	Glu13Ter	Epstein-Barr
NG12723	CSA	Male	1	Caucasian	CSA - unstated		Epstein-Barr
NA02090	XP-A	Male	16	African	389G>A	No splicing - CSA	Untransformed
NA02250	XP-A	Female	17	Caucasian	555G>C	No splicing - CSA	Epstein-Barr
NA02345	XP-A	Female	8	Asian	No splice 3' acceptor	No splice site	Epstein-Barr
NA02252	XP-B	Female	31	Caucasian	C>A At slice site	No splice site	Epstein-Barr
NA21148	XP-B	Female	NA	Caucasian	273C>T	Arg425Ter	Epstein-Barr
NG00780	WRN	Male	60	Caucasian	1336C>T	Arg368Ter	Untransformed
NG03141	WRN	Female	30	Caucasian	2476C>T	Gln748TER	Untransformed
NG03829	WRN	Male	42	Caucasian	Undetermined	Undetermined	Epstein-Barr
NG06300	WRN	Male	37	Caucasian	Various chrom issues	PHE1074LEU(	Epstein-Barr
NG07896	WRN	Female	57	Caucasian	Various chrom issues	CYS1367ARG PHE1074LEU	Epstein-Barr
NG12795	WRN	Male	19	Asian	Various chrom issues	Undeturmined	Untransformed
AG17524	RTS	Female	4	Caucasian	2626G>A 4644delAT 1551G>T	Frameshift	Untransformed
AG18373	RTS	Female	42	Hispanic			Untransformed
NG05013	RTS	Male	10	Caucasian	2-BP DEL-2492 IVS12AS	Splicing	Untransformed
NG05140	RTS	Male	10	Caucasian	Undetermined		Epstein-Barr
NG002466	RTS	Male	10	Caucasian	Undetermined		Epstein-Barr

Table 3-2: For each of the studied cohorts the average number of reads with embedded MSTs and the average depth per MST locus that passed filters.

	Read Depth		Reads per locus	
	Mean	**SE**	**Mean**	**SE**
Proficient	4686843	486655	47.4	9.98
CSA	6626761	1538204	53.0	9.86
XPA/B	3265406	264499	37.7	10.18
WRN (RecQL2)	2899448	87416	35.3	12.07
RTS (RecQL4)	3705867	1161082	39.1	13.07

Table 3-3: XPA/B, WRN and RTS patients show evidence of loss of heterozygosity (LOH) and reduced SMV. Conversely, functionally impaired CSA function leads to a significant increase in SMV rate.

		Homozygotic		Heterozgytic		MSTs with minor alleles	
		Mean	SE	Mean	SE	Mean	SE
MSI LOCI	Proficient	96.43	0.43	3.57	0.43	3.39	0.83
	CSA	96.58	0.25	3.42	0.26	4.84 *	0.59
	XPA/B	97.50 #	0.14	2.50 #	0.14	2.94	0.34
	WRN (RecQL2)	97.56 *	0.08	2.44 *	0.08	2.58 #	0.11
	RTS (RecQL4)	97.04 #	0.33	2.96 #	0.33	3.14	0.67
NON-MSI	Proficient	98.80	0.08	1.19	0.08	1.34	0.36
	CSA	98.86	0.06	1.14	0.06	1.39	0.43
	XPA/B	98.87	0.07	1.13	0.07	1.46	0.11
	WRN (RecQL2)	99.01	0.15	0.99	0.15	1.04	0.36
	RTS (RecQL4)	98.78	0.15	1.22	0.15	1.04	0.68

* - Significantly different from Proficient cells $p < 0.01$ ANOVA, Fisher's PLSD
# - Significantly different from Proficient cells $p < 0.05$ ANOVA, Fisher's PLSD

Table 3-4: WRN and XPA/B cohorts show a significant decrease in the fraction of heterozygotic loci and SMV rates for single nucleotide repeats.

	Homozygotic		Heterozgytic		MSTs with minor alleles	
	Mean	**SE**	**Mean**	**SE**	**Mean**	**SE**
Proficient	89.84	0.82	8.64	0.57	10.72	2.02
CSA	90.25	0.58	9.75	0.58	14.85	1.90
XPA/B	92.53 #	0.38	7.47 #	0.38	7.63 #	0.48
WRN (RecQL2)	92.57 #	0.19	7.43 #	0.19	8.14 #	0.33
RTS (RecQL4)	91.12	0.66	8.88	0.66	10.47	1.84

- Significantly different from Proficient cells $p < 0.05$ ANOVA, Fisher's PLSD

Table 3–5: The fraction of reads that contribute to low frequency alleles is significantly higher in the WRN and RTS cohorts, however, no difference in the average number of minor alleles per locus was uncovered.

	Percent coverage for minor alleles		Minor alleles per locus	
	Mean	SE	Mean	SE
Proficient	9.98	0.82	1.35	0.03
CSA	9.86	1.99	1.37	0.04
XPA/B	10.18	1.59	1.23	0.01
WRN (RecQL2)	12.07 *	0.53	1.26	0.01
RTS (RecQL4)	13.07 *	2.56	1.38	0.04

* - Significantly different from Proficient cells $p < 0.05$ ANOVA, Fisher's PLSD

REFERENCES:

1. Shlien A, Malkin D (2010) Copy number variations and cancer susceptibility. Curr Opin Oncol 22: 55-63.
2. Simpson AJ, Camargo AA (1998) Evolution and the inevitability of human cancer. Semin Cancer Biol 8: 439-445.
3. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.
4. Hasty P, Campisi J, Hoeijmakers J, van Steeg H, Vijg J (2003) Aging and genome maintenance: lessons from the mouse? Science 299: 1355-1359.
5. Lehmann AR, McGibbon D, Stefanini M (2011) Xeroderma pigmentosum. Orphanet J Rare Dis 6: 70.
6. Kitano K (2014) Structural mechanisms of human RecQ helicases WRN and BLM. Front Genet 5: 366.
7. Fan W, Luo J (2008) RecQ4 facilitates UV light-induced DNA damage repair through interaction with nucleotide excision repair factor xeroderma pigmentosum group A (XPA). J Biol Chem 283: 29037-29044.
8. Nardo T, Oneda R, Spivak G, Vaz B, Mortier L, et al. (2009) A UV-sensitive syndrome patient with a specific CSA mutation reveals separable roles for CSA in response to UV and oxidative DNA damage. Proc Natl Acad Sci U S A 106: 6209-6214.
9. Hile SE, Shabashev S, Eckert KA (2013) Tumor-specific microsatellite instability: do distinct mechanisms underlie the MSI-L and EMAST phenotypes? Mutat Res 743-744: 67-77.
10. Hong SP, Min BS, Kim TI, Cheon JH, Kim NK, et al. (2012) The differential impact of microsatellite instability as a marker of prognosis and tumour response between colon cancer and rectal cancer. Eur J Cancer 48: 1235-1243.
11. Gymrek M, Erlich Y (2013) Profiling short tandem repeats from short reads. Methods Mol Biol 1038: 113-135.
12. Fonville NC, Ward RM, Mittelman D (2011) Stress-induced modulators of repeat instability and genome evolution. J Mol Microbiol Biotechnol 21: 36-44.
13. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-477.
14. Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70: 431-435.
15. Abdulovic AL, Hile SE, Kunkel TA, Eckert KA (2011) The in vitro fidelity of yeast DNA polymerase delta and polymerase epsilon holoenzymes during dinucleotide microsatellite DNA synthesis. DNA Repair (Amst) 10: 497-505.
16. Viguera E, Canceill D, Ehrlich SD (2001) Replication slippage involves DNA polymerase pausing and dissociation. EMBO J 20: 2587-2595.
17. Hombauer H, Srivatsan A, Putnam CD, Kolodner RD (2011) Mismatch repair, but not heteroduplex rejection, is temporally coupled to DNA replication. Science 334: 1713-1716.
18. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, et al. (2010) Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One 5: e15661.

19. Tanskanen T, Gylfe AE, Katainen R, Taipale M, Renkonen-Sinisalo L, et al. (2013) Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients. Scand J Gastroenterol 48: 672-678.
20. Xiao H, Yoon YS, Hong SM, Roh SA, Cho DH, et al. (2013) Poorly differentiated colorectal cancers: correlation of microsatellite instability with clinicopathologic features and survival. Am J Clin Pathol 140: 341-347.
21. Sydow JF, Cramer P (2009) RNA polymerase fidelity and transcriptional proofreading. Curr Opin Struct Biol 19: 732-739.
22. Kellinger MW, Ulrich S, Chong J, Kool ET, Wang D (2012) Dissecting chemical interactions governing RNA polymerase II transcriptional fidelity. J Am Chem Soc 134: 8231-8240.
23. Berquist BR, Wilson DM, 3rd (2012) Pathways for repairing and tolerating the spectrum of oxidative DNA lesions. Cancer Lett 327: 61-72.
24. Goula AV, Berquist BR, Wilson DM, 3rd, Wheeler VC, Trottier Y, et al. (2009) Stoichiometry of base excision repair proteins correlates with increased somatic CAG instability in striatum over cerebellum in Huntington's disease transgenic mice. PLoS Genet 5: e1000749.
25. Lin Y, Hubert L, Jr., Wilson JH (2009) Transcription destabilizes triplet repeats. Mol Carcinog 48: 350-361.
26. Budworth H, McMurray CT (2013) Bidirectional transcription of trinucleotide repeats: roles for excision repair. DNA Repair (Amst) 12: 672-684.
27. Butt FM, Moshi JR, Owibingire S, Chindia ML (2010) Xeroderma pigmentosum: a review and case series. J Craniomaxillofac Surg 38: 534-537.
28. Zhao XN, Usdin K (2014) Gender and cell-type-specific effects of the transcription-coupled repair protein, ERCC6/CSB, on repeat expansion in a mouse model of the fragile X-related disorders. Hum Mutat 35: 341-349.
29. Parsons JL, Dianov GL (2013) Co-ordination of base excision repair and genome stability. DNA Repair (Amst) 12: 326-333.
30. Trego KS, Chernikova SB, Davalos AR, Perry JJ, Finger LD, et al. (2011) The DNA repair endonuclease XPG interacts directly and functionally with the WRN helicase defective in Werner syndrome. Cell Cycle 10: 1998-2007.
31. Bavarva JH, Tae H, McIver L, Karunasena E, Garner HR (2014) The dynamic exome: acquired variants as individuals age. Aging (Albany NY) 6: 511-521.
32. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754-1760.
33. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
34. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.
35. Vaksman Z, Fonville NC, Tae H, Garner HR (2014) Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies. PLoS One 9: e110263.
36. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
37. Chen TR, Dorotinsky CS, McGuire LJ, Macy ML, Hay RJ (1995) DLD-1 and HCT-15 cell lines derived separately from colorectal carcinomas have totally different

chromosome changes but the same genetic origin. Cancer Genet Cytogenet 81: 103-108.
38. Grigorova M, Staines JM, Ozdag H, Caldas C, Edwards PA (2004) Possible causes of chromosome instability: comparison of chromosomal abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. Cytogenet Genome Res 104: 333-340.
39. Russo MT, Blasi MF, Chiera F, Fortini P, Degan P, et al. (2004) The oxidized deoxynucleoside triphosphate pool is a significant contributor to genetic instability in mismatch repair-deficient cells. Mol Cell Biol 24: 465-474.
40. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, et al. (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol Evol 2: 620-635.
41. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, et al. (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 5: 606-620.
42. Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, et al. (2013) Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda) 3: 451-463.
43. Manthei KA, Keck JL (2013) The BLM dissolvasome in DNA replication and repair. Cell Mol Life Sci 70: 4067-4084.
44. Simon T, Kohlhase J, Wilhelm C, Kochanek M, De Carolis B, et al. (2010) Multiple malignant diseases in a patient with Rothmund-Thomson syndrome with RECQL4 mutations: Case report and literature review. Am J Med Genet A 152A: 1575-1579.
45. Menck CF, Munford V (2014) DNA repair diseases: What do they tell us about cancer and aging? Genet Mol Biol 37: 220-233.
46. Schindler D, Hoehn H (2007) Fanconi anemia : a paradigmatic disease for the understanding of cancer and aging. Basel ; New York: Karger. xii, 229 p. p.
47. Vidal V, Bay JO, Champomier F, Grancho M, Beauville L, et al. (1998) The 1396del A mutation and a missense mutation or a rare polymorphism of the WRN gene detected in a French Werner family with a severe phenotype and a case of an unusual vulvar cancer. Mutations in brief no. 136. Online. Hum Mutat 11: 413-414.
48. Liu Y, Wilson SH (2012) DNA base excision repair: a mechanism of trinucleotide repeat expansion. Trends Biochem Sci 37: 162-172.
49. Kozmin SG, Jinks-Robertson S (2013) The mechanism of nucleotide excision repair-mediated UV-induced mutagenesis in nonproliferating cells. Genetics 193: 803-817.
50. Lai Y, Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. Mol Biol Evol 20: 2123-2131.
51. Hile SE, Yan G, Eckert KA (2000) Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res 60: 1698-1703.
52. Hubert L, Jr., Lin Y, Dion V, Wilson JH (2011) Topoisomerase 1 and single-strand break repair modulate transcription-induced CAG repeat contraction in human cells. Mol Cell Biol 31: 3105-3112.
53. Bavarva JH, Tae H, McIver L, Garner HR (2014) Nicotine and oxidative stress induced exomic variations are concordant and overrepresented in cancer-associated genes. Oncotarget 5: 4788-4798.

