# Novel Microsatellite Detection, Microsatellite Based Biomarker Discovery in Lung Cancer and The Exome-Wide Effects of a Dysfunctional DNA Repair Mechanism

Karthik Raja Velmurugan

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

Harold R. Garner, Co-chair
David R. Bevan, Co-chair
Christopher B. Lawrence
Pawel Michalak

February 15, 2017
Blacksburg, Virginia

Keywords: Microsatellite, Next-Generation Sequencing, Lung Cancer, Fanconi Anemia, Biomarker Discovery, Bioinformatics algorithm

# Novel Microsatellite Detection, Microsatellite Based Biomarker Discovery in Lung Cancer and The Exome-Wide Effects of a Dysfunctional DNA Repair Mechanism

Karthik Raja Velmurugan

## ABSTRACT

Since the dawn of the genomics era, the genetics of numerous human disorders has been understood which has led to improvements in targeted therapeutics. However, the focus of most research has been primarily on protein coding genes, which account for only 2% of the entire genome, leaving much of the remaining genome relatively unstudied. In particular, repetitive sequences, called microsatellites (MST), which are tandem repeats of 1 to 6 bases, are known to be mutational hotspots and have been linked to diseases, such as Huntington disease and Fragile X syndrome. This work represents a significant effort towards closing this knowledge gap. Specifically, we developed a next generation sequencing based enrichment method along with the supporting computational pipeline for detecting novel MST sequences in the human genome. Using this global MST enrichment protocol, we have identified 790 novel sequences. Analysis of these novel sequences has identified previously unknown functional elements, demonstrating its potential for aiding in the completion of the euchromatic DNA.

We also developed a disease risk diagnostic using a novel target specific enrichment method that produces high resolution MST sequencing data that has the potential to validate, for the first time, the link between MST genotype variation and cancer. Combined with publically available exome datasets of non-small cell lung cancer and 1000 genomes project, the target specific MST enrichment method uncovered a

signature set of 21 MST loci that can differentiate between lung cancer and non-cancer control samples with a sensitivity ratio of 0.93.

Finally, to understand the molecular causes of MST instability, we analyzed genomic variants and gene expression data for an autosomal recessive disorder, Fanconi anemia (FA). This first of its kind study quantified the heterogeneity of FA cells and demonstrated the possibility of utilizing the DNA crosslink repair dysfunctional FA cells as a suitable system to further study the causes of MST instability.

# GENERAL AUDIENCE ABSTRACT

The field of genetics has enjoyed substantial growth since the conclusion of the human genome project, which was declared complete in the year 2003. The human genome project produced the first framework for the human DNA sequence, the human genome. With the availability of this framework, the understanding of the genetic basis for a number of diseases has significantly grown, which has resulted in better methods of clinical diagnosis and treatment. While the current focus on understanding the genomic regions that are responsible for making proteins has inarguably helped, it has also created a gap in knowledge. Protein coding regions of the human genome account only for 2% of the entire human genome and a large part (47%) of the genome is occupied by repetitive DNA. DNA sequences can be complex, with the nucleotides arranged in no particular order, e.g. ATCGTACGA, or DNA sequences can be repetitive, e.g. ATATATATAT. Repetitive sequences, which have repeating units of 1 to 6 bases, are called microsatellites (MST). MSTs have been shown to be unstable and they have been linked to diseases such as Huntington disease and Fragile X syndrome. This work helps to close this knowledge gap by developing molecular methods and computational tools focused on identifying MST variations. Research conducted with this aim has resulted in three major accomplishments. One, we developed novel molecular and computational methods which we used to detect 790 previously unknown sequences in the human genome. This work proved the ability of our method to uncover functional elements in the human genome that can potentially answer numerous biological questions. Two, we developed another novel method for the production of high resolution MST sequence data that not only can improve MST research in general but also shows the potential for the development of new genetic diagnostics and cancer therapeutics. We identified a signature set of 21 MST sequences that can differentiate between lung cancer patient genomes and non-cancer control genomes. These results represent the first potential validation for a proposed link between MST sequence length (genotype) variation and cancer. Three, we attempt to understand a possible molecular cause and consequences of MST instability in a disease called Fanconi anemia. The results from this work not only, for the first time, quantify the effects of this disease on the genome but also establishes Fanconi anemia as a suitable system for studying MST instability in detail.

## ACKNOWLEDGMENTS:

Indian tradition propounds that there are four important persons in every individual's life: 1. Mother, 2: Father, 3. Teacher and 4. The quintessence of life. I have found this list and its order to have been well meditated upon and hence have chosen to express my gratitude in a similar fashion.

Mom, thank you for the undying support and instilling in me an ache to learn. Dad, thank you for the trust that you have on me to let me conduct my own life. I have found us to be an impenetrable unit (both of you, Ananthi and I) of fellow travelers and this PhD work is but another merry consequence of our comradery.

“Teachers are not made, but born,” are Osho's words. I find this to be very true. While I have had the opportunity to study under an innumerable number of teachers, three have immeasurably contributed to my career and to my own personal growth as a responsible human being.

Dr. Pandjassarame Kangueane, your consistency and hope for the future is unbelievable. Even when we are on the opposite sides of the world I know that you are working untiringly, from wherever you are, from 5 in the morning to late in the evening. The very remembrance of that persistence gives me strength.

Dr. Jin Gyoon Park, your kindness and patience is nothing like I have ever seen. When I came to you, I only had the will to learn and you transformed it into knowledge and confidence. Please know that I immensely appreciate the opportunity I had to work with a terrific scientist like you.

Dr. Harold Garner, when I joined the GBCB program, never did I imagine that I would end up working with you and here we are with me, possibly, being your last graduate student, which I hope is a position I am worthy off. By the time I started working with you I had almost lost all the excitement I had for science but you somehow found a way to bring it all back. The working style of every student is heavily influenced by his/her teacher. I will carry with me these gems: the attention to every little detail, the incredible comprehensiveness of work, the can-do attitude and the collectedness of a martial arts teacher.

It becomes immediately impossible to adequately convey my gratitude without the mentioning of the battle Paramount that I have found myself to be engaged in. Never did I even imagine that when I started this graduate school journey I would also go through a spiritual un-conditioning. To quote one of my favorite movie lines: “My journey took me somewhat further down the rabbit-hole than I'd intended and, though I dirtied my fluffy white tail, I've emerged enlightened.” – Sherlock Holmes.

I dedicate this work to the existence
that never asks to be acknowledged,
that does through non-doing,
that exists beyond the paradigms of space and time,
that with its infinite wisdom has found me,
that which empowers and contains me and becomes the end of all questions.

Bavani, if not for you I would have gone through graduate school alone. Your accompaniment means the world to me.

Nick, Robin and Ramu: Insanity was not far away if not for you guys to keep it at bay with your steady extension of support and kind words.

Natalie: All the data that has been analyzed and presented in this thesis is based on the bench work you did. While my PhD ended up being a unique situation, it would have been phenomenally difficult for me to have continually progressed if not for the wet lab work you promptly did before you left.

To my PhD thesis committee, Drs. David Bevan, Chris Lawrence and Pawel Michalak, thank you for helping me get through the many formalities and, more importantly, for your ready willingness to help me in spite of your busy schedules.

FACILITIES:
I would like to express my appreciation to VCOM and the Bradley foundation for supporting this work.

I would also like to express my sincere thanks to Erin Cassidy at the Virginia Bioinformatics Institute for the continued support and access to VBI's supercomputing cluster (Shadowfax) and to the ever helpful VBI's computational core.

## Table of Contents:

## LIST OF FIGURES

**Chapter 1:**

**Chapter 2:**

**Chapter 3:**

**Chapter 4:**

## LIST OF TABLES

Chapter 2:

Chapter 3:

**Chapter 4:**

# Chapter 1: Introduction and overview

# INTRODUCTION

**The genome and repetitive DNA:** The human genome project was declared complete in 2004[1]. Since then, a total of 3981 eukaryotic organisms have been sequenced[2]. The human genome is approximately 3 billion base pairs long[3]. Protein coding genes occupy about 2% of the human genome while introns occupy 24% (Figure 1-1). About 47% of the human genome is found to be repetitive sequences. Other organisms, compared to human, have a higher percentage of their genome occupied by repetitive DNA: Mus musculus (mouse) – 55%, Gallus gallus (chicken) – 88%, Drosophila melanogaster (fruit fly) – 71%, Anopheles gambiae (mosquito) – 85% and Caenorhabditis elegans – 80%[4]. It is evident that a large percentage of most genomes is composed of repetitive DNA. Repeats can be long transposable elements (DNA sequences that can alter their position in the genome) such as DNA transposons and retrotransposons or tandem repeats such as minisatellites and microsatellites.

**Microsatellites and microsatellite instability:** Microsatellites (MST) are short tandem repeats of 1 to 6 bases (Minisatellites contain repeats of greater than six bases). Repetitive DNA sequences are prone to slippage (replication slippage happens when the DNA polymerase erroneously adds more nucleotides than needed) and breaks (chromosomal translocations) during DNA replication by DNA polymerase. Hence MSTs are often mutational hot spots. The mutational rates at MSTs have been shown to be higher than what is found to occur in non-repetitive DNA[5-8]. Microsatellite mutations, or instability (MSI), caused by polymerase slippage has been shown to occur at a higher rate in MSTs made up of repeating monomers, dimers and trimers. The higher rate of variability at the shorter MSTs can be explained by the greater reduction in DNA complexity (equal and more stochastic distribution of the 4 nucleotides), compared to the repeats with longer units [9-11].

**Challenges in MST research:** There are two reasons MSTs have been poorly studied compared to non-repeating coding regions. The first reason is that the large majority of MSTs occur in non-coding regions, which until very recently were considered “junk” DNA. Consequently, most of the programs for mapping reads from next generation sequencing were tuned to study the exome. Popular programs like Bowtie and BWA mask repeat regions in the reference genome and map sequencing reads only to the non-repetitive parts of the genome[12,13]. Thus, research projects that use such codes completely ignore this part of the genome and miss any relevant genetic or clinical contributions therein. The second reason is that a microsatellite-containing sequencing read needs to include both flanking regions of the repeated sequence to be correctly mapped. However, until recently the sequencing technology for reads of sufficient length was expensive and rarely used. Without a read of sufficient length, a repeat region can map to many different locations in the genome and hence a single read that contains nothing but MST will be mapped to multiple locations in the genome. Thus, the alleles or genotype of any microsatellite-containing locus with miss-mapped reads will not be called correctly. Hence, there is an urgent need to develop dedicated DNA enrichment techniques and the accompanying computational tools to further our understanding of the repeat regions of the genome.

**MST genotyping:** The accepted convention for MST genotyping from next generation sequencing reads is to calculate the number of reads that contain the entire MST along with at least 5 flanking bases on both sides. The allele with the highest number of reads is considered to be the primary allele. If an allele with a read depth larger than the half of the read depth of the primary allele is present, it is considered as the secondary allele and that MST locus is assumed to be heterozygous. If no such secondary allele is found, then the MST locus is assumed to be homozygous. After having established the genotype of an MST locus, the procedure searches any other alleles that are found with at least 3 reads, and if found the allele is assumed to be a minor allele[14]. While genotyping variants in the non-repetitive parts of the genome are not heavily dependent on read depth, genotyping MST variation depends both on length of the read and the depth of coverage for a given locus. The accuracy of MST genotyping is generally less than 94% for a sample coverage (read depth) of 40X[15]. Due to the higher possibility of allele variation and insufficient read depth, comparing MST alleles in population size studies are generally distribution oriented; beginning with defining a modal (most frequent) genotype in the control population and then looking for a significant difference in the ratio of the modal and non-modal genotypes in the disease population[14,16,17].

**Recent advances in MST research:** Figure 1-2 shows a few allele variations of a CAG trimer. Such variations in a CAG MST have been linked to Huntington's disease[18]. Figure 1-3 shows the expansion of one CAG motif to three, adding two extra glutamines to the HTT protein. Now that the association between MST allele variations and diseases is well established, efforts are being made to comprehensively examine MST regions of the genome. While typical genotyping methodologies have a less than 20% accuracy in identifying MSTs, newer computational techniques have been developed to increase the accuracy of the MST genotyping to 94% [15,19]. MST genotyping, since then, is becoming an important part of genomic variant detection[19]. Taking advantage of these newer MST genotyping methodologies, including our methodology described below, several studies have shown the connection between MSI and diseases. For example, a set of 55 MST loci was found to distinguish normal and breast cancer germline exomes with a sensitivity of 0.88[14]. Tumor grade specific signatures that can differentiate glioblastoma and lower-grade glioma were identified by comparing MST alleles from normal germline samples from the 1000 Genomes Project (1kGP) with two grades of gliomas and glioblastoma from The Cancer Genome Atlas[20,21]. Forty-eight glioblastoma specific MST loci and 42 lower-grade glioma specific MST loci were identified, and 29 MST loci were identified that can differentiate between these two cancer types[17]. Further, a set of 60 MST loci that can differentiate normal germline samples and ovarian cancer germline samples were identified at a high sensitivity of 0.90[16]. Recently, an extensive study was done to analyze 18 types of cancer using 5930 cancer exomes. Surprisingly, 14 of the 18 cancer types were identified as having MSIs[22].

**MSTs and mismatch repair:** The process of copying DNA, replication, often adds the incorrect nucleotide and thereby introducing an error into the genome. As MSTs are particularly vulnerable to replication slippage, understanding the following topics in detail is critical: 1. Occurrence of mismatch between DNA strands, 2. the cellular mechanisms that are in place to correct an erroneous nucleotide introduction and 3. the effects of an uncorrected mismatch. While erroneous extensions in a MST monomer leads to MST allele variation, extensions in a trimer MST can also form loop structures that have been shown to cause diseases[23]. When a mismatch is introduced by the polymerase, the proofreading mechanism that follows the polymerase enzyme attempts to correct the error. Two mismatch repair complexes exist to deal with mismatch[24]. Depending on the length of the mismatch, one of two mismatch repair (MMR) mechanisms is activated. Analogs of MutS protein, that is found in prokaryotic systems, are the two MMR complexes. Short mismatches of 1-

2 bases are handled by the MSH2/MSH6 complex while longer insertions and deletions are dealt by the MSH2/MSH3 complex[25,26].

Since MSTs are prone to slippage and breaks, incorrect MMR result in MSI. Recently, it has been shown that nucleotide excision repair and the cross-link repair complex also can affect the instability of MST trimers[27-29]. One of the projects described below explores these mechanisms in further detail.