Chapter 4: Somatic microsatellite variability as a predictive marker for colorectal cancer and liver cancer progression.

ABSTRACT:

Microsatellites (MSTs) are short tandem repeated genetic motifs that comprise ~3% of the genome. MST instability (MSI), defined as acquired/lost primary alleles in tumors at a small subset of microsatellite loci (e.g. Bethesda markers), is a clinically relevant marker for colorectal cancer. However, these markers are not highly accurate for other types of cancers, and no clinically actionable genetic markers have been found for liver cancer, a cancer with a very high mortality rate. Here we show that somatic MST variability (SMV), defined as the presence of additional, non-primary (aka minor) alleles at MST loci, is a complementary measure of MSI, and is a genetic marker for colorectal and liver cancer. Re-analysis of Illumina sequenced exomes from The Cancer Genome Atlas and find that SMV may distinguish a subpopulation of African American patients with colorectal cancer, which represents ~33% of the population in this study. Further, for liver cancer, a higher rate of SMV may be indicative of an earlier age of onset. In conclusion, the work presented here suggests that classical MSI should be expanded to include SMV, and no longer limited to alterations of the primary alleles at a small number of microsatellite loci. This new/complementary measure of microsatellite variation may represent a potential new diagnostic for a variety of cancers and may provide new information for colorectal cancer patients.

INTRODUCTION:

Cancer is a complex disease, and the variety and specificity of treatment options reflect this, differing based on tumor organ origin, cancer stage, malignancy status, previous response to treatment, recurrence and many other factors. To add to this complexity, even tumors that originate in the same organ or tissue can respond differently to the same treatment procedure. These challenges have led to a 'personal' approach to cancer treatment that relies on a combination of physiology and genomics to determine treatment options. To date the patient specific approach is still very limited because the majority of the known genomic markers are primarily useful for only predisposition screening. One of the few exceptions is a phenomenon called microsatellite instability (MSI). MSI is a pervasive erratic expansion of microsatellites (MSTs), tandem repeats of 1-6 nucleotide motifs, and is associated with approximately 15-20% of colorectal cancers (CRC). MSI is a clinically actionable marker in that treatment options vary in patients with tumors identified as MST unstable (MSI-low or MSI-high) compared to MST stable (MSS) tumors [1,2]. The identification of MSI, and treatment options associated with its diagnosis, is in-part responsible for the drastic improvement in treatment success rate to >65%, as measured by 5-year survival according to the CDC and NCI (http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf). MSI has also been shown be predictive of treatment outcomes and tumor recurrence in other cancers including endometrial, ovarian and breast [1-3].

Unlike CRC, similar genomic markers for liver cancer have not been found. Hepatocellular cancer (HCC) is the 4th most common cancer with ~1 million new cases worldwide and has one of the highest mortality rates of any cancer type [4]. Current 1-year survival rates for liver cancers are <50% and 5-year mortality rate ~84% [5] (CDC and NCI website, see above). Several genomic studies have attempted to find a genetic risk factor for HCC, however to-date none have been as successful [6,7], and the only known risk factors for liver cancer are exposure to toxins, cirrhosis and uncontrolled diabetes.

Somatic variations (SV), polymorphisms that arise in cell populations, often play a critical role in cellular reprogramming and cancer development [12]. SV resulting from DNA damage or inappropriate nucleotide insertion during DNA replication is often increased during stressed or rapidly dividing cell populations such as tumors. MSTs are mutational "hot-spots", meaning they experience a significantly greater rate of somatic variability and population polymorphism than adjacent non-repetitive DNA [8-11]. The unique repetitive genomic configuration of MSTs can lead to the development of complex DNA structures susceptible to polymerase slippage and DNA breaks [8,13-15]. This results in a distinct mutational profile for MSTs with a bias for indels as opposed to single nucleotide polymorphisms (SNPs), frequently observed in non-repetitive DNA [16]. Although, traditionally MST expansion of tri-nucleotide (GCC and CAG) repeats have been studied due to their connection to numerous neurological diseases, including Fragile X and Huntington's coria, recent work suggests that MSTs may also exhibit a contraction bias to which mono-nucleotide motifs are most susceptible [11,17-21]. Many MSTs, especially those in promoter and exonic regions, are under increased selective pressure and therefore

MST genomic localization is also important [19,22-24]. These Somatic MST Variability (SMV) trends or biases are significantly altered in cells with impaired mismatch repair (MMR). Cells with impaired *MUTYH, MLH1* or *MSH2/6* complexes (associated with familial colorectal cancer, Lynch or Muir-Torre syndrome), two of the three essential complexes required for removal and replacement of incorrect nucleotides, show a significant increase in SMV regardless of genomic localization [17,25]. For these disorders, although the predominant mutated motif is still mono-nucleotides, other motifs including di- and tetra-nucleotide MSTs, also show an increase in somatic mutations [16,17,25].

MSI is a measure of the frequency of altered primary alleles relative to a patient's germline within a select set of microsatellite loci. To date the only clinically approved test for MSI (Promega, Fitchburg WI) is based on five Bethesda markers (BAT-25, BAT-26, NR-21, NR-24 and MONO-27). These MSI markers have been tested for a wide variety of cancers other than CRC, gastric and endometrial tumors, but their global applicability appears to be limited [17,18,22]. Expansion of analysis of genomic instability and or microsatellite instability beyond these 5 loci may yield new markers with more general applicability. The introduction of Next-Gen sequencing (NGS) enabled detailed genomic research on a global scale. Over the past two years several papers have compared MSI results obtained from the current clinical test with NGS [17,25-28]. Results from these publications revealed deficiencies in the current clinical assay in identifying MSI in gastric, cervical and even some colorectal tumors. Our group has recently developed a novel tool that identifies all the sequenced alleles for a given MST locus in a Next-Gen sequenced sample, and was subsequently used to quantify SMV in cell lines with known repair deficiencies [16]. In that study we

demonstrated that it is possible to establish a baseline SMV profile in DNA repair proficient cell lines for comparison and that the SMV profile of cell lines with DNA repair impairment changes in a pathway dependent manner. A comparison of DNA repair proficient cell lines and DLD-1 cells, a CRC MSI cell line, demonstrated a ~70% increase in heterozygotic MSTs loci which was attributed to an increase in mutation rate. The gain in heterozygocity was also found in non-repetitive DNA [16].

Although MSI is presumably a genome-wide phenomenon, the classification of MSI is generally restricted to the small subset of loci that make up the Bethesda markers. Recent genomic studies have argued for an increased emphasis on global neoplasmic MST changes to broaden of definition of MSI [16,17,25,27,28]. This work is the first to test somatic variability of MST and non-repetitive DNA sequences in colon and Hepatocellular carcinoma (LIHC) using next-gen sequencing. In this paper we show that SMV can be used as an additional measure, yielding information that is not obtained by simply using the current Bethesda markers. The results described here imply a race dependent hypo-variability in CRC patients. Further, MST hyper-variability in LIHC patients may be associated with earlier onset.

METHODS:

TCGA samples: Tumor and extemporaneous tissue control sequences for CRC and LIHC patients were obtained from The Cancer Genome Atlas (http://cancergenome.nih.gov). Patients sequencing data to be analyzed were limited to exomic sequences, which was optimized for exome capture and paired-end 2 x 70+ illumine sequencing. Therefore all

samples selected were less then 30 months old. Further, due to the susceptibility of MST mutation, samples that have undergone genome amplification were not used. A complete list of samples analyzed can be found in supplementary materials (Suppl. Spreadsheet 1).

Sequencing analysis pipeline: The pipeline was described in more detail in Vaksman et al [16]. However, briefly, TCGA downloaded bam files were reverted to original fastq files using Picard, which were then aligned to HG19/GRCh37 (http://www.genome.ucsc.edu) using BWA-mem [29]. Output sam files were sorted, indexed and filtered for PCR duplicates using samtools, then locally realigned GATK.

Microsatellite multi-allele software: The MST multi-allele caller is described in greater detail in Vaksman et al [16] however a brief description follows.

A list of MST loci was generated by Tandem Repeat Finder (TRF) [30] using the human reference genome HG19 available on the UCSC genome browser website (http://genome.ucsc.edu). MST genotype and somatic variability for each colorectal and liver cancer patient sequencing data set were evaluated using our multi-allele caller using this MST loci list. Bam files for each patient were used to obtain reads with MSTs through an intermediate step using Samtools- view command. Reads that did not meet various quality control criteria, such as mapping score below 10% or average phred score 28 per base, were eliminated by program filters. Determination of MST sequences and sequence lengths was done by alignment of the read to the locus not by a user defined minimal length flanking sequence, for this study the defined flanker length was 7nucleotideon either side of the MST.

A MST locus was called based on a user-defined parameter of minimal coverage and an allele is called based on a minimal number of confirming reads. For this study minimal coverage was 15 reads per locus called and a minimum of 3 confirming reads per allele called. Also, the upper limit for coverage was set at 300 reads to remove loci in genomicly duplicated regions. Genotype and haplotype for each locus were called based on the following criteria; 1) loci with a single allele with a minimum coverage of 15 reads were considered homozygotic with no minor alleles. 2) For loci with the appropriate coverage and a second allele, if the allele is 25% of the total depth for the locus or greater than 50% of the depth for the most common allele, this locus would be considered heterozygotic. 3) If an allele does not meet the criteria described in rule 2 or is not the first or second most common allele, yet has at least 3 reads to substantiate this additional allele, it is considered a minor allele or a somatic variation allele. In this paper SMV rate is defined as the percent of MST loci with minor alleles [16].

In this report we also analyzed somatic variation in over 3 million non-MST loci and compared the results to SMV. Due to MSTs multi-nucleotide configuration we did not use a nucleotide by nucleotide approach as is commonly used for genotyping, instead we generated over 3 million randomly selected loci consisting of 15-nucleotide long sequences (the approximate mean length of MSTs identified in exome sequencing in our analysis). All the loci used were at least 50 nucleotides away from MSTs. The non-MST loci were analyzed for somatic variability using the multi-allele caller described above and the same user-defined parameters, as was done with MST loci [16].

Statistical analysis: All correlations and regression analyses were done using R and Excel. Plotting presented here was done using Excel table functions for ease of use.

RESULTS:

MSI instability is found in approximately 10-20% of CRC tumors and can arise either spontaneously or be associated with hereditary MMR dysfunction. This diagnosis is usually welcome since it provides vital information for treating the patient and is associated with a better patient prognosis. However, recent genome-wide studies indicate that the Bethesda markers may have a higher propensity for false negatives [17,31]. The underestimation may be due to how global MSI manifests it-self. One major assumption is that MSI will be present as a genotypic change, however in a previous publication we found that MSI can also be present as an increase in the number of non-genotypic alleles present within sequencing data from an individual, or somatic microsatellite variation (SMV) [16]. In this paper we utilize our previously published tool to evaluate SMV trends in CRC patient genomes obtained from The Cancer Genome Atlas, and compare with reported MSI results. Further, we also quantified SMV in patients with liver cancer (LIHC), a cancer not known to have classical MSI.

Genotype changes in CRC and LIHC patients: We obtained exome sequencing data from 182 CRC patients available from The Cancer Genome Atlas that matched the quality control criteria described in the methods. For the 182 genomes, all but 9 had a matched tumor and colon/GI control (non-cancer) tissue sequenced as well. For the tumor samples, on average we were able to call 128,589 MSTs per samples (SE ±1332) with an average read depth of

32 (SE ±4.1) reads per locus called. For the control samples, the mean number of loci called was 126,238 (SE ±1552) MSTs per sample with a read depth of 33 (SE ±3.3). The mean number of non-MST loci called for tumor samples is 129,101 (SE ±1620) and 128,593 (SE ±1613) for control tissue. The average coverage depth, the average number of reads that met our criteria was 36 and 29 (SE ±3.8 and 3.2), respectively. In addition, 82 subjects with liver cancer LIHC were available from The Cancer Genome Atlas, 76 of which had both tumor and tissue control samples sequenced. The average loci called for the LIHC samples were 123,485 and 126,946 (SE ±1864 and 2055) with a depth of 36 and 32 (SE ±4.2 and 4.4) for tumor and control samples, respectively. For non-MST loci, we were able to call 111,733 and 114,549 (SE ±1295 and 1465) with a depth of 32 and 33 (SE ±2.9 and 3.7) for tumor and control samples, respectively.

Genomic instability is known to lead to somatically variant DNA sequences that can be detected as changes in genotype. A breakdown of haplotype distribution for CRC cancer and controls shows that 93.6% and 94.3% (SE ±0.13 and 0.09) of the MST loci were homozygotic while 6.4% and 5.7% (SE ±0.13 and 0.09) were found to be heterozygotic (table 4-1). In LIHC patients 95.1% and 94.9% (SE ±0.18 and 0.20 respectively) of the MST loci were homozygotic in tumor and control tissues respectively. As a comparison, the homozygosity rate for non-MST loci that were tested in the same method as MST loci (see methods) were significantly lower. In non-MST loci 98.6% were homozygotic in CRC tumors and controls, and 98.7% for both tissue types in LIHC tumors (table 4-1). As anticipated, these results show that MSTs have a higher rate of polymorphism then non-

repetitive DNA sequences. These data also suggest a greater discordance rate in MSTs than is found at non-MST loci.