**Fanconi Anemia:** Fanconi anemia (FA) is a congenital autosomal recessive disorder found predominantly in Jewish populations[30,31]. An autosomal recessive disorder is when the offspring inherits a mutated copy of the gene from both parents. FA cells have been found to be enriched with chromosomal aberrations that are caused by unrepaired DNA crosslinks[32-34]. Thirteen genes/proteins have been reported to form the core ingredients of the FA pathway involved in repair. They are also called complementation groups and they are identified as A, B, C, D1, D2, E, F, G, I. J. L, M and N[35-45]. DNA crosslinking agents such as mitomycin have been used to disrupt the cell division in cancer but the cell's response to inter-strand crosslinks has not been understood[46]. The FA pathway is known to gather crosslink repair proteins but it is also suspected to be involved in general upkeep for genomic stability[47]. Hence comprehensive understanding of the FA pathway and its full range of functionalities is important. Thus, exploiting this disease will give us important insights into mismatch repair. One of the projects below studies the role of microsatellite instability on FA.

## Overview of the 3 projects, hypothesis and approach

**1. Novel MST detection and the possibility of completing the human genome:** The human genome project was focused on the completion of the euchromatic portion of the genome and the first completed framework contained about 94% of the euchromatin[3,48]. Earlier attempts have been made to complete the last 1% of the incomplete reference genome. One such recent effort found 309 missing sequences in the human reference genome and also confirmed about 76% of these sequences to be commonly present in primate genomes[49]. Other researchers and we believe that the last 1% of the euchromatic DNA could be embedded in the repeat rich regions of the genome which are usually ignored during sequencing due to the technical reasons that were previously discussed[50]. Equipped with the necessary experience and tools to deal with challenges of studying the repeat regions of genome, our hypothesis is that the last 1% of the euchromatin (possibly including protein coding genes and/or functional elements) could be found hidden in the usually ignored MST regions of the genome. To accomplish this, we developed a novel target enrichment system to specifically sequence repeat regions genome-wide, and developed the necessary computational tools to extract possible novel MSTs and functional elements embedded within repeat rich regions.

A previous attempt to complete the genome by Liu *et al.* was done entirely on the exome data acquired from the 1000 genomes projects (1kGP)[49]. The exome data downloaded from the 1kGP are sequenced using baits (hybridizing sequence adapters that can pull down target sequences in a given sample) that are designed to capture the known exome regions and genomic regions that are found to flank these exomic sequences.  It should be noted that these baits do not capture the repeat regions and by that have no access to pull down and enrich any possible functional element/complex DNA that is embedded within the repeat rich regions of the genome. To overcome this hurdle, we generated, through computational means, a set of baits that will specifically hybridize to the repeat regions of the genome.

To overcome the inefficiency of mapping MST reads, we performed analysis only on the reads that were found to be unmapped to the reference genome. The computational pipeline specifically designed to extract possible novel MSTs and hidden functional elements utilizes the paired-end sequencing method and advances in sequence assembling algorithms to overcome the technical issues related to MST computation.

**MST based biomarker discovery in non-small cell lung cancer:** The relevance of the MST instability in diseases such as Huntington's disease is well established[18]. A normal HTT gene (gene that codes for the huntingtin protein) contains 35 or fewer CAG repeats while a gene with 36 or more CAG repeats has been linked to the onset of Huntington's disease. An array of other neurological diseases, such as Fragile X syndrome, have been established as trinucleotide repeat disorders[51]. The Garner lab has worked on MST instability and its influence on cancer for two decades. The lab has provided comprehensive computational insight into the positive influence of MST allele variability and cancer occurrences[14,16,17]. For example, McIver et al present a set of 55 MST markers that correlate positively with breast cancer and any patient with 76% or more of the 55 MST loci with the cancer genotype will be determined as cancer-like, or at enhanced risk of developing breast cancer. These research projects take full advantage of publically available genomic datasets such as The Cancer Genome Atlas (TCGA) and 1kGP. While utilizing publically downloaded datasets come with the challenge of low depth of coverage, appropriate statistical methods of using genotype distributions have been employed to counter this challenge[14]. These comprehensive computational projects have established statistical models that demonstrate the potential power of MST markers as risk, sub-typing and companion diagnostics.

To validate this research, it is necessary to study the consistency of these findings with high depth sequencing. As mentioned earlier, research conducted through public datasets are usually limited by their depth of coverage. The ultra-high-depth genotyping of the MST markers will validate MST genotyping, thus confirming the value of MST biomarker research. For the purpose of this project we focused on non-small cell lung cancer[52]. We developed a target enrichment system to enrich a set of 300 disease related MST loci along with 90 control MST loci and sequenced 30 lung cancer samples and 90 1kGP samples. Recent studies show highly convincing results connecting MST instability with a large (14 out of 18) varieties of cancers[22]. This large-scale study, again, emphasizes the need of further validation of the influence of MSI on cancer.

About 85% of the lung cancer cases are non-small cell lung cancer[52]. The well-established fact that never smokers are susceptible to lung cancer too, makes genetic biomarker screening a priority[53]. More importantly, recent studies show that 20% of people who get lung cancer are non-smokers[54]. This not only emphasizes the need for a comprehensive understanding of genetic biomarkers in lung cancer but in all other cancer types. For this project, we also try to identify genetic biomarkers for prescreening those at risk of lung cancer (including never smokers) through the development of the target specific MST enrichment kit.

**Exploring the molecular sources of MST instability:** MST instability is primarily caused by slippage during replication[27,28]. The proofreading mechanism that follows the DNA polymerase-driven replication initiates one of two mismatch repair pathways depending on the length of the error that has been introduced by the polymerase enzyme[24-26]. It has been largely known that MST instability is caused by mismatch repair mechanism dysfunction, but recent studies show that nucleotide excision repair and inter-strand crosslink repair also contributes to MST instability[28,29].

The Garner lab has studied MSTs extensively: 1. Understanding Mendelian inheritance in MST, 2. Exploring MST genotyping from 1kGP samples, 3. Developing computational tools to improve accuracy of MST genotyping, 4. Developing diagnostic models and statistical methods to understand the influence of MST instability on breast, ovarian, brain and lung cancers[14-17,19,55]. This project builds on this work by exploring the mechanisms that cause MST instability.

While nucleotide excision repair and crosslink repair mechanisms have been shown to affect MST instability, this project will investigate the mechanism by which crosslink repair results in MSI. To understand inter-strand crosslink caused MSI, the Fanconi anemia system has been chosen. PD20 cells are a cell line that have a dysfunctional Fanconi anemia (FA) pathway gene called FANCD2. PD20 cells which are retrovirally corrected for FANCD2 expression have also been obtained to study and compare the effects of a functional and dysfunctional FA pathway and their corresponding effects of MST instability.

While a list of 16 genes have been linked to the dysfunction of the FA pathway, very little is known about the downstream effects of a FA pathway affected by the failure to repair inter-strand crosslinks. DNA lesions causing chromosomal translocations have been identified to occur in FA patients[56]. This work aims to address the need to explore in detail the effect of crosslink repair and its effect on MST instability and the urgent demand to quantify and understand the types and amount of genomic damage that is caused by a dysfunctional FA pathway.

**Figure 1-1: Composition of the human genome.** About 47% of human genome is composed of a variety of repetitive DNA of which microsatellites have known to significantly change DNA complexity and thereby are mutational hot spots.

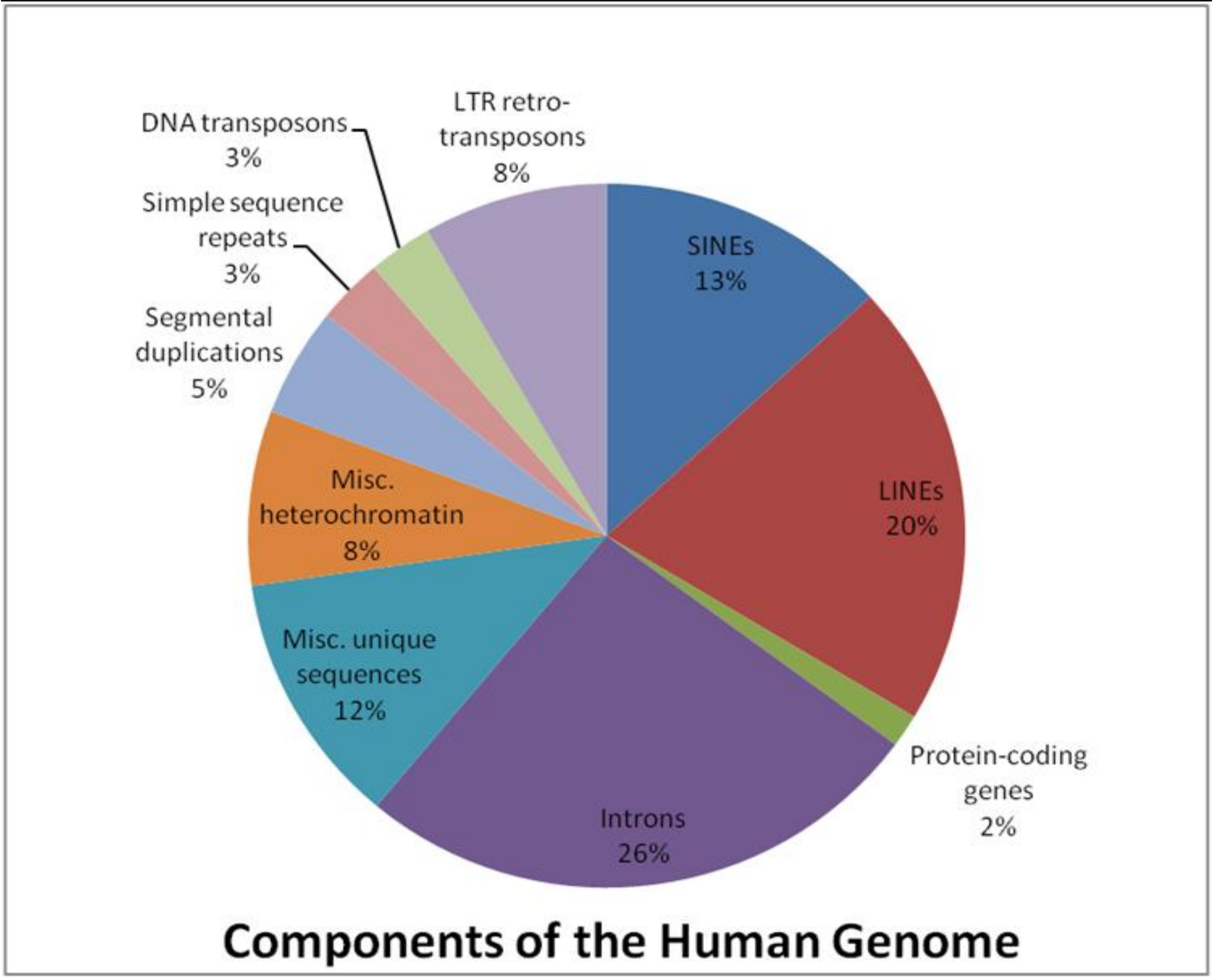
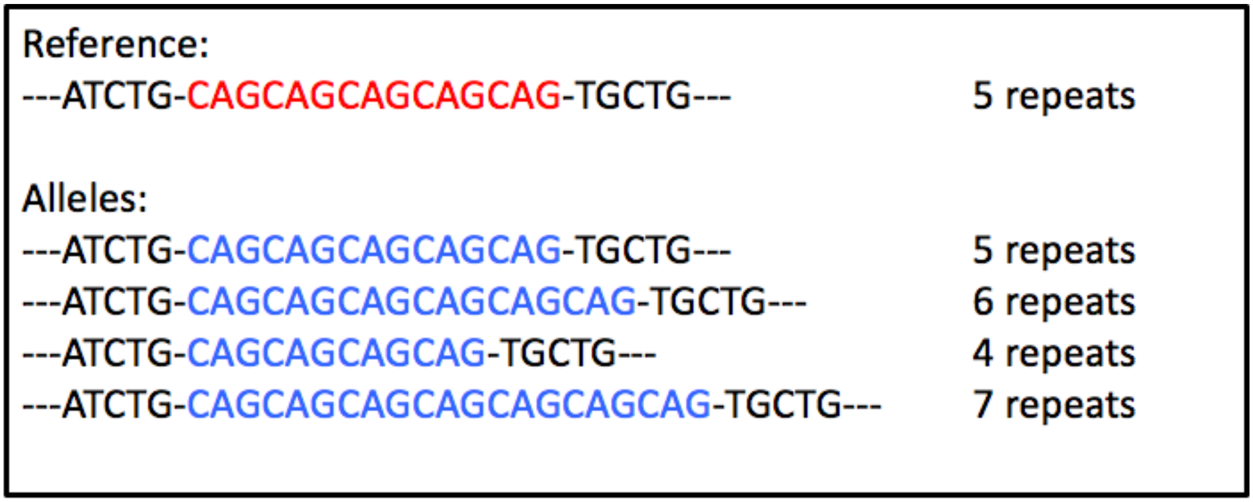


**Figure 1-2: MST allele variation.** This figure illustrates a few of the many possibilities of allele length variation of the MST shown as reference. It should be noted that without the flanking region, this CAG microsatellite can be placed in the genome in at least a thousand locations, making the identification of a single MST impossible. By considering the flanking regions, the genotyping programs identify a MST from the unmapped reads that are correctly mapped to the appropriate locus.



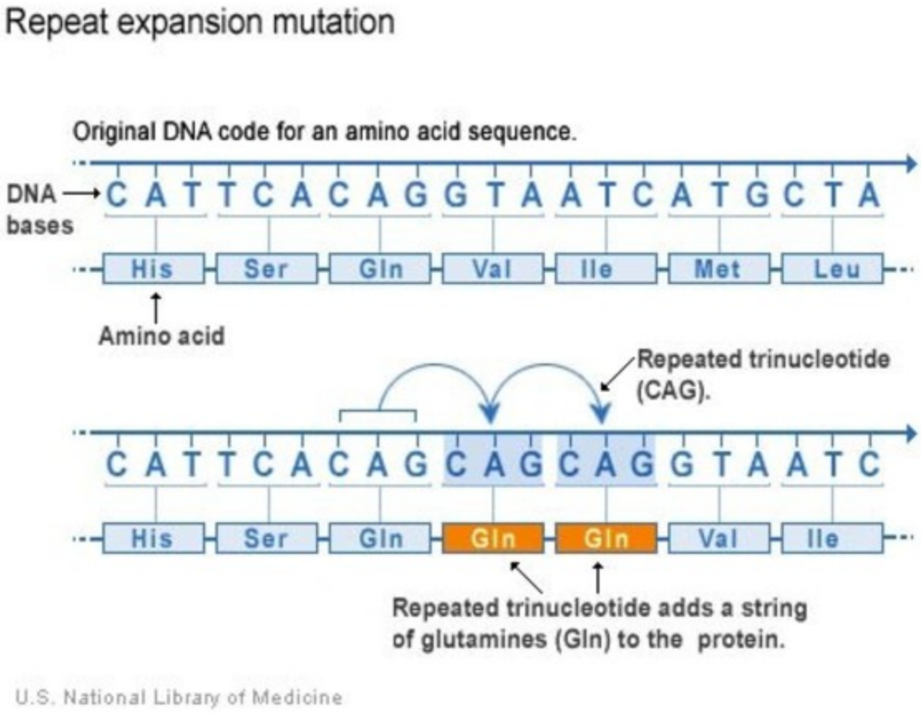
Reference:

| | |
|---|---|
| ---ATCTG-CAGCAGCAGCAGCAG-TGCTG--- | 5 repeats |

Alleles:

| | |
|---|---|
| ---ATCTG-CAGCAGCAGCAGCAG-TGCTG--- | 5 repeats |
| ---ATCTG-CAGCAGCAGCAGCAGCAG-TGCTG--- | 6 repeats |
| ---ATCTG-CAGCAGCAGCAG-TGCTG--- | 4 repeats |
| ---ATCTG-CAGCAGCAGCAGCAGCAGCAG-TGCTG--- | 7 repeats |

**Figure 1-3: CAG repeat expansion causes Huntington’s disease.** Huntington’s disease is one of the diseases that is caused by a CAG repeat expansion. Huntington’s disease is not caused when the MST contains 35 or fewer repeats while 36 or more has been shown to be strongly associated with the disease condition.


Repeat expansion mutation
Original DNA code for an amino acid sequence.
DNA bases
C A T T C A C A G G T A A T C A T G C T A
His Ser Gln Val Ile Met Leu
Amino acid
Repeated trinucleotide (CAG).
C A T T C A C A G C A G C A G G T A A T C
His Ser Gln Gln Gln Val Ile
Repeated trinucleotide adds a string of glutamines (Gln) to the protein.
U.S. National Library of Medicine

REFERENCES:

1 *The Human Genome Project*, <https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/> (

2 NCBI. *Genome list*, <https://www.ncbi.nlm.nih.gov/genome/browse/> (

3 Institute, T. N. H. G. R. *The Human Genome Project Completion*, <https://www.genome.gov/11006943/human-genome-project-completion-frequently-asked-questions/> (

4 RepeatMasker. *RepeatMasker Genomic Datasets*, <http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html> (

5 Ananda, G. *et al.* Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome. *Genome Biol Evol* **5**, 606-620, doi:10.1093/gbe/evs116 (2013).

6 Baptiste, B. A. *et al.* Mature microsatellites: mechanisms underlying dinucleotide microsatellite mutational biases in human cells. *G3 (Bethesda)* **3**, 451-463, doi:10.1534/g3.112.005173 (2013).

7 Barros, P., Boan, F., Blanco, M. G. & Gomez-Marquez, J. Effect of monovalent cations and G-quadruplex structures on the outcome of intramolecular homologous recombination. *FEBS J* **276**, 2983-2993, doi:10.1111/j.1742-4658.2009.07013.x (2009).

8 Gemayel, R., Vinces, M. D., Legendre, M. & Verstrepen, K. J. Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu Rev Genet* **44**, 445-477, doi:10.1146/annurev-genet-072610-155046 (2010).

9 Abdulovic, A. L., Hile, S. E., Kunkel, T. A. & Eckert, K. A. The in vitro fidelity of yeast DNA polymerase delta and polymerase epsilon holoenzymes during dinucleotide microsatellite DNA synthesis. *DNA Repair (Amst)* **10**, 497-505, doi:10.1016/j.dnarep.2011.02.003 (2011).

10 Bagshaw, A. T., Pitt, J. P. & Gemmell, N. J. High frequency of microsatellites in S. cerevisiae meiotic recombination hotspots. *BMC Genomics* **9**, 49, doi:10.1186/1471-2164-9-49 (2008).

11 Hile, S. E., Yan, G. & Eckert, K. A. Somatic mutation rates and specificities at TC/AG and GT/CA microsatellite sequences in nontumorigenic human lymphoblastoid cells. *Cancer research* **60**, 1698-1703 (2000).

12 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9**, 357-359, doi:10.1038/nmeth.1923 (2012).

13 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754-1760, doi:10.1093/bioinformatics/btp324 (2009).

14 McIver, L. J., Fonville, N. C., Karunasena, E. & Garner, H. R. Microsatellite genotyping reveals a signature in breast cancer exomes. *Breast Cancer Res Treat* **145**, 791-798, doi:10.1007/s10549-014-2908-8 (2014).

15 McIver, L. J., Fondon, J. W., 3rd, Skinner, M. A. & Garner, H. R. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**, 193-199, doi:10.1016/j.ygeno.2011.01.001 (2011).

16 Fonville, N. C., Vaksman, Z., McIver, L. J. & Garner, H. R. Population analysis of microsatellite genotypes reveals a signature associated with ovarian cancer. *Oncotarget* **6**, 11407-11420, doi:10.18632/oncotarget.2933 (2015).

17 Karunasena, E. *et al.* Somatic intronic microsatellite loci differentiate glioblastoma from lower-grade gliomas. *Oncotarget* **5**, 6003-6014, doi:10.18632/oncotarget.2076 (2014).

18 Science, U. S. N. L. o. *Huntington Disease*, <https://ghr.nlm.nih.gov/condition/huntington-disease> (

19 McIver, L. J., McCormick, J. F., Martin, A., Fondon, J. W., 3rd & Garner, H. R. Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene* **516**, 328-334, doi:10.1016/j.gene.2012.12.068 (2013).

20 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74, doi:10.1038/nature15393 (2015).

21 Institute, T. N. C. The Cancer Genome Atlas.

22 Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* **22**, 1342-1350, doi:10.1038/nm.4191 (2016).

23 Hite, J. M., Eckert, K. A. & Cheng, K. C. 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 (1996).

24 McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* **11**, 786-799, doi:10.1038/nrg2828 (2010).

25 Huang, J. *et al.* MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. *Cancer research* **61**, 1619-1623 (2001).

26 Iaccarino, I., Marra, G., Palombo, F. & Jiricny, J. hMSH2 and hMSH6 play distinct roles in mismatch binding and contribute differently to the ATPase activity of hMutSalpha. *EMBO J* **17**, 2677-2686, doi:10.1093/emboj/17.9.2677 (1998).

27 Hubert, L., Jr., Lin, Y., Dion, V. & Wilson, J. H. Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. *Hum Mol Genet* **20**, 4822-4830, doi:10.1093/hmg/ddr421 (2011).

28 Lin, Y., Hubert, L., Jr. & Wilson, J. H. Transcription destabilizes triplet repeats. *Mol Carcinog* **48**, 350-361, doi:10.1002/mc.20488 (2009).

29 Concannon, C. & Lahue, R. S. Nucleotide excision repair and the 26S proteasome function together to promote trinucleotide repeat expansions. *DNA Repair (Amst)* **13**, 42-49, doi:10.1016/j.dnarep.2013.11.004 (2014).

30 Kutler, D. I. *et al.* A 20-year perspective on the International Fanconi Anemia Registry (IFAR). *Blood* **101**, 1249-1256, doi:10.1182/blood-2002-07-2170 (2003).

31 Moldovan, G. L. & D'Andrea, A. D. How the fanconi anemia pathway guards the genome. *Annu Rev Genet* **43**, 223-249, doi:10.1146/annurev-genet-102108-134222 (2009).

32 Donahue, S. L. & Campbell, C. A Rad50-dependent pathway of DNA repair is deficient in Fanconi anemia fibroblasts. *Nucleic Acids Res* **32**, 3248-3257, doi:10.1093/nar/gkh649 (2004).

33 Mace-Aime, G., Couve, S., Khassenov, B., Rosselli, F. & Saparbaev, M. K. The Fanconi anemia pathway promotes DNA glycosylase-dependent excision of interstrand DNA crosslinks. *Environ Mol Mutagen* **51**, 508-519, doi:10.1002/em.20548 (2010).

34 Meyer, S., Neitzel, H. & Tonnies, H. Chromosomal aberrations associated with clonal evolution and leukemic transformation in fanconi anemia: clinical and biological implications. *Anemia* **2012**, 349837, doi:10.1155/2012/349837 (2012).

35 de Winter, J. P. *et al.* Isolation of a cDNA representing the Fanconi anemia complementation group E gene. *Am J Hum Genet* **67**, 1306-1308, doi:10.1016/S0002-9297(07)62959-0 (2000).

36 de Winter, J. P. *et al.* The Fanconi anaemia group G gene FANCG is identical with XRCC9. *Nat Genet* **20**, 281-283, doi:10.1038/3093 (1998).

37 Dorsman, J. C. *et al.* Identification of the Fanconi anemia complementation group I gene, FANCI. *Cell Oncol* **29**, 211-218 (2007).

38 Howlett, N. G. *et al.* Biallelic inactivation of BRCA2 in Fanconi anemia. *Science* **297**, 606-609, doi:10.1126/science.1073834 (2002).

39 Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. *Nat Genet* **37**, 934-935, doi:10.1038/ng1625 (2005).

40 Meetei, A. R. *et al.* X-linked inheritance of Fanconi anemia complementation group B. *Nat Genet* **36**, 1219-1224, doi:10.1038/ng1458 (2004).

41 Meetei, A. R. *et al.* A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat Genet* **37**, 958-963, doi:10.1038/ng1626 (2005).

42 Reid, S. *et al.* Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat Genet* **39**, 162-164, doi:10.1038/ng1947 (2007).

43 Smogorzewska, A. *et al.* Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. *Cell* **129**, 289-301, doi:10.1016/j.cell.2007.03.009 (2007).

44 Strathdee, C. A., Gavish, H., Shannon, W. R. & Buchwald, M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. *Nature* **356**, 763-767, doi:10.1038/356763a0 (1992).

45 Timmers, C. *et al.* Positional cloning of a novel Fanconi anemia gene, FANCD2. *Mol Cell* **7**, 241-248 (2001).

46 Deans, A. J. & West, S. C. DNA interstrand crosslink repair and cancer. *Nat Rev Cancer* **11**, 467-480, doi:10.1038/nrc3088 (2011).

47 Michl, J., Zimmer, J. & Tarsounas, M. Interplay between Fanconi anemia and homologous recombination pathways in genome integrity. *EMBO J* **35**, 909-923, doi:10.15252/embj.201693860 (2016).

48 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

49 Liu, Y. *et al.* Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. *BMC Genomics* **15**, 685, doi:10.1186/1471-2164-15-685 (2014).

50 Miga, K. H., Eisenhart, C. & Kent, W. J. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* **43**, e133, doi:10.1093/nar/gkv671 (2015).

51 Budworth, H. & McMurray, C. T. A brief history of triplet repeat diseases. *Methods Mol Biol* **1010**, 3-17, doi:10.1007/978-1-62703-411-1_1 (2013).

52 Society, A. C. *What is non-small cell lung cancer?*, <http://www.cancer.org/cancer/lungcancer-non-smallcell/detailedguide/non-small-cell-lung-cancer-what-is-non-small-cell-lung-cancer> (2016).

53 Li, Y. *et al.* Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *The Lancet. Oncology* **11**, 321-330, doi:10.1016/S1470-2045(10)70042-5 (2010).

54 Rivera, G. A. & Wakelee, H. Lung Cancer in Never Smokers. *Advances in experimental medicine and biology* **893**, 43-57, doi:10.1007/978-3-319-24223-1_3 (2016).

55 Karunasena, E. *et al.* 'Cut from the same cloth': Shared microsatellite variants among cancers link to ectodermal tissues-neural tube and crest cells. *Oncotarget* **6**, 22038-22047, doi:10.18632/oncotarget.4194 (2015).

56 Medicine, U. S. N. L. o. *Fanconi Anemia*, <https://ghr.nlm.nih.gov/condition/fanconi-anemia - genes> (

# Chapter 2: Genomic leftovers: identifying novel microsatellites, over represented motifs and functional elements in the human genome

Fonville, N. C. et al. Genomic leftovers: identifying novel microsatellites, overrepresented motifs and functional elements in the human genome. Sci. Rep. 6, 27722; doi: 10.1038/srep27722 (2016).

## ABSTRACT

The human genome is 99% complete. This study contributes to filling the 1% gap by enriching previously unknown repeat regions called microsatellites (MST). We devised a Global MST Enrichment (GME) kit to enrich and nextgen sequence 2 colorectal cell lines and 16 normal human samples to illustrate its utility in identifying contigs from reads that do not map to the genome reference. The analysis of these samples yielded 790 novel extra-referential concordant contigs that are observed in more than one sample. We searched for evidence of functional elements in the concordant contigs in two ways: (1) BLAST-ing each contig against normal RNA-Seq samples, (2) Checking for predicted functional elements using GlimmerHMM. Of the 790 concordant contigs, 37 had an exact match to at least one RNA-Seq read; 15 aligned to more than 100 RNA-Seq reads. Of the 249 concordant contigs predicted by GlimmerHMM to have functional elements, 6 had at least one exact RNA-Seq match. BLAST-ing these novel contigs against all publically available sequences confirmed that they were found in human and chimpanzee BAC and FOSMID clones sequenced as part of the original human genome project. These extra-referential contigs predominantly contained pentameric repeats, especially two motifs: AATGG and GTGGA.

# INTRODUCTION

In April of 2003 the Human Genome Project was declared complete, and from it we gained a framework to build the reference genome upon which the majority of analyses are anchored. The scope of the Human Genome Project was focused on the 94% of the genome that is euchromatin[1], now sequenced to 99% completion[2]. Attempts are being made to complete the 1% of the incomplete "complete" human reference[3], however we and others have hypothesized that some genomic sequence regions (which may contain functional elements, genes) may be missing from the human reference because they are embedded in refractory repetitive DNA sequence, e.g. microsatellites (MSTs)[4]. MST sequences, regions of repeated 1- to 6-mer DNA motifs, are abundant throughout the genome and are a source of significant genomic variation[5]. However, to date, analysis of microsatellite-containing loci has been limited because standard exome enrichment and whole genome sequencing uses software to mask out repeats[6], focuses on capturing non-repetitive DNA, or is designed to capture only a small subset of the known MST loci[7]. In this paper we present a novel target enrichment strategy specifically designed to enrich for all microsatellite loci based on the repeat motif, rather than the flanking sequence, as baits, and have paired this technique with our recently developed method for analysis of unmapped reads[8]. Our analysis has revealed: 1) assembly of contigs from unmapped genome sequences and high-depth sequences from this novel target enrichment system that specifically selects for repetitive elements enables the quantification and characterization of these regions; 2) concordant contigs, those that appear in multiple samples, contain new structural elements (potential genes/pseudogenes, etc.), a subset of which have high similarity to expressed mRNAs; 3) these extra-referential genome regions are dominated by 5-mer repeats, in particular, an AATGG and a GTGGA centromeric repeat. This platform technology has the potential to extend "reference genomes" and identify new functional elements.