To test the if MSTs do indeed have a greater mutation rate then non-repetitive DNA sequences we measured the discordance rates between somatic and control tissues. Discordance was measured by comparing genotypes for each locus in somatic and control tissues for every individual in our sample set (spreadsheet 1). For each locus that had a difference in genotype we determined if the difference was a loss of allele (shift from heterozygous to homozygous or loss of heterozygosity (LOH)), gain of allele (shift from homozygotic to heterozygotic, aka gain of heterozygosity (GOH)) or if the locus had the same haplotype but a difference in genotype (no allele is the same) [16]. As anticipated, on average, MSTs had a >10-fold increase in genotype discordance rates over non-MST loci (table 4-2). The average discordance rate for MST loci was 4.7% (0.15% SE) in CRC patients and 3.5% (0.12% SE) for LIHC patients, whereas non-MST loci showed only 0.39% (0.01% and 0.05% SE) discordance for both CRC and LIHC patients (table 4-2). Both CRC and LIHC patients showed a similar distribution of potential discordance outcomes in MST loci (LOH, GOH or change in genotype but not haplotype). Genotype but not haplotype changes made up an average of 51.3% and 54.7% percent (0.4% and 0.9% SE) for CRC and LIHC patients respectively, while in non-MST loci this was only 15.4 and 16.6 (0.32% and 0.64% SE) percent of total discordance loci. An actual change in haplotype, as indicated by LOH or GOH, accounted for only 48.7% and 45.3% of discordant MST loci while accounting for over 83% for non-repetitive DNA sequences. These results confirm that MST loci have a significantly greater mutation rate than non-MST loci, and that MST associated

mutations are maintained in cancer subpopulations. Further, these results show that in these two cancer types the majority of loci will maintain their haplotype alleles and that of those that do have an altered haplotype, they show an equal likelihood for the gain or loss of a haplotype allele.. However, this common method of measuring genomic instability lacks the ability to determine if 'lost' alleles (for LOH or loci with genotype but not haplotype differences) disappear completely or are present but below the threshold that would normally be expected of a haplotype allele, suggesting that they are present in a subpopulation of the cells whose genomic content was sequenced.

SMV in CRC and LIHC cancer patients: The term SMV here is used to describe the prevalence of minor alleles in MST loci for a given patient. To quantify SMV we analyzed the percent of MST loci with minor alleles, those alleles that do not contribute to haplotype. The mean rate of minor alleles for CRC tumor MST loci is significantly greater than control tissues (14.3% and 12.8%, SE ±0.4 and 0.4 respectively, $p < 0.01$) (table 4-1). Similarly, for LIHC, tumor samples displayed a greater, but not statistically significant, SMV rate compared to non-tumor control tissue with 12.5% and 11.3% (SE ± 0.9 and 0.8% respectively) of loci having minor alleles. For both cancers and tissue types, the rate of somatic variability in MST loci was significantly greater ($p < 0.01$) than in non-MST loci. As shown in table 4-1, the fraction of non-MST loci with minor alleles is 7.5% and 6.2% for CRC, and 7.5% and 5.7% for LIHC tumor and control tissues, respectively.

Motif length and nucleotide makeup have both previously been shown to play a key role in the stability of MSTs [10,11,17]. To evaluate the contribution that the various motifs carry,

we determined the MST motif makeup of the loci that have one or more minor alleles. Results depicted in figure 4-1A show that over 55% of MST loci that have at least one minor allele are single nucleotide runs for both CRC and LIHC tumor and control tissues. The next most common MST motif lengths displaying SMV were tri-nucleotide and di-nucleotide motifs, making up ~20% and 12% of the total loci, respectively, (figure 4-1A) for all the cancers and tissue types. These results are significant since single nucleotide repeats make up only 21% of the total MSTs we analyzed while tri-nucleotide repeats make up 36% of the total. Interestingly, using a t-test comparison no significant differences were present for any of the individual motif lengths when comparing the two tissue types within each cancer. To explore the reason for the overabundance of single nucleotide repeats we calculated the percentage of loci displaying SMV for each MST motif length. The results definitively show that single-nucleotide repeats display a significantly greater rate of SMV (35.8 and 34.8% for CRC and 30.1 and 29.7% for LIHC tumor and control tissue respectively) than the rates for other MST motif lengths (figure 4-1B). An ANOVA comparison shows no significant difference of the two single nucleotide motifs, A/T and C/G runs, with MST size as the repeated measures variable (figure 4-2) Taken together these results suggest that single-nucleotide repeats play a disproportionate role in SMV in the two cancer types, and are consistent with previous MST work with various cancers, including CRC, by our group and others [17,21,32].

SMV in CRC: MSI is most commonly associated with hereditary CRC and therefore MSI testing is commonly conducted on these patients. Within the CRC dataset we analyzed, MSI metadata testing results are given for 155 patients. Of the patients for which the data is

supplied 102 are considered MS-S (MST stable), 24 are MSI-L (with 1 – 3 of the 5 loci showing different primary alleles, and 29 were found to be MSI-H (with 4 or more of the loci showing different primary alleles). Since MSI is considered a genome wide phenomenon we hypothesized that patients testing MSI-H may also show an increase in SMV, that is in addition to acquiring/loosing primary alleles, that they would show an increase in the number of robust minor alleles. We compared MSI status with haplotype and found that heterozygosity was significantly increased in MSI-H tumors (9.0%) as compared to CRC tumors testing MSI-L and MS-S (6.1% and 6.2% MSS and MSI-L respectively, figure 4-3A). This was not the case for control tissue, where no significant difference emerged between the three MSI groups (figure 4-3B), confirming that heterozygosity changes were predominantly introduced in MSI-H tumors. We next compared the fraction of loci with minor alleles with MSI status in tumor samples and found no significant difference between the MSS, MSI-L and MSI-H subgroups (14.6, 14.7 and 15.1% respectively, figure 4-3C). These results suggest that haplotype, but not SMV measures can be used to predict tumor MSI status in CRC patients.

In the previous section we demonstrated that single-nucleotide repeats contributed disproportionally to overall SMV in both CRC and LIHC patients. Further, a comparison of the fraction of single-nucleotide loci with minor alleles and overall SMV revealed a significant positive correlation (figure 4-4A). To confirm that the disproportionate contribution does not bias overall SMV we evaluated the relationship between the fraction single-nucleotide loci contribution with the overall SMV rate. Surprisingly, the results yielded a negative correlation of the two factors (figure 4-4B); meaning that as the overall

SMV rate increases the influence of single-nucleotide loci on the total SMV is reduced. Two aspects should be noted in both of the described correlations: first is that a binomial is a much better fit then a linear regression, with a biphasic inflection point at 16% SMV rate (figure 4-4C and D, after removal of outliers), second is that the small group of 11 outliers, the subset that were encircled in figures 3A and B, consisted solely of African American/Black CRC patients and represents 38% (11 of 38) of African American/Black subjects analyzed in this study. The 11 patient specimens consisted of 5 males and 7 females, and although they were acquired in 5 different centers (Christiana Healthcare, Candler, International Tissue Consortium, University of Pittsburgh and Fondazione-Besta) all were sequenced at Baylor College of Medicine, thus minimizing sequence acquisition bias. To eliminated differences in coverage as the reason for the outliers we compared by finding no significant correlation between total loci called and percent single-nucleotide contribution to total SMV (figure 4-5). Further, no difference was observed when comparing mean coverage per called MST locus for the 11 patients (23.3 (2.2 SE) for the 11 patients and 24.8 (1.3 SE) for the rest of the CRC population tested). These results suggest a potential predisposition for CRC in African American patients that is not currently known or tested in this patient population. Interestingly, MMR may be associated with CRC in this population however none of these 11 patients have been tested (80% of the untested CRC subjects were African Americans).

SMV and cancer onset – CRC and LIHC: As with CRC patient data, when we analyzed the LIHC samples, single-nucleotide loci made up over 55% of the total loci with minor alleles (while they make up only 21% of the total MSTs called) in both tumor and tissue control

samples (figure 4-4A and B). Also, as with CRC patients, we used this data to determine a cutoff for SMV-high and SMV-stable groups. A regression analysis, results of which were plotted in figure 4-6A, shows a reduction in the fraction of single-nucleotide loci that make up the overall number of loci with minor alleles. These results are consistent with CRC patients. As seen in figure 4-6A and also 4B, there are 5 patients that are clearly outliers, however, unlike with the CRC data, the make-up of this group is 4 Caucasian, 1 African decent; 3 males, 2 females; and 4 of the 5 had a predisposition (either alcohol abuse or infection). Further, all the samples were sequenced in the same center. Therefore, unlike with the CRC individuals, there is no indication of why these samples might be outliers, however they were removed from the following analysis.

A comparison of the fraction of single-nucleotide loci with minor alleles with overall SMV rate as depicted in figure 4-6B, shows a concurrent increase in the rates of both (a positive correlation) in a biphasic manner similar to CRC data. When we overlayed and statistically compared the patient data from both cancer types, CRC and LIHC, we found no difference in the distribution (Figure 4-7). We found the point of inflection to be at 14%, which we used as the cutoff for classifying an individual as SMV-stable or SMV-high. Using the inflection points for both CRC and LIHC tumors we compared age of onset for SMV-high and SMV-stable. For CRC tumor samples, no significant difference was found in the age of initial diagnosis between SMV-stable and SMV-high (66 and 65 years of age respectively). Similarly, age of diagnosis was not significantly affected by MSI status (data not shown) in this dataset. However, for LIHC a significant difference did emerge between the two subgroups with the mean age of diagnosis for SMV-stable as 66.0 (±1.5) and 59.1 (±2.9) for

SMV-high (table 4-3). Similar results were found when the cut-offs were used with non-tumor control tissue sequencing data for both LIHC and CRC patients (table 4-3). Again, only SMV-high LIHC patients were found to have a significantly lower age of onset as compared to LIHC SMV-stable. These results indicate that SMV may be a valuable measure of MST instability and may serve to expand the role of MSI to other cancers, outside of CRC and endometrial cancer.

DISCUSSION:

The importance of MST instability in cancer cannot be overstated. Identification of individuals with impaired MMR, Lynch-syndrome, a set of mutations leading to early onset MST unstable CRC or endometrial cancer has lead to earlier detection and a significant decrease in mortality in this subset of CRC patients [1,2]. Unlike CRC no genomic/genetic markers currently exist to determine predisposition or treatment options for liver cancer. To date the only known markers include life-style, disease or infections causing cirrhosis or other liver damage [33,34] however according to the national cancer institute approximately 30% of patients show no predisposition markers (www.cancer.gov/cancertopics/pdq/treatment/adult-primary-liver). The importance in finding a marker for predisposition or treatment is understated by the fact that liver cancer has the second highest mortality rate of all cancers, a 5-year rate of over 84% as cited by the national cancer center (NCI) (http://www.cancer.org/acs/groups/content/@research/documents/webcontent/acspc-042151.pdf). Although several genomic studies have found gene markers associated with

liver cancer tumor none of these yield no information for age of onset or treatment [34-37]. In this paper we used SMV, a measure of MST stability as measured by somatic variability, to assess its use in predicting early onset of LIHC or CRC.

MSI is defined based on markers found that were specific to CRC and molecular identification methods. With the reduction in cost in genomics and its increased use in clinical settings an expansion of how MSI is defined may allow it to be used as a predictive tool for more cancers. Here we used MST somatic variability/SMV as an alternative measure of MSI in LIHC and CRC. To use SMV as a diagnostic tool, a cut-off for instability has to be set which differentiates 2 or more patient populations. SMV had to be defined within our dataset in order to use it as a measure with these two cancer types. Here we defined SMV status based on overall SMV rate as a product of single-nucleotide SMV. This method was partially based on previous work by Yoon et al [17] in which they used single nucleotide repeat genotype changes as a cutoff measure to determine MSI status based on next-gen sequencing. For our study the cutoff for determining SMV-high or SMV-stable patient populations was selected at the point of inflection in the binomial distribution (figure 4-4C and D, and 4-6C) for each type of cancer. This cutoff was selected because it was associated with a stabilization of single-nucleotide run SMV while overall SMV, as well as the SMV for other motifs, was still increasing. Since mutation rates for single-nucleotide MSTs are known to be significantly higher than other motifs but here the rates plateaued only at 40-50%, lower then anticipated, we believe that this point may represent a change either in mechanism associated with mutation accumulation or the maximum SMV rate for single nucleotide repeats.

Using the cutoffs to distinguish SMV-high and SMV-stable patients, described previously, we assessed age of first diagnosis for each cancer type. Counter to several previous studies on MSI [27,38], in this population of CRC patients MSI status was not associated with early initial diagnosis. Using the SMV measure we found the same result, SMV high was not associated with an earlier age of diagnosis. However, LIHC SMV-high patients in this study have been on average diagnosed 6 – 7 years earlier than SMV stable patients. Although onset and diagnosis can be separated by years in LIHC the fact that cancer stage at diagnosis did not differ significantly between the two SMV groups gives support the fact that onset data would parallel detection. It would have also been beneficial to compare SMV rates with the outcomes of various treatments, especially for LIHC patients, nonetheless the variability in the treatments used and inconsistent outcome reports made that comparison not possible.