## METHODS

Standard exome enrichment sequencing is designed using a bait set that contains the sequence of the known high complexity exomic regions. However, portions of the human genome remain unknown and as such are not captured and evaluated by current enrichment technologies. In addition, whole genome sequencing, which can be used to sequence these additional unknown regions is limited in its ability to evaluate these regions because sequencing reads are aligned to the known reference genome, and they lack sufficient sequencing depth for reliable assembly. Although these methods (WGS and exome enrichment) are excellent for evaluating a large portion of the genome, they are not optimal for identifying and aligning novel genomic sequence (i.e. gap filling, finishing genomes containing highly repetitive regions). Similarly, only reads in RNA-Seq data that are aligned to known reference genes are quantified, thus, an incomplete reference genome also impacts expression studies.  One potential reason that sections of the human genome remain unknown, or are not included in the reference, is that they contain highly repetitive DNA that makes it difficult to sequence and align properly. We have created a reference-independent enrichment method that is designed to specifically enrich for repetitive DNA. This global microsatellite enrichment (GME) assay uses a bait design in which each 120 nt bait is composed of 4 x 30nt segments, selected to minimize the potential for intra-bait hairpin formation. Every possible 1-6 nt repetitive motif is represented within the bait set.

**Design of global microsatellite enrichment (GME) bait set:** We designed a custom bait set that target all 1- to 6-mer microsatellite motifs. Each120nt bait is broken into four 30nt regions, each of which targets a different motif sequence. We programmed and ran a custom PERL script to design the baits to maintain approximately a 40% G/C content along the full length of the bait (across all four motifs on each bait).  The custom script also evaluated the potential for hairpin formation of the baits and selected motifs for each bait to have a lower probability of internal hairpin formation. The baits were uploaded into Agilent's eArray.

**Enrichment and Sequencing:** Agilent exome enrichment, and our custom enrichment were performed according to the manufacturer's directions. For the combined GME + Exome enrichment, the bait sets were combined in house. All enrichments were sequenced using the 150bp Illumina HiSeq Rapid-Run.

**Cell culture:** DLD1 (ATCC® Number: CCL-221™) and SW403 (ATCC® Number: CCL-230™) human cell lines were purchased from ATCC (http://www.atcc.org Date of access: 01/06/2014). Cells were grown to confluence in DMEM + 10% FBS at 37C with 5% CO2 (DLD1) or Lebovitz media at 37C with no CO2 (SW403). Genomic DNA was isolated using Qiagen DNA Blood and Tissue kit according to the manufacturer's protocol. DNA for the 16 normal samples was purchased from Coriell.

**Novel microsatellites prediction from unmapped reads:**  Sequenced reads of all samples were obtained in the form of fastq files. The sequenced reads were paired-end with a sequence length of 150 bases. The reads were quality checked and trimmed using Trimmomatic. The program uses sliding windows to check for the quality and a window of 10 bases was used. A quality score threshold of 20 was used; a 20 quality score means an error probability of 1 in 100 bases. The length threshold was set to 70 bases so any sequence read that is shorter than 70 bases after the quality trimming will be filtered out.

The BWA aligner was used to align the paired-end reads to the human reference genome, Hg19. A custom written python program was used to extract the unmapped reads from the sam files and output them in fasta format. The SAMTOOLS view and index programs were used to sort the sam files and convert sam file to bam format. The ‘add-read-groups’ program in the Picard software suite was used to add read groups to the bam files. The GATK program was used for indel realignment. The bam file that was indel realigned was then used to collect the unmapped reads in fasta format using a custom written PERL script for further processing.

The fasta file with unmapped reads was passed on to a shell program to check and remove recurring “N”s in the unmapped reads. Any read shorter than 50 bases after the removal of “N”s was discarded. The “N” filtered unmapped reads were used to form contigs using the Velvet program. The kmer length (hash length) used for the velveth program was 71. The choice of an odd number for hash length is a requirement of the program so as to avoid palindromes. The resultant contig file is then passed on to the Tandem Repeat Finder (TRF) program for microsatellites identification. Note that there are a total of 1811360 known microsatellite loci in the Hg19 reference genome as identified by TRF.  The match weight, mismatch penalty, indel penalty, match probability, indel probability, minimum alignment score and period size used are 2, 7, 5, 80, 10, 14 and 6, respectively. The ‘-h’ parameter was used to suppress HTML output. A custom written PERL program was used to extract the predicted microsatellite list from binary output provided by TRF. In order to check if there are known microsatellites in the list generated by TRF, the TRF identified microsatellites are BLASTed against known genomes. For this purpose, a PERL program was written to add flanking regions (30 bases on each side) to the microsatellite list by referring back to the contigs file generated by Velvet. The microsatellites with flanking sequences are then BLASTed using the blastn program against the blast formatted nucleotide sequence database and the human genome database downloaded from https://ncisf.org/software-databases/blast-databases (Date of access: 08/09/2014). An e-value of 0.001 was used for the blastn program. A PERL program was written to separate the microsatellites that were either found in the human or the nucleotide sequence database hence leaving only novel microsatellites.

**Read number calculation:** The read depth is calculated from the contig depth using the formula provided by Velvet. For a given contig the information about the contig is provided in the contig ID by the Velvet program (e.g. NODE_35_length_226_cov_17.079645). Using this information the contig coverage can be converted into read depth: C=Ck*L/(L-k+1). C is the read depth, Ck is the contig coverage, L is the average read length, and K is the read kmer length used.

**Identifying novel contigs/MSTs:** Any assembled contig that has 100% identity match with a human or a NT database sequence for more than half the length of the query sequence, then the contig is considered known. The remaining contigs are considered novel. All TRF predicted MSTs in the novel contigs are considered to be novel MSTs.

**Kmer calculation:** The kmers are divided into six categories: 1-mer, 2-mer, 3-mer, 4-mer, 5-mer and 6-mer. The main input file to calculate kmer frequency information is the final output from the previous section that contains a predicted novel microsatellite list. The kmers were grouped into families. A kmer is added to a family if a) they are cyclically same or b) they are cyclical reverse complements. For example, if AACT is a family name, then TAAC, CTAA and ACTA belong to the same family. According to reverse complementarity, AGTT, TAGT, TTAG and GTTA also belong to the AACT family. So an N-mer family has N*2 possible family members. All the read depths in a family are summed and hence each family has one kmer read depth value. To aid in the comparison of kmer families across samples, the name of the kmer families are kept similar. After grouping the cyclical kmers and cyclical reverse complements kmers, measures were taken to unify the family names across samples. The entire kmer calculation procedure was done using a series of custom python scripts.

**Concordant contig calculation:** The novel contigs from all the 16 + 6 samples were pooled together to find contigs that are observed in two or more samples, i.e. concordant. The pooled contigs were converted into a fasta file and was formatted into a BLAST database. All the individual contigs were then BLASTed against the database. Alignments that are more than 70% of the query length and aligned with 0 or 1 mismatch (to allow for potential individual variation) were considered for further analysis. A python script was written to generate two lists (with no mismatch and 1 mismatch) of concordant contigs that were found in more than one sample. The contigs in each concordant group were assembled into one contig sequence using the CAP3 program.

**Evaluation of contigs for gene-like structure:** All concordant contigs were fed into the GlimmerHMM program for Gene-Like Structure (GLS) prediction. The program was trained using the human training data provided with the GlimmerHMM package. The program can predict one or more exons in a given contig. The predicted exons in a GLS can be of three types; initial, internal and final. All the GLS with more than one exon were considered for further processing.

**Comparison of putative cDNA to RNA-seq data:** The start and end positions of the exons were obtained from the GlimmerHMM output to extract putative cDNA sequences. The cDNA sequences of all concordant contigs from all the samples were combined to make a single cDNA database. Ten lymphoblastoid RNA-Seq samples were obtained in the form of FASTQ files (http://www.ebi.ac.uk/arrayexpress/experiments/E-GEUV-1/samples/ Date of access: 15/01/2015) and were BLAST searched against the cDNA database. A cDNA query sequence that aligned to an RNA-Seq read with an identity of 100% and with an alignment length longer than 70% of the query length was considered to be a highly similar match. The RNA-Seq reads evaluated in this study were on average 75 bases long. The results of the both the BLAST searches were combined and are provided in Extended Data Table 7. Analysis of the BLAST search output was done using custom written python scripts.

Similar to the paragraph above, the concordant contigs were searched against RNA-Seq sample reads. The results of this BLAST search are provided in Extended Data Table 5.

**Whole genome analysis:** The k-mer analysis graph presented in Figure 1 contains the k-mer distribution of all the 22 samples along with the human reference genome and a whole genome sample. The whole genome sample was added to compare the k-mer levels of the six samples with a sample where the entire genome is sequenced without enrichment bias and not just the exome.

A B-Lymphocyte whole genome DNA sequencing sample (HG00106) was downloaded from the 1000 genomes project for this purpose. The paired-end sequencing reads were downloaded in the form of a FASTQ file. The reads were N filtered and mapped to the human reference genome (Hg19) using BWA. The unmapped reads were separated from the SAM file. The unmapped reads, in the form of a FASTA file, was used as the input for the TRF program to predict MST loci. The same parameters used for the other samples for TRF was used here too. A custom written PERL program was used to analyze a list of known MSTs and the TRF output data file to generate a list of predicted MST loci along with their read depths. This list of MST loci was used in the preparation of Figure 1.

## RESULTS AND DISCUSSION

To verify the performance of our GME, we selected two colorectal cell lines: MST stable SW403 cells, and MST unstable DLD-1 cells[9,10] (Extended Data Table 1). For each cell line, DNA isolated from cells grown as a single large cell culture was split and enriched using: (1) the Agilent exome enrichment kit; (2) our GME enrichment kit; or (3) an admixture of exome and our enrichment where the baits were combined prior to the enrichment. This enabled us to quantify the relative enrichment of repetitive regions for each enrichment system. Once enrichment performance was optimized and verified, we then sequenced 16 additional normal human DNA samples (Extended Data Table 1) from the 1000 Genomes Project using this enrichment system, and then processed the data to identify novel contigs that reproducibly appear in multiple samples. These concordant contigs were then analyzed for potential functional elements, as predicted by GlimmerHHM and supported by high similarity to RNA-Seq reads found in a variety of tissues. All enrichments, on average, had 99% of the high-quality reads map to the known human reference Hg19 (Table 1), consistent with what is expected given that 99% of the genome is "complete". A substantial fraction of the reads from our GME enrichment, as opposed to those reads from exome enrichment, fell outside the exome (Table 1, Exome Overlap %), consistent with our goal of global enrichment for genomic microsatellite loci. On average, 0.45% of the total reads were found to be unmapped in the GME and the combined samples while only 0.1% of the exome enrichment samples were unmapped (Table 1). This is, again, consistent with the GME system specifically targeting repetitive regions. The GME of the DLD-1 cell line and the SW403 cell line were both enriched for microsatellite loci, but in different manners. DLD-1-GME captured a greater fraction of the known MSTs (identified from the Hg19 reference genome using Tandem Repeat Finder (TRF)[11]), whereas SW403-

GME captured fewer of the known MSTs (Table 1) but had greater depth for a greater fraction of the MSTs that were captured (Extended Data Fig. 1). That this was due to variation in hybridization temperature is supported by an increase in C/G nucleotide sequences captured in this sample (Extended Data Fig. 2). While the MST loci amounts in the colorectal samples varied, the additional normal samples processed at the optimum temperature produced consistent results. On average, 87% of the mapped reads in the normal samples contained MST loci. (Table 1)

Any reads containing novel DNA/MST sequences are by definition unmapped relative to the reference genome, therefore we analyzed the content of the unmapped reads using our unmapped read analysis pipeline ([8] and methods). We built contigs from those reads that did not align to the reference (Table 2), and then re-aligned these contigs to the reference to further eliminate any contigs with known sequence from further analysis. Not only were a higher number of contigs built from the unmapped read analysis on the GME samples, but there were also a higher number of contigs containing MSTs (Table 2), as expected. We identified between 162 and 1469 novel contigs per enrichment, of which between 8% and 56% contained MSTs (Table 2). These MST-containing loci were found to be covered at a significant read depth. On average, 65% of MSTs found in novel contigs of the normal samples and the colorectal cell lines were supported by more than 10 reads (Extended Data Fig. 1 and 3). The analysis of all 20 independent GME samples (2 colorectal GME only, 2 colorectal GME combined with exome, and 16 normal) has yielded 790 concordant contigs (283 bp average length) that were observed in more than one sample (on average, each contig is seen in 5.6 samples, Extended Data Table 2). The concordant contigs were observed in as many as all 20 samples (Extended Data Table 3). This high reliability data (as confirmed in multiple samples) can reveal robust new MST-containing loci and high complexity sequence in unmapped regions. The distribution of the lengths of all the concordant contigs and those that contain a MST (Extended Data Fig. 4) shows that more than 90% of concordant contigs are shorter than 500 nt. As nextgen raw read lengths grow, longer contigs in these repeat-containing regions could be assembled.

Extended Data Fig. 5 shows that the GME approach can capture MST-containing loci of varying length; as short as 10nt, but also longer than 150nt (i.e. as observed in assembled contigs). Given that 14% of the known MSTs identified in Hg19 using TRF are >50nt, our approach of enrichment plus contig building will, as raw read lengths grow, give access to longer MST loci.

To assess the potential value of these novel contigs, we searched the 790 concordant contigs for functional elements by aligning the contigs to 10 normal Lymphoblastoid RNA-Seq samples (Extended Data Table 4). We found that 37 concordant contigs aligned to at least one RNA-Seq read (Extended Data Table 5); and 22 of these 37 were supported by at least 10 RNA-Seq reads. BLASTing these 37 concordant contigs against the all known sequences in GenBank confirmed that the vast majority of these novel sequences (i.e. less than 50% identity to any portion of the known human reference) appear to be in human and chimpanzee subclones sequenced as part of the earliest human genome sequencing efforts, but not mapped to the current human genome reference, again confirming they are part of the missing human genome reference. This raises the possibility that sequences that had been relegated to the unaligned or pre-nextgen clone sequence trash heap may eventually be captured at a depth and with a reliability that allow them to be integrated into the reference. Table 3 illustrates the top 10 concordant contigs aligned to at least 235 RNA-Seq reads. Interestingly, the contig with the most hits to RNA-Seq reads was confirmed by BLASTing against GenBank to have 99% identity to abundant

Ribosomal RNA sequences which may indicate a coding region for a new rRNA family member.

We also examined the 790 concordant contigs for the presence of gene-like structures (GLS) (e.g. putative exon, intron followed by another in-frame exon) using GlimmerHMM[12] (Table 2). We found that 249 concordant contigs contained potential GLSs (Table 2). The majority of the contigs with GLS had 2-3 exons with no contig identified as having over 8 exons (Extended Data Table 6). Extended Data Fig. 4 shows that there was no significant difference in average contig length for contigs containing a MST and/or GLS. We then generated putative cDNAs from the concordant contigs identified as having GLS and compared them to 10 RNA-Seq data sets discussed above. Six of these putative cDNAs matched to the RNA-Seq data with between 1 and 90 RNA-Seq reads aligned to each GLS (Extended Data Table 7), demonstrating that these novel contigs not only have potential GLS, but some were identified as low abundance mRNAs. It further indicates that many new potential coding regions that robustly align to RNA-Seq reads contain functional elements not recognized by GlimmerHHM.