MSI is defined as an increase in MST variation however our data indicates that both increases and decreases in mutation rates can be informative by the use of SMV as a measure of MST stability. In this study hypo-variability was found to be associated with race, as a specific marker for African American CRC patients. A subset of ~30% of African American patients in this study, show a distinct pattern of MST stability with low SMV and lower heterozygositye rate. These 11 are in contrast to MSI-high patients which have a significantly higher heterozygosity rate compared to the rest of the CRC population (figure 4-8A and 4-3A). The hypo-variability found in these 11 CRC patients is mainly due to the low single nucleotide SMV rate rather. In-fact these patients would not be considered outliers if we used

SMV rates for other motif lengths (figure 4-8B) even though they show a very low SMV rate at all other MST motif lengths.

One of the most unexpected observations in this paper is the inverse relationship between the contributions of single nucleotide SMV to the total SMV when regressed against total SMV (figure 4-4B and 4-6A). This means that as the overall SMV increases the percent of single nucleotide loci that make up total SMV is reduced. This is surprising because the rate of single nucleotide SMV is still increasing until the inflection point shown in figure 4-4C and 4-6C. Although surprising, this may be the result of the overall mutability of single nucleotide repeats. Due to their high polymorphism rate in non-stress conditions, which can be greater then $x * 10^3$ per nucleotide as compared to $x * 10^4$ or greater di and tri – nucleotide MSTs [10,39-41], any systemic increase in overall MST mutability such as impaired MMR will have a greater effect on more stable motifs [39,40], while having a more blunted effect on single nucleotide MSTs. This was shown in part by the Eckert group when they found that impaired MMR causes a similar mutation rate in single di and for some tetra- nucleotide motifs between $x * 10^3$ and $x * 10^2$ per nucleotide [10,39-43]. This is underscored by the data when correlating SMV in single nucleotide runs and other MST motifs. Figure 4-8C shows a positive slop when plotting single nucleotide run SMV with di and tri-nucleotide SMV however the rate increase in single nucleotide repeats is less then the other motifs show based on the slope, which is less then 1 for both motif lengths.

In conclusion, MMR dysfunction leading to MSI is one of the few markers associated with cancer that are predictive for onset and are clinically actionable. Currently, MSI is based on a

small number of markers whose use is informative to only CRC and endometrial cancers. The data presented in this paper suggests that in addition to MSI, somatic variability may increase its effectiveness in CRC and endometrial cancers, and independently may be an effective marker for other cancers. Specifically, here we found that our definition of MSI, one that includes SMV, may have relevance for liver cancer, a cancer type with no known genetic treatment markers. However, the various implications of SMV on these and other cancers have not been explored in this paper and require further studies.

ACKNOWLEDGEMENTS: This work was funded by the Virginia Bioinformatics Institute Medical Informatics Systems Division director's funds, Virginia Bioinformatics Institute Genomics Research Lab Small Grant (CLF-1172), high performance computing was supported by a grant from the National Science Foundation (OCI-1124123) and NSF S-STEM grant (DUE-0850198). We thank the system administrators in the VBI computational core (Michael Snow, Dominik Borkowski, David Bynum, Douglas McMaster, and Vedavyas Duggirala) for technical support. We also acknowledge members of the VBI Genomics Research Lab (Saikumar Karyala, Jennifer Jenrette, Megan Friar, and Kris Lee) for the library prep, and sequencing of genomic and Sanger validation samples. ZV designed and ran the study, analyzed the data and prepared the manuscript. HRG directed and coordinated the study.

DISCLOSURE DECLARATION: ZV declares that he have no competing interests. HRG is owner and founder of Genomeon, LLC, which has licensed these findings, however

Genomeon was not involved in funding or directing this work. HRG receives no salary or other compensation from Genomeon.

TABLES:

Table 4-1: Mean (and SE) SMV and somatic variability (SV) in colorectal cancer tumor samples is significantly greater than in control tissue.

		Tumor MST		Control MST		Tumor Non-MST		Control Non-MST	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE
CRC patients	Homo-zyg	93.64 #,*	0.13	94.29 *	0.09	98.63	0.02	98.63	0.02
	Hetero-zyg	6.36 #,*	0.13	5.71 *	0.09	1.36	0.02	1.37	0.02
	Multi-alleles	14.30 #,*	0.38	12.79 *	0.36	7.49 #	0.37	6.16	0.34
LIHC patients	Homo-zyg	95.14*	0.18	94.95 *	0.20	98.74	0.03	98.68	0.03
	Hetero-zyg	4.86 *	0.18	5.05 *	0.20	1.26	0.03	1.32	0.03
	Multi-alleles	12.45 #,*	0.92	11.28 *	0.78	7.53 #	0.99	5.68	0.70

# $p < 0.01$ compared to control tissue for MST or non-MST
* $p < 0.01$ compared to equivalent tissue for non-MST

Table 4-2: Concordance and types of genotypic changes between tumor and control tissue for CRC and Liver cancer.

		Total loci observed		Percent discordance		% genotype		% LOH		% GOH	
		Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE
CRC	MST loci	113173.57	1510.97	4.69 *	0.15	51.25 *	0.47	22.43 *	0.71	26.32 *	0.83
	non-MST loci	129101.94	1620.50	0.39	0.01	15.38	0.32	47.86	1.01	36.76	0.89
LIHC	MST loci	113401.38	1744.93	3.51 *	0.12	54.57 *	0.90	23.12 *	0.52	22.31 *	0.63
	non-MST loci	130541.96	1852.50	0.39	0.05	16.58	0.64	47.73	1.13	35.69	0.99

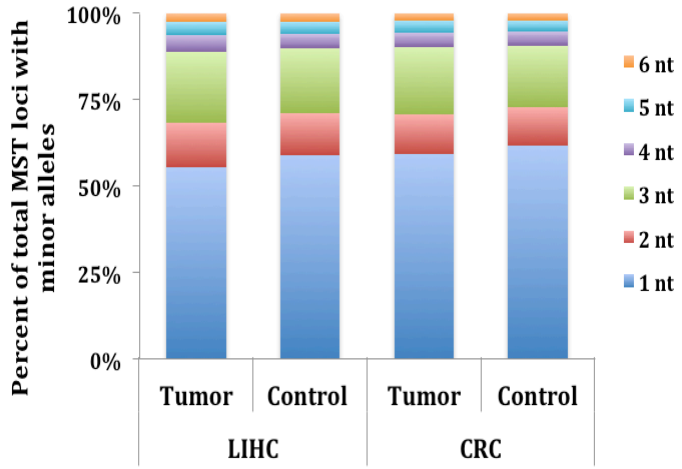
* $p < 0.05$ compared to non-MST loci

Table 4-3: SMV-H in both tumor and controls tissue is correlated to lower age of on-set for liver cancer, but not for colorectal cancer.

	Cancer	SMV-S Mean (SE)	SMV-H Mean (SE)	T-test (p <)
LIHC	**Tumor tissue**	66.0 (1.5)	59.12 (2.9)	0.038
	Control tissue	66.7 (1.4)	58.4 (3.0)	0.013
CRC	**Tumor tissue**	66.0 (1.1)	65.0 (1.7)	0.28
	Control tissue	63.1 (1.9)	65.8 (1.1)	0.133

Figure 4-1: Single nucleotide MSTs show the highest rate of somatic variability and make up over 55% of MST loci with minor alleles. The total number of loci with minor alleles in both tumor and control tissue types for each CRC and LIHC patient were calculated and the percent contribution for of each MST motif length was compared. Figure A) shows that single nucleotide motifs makeup on average over 55% of the total MST loci with minor alleles while tri-nucleotide MSTs, making up the second highest percentage only makeup approximately 21% of the total. Figure 1B) shows that the reason for this disparity is most likely due to the fact that single nucleotide motifs 5 – 6 times more likely to show minor alleles then other MST motif lengths. This table does have SE bars however they are too small to be seen as they are less then 2 percent for 1A and 1 percent for 1B.

A)



B)

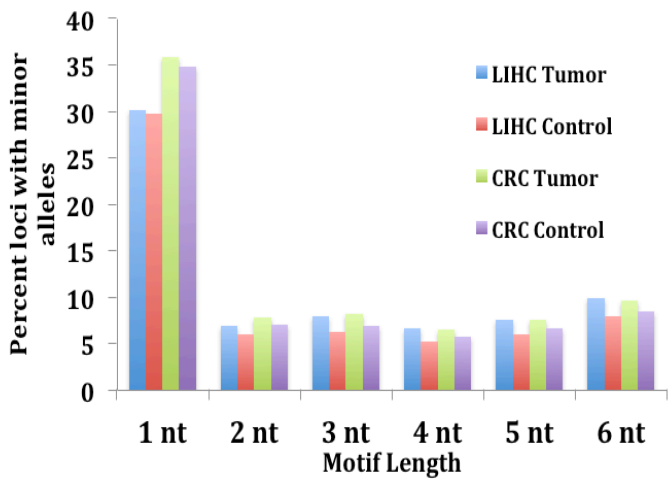


Figure 4-2: No difference is seen when comparing the percent SMV between the two single nucleotide motifs, A/T and C/G runs.

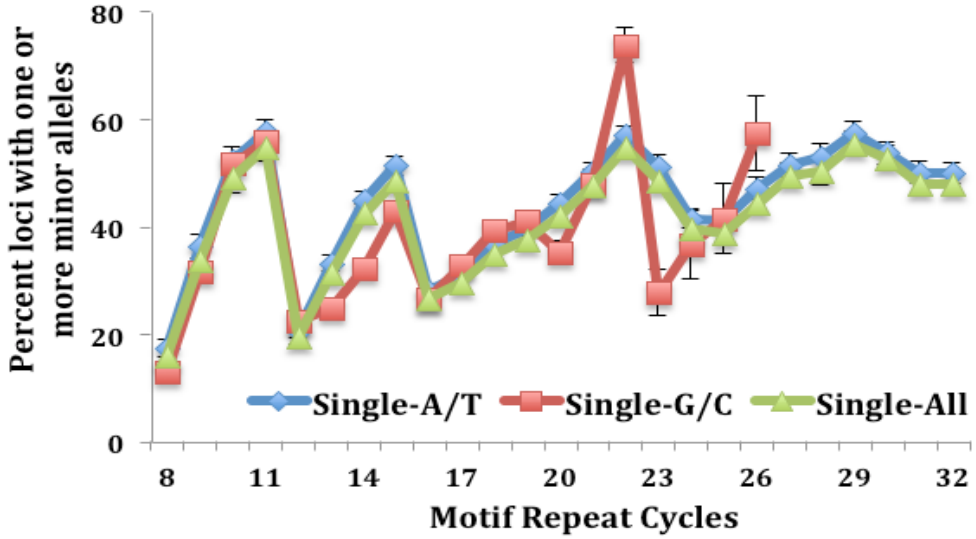


Figure 4-3: Microsatellite instability is not correlated to SMV in colorectal cancer. CRC patients were grouped by MSI status (classified as MSI stable (MSS), SMI-L or MSI-H) and analyzed for differences in genomic stability by comparing A) percent heterozygotic loci in tumor tissue, B) percent heterozygotic loci in control tissue and C) SMV in tumor tissue. The MSI high group displayed a significant increase in heterozygosity as compared to MSS and MSI-L group (* - $p < 0.01$, ANOVA followed by a Fishers PLSD) while no statistical difference was found when observing SMV in tumor tissues (ANOVA $p > 0.24$).

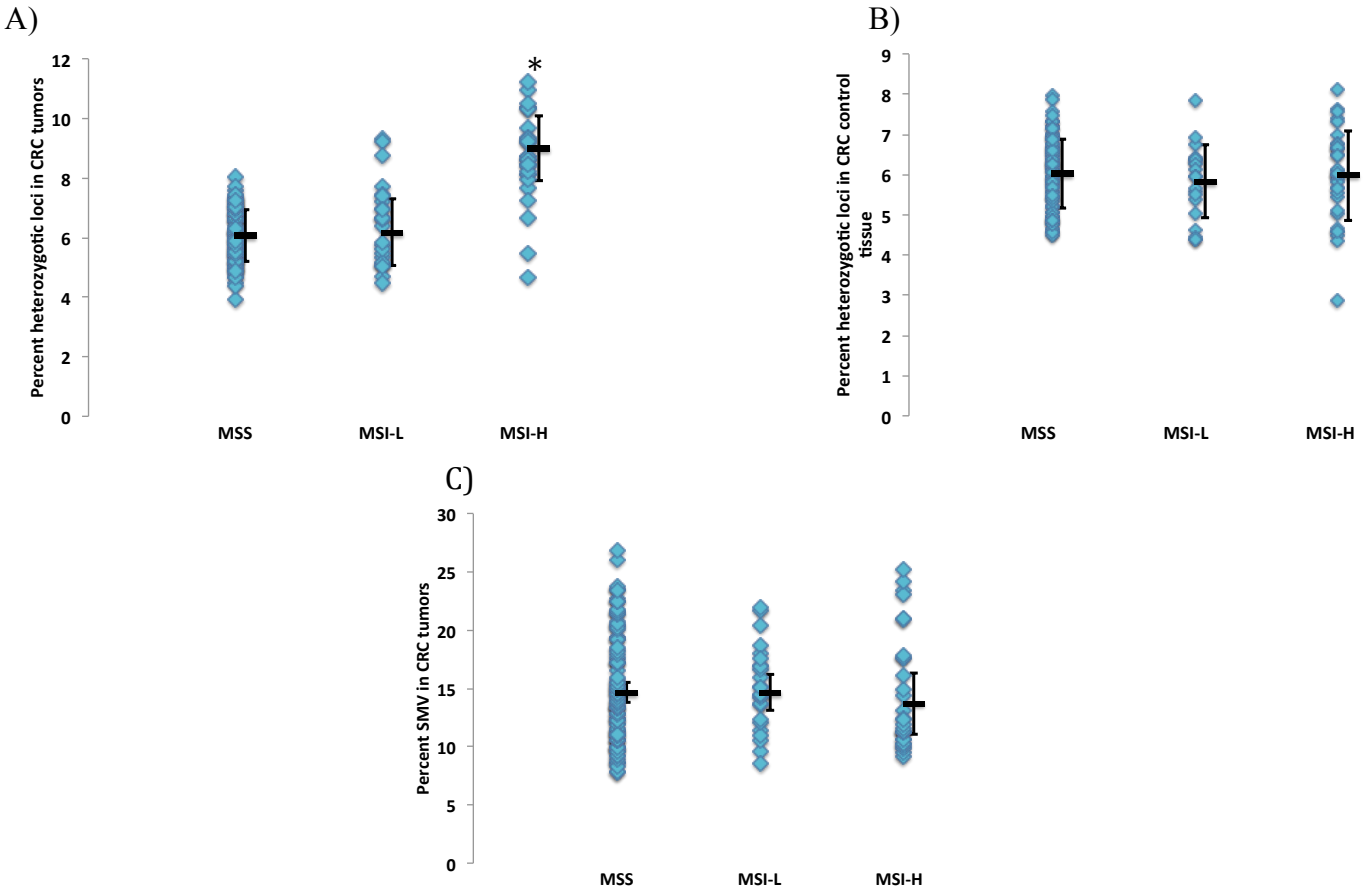
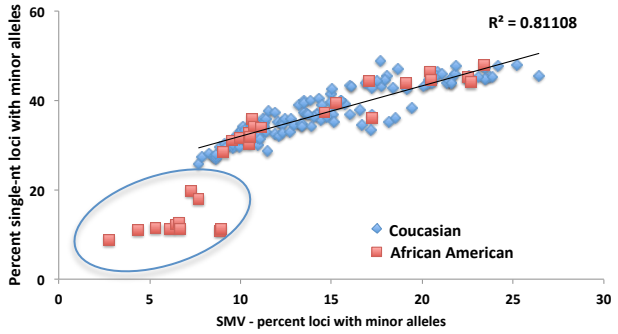
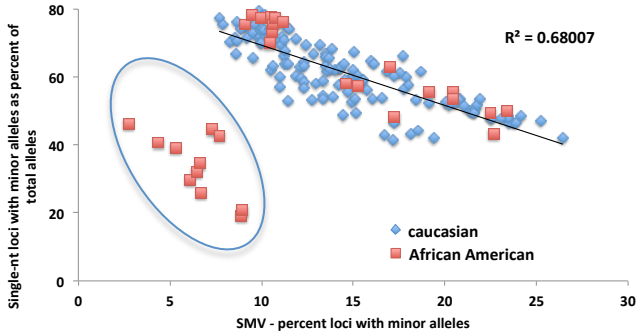


Figure 4-4: The distribution of the contribution of single nucleotide SMV to overall SMV in CRC patients. A) A linear fit of the total SMV against single nucleotide SMV distribution and B) total SMV against single nucleotide as a percent of total SMV. In both figures a clear set of outliers entirely consisting of 32% of the total African American (orange squares) CRC patient population is encircled. C) and D) show the same distributions with the binomial fit inflection point used as a cutoff between SMV-high (blue diamonds) and SMV stable (orange square) patient groups.

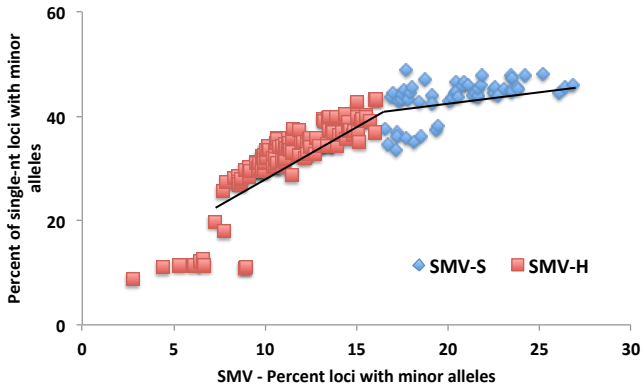
A)



B)



C)



D)

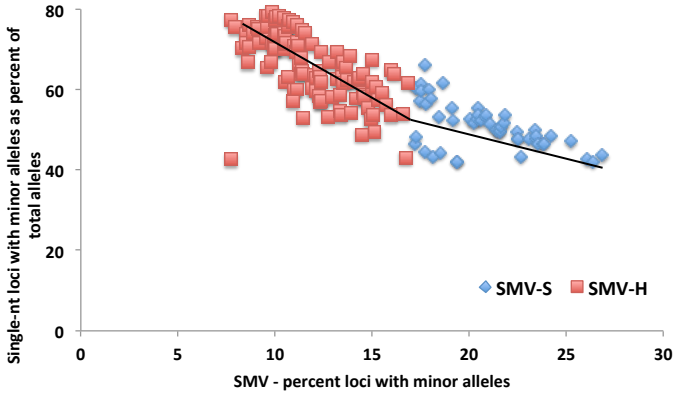


Figure 4-5: The total number of loci called for the 11 outlier African American CRC patients does not explain their low SMV. Although the mean for the 11 patients was lower all the patients were found to be within the distribution for all the CRC patient.

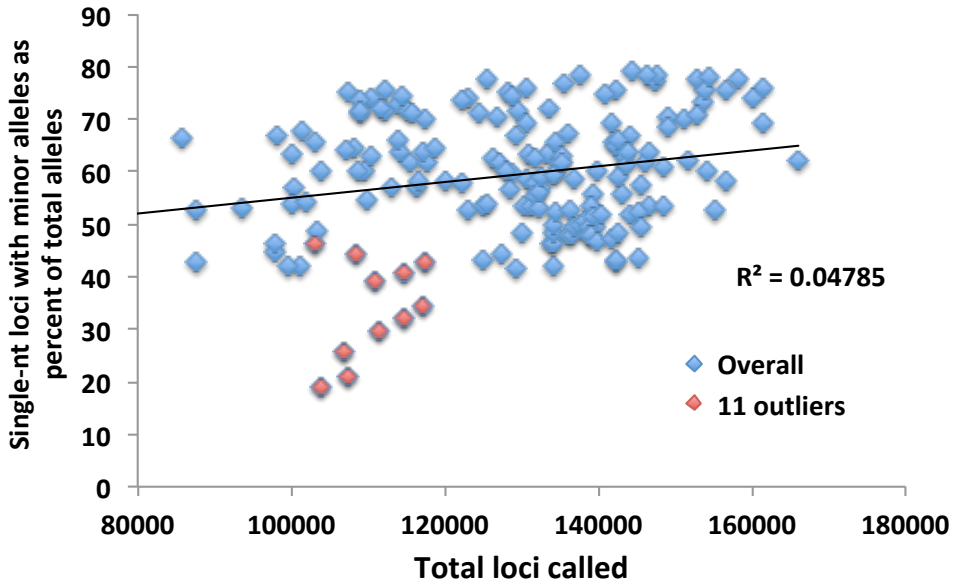


Figure 4-6: A binomial distribution is the best fit model for the comparison of single nucleotide SMV and total SMV for LIHC patients, with the inflection point serving as a break point between SMV-high and SMV stable. A) A binomial fit for the percent contribution of single nucleotide to total SMV. B) The percent of single nucleotide SMV as compared to total SMV. C) When the 5 statistical outliers (z transformation and Grubbs test), are omitted from the distribution the inflection point at 14% SMV, presented by the break in the line, serves as the break for SMV-high and SMV-stable.

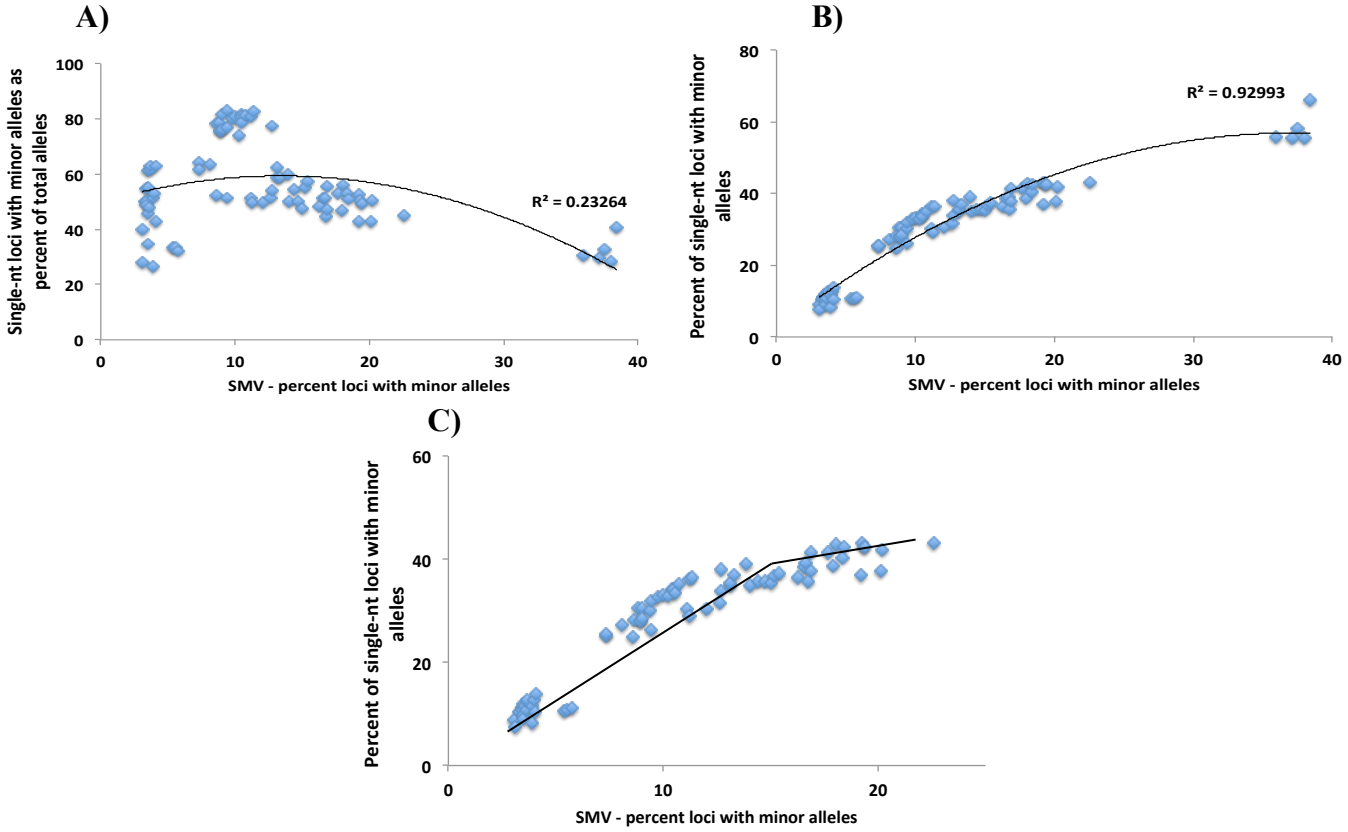


Figure 4-7: No difference in the distributions of total SMV against single nucleotide SMV between CRC and LIHC patients. An overlay of CRC and LIHC patient data.