Analysis of the microsatellite motifs represented in the unmapped reads containing novel MSTs revealed an over-abundance of reads containing pentameric repeats (Fig. 1). A comparison of the relative abundance of the MST-containing contigs to the expected “known” motif ratios present from the human reference (Hg19) identified with Tandem Repeat Finder (TRF) and those present in a whole genome sequenced sample from the 1000 genomes project shows that the abundance of pentameric MST loci repeats are >3.5 fold more abundant in the whole genome sample than expected from the known reference (Fig. 1). Pentameric MSTs were present in the exome enrichment at a similar abundance as in the whole genome sequenced sample, however, in our GME the fraction of novel pentameric MSTs that were captured was an average of  9.2 fold greater than the fraction found in the known reference and 2.6 fold greater than the whole genome sequenced sample (a list of the top 5 most abundant motifs is given in Extended Data Table 8), consistent with these pentameric MSTs being localized extra-referentially.

Examination of the specific sequences of the MST containing loci revealed that the motifs AATGG and GTGGA were abundant in all samples, including the analysis of the whole genome sample, but not in the known human reference data (Extended Data Table 8). AATGG has been identified as a human centromeric[13,14] sequence motif and therefore the overabundance of this sequence in the data, but its absence from the human reference assembly, is consistent with it being a component of centromeric heterchromatin. The motif GTGGA differs from AATGG by one nucleotide, and based on its similarity to a known centromeric repeat and overabundance in the data also makes it a candidate centromeric repeat as well.  The overabundance of reads that contain the AATGG sequence was unexpected, and an average of 53% of unmapped reads with MSTs contained the AATGG motif, whereas, reads were devoid of the often-studied telomeric repeat, GGGTTA.

# CONCLUSION

Our GME provides the possibility for discovery of new/non-reference DNA sequence through the non-specific (i.e. not linked to a single reference locus) capture of repetitive DNA. Applying this target enrichment strategy broadly and pairing it with our algorithmic approach for identifying novel concordant contigs from unmapped reads can drive the human genome towards true completion; identify and annotate new potential functional elements therein; finish genomes containing even more repetitive sequence, such as plants; and be key to quantifying and comprehending the importance of telomeric and centromeric structure.

# FIGURES

**Figure 2-1: Frequency of MSTs motif classes in unmapped reads relative to those found in the reference genome (HG19) and the whole genome.** The majority of novel microsatellite loci captured using our method contained pentameric repeats. For comparison, most loci found in the reference genome are not pentameric; and analysis of whole genome data confirms that much of the missing genome is associated with pentamer repeat regions.

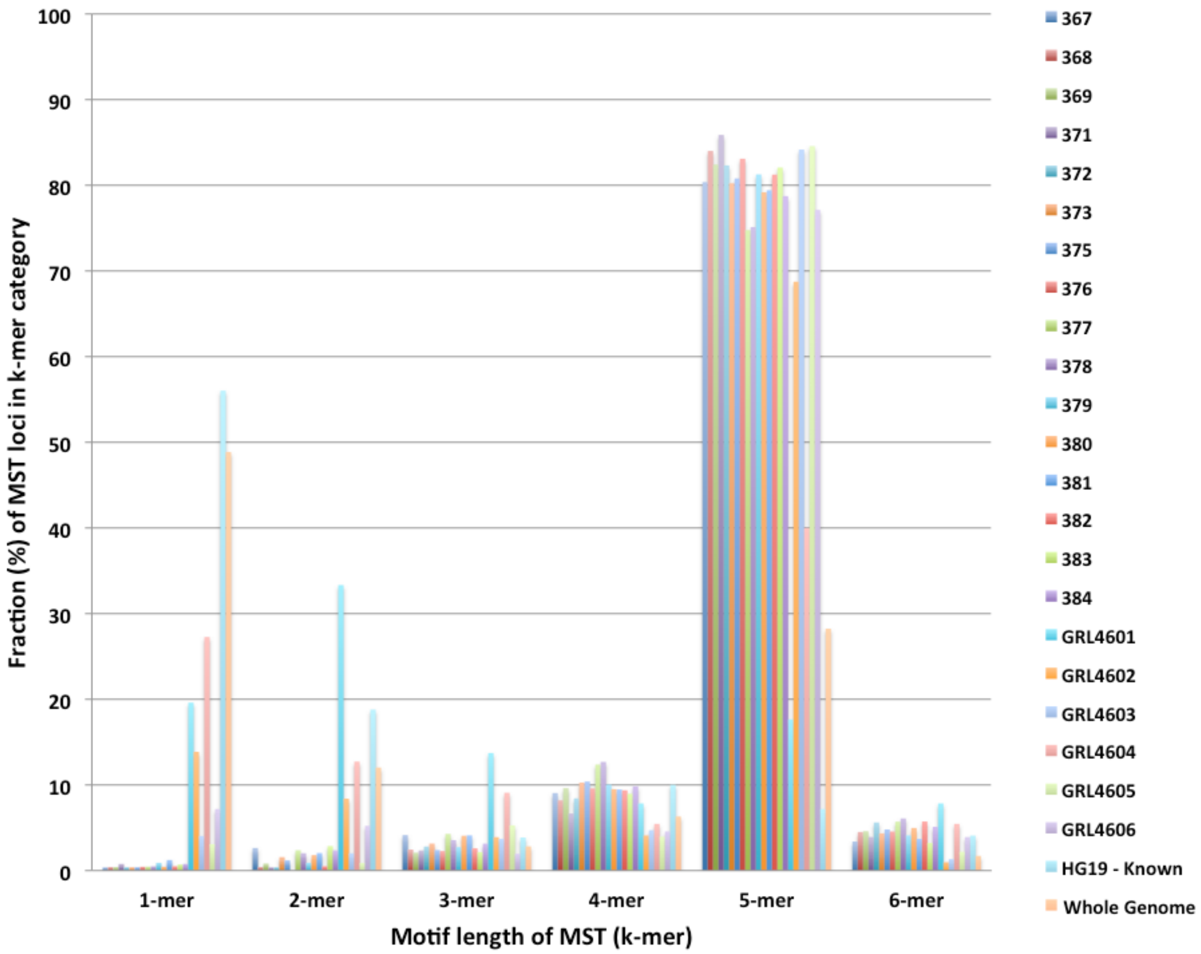


**Table 2-1: Mapping of enrichment reads.**

| Sample & Enrichment | Total Reads | Mapped % | Exome Overlap % | Mapped with MST % | Known MST Loci called % | Unmapped % | Unmapped with MST% |
|---|---|---|---|---|---|---|---|
| **DLD-1- Exome** | 107763705 | 99.9 | 53.7 | 7.3 | 44.4 | 0.1 | 1.6 |
| **DLD-1 - GME** | 99869840 | 99.6 | 2.9 | 15.7 | 61.7 | 0.4 | 2.8 |
| **DLD-1 - Comb** | 86571454 | 99.8 | 52.3 | 12.1 | 41.1 | 0.2 | 4.2 |
| **SW403 - Exome** | 99392535 | 99.9 | 57.3 | 6.8 | 41.2 | 0.1 | 1.8 |
| **SW403 - GME** | 93042396 | 99.1 | 2.3 | 67.4 | 11.7 | 0.9 | 3.8 |
| **SW403 - Comb** | 95846052 | 99.7 | 51.7 | 12.0 | 27.3 | 0.3 | 3.3 |
| **Normal-1-GME** | 107184846 | 99.2 | 1.8 | 86.7 | 33.3 | 0.8 | 1.1 |
| **Normal-2-GME** | 86039208 | 98.4 | 1.9 | 87.5 | 26.3 | 1.6 | 0.5 |
| **Normal-3-GME** | 77776824 | 98.1 | 1.9 | 88.6 | 23.3 | 1.9 | 0.4 |
| **Normal-4-GME** | 76888422 | 99.0 | 2.0 | 89.1 | 23.1 | 1.0 | 0.9 |
| **Normal-5-GME** | 88088498 | 97.6 | 2.1 | 87.1 | 26.1 | 2.4 | 0.5 |
| **Normal-6-GME** | 87529722 | 97.7 | 2.0 | 87.7 | 25.9 | 2.3 | 0.4 |
| **Normal-7-GME** | 85362982 | 98.8 | 2.0 | 86.5 | 26.7 | 1.2 | 0.7 |
| **Normal-8-GME** | 69912104 | 98.3 | 1.8 | 88.7 | 24.1 | 1.7 | 0.6 |
| **Normal-9-GME** | 89072202 | 99.1 | 1.8 | 83.5 | 36.8 | 0.9 | 0.7 |
| **Normal-10-GME** | 88599848 | 99.1 | 1.9 | 87.0 | 30.6 | 0.9 | 0.7 |
| **Normal-11-GME** | 67477542 | 99.1 | 1.9 | 87.1 | 27.0 | 0.9 | 1.1 |
| **Normal-12-GME** | 74895624 | 98.4 | 2.1 | 87.6 | 27.3 | 1.6 | 0.5 |
| **Normal-13-GME** | 102597820 | 98.3 | 2.0 | 89.1 | 29.3 | 1.7 | 0.5 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Normal-14-GME** | 39497576 | 98.5 | 1.8 | 89.6 | 22.4 | 1.5 | 0.6 |
| **Normal-15-GME** | 97582298 | 99.3 | 1.8 | 88.1 | 29.3 | 0.7 | 1.3 |
| **Normal-16-GME** | 91613940 | 98.1 | 1.8 | 86.7 | 32.0 | 1.9 | 0.6 |

Table 2-2: Novel contigs and MSTs from unmapped reads.

| Sample & Enrichment | Novel Contigs | % Contigs with MSTs | % Contigs with GLS | % Contigs with both MST and GLS | Total MSTs | % Novel MSTs |
|---|---|---|---|---|---|---|
| **DLD1-Exome** | 224 | 8 | 16 | 2 | 48 | 35 |
| **DLD1-GME** | 1469 | 20 | 23 | 7 | 510 | 71 |
| **DLD1-Comb** | 515 | 36 | 25 | 9 | 297 | 79 |
| **SW403-Exome** | 162 | 19 | 16 | 3 | 55 | 62 |
| **SW403-GME** | 372 | 46 | 34 | 19 | 227 | 92 |
| **SW403-Comb** | 267 | 38 | 34 | 15 | 153 | 88 |
| **Normal-1-GME** | 312 | 52 | 34 | 20 | 265 | 78 |
| **Normal-2-GME** | 278 | 55 | 28 | 18 | 244 | 82 |
| **Normal-3-GME** | 261 | 54 | 30 | 17 | 239 | 74 |
| **Normal-4-GME** | 289 | 53 | 30 | 17 | 255 | 75 |
| **Normal-5-GME** | 257 | 54 | 34 | 18 | 249 | 70 |
| **Normal-6-GME** | 316 | 50 | 33 | 17 | 253 | 79 |
| **Normal-7-GME** | 275 | 52 | 36 | 20 | 250 | 71 |
| **Normal-8-GME** | 255 | 56 | 31 | 18 | 219 | 82 |
| **Normal-9-GME** | 245 | 52 | 36 | 18 | 210 | 74 |
| **Normal-10-GME** | 221 | 54 | 31 | 17 | 197 | 75 |
| **Normal-11-GME** | 248 | 52 | 33 | 20 | 219 | 79 |
| **Normal-12-GME** | 265 | 51 | 38 | 21 | 221 | 76 |
| **Normal-13-GME** | 308 | 51 | 32 | 18 | 243 | 79 |
| **Normal-14-GME** | 198 | 52 | 29 | 16 | 192 | 68 |
| **Normal-15-GME** | 351 | 52 | 34 | 19 | 279 | 81 |
| **Normal-16-GME** | 307 | 51 | 28 | 16 | 254 | 79 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **Total** | 7395 | 42 | 29 | 14 | 5079 | 77 |
| **Concordant** | 790 | 42 | 32 | 16 | 533 | 100 |

**Table 2-3: Alignment analysis of concordant contigs with normal lymphoblastoid RNA-Seq samples.** The top 10 out of 37 concordant contigs that had RNA-Seq hits are presented in this table. HS: Homo sapiens; PT: Pantroglodytes; Chr: chromosome.

| # | RNA-Seq samples | | | | | | | | | | Total aligned reads | Contig length | BLAST hit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | | |
| 1 | 9441 | 17648 | 208 | 31 | 6957 | 15064 | 7134 | 2005 | 12173 | 7 | 70668 | 370 | HS clone. Chr21 |
| 2 | 24 | 76 | 77 | 7 | 26 | 33 | 19 | 63 | 32 | 11 | 368 | 615 | HS FOSMID clone. Chr7 |
| 3 | 24 | 67 | 75 | 7 | 21 | 33 | 20 | 61 | 32 | 11 | 351 | 628 | HS FOSMID clone. Chr7 |
| 4 | 24 | 69 | 75 | 7 | 21 | 33 | 19 | 60 | 32 | 11 | 351 | 624 | HS FOSMID clone. Chr7 |
| 5 | 55 | 38 | 28 | 11 | 4 | 48 | 0 | 48 | 46 | 26 | 304 | 531 | HS FOSMID clone. Chr17 |
| 6 | 25 | 67 | 43 | 4 | 23 | 23 | 16 | 52 | 29 | 14 | 296 | 310 | PT BAC clone. Chr7 |
| 7 | 18 | 25 | 38 | 5 | 57 | 36 | 35 | 7 | 30 | 24 | 275 | 303 | HS BAC clone. Chr17 |
| 8 | 22 | 66 | 42 | 4 | 19 | 18 | 13 | 52 | 22 | 13 | 271 | 314 | PT BAC clone. Chr7 |
| 9 | 21 | 66 | 41 | 4 | 16 | 18 | 13 | 51 | 21 | 13 | 264 | 323 | PT BAC clone. Chr7 |
| 10 | 19 | 61 | 39 | 4 | 18 | 12 | 11 | 44 | 16 | 11 | 235 | 289 | PT BAC clone. Chr7 |

# REFERENCES

1 Consortium, I. H. G. S. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945 (2004).

2 Project, T. H. G. https://www.genome.gov/10001772/all-about-the--human-genome-project-hgp/ (2008) (Date of access: 08/08/2014).

3 Liu, Y. *et al.* Discovery of common sequences absent in the human reference genome using pooled samples from next generation sequencing. *BMC Genomics* **15**, 685 (2014).

4 Miga, K. H., Eisenhart, C. & Kent, W. J. Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments. *Nucleic Acids Res* **43**, e133 (2015).

5 Fonville, N. C., Ward, R. M. & Mittelman, D. Stress-induced modulators of repeat instability and genome evolution. *J Mol Microb Biotech* **21**, 36-44 (2011).

6 Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-3.0*. http://www.repeatmasker.org (1996-2010) (Date of access: 08/08/2014).