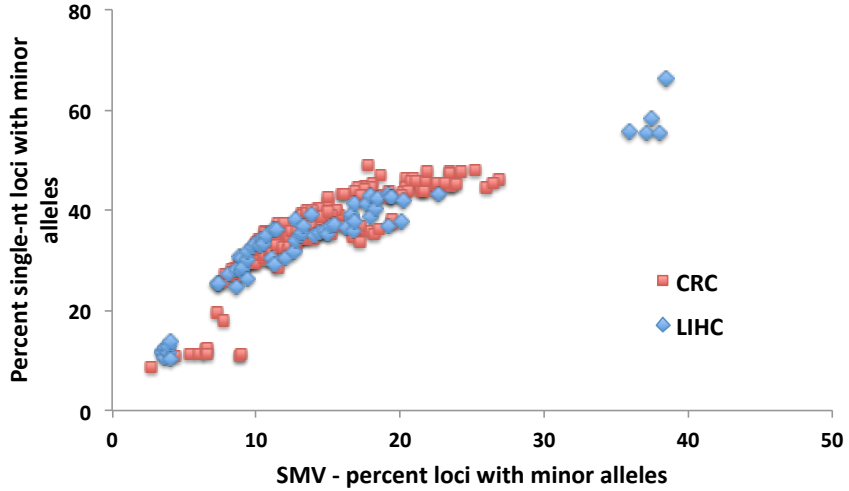
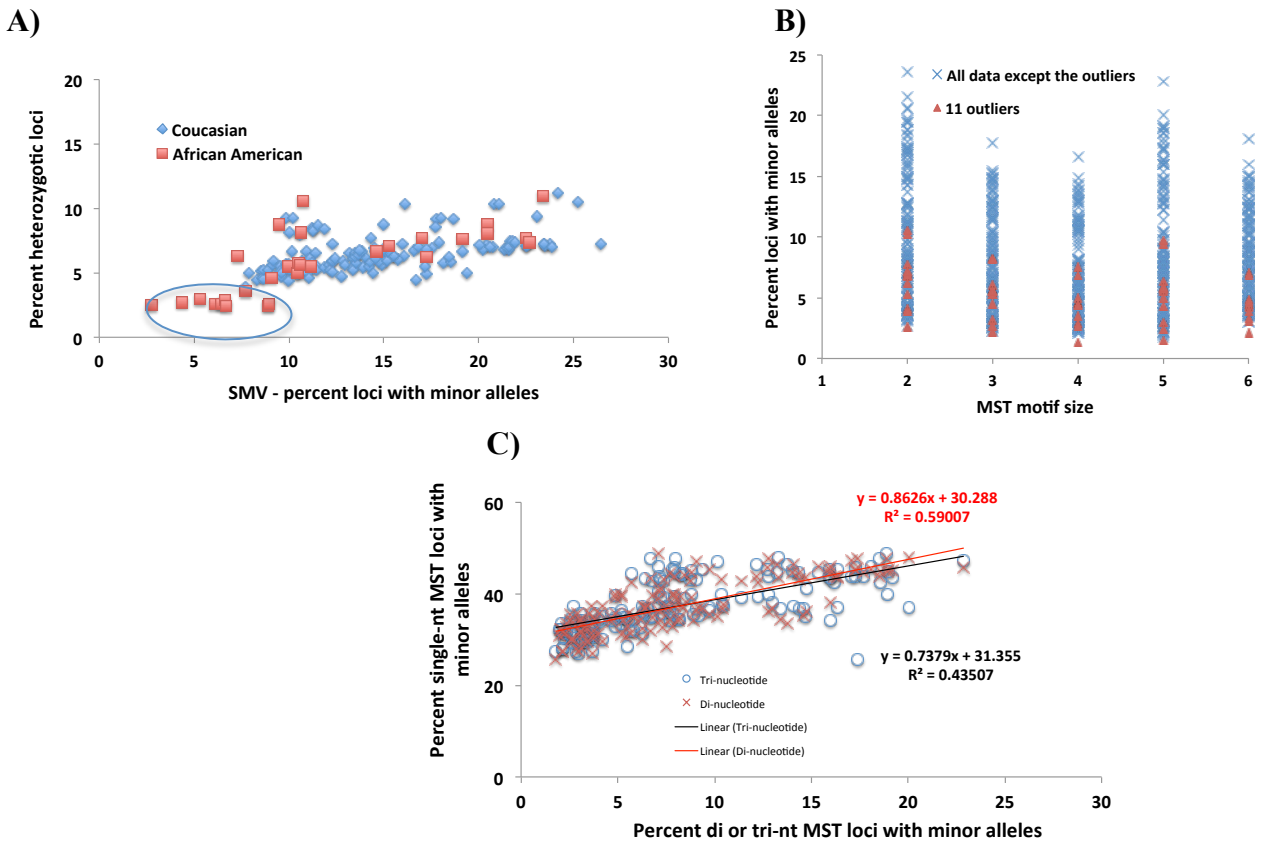


Figure 4-8: The 11 CRC outliers based on the previously described distribution are not outliers in any other MST motifs. A) A distribution of the percent heterozygotic loci as a function of total SMV rate. Encircled are the 11 outleirs who, as in figure 4-4A, remain outliers in as a function of both parameters, not just SMV. B) The 11 outliers (orange triangles) would are not outside of the distribution for any other motif length, other then single nucleotide repeats. C) The distributions of single nucleotide repeats as a function of di and tri nucleotide repeats SMV. The regression line slope is below one meaning di and tri nucleotide repeats increase at a higher rate then single nucleotide repeats



REFERENCES:

1. Caliman LP, Tavares RL, Piedade JB, AC DEA, K DEJDDC, et al. (2012) Evaluation of microsatellite instability in women with epithelial ovarian cancer. Oncol Lett 4: 556-560.
2. Adem C, Soderberg CL, Cunningham JM, Reynolds C, Sebo TJ, et al. (2003) Microsatellite instability in hereditary and sporadic breast cancers. Int J Cancer 107: 580-582.
3. Regitnig P, Moser R, Thalhammer M, Luschin-Ebengreuth G, Ploner F, et al. (2002) Microsatellite analysis of breast carcinoma and corresponding local recurrences. J Pathol 198: 190-197.
4. Jemal A, Bray F, Center MM, Ferlay J, Ward E, et al. (2011) Global cancer statistics. CA Cancer J Clin 61: 69-90.
5. Shah SA, Cleary SP, Wei AC, Yang I, Taylor BR, et al. (2007) Recurrence after liver resection for hepatocellular carcinoma: risk factors, treatment, and outcomes. Surgery 141: 330-339.
6. Pedica F, Ruzzenente A, Bagante F, Capelli P, Cataldo I, et al. (2013) A re-emerging marker for prognosis in hepatocellular carcinoma: the add-value of fishing c-myc gene for early relapse. PLoS One 8: e68203.
7. Lu X, Ye K, Zou K, Chen J (2014) Identification of copy number variation-driven genes for liver cancer via bioinformatics analysis. Oncol Rep 32: 1845-1852.
8. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-477.
9. Baptiste BA, Ananda G, Strubczewski N, Lutzkanin A, Khoo SJ, et al. (2013) Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. G3 (Bethesda) 3: 451-463.
10. Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, et al. (2010) What is a microsatellite: a computational and experimental definition based upon repeat mutational behavior at A/T and GT/AC repeats. Genome Biol Evol 2: 620-635.
11. Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, et al. (2013) Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. Genome Biol Evol 5: 606-620.
12. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.
13. Volker J, Gindikin V, Klump HH, Plum GE, Breslauer KJ (2012) Energy landscapes of dynamic ensembles of rolling triplet repeat bulge loops: implications for DNA expansion associated with disease states. J Am Chem Soc 134: 6033-6044.
14. Barros P, Boan F, Blanco MG, Gomez-Marquez J (2009) Effect of monovalent cations and G-quadruplex structures on the outcome of intramolecular homologous recombination. FEBS J 276: 2983-2993.
15. Grzeskowiak K, Ohishi H, Ivanov V (2005) Circular dichroism spectra of d(CGCGCGCGCGCG): evidence for intermediate models in the B-to-Z transition. Nucleic Acids Symp Ser (Oxf): 249-250.
16. Vaksman Z, Fonville NC, Tae H, Garner HR (2014) Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies. PLoS One 9: e110263.

17. Yoon K, Lee S, Han TS, Moon SY, Yun SM, et al. (2013) Comprehensive genome- and transcriptome-wide analyses of mutations associated with microsatellite instability in Korean gastric cancers. Genome Res 23: 1109-1117.
18. Siddle KJ, Goodship JA, Keavney B, Santibanez-Koref MF (2011) Bases adjacent to mononucleotide repeats show an increased single nucleotide polymorphism frequency in the human genome. Bioinformatics 27: 895-898.
19. Denver DR, Morris K, Kewalramani A, Harris KE, Chow A, et al. (2004) Abundance, distribution, and mutation rates of homopolymeric nucleotide runs in the genome of Caenorhabditis elegans. J Mol Evol 58: 584-595.
20. Fonville NC, Ward RM, Mittelman D (2011) Stress-induced modulators of repeat instability and genome evolution. J Mol Microbiol Biotechnol 21: 36-44.
21. Lauren J McIver NCF, Enusha Karunasena, Harold R Garner (Submitted) Microsatellite genotyping reveals a signature in breast cancer exomes. Breast Cancer Research and Treatment.
22. Mestrovic N, Castagnone-Sereno P, Plohl M (2006) Interplay of selective pressure and stochastic events directs evolution of the MEL172 satellite DNA library in root-knot nematodes. Mol Biol Evol 23: 2316-2325.
23. Williams LE, Wernegreen JJ (2013) Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont. Genome Biol Evol 5: 599-605.
24. Payseur BA, Jing P, Haasl RJ (2011) A genomic portrait of human microsatellite variation. Mol Biol Evol 28: 303-312.
25. Kim TM, Laird PW, Park PJ (2013) The landscape of microsatellite instability in colorectal and endometrial cancer genomes. Cell 155: 858-868.
26. Neveling K, Feenstra I, Gilissen C, Hoefsloot LH, Kamsteeg EJ, et al. (2013) A post-hoc comparison of the utility of sanger sequencing and exome sequencing for the diagnosis of heterogeneous diseases. Hum Mutat 34: 1721-1726.
27. Tanskanen T, Gylfe AE, Katainen R, Taipale M, Renkonen-Sinisalo L, et al. (2013) Exome sequencing in diagnostic evaluation of colorectal cancer predisposition in young patients. Scand J Gastroenterol 48: 672-678.
28. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, et al. (2010) Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One 5: e15661.
29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
30. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.
31. Kanchi KL, Johnson KJ, Lu C, McLellan MD, Leiserson MD, et al. (2014) Integrated analysis of germline and somatic variants in ovarian cancer. Nat Commun 5: 3156.
32. Natalie C Fonville LJM, Zalman Vaksman, Harold R Garner (Submitted) Microsatellites in the exome are predominantly single-allelic and invariant. . Genome Biology.
33. Haga H, Patel T (2014) Molecular diagnosis of intrahepatic cholangiocarcinoma. J Hepatobiliary Pancreat Sci.
34. Li S, Mao M (2013) Next generation sequencing reveals genetic landscape of hepatocellular carcinomas. Cancer Lett 340: 247-253.
35. Xu L, Hazard FK, Zmoos AF, Jahchan N, Chaib H, et al. (2014) Genomic analysis of fibrolamellar hepatocellular carcinoma. Hum Mol Genet.

36. Kim H, Park YN (2014) Hepatocellular carcinomas expressing 'stemness'-related markers: clinicopathological characteristics. Dig Dis 32: 778-785.
37. Kim DC, Chung WJ, Lee JH, Jang BK, Hwang JS, et al. (2014) Clinicopathological characteristics of PIK3CA and HBx mutations in Korean patients with hepatocellular carcinomas. APMIS 122: 1001-1006.
38. Steinke V, Holzapfel S, Loeffler M, Holinski-Feder E, Morak M, et al. (2014) Evaluating the performance of clinical criteria for predicting mismatch repair gene mutations in Lynch syndrome: a comprehensive analysis of 3,671 families. Int J Cancer 135: 69-77.
39. Abdulovic AL, Hile SE, Kunkel TA, Eckert KA (2011) The in vitro fidelity of yeast DNA polymerase delta and polymerase epsilon holoenzymes during dinucleotide microsatellite DNA synthesis. DNA Repair (Amst) 10: 497-505.
40. Hile SE, Shabashev S, Eckert KA (2013) Tumor-specific microsatellite instability: do distinct mechanisms underlie the MSI-L and EMAST phenotypes? Mutat Res 743-744: 67-77.
41. Hile SE, Yan G, Eckert KA (2000) Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. Cancer Res 60: 1698-1703.
42. Eckert KA, Mowery A, Hile SE (2002) Misalignment-mediated DNA polymerase beta mutations: comparison of microsatellite and frame-shift error rates using a forward mutation assay. Biochemistry 41: 10490-10498.
43. Shah SN, Hile SE, Eckert KA (2010) Defective mismatch repair, microsatellite mutation bias, and variability in clinical cancer phenotypes. Cancer Res 70: 431-435.

Chapter 5: Concluding remarks and future direction

RESEARCH OVERVIEW AND FUTURE DIRECTION:

Here we summarize this work, which falls into basically two sections: 1) the development of basic software technology to properly call all alleles at MSTs, and 2) to deploy this new enabling technology to a host of applications to advance the understanding of the roles of MSTs in biology and medicine. Lastly, throughout this last chapter, various ways in which this work can and should be advanced to continue to advance the science that is contained therein.

Enabling technology development:

Prior to this work, there was no software or even algorithm concepts directed at measuring all robust alleles within a nextgen sequence data set. Effort was at most focused on measuring the genotype at MSTs, which itself was a recent advance by our lab advancing all previous software that only called a single allele per locus, and those software packages did a poor job at that. My goal is to completely evaluate MSTs loci, calling all alleles, length change alleles, alleles with embedded SNPs; and calling those alleles at high accuracy/reliability.

Minor allele caller: Since mathematically MSTs are treated as a recurrent number series, computational string manipulation algorithms have difficulty aligning these sequences [1,2] and therefore, historically, in biology/genomic investigations MSTs are masked during analysis [2,3] because it has been difficult or impossible to accurately analyze these sequences. The current approach of handling MSTs, which was pioneered in the Garner Lab, is by aligning unique flanking regions adjacent to the MST and calling genotypes based on the sequence in between these flankers. This technique has the advantage of accuracy [3,4] but with a major drawback, alleles are called based on MST length but not sequence (ex.

Allele 1 – AAAAAA and allele 2 – AATAAAA, the program will call it homozygotic with 6 nt). By using this strategy alone we lose all the rich sequence variation that is due to embedded SNPs which is approximately 15-35% of the total diversity observed (chapter 2 [5]). Therefore, the initial goal of this work was to develop a novel approach for MST analysis that enables us to identify all the robust polymorphic alleles for each locus within a single sample, what we termed somatic MST variability. Our newly developed multi-allele caller captures low and high frequency alleles and bins them based on sequences and size differences [5].

The unique features of the multi-allele caller, namely the identification of low frequency, yet reliable alleles, establish a novel technique to study MST stability. In research, MSTs have generally been underexplored due to the difficulty and expense of working with them. This is most pronounced in single nucleotide runs longer then 14 nucleotides, for which accurate Sanger-sequencing based genotyping is nearly impossible. MSTs are highly sensitive to changes in genomic integrity and therefore can serve as genomic sensors as is currently done for MMR deficient Lynch syndrome patients. Although MSTs are under selective pressure, the polymorphism rate is over 3 times greater for MSTs, and over 5X for single nucleotide repeats, than that of non-repetitive DNA sequences (chapter 2 and 4). The balance between MSTs error/damage induction and error correction enables a sufficient signal to noise ratio to observe slight changes in genomic dynamics. Further, because the noise (normal variation rate) is so much more pronounced than in non-repetitive sequences the changes observed can be bi-directional. Both data presented in this thesis and an unpublished dataset we analyzed supports this claim. In Vaksman et al. (chapter 2 [5]) we showed both a loss of

heterozygosity and a decrease in the percent of loci with minor alleles in FANCD2 cells as compared to their derived retrovirally corrected line. However, in HEK293 cells the addition of HSP90 inhibitor causes a significant increase in SMV with no change in genotype (unpublished data).

To date, the only clinical application that employs MST stability analysis is to determine the MSI status of tumors in spontaneous or familial colorectal cancer. However, MSTs are the causative factor for over 40 tri-nucleotide expansion disorders such as Huntington's chorea and Friedreich's ataxia [6]. Although these are heritable disorders, for some such as fragile X, the causative MST expansion is not present in parents but is caused by somatic mutation in meiotic cells. Little is known of the susceptibility for expansion nor is there a clinical test to determine predisposition for MST expansion. Since the approach presented here is a global genomic platform that is sensitive to the presence of low frequency alleles this method may be employed to establish a predisposition to tri-nucleotide, indeed all repeat motif, expansions in unaffected proband family members.

Applications to targeted sequencing: A natural expansion of the multi-allele caller is its use with targeted capture enrichment data. The targeted capture enrichment and targeted hybridization methods such as the SureSelect, (Agilent Santa Clara, CA), SeqCap (Roche NimbleGen, Madison, WI) kits use targeted DNA "bait" sequences in order to isolate DNA regions of interest for next-gen sequencing. This low coverage high depth procedure results in hundreds to thousands reads per target and the theoretical possibility of capturing large numbers of minor alleles. However, due to several limitations of pipeline and more

specifically the minor allele caller, this method is not optimized for use with these types of studies. However, with recent data captured by the lab, work is underway to understand the applicability and accuracy of such algorithms when applied to extremely high depth data.

The minor-allele caller was optimized for use in a variety of parameter spaces typical of nextgen sequencing projects: high/low coverage/depth (tens to few hundreds reads per locus), millions of loci, HiSeq exome, RNAseq, customized microsatellite-specific target enrichment kits and whole genome sequencing. However, the parsing and genotyping/allele-calling method currently employed by the minor-allele caller, as a stand-alone program, must still be characterized and adjusted for the error rates associated high depth MST targeted sequencing. Undergoing development is an expansion module which uses globally aligned .sam files and a built-in custom Waterman-Smith algorithm for local realignment of repeat sequences. The most difficult problem to overcome is to statistically adjust for the Hi/MiSeq (Illumina, San Diego, CA) single nucleotide repeat embedded read length distribution. In a current dataset, with over 4000 targeted single nucleotide repeat sequences, the typical distribution which comprises over 70-90% of the data is N_i, $N\text{-}1_j$, $N\text{-}2_k$, $N+1_l$….. (where N is the most common length N-/+n is the change in length and i corresponds to the read depth for the most common allele and j,k,l are a stepwise percent reduction) with 3 – 10 times more alleles identified than the any other MST motif size regardless of repeat cycle number (total length). Optimally, since mathematically the data can be inspected as a directional change distribution, however, the current iteration is attempting to handled it with vector analysis using special statistics methods [7]. One of my goals is to continue to advance this research, post-graduation, to meet the needs of the lab and the needs of researchers as they respond to

new nextgen technology which will produce even higher depth data at longer read lengths.

Application to biology and medicine; SMV in DNA repair disorders:

Somatic mutations, novel genomic polymorphisms that arise within a cell population, play a critical role in cellular reprogramming leading to the development of cancer [8]. Somatic mutations result from DNA damage or inappropriate nucleotide insertion during replication. Suppression of somatic mutations is essential for genomic stability, therefore, to survive, cells have evolved multiple mechanisms to repair damaged or unpaired nucleotides [9-11]. Congenital DNA repair deficiency disorders (DRDDs) is a conglomerate of inherited conditions whereby affected individuals are unable to appropriately resolve DNA damage, leading to cell cycle arrest or genomic instability [12-16]. These disorders are characterized by neurological deficiencies, poor skin pigmentation, short stature, photosensitive and a predisposition to cancer at various stages of life. Although genomic instability is a hallmark of DRDDs it is often measured with regards to loss of heterozygosity (LOH) or chromosomal mosaicism rather than the increase in somatic variability [17-19]. As a result, little is known about the prevalence of somatic mutation or changes in somatic variability in these patients, including those that have developed cancer. This work is a significant advancement, in that it directly measures and quantifies the increased diversity of cellular populations, and has enabled us to better understand these disorders.

Single and interstrand DNA damage includes radiation or chemically induced cross-linking, loss of nucleotides, oxidation, deamination, nucleotide mis-insertion or simply mechanical severing. Cells have evolved highly specialized repair mechanisms to handle each of the

aforementioned types of DNA damage [20,21]. The machinery required for each repair pathway is very specific with few known overlapping proteins. All DNA repair mechanisms are grouped into either single or double strand repair with mainly polymerases and helicases redundantly overlapping between the two types of damage repair. Single strand DNA repair (SSDR) is further sub-grouped into mismatch, nucleotide excision and base excision repair (MMR, NER and BER) [22,23]. MMR and NER are bulk adduct repair pathways involved in gross single strand aberrations due to photoreactivity or DNA malformation while BER is responsible for the repair of chemical modifications to the DNA. Double strand DNA repair (DSDR) is divided into homologous recombination (HR), nonhomologous end joining (NHEJ) and cross-link repair (FANC). FANC and HR are integrally linked with many various HR proteins also having FANC names (BRCA2 is also known as FANCD1) [15,24-26]. Without a new tool, like the measurement of the SMV, the proper differentiation of each of these different repair deficiencies was not possible. My goal as to change this.

DRDDs are disorders for which one or more of these pathways are disrupted. Individuals with these disorders accumulate somatic mutations at a much higher rate than a healthy individual, leading to progeria-like symptoms and a cancer predisposing. The types of mutations normally anticipated from each pathway are fairly defined and are based on the function of the protein conglomerate when attached to the DNA. For example, in Ruthmond-Thumson (RecQL4) or Fanconi anemia D1 (BRCA2 gene) patients, double strand damage leads to a loss of genomic segments as large as 5kb due to repair being initiated by NHEJ or strand annealing machinery [27-29]. However, for Bloom and Werner syndromes, chromosomal instability is common due to the lack of Holiday junction resolution [12,24].

SSDR pathways - a comparison of cell line and patient DNA sequences: Cell lines are considered models for human conditions and often display similar (but not the same) response patterns to the modeled patient population. Our initial study, presented in chapter 2, included a colorectal cancer tumor cell line lacking a functional MutL/S component of MMR, the only DNA repair pathway that has been directly associated with MSI. The DLD-1 cell line is commonly used to characterize MMR function associated with Lynch syndrome [30-32]. These cells are MST unstable and are primarily diploid for all chromosomes, an unusual characteristic for a cancer cell line [33,34]. Due to the lack of a functional MMR pathway in the DLD-1 cells we anticipated an increase in the mutation rate leading to an increase in the number of loci that have minor alleles. However, results depicted in chapter 2 show that although DLD-1 cells displayed a greater number of heterozygotic loci as compared to "normal" cell lines, the fraction of loci with minor alleles was significantly less than seen in matching "normal" cell lines. Interestingly, a similar pattern emerged in our cohort of MSI-high (MMR impaired) colorectal cancer patient tumor samples. Depicted in chapter 4 table 1, individuals showing MST instability, MSI-high, were shown to present a similar pattern of SMV, with a significantly higher fraction of heterozygotic loci than the MST stable or MSI-low colorectal cancer patients.

Unlike the clear-cut effects of MMR on MST stability, the consequences of impairment of either of the NER pathways, GG-NER and TC-NER, on MST stability are unknown. Recently, evidence of an indirect effect of NER through ineracting cross-pathway proteins has surfaced [35,36]. Exome sequencing results presented in chapter 3 indicated that both NER

pathways play a role in the stability of MSTs. Impairment of CSA function (ERCC8, Cockayene's syndrome A) significantly increases overall SMV by over 30% however XP subgroup A patients were no different from the control group. However, in exonic regions, SMV in XP-A patients was 55% greater than the control group. Similarly, CSA deficient patients also exhibited a significantly greater SMV rate, ~ 40% greater, than the control group. Further, for both CSA and XP-A patients, the concordance in loci containing minor alleles was significantly lower than for the control subjects. Concordance was calculated as the fraction of loci with minor alleles that were found in 75% of the samples. Only loci that were found in 100% of the samples regardless of minor alleles status were used to calculate concordance. Based on results from the cell lines [5] concordance between subjects ranges between 10 and 20%. For the control group concordance was over 10% of the loci while for CSA and XP-A concordance was 4.6 and 3.6% respectively. Based on these data we can conclude that NER impairments affect MST stability in various ways, including increasing MST polymorphism rates that exhibit a more random distribution of mutations. These results are significant since they are the first clear evidence that NER does play a role in MST stability by either directly or indirectly stabilizing MSTs. However, because we do not as of yet have data for patients with CSB deficiency or other XP genes, we cannot determine the exact mechanism responsible for this result. However, a future goal of this line of research is to measure this as more well-characterized model cell lines become available from collaborators.

Double strand DNA repair – multi gene disorders: Damage resulting in double strand breaks requires a different approach to repair. In order to make an exact duplicate of the damaged

DNA a cell requires a second, identical strand and a mechanism for filling in the damaged region, HR [37,38]. Impairments in HR leads to either LOH or loss of chromosomal regions. BRCA2 is an essential component of HR. BRCA2 associates with the damaged DNA and Rad51 in order to maintain the close proximity of the broken ends, initiate strand invasion and recruit the necessary repair mechanisms [37]. Capan-1 is the only known cell line to have a completely non-functional BRCA2. Our exome analysis of these cells showed that, as anticipated, a significant loss of heterozygosity. Further, we also detected a significant increase in the fraction of loci with minor alleles, 6.2% as compared to 5.1% (Capan-1 cells and "normal" cell line mean respectively) (Table 3). We also found a difference between the SNP to indel ratio, nearly a 1:1 ratio, for Capan-1 cells. This was significantly different from all other cell lines tested (Table 4).

The FANC complex and pathway is comprised of >14 genes responsible for interstrand cross-link repair [39]. It is directly associated with HR, with a subgroup of the disorder involving the BRCA2 gene, also known as FANC group D1. The severity of the disorder ranges greatly and is based on the subgroups/proteins that carry the mutation [14,40]. Unlike NER and base excision repair, sequencing and data analysis of patients with an impaired BRCA2 gene, mainly those with familial breast cancer, have been done, however, because the major focus is development of disease predictive secondary mutations, little exploration has been done on genomic integrity and structural stability involved. A targeted analysis to study this genomic stability phenomenon, such as structural variation, loss of intermediate length sequences (2-5kb), changes in telomeres and microsatellite stability should yield important information about the function of BRCA2 on a genome wide scale.