7 Guilmatre, A., Highnam, G., Borel, C., Mittelman, D. & Sharp, A. J. Rapid multiplexed genotyping of simple tandem repeats using capture and high-throughput sequencing. *Hum Mutat* **34**, 1304-1311 (2013).

8 Tae, H., Karunasena, E., Bavarva, J. H., McIver, L. J. & Garner, H. R. Large scale comparison of non-human sequences in human sequencing data. *Genomics* **104**, 453-458 (2014).

9 Ahmed, D. *et al.* Epigenetic and genetic features of 24 colon cancer cell lines. *Oncogenesis* **2**, e71 (2013).

10 Vilar, E. *et al.* MRE11 deficiency increases sensitivity to poly(ADP-ribose) polymerase inhibition in microsatellite unstable colorectal cancers. *Cancer Res* **71**, 2632-2642 (2011).

11 Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**, 573-580 (1999).

12 Majoros, W. H., Pertea, M. & Salzberg, S. L. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics (Oxford, England)* **20**, 2878-2879 (2004).

13 Grady, D. L. *et al.* Highly conserved repetitive DNA sequences are present at human centromeres. *Proc. Natl. Acad. Sci. USA* **89**, 1695-1699 (1992).

14 Subramanian, S., Mishra, R. K. & Singh, L. Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* **4**, R13 (2003).

**Figure S2-1: Read coverage analysis for MST motifs.** For every sample >15% of known microsatellite loci (MSTs in mapped reads) (A) and >50% of novel microsatellite loci (MSTs in novel contigs) (B) were covered by more than 10 sequencing reads.

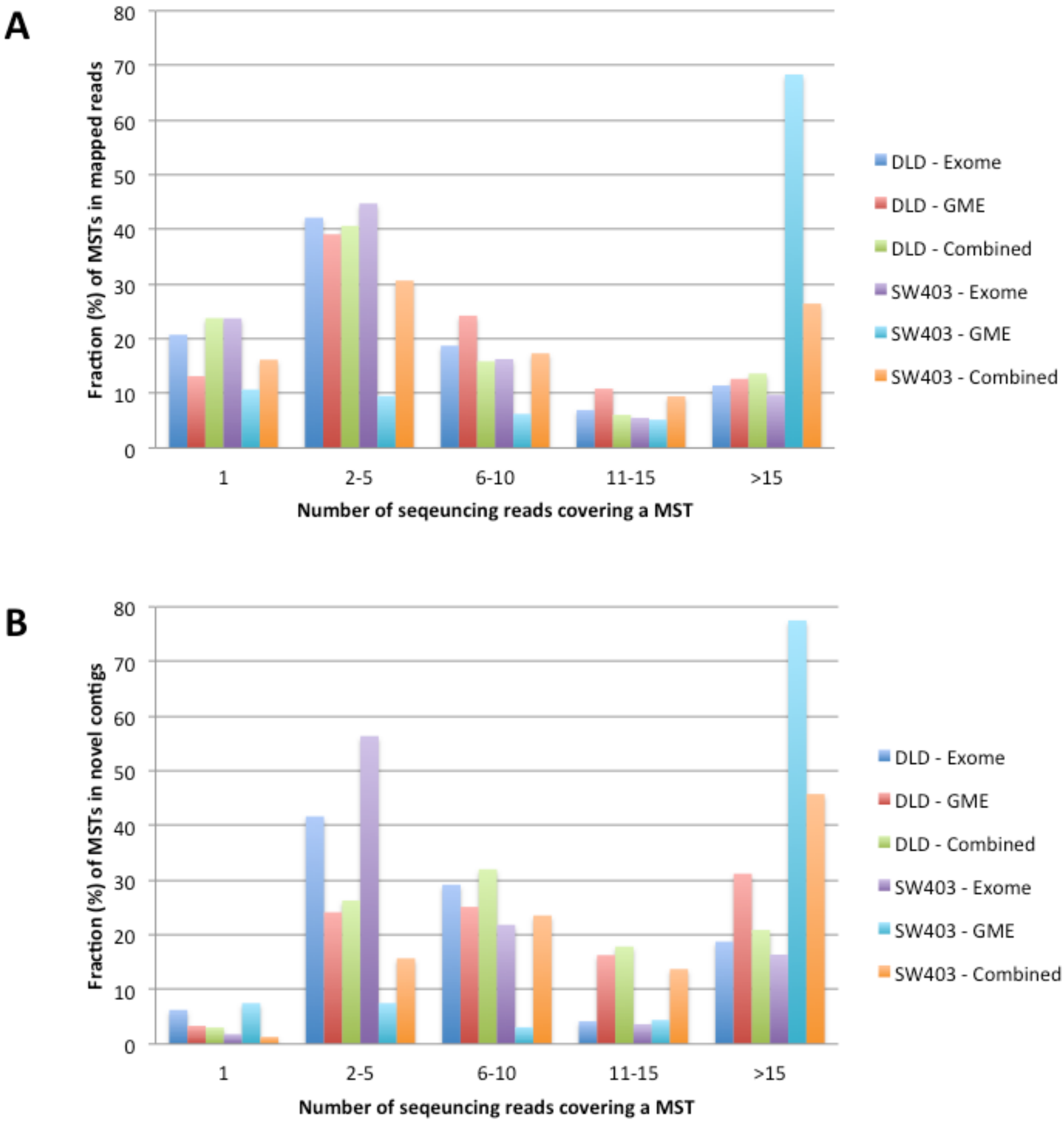


**Figure S2-2: Homopolymer Nucleotide Distribution.** As our GME captures repetitive motifs based on 30nt repetitive regions, it was possible that C/G nucleotides were captured in higher abundance than with standard exome capture. We found that the SW403-GME sample, but not the DLD1-GME enriched sample was enriched for C/G nucleotides. This difference in capture between the two GME samples may be due to slight differences in hybridization temperatures used while optimizing the capture process, and may indicate a use for this technology in specifically capturing C/G nucleotide regions in the future.

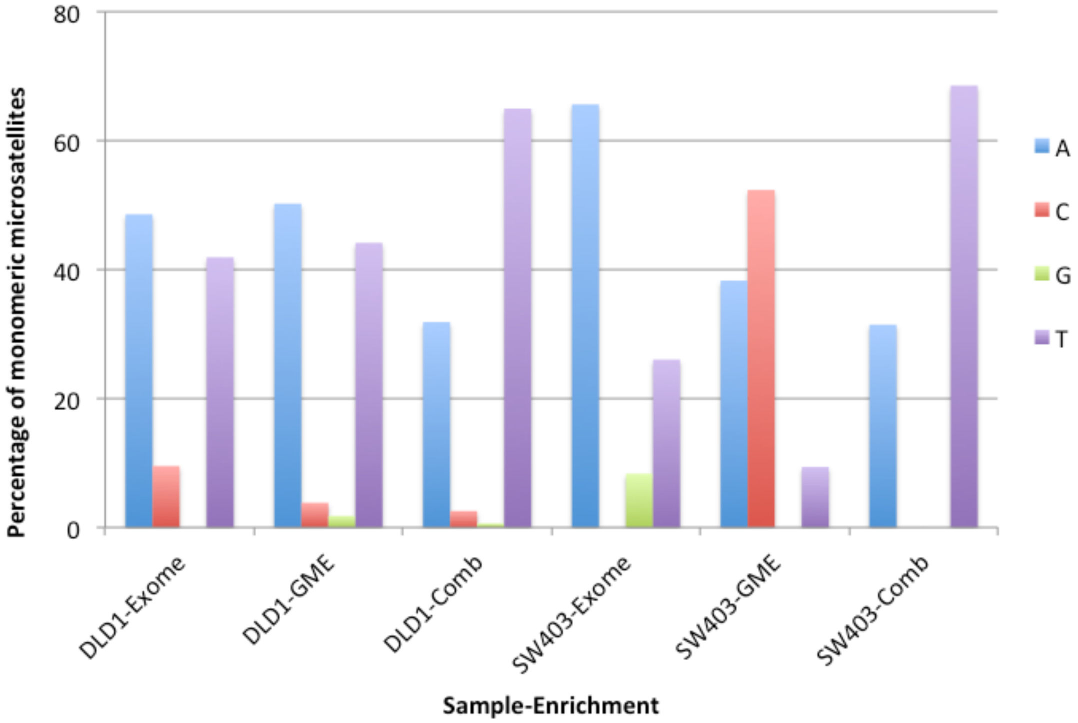


**Figure S2-3: Read coverage analysis of MST motifs in normal human samples.** (A) In every individual sample more than 30% of the MSTs are covered by more than 15 reads while an equal percentage are covered by 2 to 5 reads. (B) Approximately 60% of MSTs found in the novel contigs are covered by more than 15 reads.

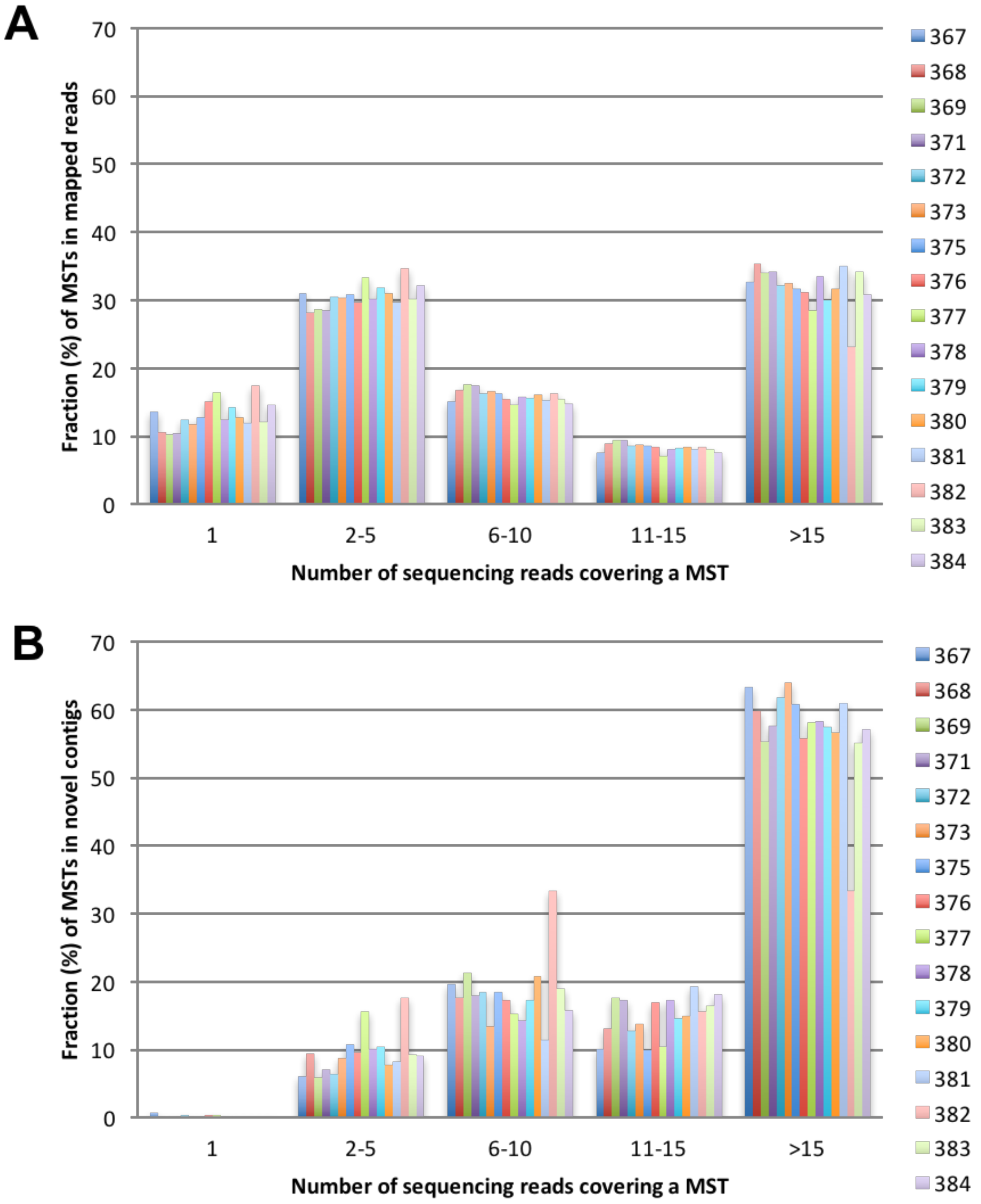


**Figure S2-4: Concordant contig length distribution.** This histogram shows how the concordant contigs with and without MSTs and GLS are distributed according to their lengths. Each category of contigs represents the combined information of all the samples (i.e. 16 GME sequenced normal samples, 3 DLD1 samples and 3 SW403 samples). It should be noted that except for the "less than 250 bases" bin, all other contigs contain two or more reads.

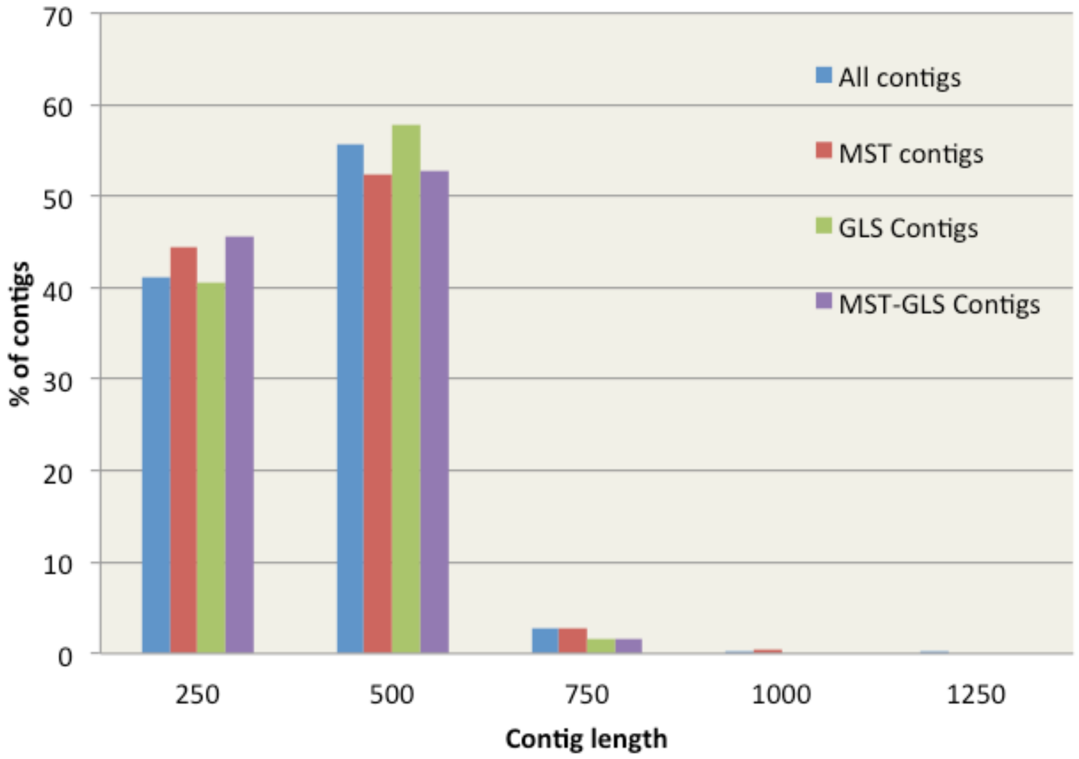


**Figure S2-5: Length distribution of novel microsatellites.** The following histograms show the MSTs according to their length in the colorectal cell lines (A) and in the normal samples (B).

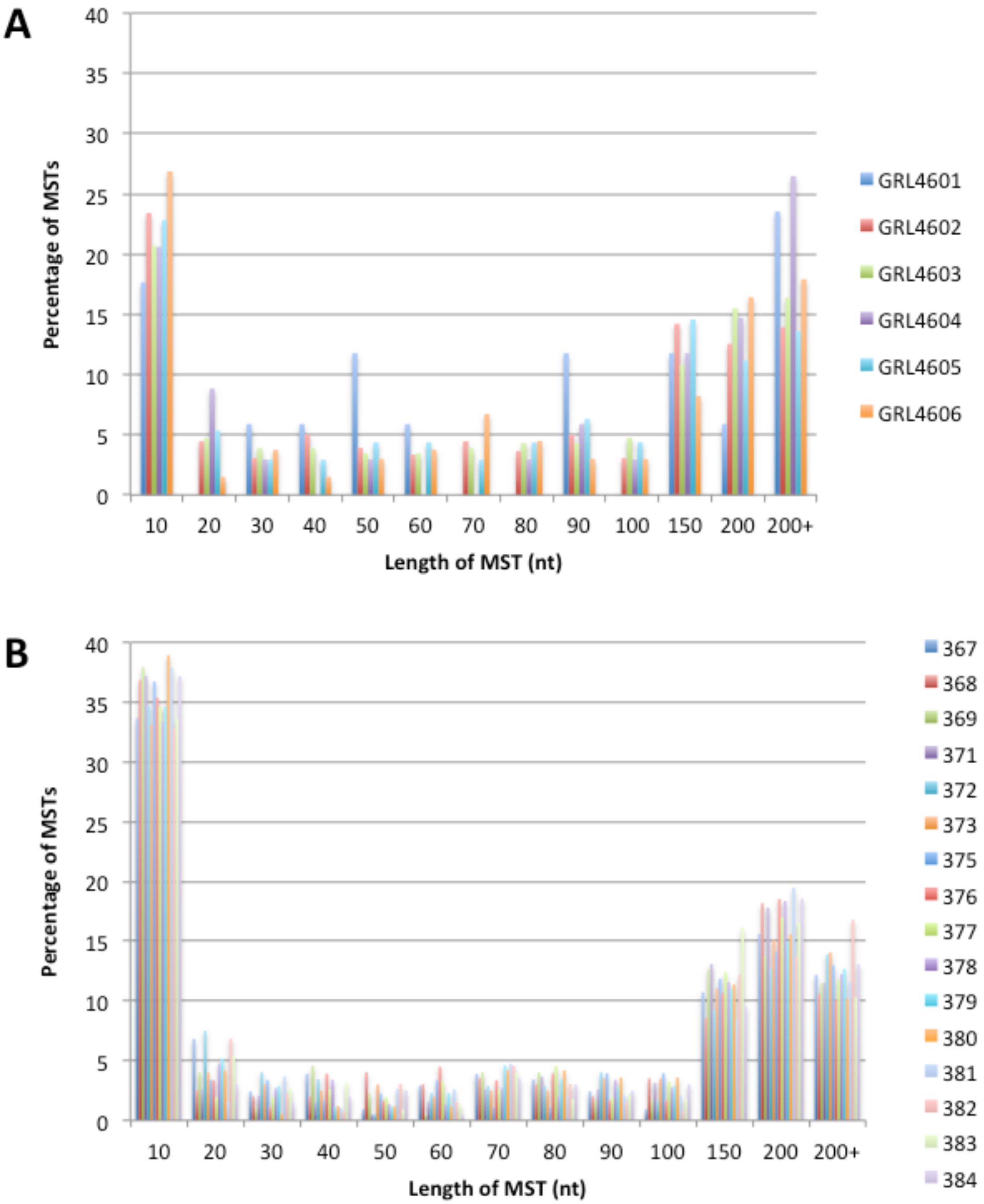


**Table S2-1: Details of the samples on which GME was performed.**

| Sample # | Sample | Sample Type | Sample Source |
|---|---|---|---|
| 1 | DLD1-Exome | Colorectal | ATCC CCL-221 |
| 2 | DLD1-GME | Colorectal | ATCC CCL-221 |
| 3 | DLD1-Comb | Colorectal | ATCC CCL-221 |
| 4 | SW403-Exome | Colorectal | ATCC CCL-230 |
| 5 | SW403-GME | Colorectal | ATCC CCL-230 |
| 6 | SW403-Comb | Colorectal | ATCC CCL-230 |
| 7 | Normal-1-GME | Lymphoblastoid | Coriell HG00384 |
| 8 | Normal-2-GME | Lymphoblastoid | Coriell HG00383 |
| 9 | Normal-3-GME | Lymphoblastoid | Coriell HG00382 |
| 10 | Normal-4-GME | Lymphoblastoid | Coriell HG00381 |
| 11 | Normal-5-GME | Lymphoblastoid | Coriell HG00380 |
| 12 | Normal-6-GME | Lymphoblastoid | Coriell HG00379 |
| 13 | Normal-7-GME | Lymphoblastoid | Coriell HG00378 |
| 14 | Normal-8-GME | Lymphoblastoid | Coriell HG00377 |
| 15 | Normal-9-GME | Lymphoblastoid | Coriell HG00376 |
| 16 | Normal-10-GME | Lymphoblastoid | Coriell HG00375 |
| 17 | Normal-11-GME | Lymphoblastoid | Coriell HG00373 |
| 18 | Normal-12-GME | Lymphoblastoid | Coriell HG00372 |
| 19 | Normal-13-GME | Lymphoblastoid | Coriell HG00371 |
| 20 | Normal-14-GME | Lymphoblastoid | Coriell HG00369 |
| 21 | Normal-15-GME | Lymphoblastoid | Coriell HG00368 |
| 22 | Normal-16-GME | Lymphoblastoid | Coriell HG00367 |

**Table S2-2: The statistics of the concordant contigs are furnished in the below table.** The data is divided according to the number of mismatches allowed for the concordant contig calculation to illustrate that if some mismatch is allowed (due to possible variations among individuals) the additional contigs may be further assembled. However, for further analysis, only those concordant contigs assembled using the strict 0 mismatch value were used.

| **Statistics** | **Mismatch** | |
|---|---|---|
| | **0** | **1** |
| Total concordant contigs (observed in at least 2 samples) | 790 | 747 |
| Total sample contigs found concordant | 4419 | 4915 |
| Maximum # of sample contigs assembled into a concordant contig | 25 | 29 |
| Minimum # of sample contigs assembled into a concordant contig | 2 | 2 |
| Average # of samples a concordant contig was found | 5.6 | 6.6 |

**Table S2-3: Concordant contig sample distribution.** Novel contigs that were present in multiple samples are considered to be concordant. The table groups the concordant contigs according to the number of samples in which they were found. The data is divided by mismatch 0 and mismatch 1 (Number of mismatches allowed while generating the concordant contigs).

| # of | Concordant contigs | |
|---|---|---|
| samples | 0mismatch | 1mismatch |
| 2 | 243 | 188 |
| 3 | 125 | 99 |
| 4 | 85 | 80 |
| 5 | 59 | 48 |
| 6 | 36 | 50 |
| 7 | 44 | 38 |
| 8 | 29 | 30 |
| 9 | 31 | 28 |
| 10 | 20 | 24 |
| 11 | 26 | 30 |
| 12 | 14 | 21 |
| 13 | 18 | 29 |
| 14 | 13 | 17 |
| 15 | 9 | 13 |
| 16 | 8 | 11 |
| 17 | 8 | 9 |
| 18 | 15 | 18 |
| 19 | 7 | 11 |
| 20 | 0 | 2 |

**Table S2-4:** Details about the RNA-Seq samples that were used to confirm that the concordant contigs contain potential functional elements.

| Sample # | SampleID | Sample Type | Sample Source |
|---:|---|---|---|
| 1 | ERR188040 | Lymphoblastoid | ArrayExpress - EBI |
| 2 | ERR188231 | Lymphoblastoid | ArrayExpress - EBI |
| 3 | ERR188043 | Lymphoblastoid | ArrayExpress - EBI |
| 4 | ERR188280 | Lymphoblastoid | ArrayExpress - EBI |
| 5 | ERR188325 | Lymphoblastoid | ArrayExpress - EBI |
| 6 | ERR188327 | Lymphoblastoid | ArrayExpress - EBI |
| 7 | ERR188373 | Lymphoblastoid | ArrayExpress - EBI |
| 8 | ERR188313 | Lymphoblastoid | ArrayExpress - EBI |
| 9 | ERR188382 | Lymphoblastoid | ArrayExpress - EBI |
| 10 | ERR188359 | Lymphoblastoid | ArrayExpress - EBI |

**Table S2-5: Thirty-seven concordant contigs aligned to at least one RNA-Seq read.** As indicated in methods, a RNA-Seq hit must align with 0 mismatches to at least 70% of the length of the contig.

| # | RNA-Seq samples | | | | | | | | | | Total aligned reads | Contig length | BLAST hit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | | | |
| 1 | 9441 | 17648 | 208 | 31 | 6957 | 15064 | 7134 | 2005 | 12173 | 7 | 70668 | 370 | HS clone. Chr21 |
| 2 | 24 | 76 | 77 | 7 | 26 | 33 | 19 | 63 | 32 | 11 | 368 | 615 | HS FOSMID clone. Chr7 |
| 3 | 24 | 67 | 75 | 7 | 21 | 33 | 20 | 61 | 32 | 11 | 351 | 628 | HS FOSMID clone. Chr7 |
| 4 | 24 | 69 | 75 | 7 | 21 | 33 | 19 | 60 | 32 | 11 | 351 | 624 | HS FOSMID clone. Chr7 |
| 5 | 55 | 38 | 28 | 11 | 4 | 48 | 0 | 48 | 46 | 26 | 304 | 531 | HS FOSMID clone. Chr17 |
| 6 | 25 | 67 | 43 | 4 | 23 | 23 | 16 | 52 | 29 | 14 | 296 | 310 | PT BAC clone. Chr7 |
| 7 | 18 | 25 | 38 | 5 | 57 | 36 | 35 | 7 | 30 | 24 | 275 | 303 | HS BAC clone. Chr17 |
| 8 | 22 | 66 | 42 | 4 | 19 | 18 | 13 | 52 | 22 | 13 | 271 | 314 | PT BAC clone. Chr7 |
| 9 | 21 | 66 | 41 | 4 | 16 | 18 | 13 | 51 | 21 | 13 | 264 | 323 | PT BAC clone. Chr7 |
| 10 | 19 | 61 | 39 | 4 | 18 | 12 | 11 | 44 | 16 | 11 | 235 | 289 | PT BAC clone. Chr7 |
| 11 | 21 | 61 | 37 | 4 | 18 | 11 | 9 | 43 | 15 | 11 | 230 | 294 | PT uncharacterized |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 14 | 54 | 38 | 2 | 13 | 13 | 8 | 37 | 14 | 11 | 204 | 274 | PT uncharacterized |
| 13 | 5 | 11 | 37 | 2 | 67 | 4 | 3 | 7 | 0 | 2 | 138 | 327 | HS clone. |
| 14 | 11 | 18 | 2 | 0 | 66 | 13 | 3 | 12 | 10 | 0 | 135 | 242 | HS rRNA gene |
| 15 | 13 | 15 | 8 | 1 | 13 | 11 | 0 | 27 | 18 | 9 | 115 | 387 | HS clone. Chr17 |
| 16 | 11 | 18 | 13 | 0 | 9 | 7 | 6 | 12 | 11 | 8 | 95 | 466 | HS clone. Chr21 |
| 17 | 5 | 8 | 8 | 0 | 28 | 0 | 2 | 9 | 3 | 9 | 72 | 363 | HS FOSMID. |
| 18 | 7 | 13 | 13 | 0 | 22 | 6 | 1 | 3 | 1 | 5 | 71 | 285 | PT BAC clone. ChrY |
| 19 | 1 | 3 | 9 | 2 | 14 | 0 | 0 | 2 | 0 | 1 | 32 | 437 | PA BAC clone. Chr16 |
| 20 | 3 | 11 | 4 | 0 | 2 | 0 | 0 | 6 | 2 | 1 | 29 | 324 | HS clone. Chr21 |
| 21 | 0 | 2 | 0 | 6 | 0 | 0 | 2 | 8 | 3 | 0 | 21 | 1016 | BB genome scaffold |
| 22 | 1 | 5 | 1 | 4 | 0 | 0 | 0 | 1 | 0 | 3 | 15 | 512 | No hits |
| 23 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 3 | 0 | 3 | 9 | 370 | HS clone. Chr21 |
| 24 | 1 | 2 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 7 | 368 | HS FOSMID clone. Chr11 |
| 25 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2 | 0 | 2 | 6 | 372 | HS clone. Chr21 |
| 26 | 1 | 0 | 2 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 423 | HS BAC clone. |
| 27 | 1 | 1 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 5 | 495 | HS clone. Chr9 |
| 28 | 0 | 0 | 0 | 0 | 3 | 2 | 0 | 0 | 0 | 0 | 5 | 346 | HS clone. Chr21 |

| | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 3 | 381 | PT BAC clone. Chr7 |
| 30 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 2 | 259 | OF genome scaffold |
| 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 449 | PT BAC clone. ChrY |
| 32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 314 | HS FOSMID clone. Chr7 |
| 33 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 232 | No hits |
| 34 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 364 | No hits |
| 35 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 277 | HS contig. |
| 36 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 283 | HS clone. ChrX |
| 37 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 295 | HS clone. Chr17 |

HS: Homo sapiens; PT: Pan troglodytes; BB: Babesia bigemina; OF: Onchocerca flexuosa; Chr: Chromosome.

**Table S2-6: Distribution of predicted exon number within the concordant contigs predicted by GlimmerHMM to have Gene-Like Structures (GLS).**

| #Exons | #GLS |
|---|---|
| 2 | 73 |
| 3 | 105 |
| 4 | 44 |
| 5 | 24 |
| 6 | 2 |
| 8 | 1 |
| Total | 249 |

**Table S2-7: Six putative cDNAs found in the concordant contigs had at least one read hit in the RNA-Seq samples.**

| # | Aligned RNA-Seq reads | Contig length | cDNA length | BLAST hit |
|---|---|---|---|---|
| 1 | 90 | 314 | 219 | PT BAC clone. Chr7 |
| 2 | 10 | 323 | 108 | PT BAC clone. Chr7 |
| 3 | 5 | 495 | 384 | HS clone. Chr9 |
| 4 | 1 | 370 | 159 | HS clone. Chr21 |
| 5 | 1 | 259 | 138 | OF genome scaffold |
| 6 | 1 | 368 | 159 | HS FOSMID clone. Chr11 |

PT: Pan troglodytes; HS: Homo sapiens; OF: Onchocerca flexuosa; Chr: Chromosome.