An important aspect to consider is a correlation between mutation/protein modification and the severity of the disorder. In the case of a disorder caused by a large number of proteins such as that anticipated for Fanconi anemia, the severity varies based on the particular gene impacted. For example, patients with FANCA or B can survive to mid-life, however, a deficiency in FANCD1 function leads to immediate cancer development and death during childhood. Preliminary data for Fanconi anemia was obtained by sequencing the exomes of a FANCD2 cell line (PD20) and two patient samples, FANCC and FANCG. We are collaborating with others to extend this preliminary data to investigate the mechanism though insights into SMV quantification. The data show that a significant LOH, as compared to controls (3.3, 2.8, 2.4, and 2.1% "normal" mean, FANCD2, C and G respectively, table 5-1). Although inactivation of FANCD2 also showed a reduction in the number of heterozygotic loci, it was much less than FANCC and G samples (table 5 – 1). Unexpectedly, we found that the propensity of loci to acquire minor alleles was similarly reduced from "normal", for all 3 samples (combined data from chapter 2 and table 5 – 1). This is most likely due to population bottlenecks caused by massive cell death from the accumulation of mutations at a high rate. The hypothesis is substantiated by the slow growth rates these cells exhibited even under optimal conditions with no stressors present.

BRCA2 and FANCD1 are 2 sets of mutations in the same gene, (for which the official nomenclature now is) BRCA2, with similar effects on the protein but very different clinical outcomes. Individuals with BRCA2 mutations have a predisposition to breast, cervical and others cancers by mid- to late- life, while FANCD1 mutations are fatal by age 6 - 10

(http://www.ncbi.nlm.nih.gov/books/NBK1401/). The reason for the disparity and differences in the effects on the genome instability is unknown. As a future direction of this project our goal is to explore differences in genomic abnormalities and somatic variation, including SMV, in these two patients sets. The mutations introduced in both disorders cause either a truncation or a loss of a large segment of the 3' end of the gene. The result is a ~70% reduction in nuclear localization of BRCA2 or an inappropriate protein modification and loss of capacity to interact with other HR proteins [41,42]. BRCA2 mutations lead to a predisposition for breast and cervical cancer with a median unset by the age of 40 or 50 [43,44]. FANCD1 mutations, Fanconi anemia subgroup D1, are associated with early-onset (before 6 years of age) of untreatable leukemia, medulloblastomas and other solid tumors (www.ncbi.nlm.nih.gov/books/NBK1401). The reason for the disparity in disease severity is unknown. The most plausible hypothesis to explain the disparity is that the FANCD1 mutation causes greater general genomic instability than the BRCA2 mutation in exactly the same gene. Because of the rarity of FANCD1 samples, no publication currently exists which describe an in depth genomic analysis of the FANCD1 genome.

Understanding the differences in the effects exhorted on the genome will most likely shed light on various unknown functions and maybe association/interactions of the BRCA2 gene. The differences that emerged between Capan-1 cells and the FANC samples (chapter 2) leads to the expectation of that the pattern of SMV for FANCD1 samples will resemble Capan-1 cells more than the rest of the FANC subgroups. Both Capan-1 and FANCD1 mutations nearly completely abolish BRCA2 nuclear localization [42]. Further, this mutation causes one of the most severe forms of Fanconi anemia, with severe developmental retardation, short

stature, extreme photosensitivity, very early onset of cancer and death before age 10 [40]. Based on the severity of symptoms we would expect a rapid accumulation of mutations, mosaicism and an increase in somatic variability. However, lack of progress in these studies occur for the same reasons; the lack of viable BRCA2 cell lines from these severely affected patients as well as breast cancer patients and because attempts to create such cell lines yields very slow growing lines, i.e. the Capan-1 cell line proliferation puzzle.

Summarizing the future:

This work has established new technological methods to analyze the new and ever increasing amount of nextgen sequence, whether it is coming from patients or model cell lines established to understand the aberrant function and mechanism resulting in disease. We then went on to apply this new analysis approach to make advances in understanding some particularly resilient disease examples. It is my goal to continue to apply this technology, and develop more advanced forms of this technology, to further understand the wide variety of diseases, especially cancer. In addition, as touched upon in this chapter, we have begun to apply this to an ever increasing set of example disorders. Preliminary data has indicated that this approach will be informative in a host of diseases and disease variants. We have also established new collaborations to obtain the necessary materials, cell lines and patient samples, to pursue these investigations. Data, observations and manuscripts are forthcoming.

TABLES:

Table 5-1: MST and non-MST from standard exome sequencing of 'normal' cells, but not from sequencing of a single cell after whole genome amplification, show the expected high ratio of INDELs (expansions and contractions) to SNPs.

	PD20 RV:D2-1	PD20 RV:D2-2	MCF10A	HEK293	"Normal" cells Mean	"Normal" cells SD	PD20 FANCD2	FANCG	FANCC	Capan-1	DLD-1
Homo-zyg	96.8	96.8	96.4	97.0	96.7	0.3	97.2 #	97.5 #	97.9 #	97.9 #	94.5 #
Hetero-zyg	3.2	3.2	3.6	3.0	3.3	0.3	2.8 #	2.4 #	2.1 #	2.1 #	5.5 #
Minor alleles	5.4	5.3	4.5	5.3	5.1	0.4	3.1 #	3.8 #	2.9 #	6.2 #	3.2 #

REFERENCES

1. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.
2. Leclercq S, Rivals E, Jarne P (2007) Detecting microsatellites within genomes: significant variation among algorithms. BMC Bioinformatics 8: 125.
3. Highnam G, Franck C, Martin A, Stephens C, Puthige A, et al. (2013) Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. Nucleic Acids Res 41: e32.
4. McIver LJ, McCormick JF, Martin A, Fondon JW, 3rd, Garner HR (2013) Population-scale analysis of human microsatellites reveals novel sources of exonic variation. Gene 516: 328-334.
5. Vaksman Z, Fonville NC, Tae H, Garner HR (2014) Exome-wide somatic microsatellite variation is altered in cells with DNA repair deficiencies. PLoS One 9: e110263.
6. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet 44: 445-477.
7. Maguire DJ, Batty M, Goodchild MF (2005) GIS, spatial analysis, and modeling. Redlands, Calif.: ESRI Press. xiii, 480 p. p.
8. Poduri A, Evrony GD, Cai X, Walsh CA (2013) Somatic mutation, genomic variation, and neurological disease. Science 341: 1237758.
9. Harris RS, Kong Q, Maizels N (1999) Somatic hypermutation and the three R's: repair, replication and recombination. Mutat Res 436: 157-178.
10. Stratton MR (2011) Exploring the genomes of cancer cells: progress and promise. Science 331: 1553-1558.
11. Kunz C, Saito Y, Schar P (2009) DNA Repair in mammalian cells: Mismatched repair: variations on a theme. Cell Mol Life Sci 66: 1021-1038.
12. Manthei KA, Keck JL (2013) The BLM dissolvasome in DNA replication and repair. Cell Mol Life Sci 70: 4067-4084.
13. Lin Y, Wilson JH (2012) Nucleotide excision repair, mismatch repair, and R-loops modulate convergent transcription-induced cell death and repeat instability. PLoS One 7: e46807.
14. Kottemann MC, Smogorzewska A (2013) Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. Nature 493: 356-363.
15. Fan W, Luo J (2008) RecQ4 facilitates UV light-induced DNA damage repair through interaction with nucleotide excision repair factor xeroderma pigmentosum group A (XPA). J Biol Chem 283: 29037-29044.
16. Trego KS, Chernikova SB, Davalos AR, Perry JJ, Finger LD, et al. (2011) The DNA repair endonuclease XPG interacts directly and functionally with the WRN helicase defective in Werner syndrome. Cell Cycle 10: 1998-2007.
17. Butz J, Wickstrom E, Edwards J (2003) Characterization of mutations and loss of heterozygosity of p53 and K-ras2 in pancreatic cancer cell lines by immobilized polymerase chain reaction. BMC Biotechnol 3: 11.

18. Nardo T, Oneda R, Spivak G, Vaz B, Mortier L, et al. (2009) A UV-sensitive syndrome patient with a specific CSA mutation reveals separable roles for CSA in response to UV and oxidative DNA damage. Proc Natl Acad Sci U S A 106: 6209-6214.
19. Pedroni M, Di Gregorio C, Cortesi L, Reggiani Bonetti L, Magnani G, et al. (2013) Double heterozygosity for BRCA1 and hMLH1 gene mutations in a 46-year-old woman with five primary tumors. Tech Coloproctol.
20. Best BP (2009) Nuclear DNA damage as a direct cause of aging. Rejuvenation Res 12: 199-208.
21. Hou Y, Song L, Zhu P, Zhang B, Tao Y, et al. (2012) Single-cell exome sequencing and monoclonal evolution of a JAK2-negative myeloproliferative neoplasm. Cell 148: 873-885.
22. Menck CF, Munford V (2014) DNA repair diseases: What do they tell us about cancer and aging? Genet Mol Biol 37: 220-233.
23. Freitas AA, de Magalhaes JP (2011) A review and appraisal of the DNA damage theory of ageing. Mutat Res 728: 12-22.
24. Kitano K (2014) Structural mechanisms of human RecQ helicases WRN and BLM. Front Genet 5: 366.
25. Ohashi A, Zdzienicka MZ, Chen J, Couch FJ (2005) Fanconi anemia complementation group D2 (FANCD2) functions independently of BRCA2- and RAD51-associated homologous recombination in response to DNA damage. J Biol Chem 280: 14877-14883.
26. Hinz JM, Nham PB, Salazar EP, Thompson LH (2006) The Fanconi anemia pathway limits the severity of mutagenesis. DNA Repair (Amst) 5: 875-884.
27. Kass EM, Jasin M (2010) Collaboration and competition between DNA double-strand break repair pathways. FEBS Lett 584: 3703-3708.
28. Patel AG, Sarkaria JN, Kaufmann SH (2011) Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. Proc Natl Acad Sci U S A 108: 3406-3411.
29. Wang H, Zeng ZC, Bui TA, DiBiase SJ, Qin W, et al. (2001) Nonhomologous end-joining of ionizing radiation-induced DNA double-stranded breaks in human tumor cells deficient in BRCA1 or BRCA2. Cancer Res 61: 270-277.
30. Glaab WE, Tindall KR, Skopek TR (1999) Specificity of mutations induced by methyl methanesulfonate in mismatch repair-deficient human cancer cell lines. Mutat Res 427: 67-78.
31. Dexter DL, Spremulli EN, Fligiel Z, Barbosa JA, Vogel R, et al. (1981) Heterogeneity of cancer cells from a single human colon carcinoma. Am J Med 71: 949-956.
32. Russo MT, Blasi MF, Chiera F, Fortini P, Degan P, et al. (2004) The oxidized deoxynucleoside triphosphate pool is a significant contributor to genetic instability in mismatch repair-deficient cells. Mol Cell Biol 24: 465-474.
33. Chen TR, Dorotinsky CS, McGuire LJ, Macy ML, Hay RJ (1995) DLD-1 and HCT-15 cell lines derived separately from colorectal carcinomas have totally different chromosome changes but the same genetic origin. Cancer Genet Cytogenet 81: 103-108.
34. Ghadimi BM, Sackett DL, Difilippantonio MJ, Schrock E, Neumann T, et al. (2000) Centrosome amplification and instability occurs exclusively in aneuploid, but not

in diploid colorectal cancer cell lines, and correlates with numerical chromosomal aberrations. Genes Chromosomes Cancer 27: 183-190.
35. Lin Y, Hubert L, Jr., Wilson JH (2009) Transcription destabilizes triplet repeats. Mol Carcinog 48: 350-361.
36. Hubert L, Jr., Lin Y, Dion V, Wilson JH (2011) Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. Hum Mol Genet 20: 4822-4830.
37. Costanzo V (2011) Brca2, Rad51 and Mre11: performing balancing acts on replication forks. DNA Repair (Amst) 10: 1060-1065.
38. Grigorova M, Staines JM, Ozdag H, Caldas C, Edwards PA (2004) Possible causes of chromosome instability: comparison of chromosomal abnormalities in cancer cell lines with mutations in BRCA1, BRCA2, CHK2 and BUB1. Cytogenet Genome Res 104: 333-340.
39. Pickering A, Zhang J, Panneerselvam J, Fei P (2013) Advances in the understanding of Fanconi anemia tumor suppressor pathway. Cancer Biol Ther 14.
40. Schindler D, Hoehn H (2007) Fanconi anemia : a paradigmatic disease for the understanding of cancer and aging. Basel ; New York: Karger. xii, 229 p. p.
41. Holt JT, Toole WP, Patel VR, Hwang H, Brown ET (2008) Restoration of CAPAN-1 cells with functional BRCA2 provides insight into the DNA repair activity of individuals who are heterozygous for BRCA2 mutations. Cancer Genet Cytogenet 186: 85-94.
42. Feng Z, Scott SP, Bussen W, Sharma GG, Guo G, et al. (2011) Rad52 inactivation is synthetically lethal with BRCA2 deficiency. Proc Natl Acad Sci U S A 108: 686-691.
43. Chen S, Parmigiani G (2007) Meta-analysis of BRCA1 and BRCA2 penetrance. J Clin Oncol 25: 1329-1333.
44. Easton DF (1999) How many more breast cancer predisposition genes are there? Breast Cancer Res 1: 14-17.

APPENDIX: Copyrights

Below is the statement on the use or reuse of work published by PLoS One. The work this refers to is chapter 2 in this dissertation.