**Table S2-8: Top five pentameric and hexameric motif families.** The top five pentameric and hexameric motif families are shown for the DLD-1 and SW404 samples, a representative split (50,000 unmapped reads) from a whole genome sequenced sample and the known human reference hg19 where MSTs were identified using Tandem Repeat Finder (Hg19-TRF). The telomeric repeat is underlined.

| | DLD-1 Exome | | | DLD-1 GME | | | DLD-1 Combined | | |
|---|---|---|---|---|---|---|---|---|---|
| | Motif | Loci | Percent | Motif | Loci | Percent | Motif | Loci | Percent |
| Pentamer | AATGG | 2 | 22 | AATGG | 20 | 20 | GTGGA | 118 | 47.2 |
| | ATATA | 2 | 22 | GTGGA | 20 | 20 | AATGG | 116 | 46.4 |
| | AGGGG | 2 | 22 | ATATA | 6 | 6 | GAATT | 1 | 0.4 |
| | GGGAT | 1 | 11 | GGGGC | 4 | 4 | GGATT | 1 | 0.4 |
| | TTTTC | 1 | 11 | GCCCT | 4 | 4 | GGGAT | 1 | 0.4 |
| | | | | | | | | | |
| Hexamer | TCCTCT | 1 | 25 | CATCAC | 4 | 40 | CCCCGG | 1 | 25 |
| | GGGGGA | 1 | 25 | CCCTCA | 2 | 20 | AGGTCC | 1 | 25 |
| | CAGCAA | 1 | 25 | GGCCCA | 2 | 20 | CCTGGC | 1 | 25 |
| | CCTGGC | 1 | 25 | ATAAAA | 2 | 20 | CCCTCA | 1 | 25 |

| | SW404 Exome | | | SW403 GME | | | SW403 Combined | | |
|---|---|---|---|---|---|---|---|---|---|
| | Motif | Loci | Percent | Motif | Loci | Percent | Motif | Loci | Percent |
| Pentamer | GTGGA | 8 | 36.4 | AATGG | 107 | 55.7 | AATGG | 57 | 48.3 |
| | AATGG | 5 | 22.7 | GTGGA | 64 | 33.3 | GTGGA | 55 | 46.6 |
| | ATATA | 2 | 9.1 | CCCAC | 11 | 5.7 | CCCAC | 2 | 1.7 |
| | CAAAA | 1 | 4.5 | ATTCG | 2 | 1 | GGATT | 1 | 0.8 |
| | AGGGG | 1 | 4.5 | CCTTC | 1 | 0.5 | ATTTT | 1 | 0.8 |
| | | | | | | | | | |
| Hexamer | TCCTCT | 1 | 33.3 | CAGCAA | 1 | 20 | CTGGGG | 1 | 16.7 |
| | GGCCCA | 1 | 33.3 | CAACGA | 1 | 20 | TCCTCT | 1 | 16.7 |
| | CCTGGC | 1 | 33.3 | CTGGGG | 1 | 20 | CCCCGG | 1 | 16.7 |
| | | | | AAAATG | 1 | 20 | AGGGTC | 1 | 16.7 |
| | | | | CCCTCA | 1 | 20 | AAAATG | 1 | 16.7 |

| | HG19 TRF | | | Whole Genome | | |
|---|---|---|---|---|---|---|
| | Motif | Loci | Percent | Motif | Loci | Percent |
| Pentamer | CAAAA | 51763 | 28 | AATGG | 2192279 | 82.1 |
| | ATTTT | 39451 | 21.3 | CAAAA | 159237 | 6 |
| | TTTTC | 26684 | 14.4 | ATTTT | 89791 | 3.4 |
| | ATTAA | 6162 | 3.3 | TTTTC | 52059 | 2 |
| | CTTTC | 4678 | 2.5 | GTGGA | 50117 | 1.9 |
| | | | | | | |
| Hexamer | ACAAAA | 20644 | 17 | ACAAAA | 41995 | 25.9 |
| | TCTTTT | 16995 | 14 | ATAAAA | 21847 | 13.5 |
| | ATAAAA | 16293 | 13.4 | GGGTTA | 14626 | 9 |
| | GTATAT | 4297 | 3.5 | TCTTTT | 13237 | 8.2 |
| | CATACA | 3557 | 2.9 | TGTCTC | 4982 | 3.1 |

# Chapter 3: High-depth, high-accuracy microsatellite genotyping enables precision lung cancer risk classification

## ABSTRACT

There remains a large discrepancy between the known genetic contributions to cancer and that which can be explained by genomic variants, both inherited and somatic. Recently, understudied repetitive DNA regions called microsatellites have been identified as genetic risk markers for a number of diseases including various cancers (breast, ovarian and brain). In this study we demonstrate an integrated process for identifying and validating microsatellite based risk markers for lung cancer using data from the cancer genome atlas (TCGA) and the 1000 genomes project. Comparing whole exome germline sequencing data from 488 TCGA lung cancer samples to germline exome data from 390 control samples from the 1000 genomes project, we identified 119 potentially informative microsatellite loci. These loci (risk markers) were found to be able to distinguish between cancer and control samples with sensitivity and specificity ratios over 0.8. Then these loci, supplemented with additional loci from other cancers and controls, were evaluated using a custom target enrichment kit and sample-multiplexed nextgen sequencing. Thirteen of the 119 risk markers were validated using high-depth (579x±315) nextgen sequencing of 30 lung cancer and 89 control samples, resulting in sensitivity and specificity ratios were 0.90 and 0.94, respectively. When 8 loci harvested from the bioinformatic analysis of other cancers are added to the classifier, then the sensitivity and specificity rise to 0.93 and 0.97, respectively. Analysis of the genes harboring these loci revealed two genes (ARID1B and REL) and two significantly enriched pathways (chromatin organization and cellular response to stress) which suggest that the process of lung carcinogenesis is linked to chromatin remodeling, inflammation, and tumor microenvironment restructuring. We illustrate that high-depth nextgen sequencing enables a high-precision microsatellite-based risk classifier analysis approach. This microsatellite-based platform approach confirms the potential to create clinically actionable diagnostics for lung cancer.

## INTRODUCTION

Lung cancer has a high rate of incidence with 224,000 new cases projected this year alone: more than the next four cancers (colorectal, pancreatic, breast, and prostate) combined. Only 18% of those diagnosed with lung cancer will survive 5 years[1, 2]; however, early detection can dramatically improve outcomes. About 80% to 85% of lung cancers are found to be non-small cell lung cancer [3]. If found early the 5 year survival rate of non small cell lung cancer improves significantly: stage IA – 49%, stage IB – 54%, stage IIA – 30%, stage IIB – 31%, stage IIIA – 14%, stage IIIB – 5% and stage IV – 1%[4]. The differing stage dependent survival rate and varying provenance of new lung cancers underscores the value of developing a lung cancer genetic risk diagnostic – especially for screening of "at risk" populations (family members with lung cancer, second hand exposure to smoke or other hazards) which could be tested, with subsequent adjustments made to clinical observation or lifestyle. Interestingly, as the smoking rate has dropped in the US, it has been observed that ~20% of lung cancer deaths are from never smokers, attributable to other environmental exposures and genetic mutations[5, 6].

Studies of disease specific variation have largely neglected repetitive DNA in favor of Single Nucleotide Variants (SNVs). However, an abundance of neurological disorders have been linked to length specific variations in repetitive DNA microsatellites (MST)[7]. These microsatellites consist of short (1-6bp) units repeated in tandem. Recent studies have shown that microsatellites contribute to the genetic complexity of various cancers[8-10]. Based on these previous findings it is hypothesized that microsatellites may play a role in the genetics of lung cancer[11].

Our recent population-scale studies of MST loci and their repeat length variations have shown that MSTs can stratify risk, provide clinical decision support, and be potential therapeutic targets[8-10, 12-14]. These observations were made possible by building robust computational pipelines to accurately genotype MST loci based repeat length variation[8-10, 12]. Our previous work in computationally discovering clinically informative MST loci from publically available data sets (The Cancer Genome Atlas of affected individuals, the 1000 Genomes Project of healthy "normal" individuals) have yielded disease specific MST loci variations for breast cancer, ovarian cancer, glioblastoma, and lower-grade glioma[8-10, 12]. We have also shown somatic MST variability (SMV) and the presence of minor alleles can act as indicative disease markers for colorectal and liver cancer[15]. Furthermore, microsatellite variations are somatically acquired in normal tissues as one ages at rates higher than SNVs and that they are a sensitive measure of toxic environmental exposures[16, 17].

The goal of this manuscript is to discover and validate a set of microsatellite markers for lung cancer risk via comparison of patient germline and normal control germline exome sequences. Validation utilizes a custom target enrichment kit for high depth next-gen targeted sequencing. The focused, ultra-high read depth multiplexed sequencing approach used here enables accurate economical genotyping, validation and final selection of informative loci to ultimately create a high sensitivity and specificity risk classifier assay.

## METHODS

### Computational identification of LUAD and LUSC specific MST loci

A total of 266 LUAD and 222 LUSC germline exome samples were downloaded from TCGA. For the non-tumor control population, 390 germline exome samples were downloaded from the 1kGP. A set of 1.8 million MST loci, that were extracted from the human genome (38 build) using the tandem repeat finder, were genotyped in these samples. A modal genotype was computed for each MST locus using the 1kGP samples. A 2 X 2 Fisher's exact test was computed for each locus comparing the modal and non-modal genotype distributions in these two samples groups[10]. A Benjamini-Hochberg cut-off of 0.01% was used as a false discovery rate cut-off. A binary classifier was generated using ROCR library in R for the two MST loci lists to determine their potential to differentiate their corresponding lung cancer subtype from the normal control samples[42, 43].

### Assembling informative MST loci set to target enrich and validate

A set of 347 loci was assembled into the Illumina TruSeq Amplicon V1.5 kit target enrichment kit (Supplemental table 1). Of the 347, 119 were found to be specific for lung cancer, 144 were found in similar manner by analyzing other cancer datasets (Supplemental table 2) and 84 were included as controls (Supplemental tables 6, 7 and 8).

### Genomic DNA library prep and sequencing

Thirty lung cancer samples and 89 B-Lymphocyte non-tumor samples were obtained from Origene and Coriell cell repositories. The cells were cultured following the suppliers' recommended conditions (Supplemental tables 3 and 4). Isolation of genomic DNA was done using the Qiagen DNA Blood and Tissue kit following the manufacturer's protocol.

### Target sequencing, SMTEK

The assembled set of 347 MST loci was uploaded to Illumina's Design Studio tool, obtained in the form of a target enrichment kit and was used to target enrich and sequence the 30 lung cancer samples and 89 control samples.

### Genotyping of target enriched samples

The 347 MST loci were genotyped in the target enriched samples using custom written scripts, after performing quality control steps using the Trimmomatic tool to ensure only high quality reads are used in genotyping[44]. For each loci, a modal genotype and a predominant cancer genotype was computed. A modal genotype is the genotype that is found in more than 50% of the control samples and the predominant cancer genotype is the genotype that is found in more than 50% of the lung cancer samples. Any locus that has differing modal genotype and predominant cancer genotype was considered as a risk classifier (Supplemental table 5).

### Statistical procedure to assess differentiating power of the validated risk markers

Of the 119 computationally found lung cancer specific MST loci, 13 were found to differentiate lung cancer and normal control validation samples. A binary classifier was generated using the ROCR library in R using the 13 MST loci to assess their statistical capacity to call lung cancer samples from normal control samples. The sensitivity, specificity and other ROC related calculations were computed using the ROCR library in R. Odds ratio was calculated using the formula: (TP/FP)/(FN/TN), where TP, FP, FN and TN are true positive, false positive, false negative and true negative, respectively[45]. A set of 8 MST loci that were computationally found to be specific for other cancers were also found to differentiate the lung cancer samples from the normal control samples. This set was added to the 13 MST loci to form a

21 MST set. A similar statistical assessment was performed with this loci set. A leave one out cross validation was performed to quantify the consistency of the predictive power of the 21 loci classifier.

## Mechanistic analysis

Genes for each marker were identified from the UCSC genome browser referencing HG38. Functional enrichment analysis of genes harboring microsatellite markers and gene ontologies were obtained through the David Bioinformatics 6.8 Database[46]. Pathway analyses were performed using the Reactome database[26]. Alterations and co-occurrence/mutual exclusivity of genes in gene set were analyzed in TCGA lung cancer studies using cbioportal[47]. Studies included in cbioportal analyses were: Lung Adenocarcinoma[48], Lung Adenocarcinoma (TCGA, Provisional), Lung Squamous Cell Carcinoma[49], Lung Squamous Cell Carcinoma (TCGA, Provisional), and Pan-Lung Cancer[50]. Drug-ability of gene set was analyzed using the DGIdb database[24].

## RESULTS

### Cancer risk classification pipeline

A computational pipeline was created for both candidate marker discovery and validation (Figure 1). The process to identify statistically informative individual MST loci and develop a classification signature (including Receiver Operating Characteristic (ROC) curves and sensitivity and specificity calculations), follows the approach we have used previously for other cancers studies [8-10, 12]. We applied part (depicted on the left side of Figure 1) of the pipeline to compute classifiers for Lung Adenocarcinoma (LUAD) germline samples vs normal germline controls and Lung Squamous Cell carcinoma (LUSC) germline samples vs normal germline controls. Each of the individual loci found to be informative, i.e. which passed statistical and false discovery tests were harvested for inclusion on a custom nextgen target enrichment kit. Those loci were supplemented with additional informative loci gathered from an analysis of additional cancer types (breast, ovarian, melanoma and 3 different brain cancers) to identify potential pan-cancer markers. To the full set of informative loci were also added control loci that included random exon microsatellite loci, forensics/paternity testing loci, and MSI (microsatellite stability) loci to verify performance of the enrichment kit.

Once a set of potentially informative microsatellite loci were identified (left side of figure 1), lung cancer and control DNA samples were enriched for these markers using a custom Specific Microsatellite Target Enrichment Kit (SMTEK) and sequenced at high depth with 16 to 48 samples multiplexed on each sequencing run (right side of figure 1). The high depth sequencing of these regions enabled calling of high accuracy genotypes at each of the enriched loci. These genotypes were in turn used to validate those loci that could differentiate cancer from controls. Receiver Operating Characteristic curves were computed for the validated loci from the lung cancer sets and for the lung cancer set plus informative loci from the other cancer types. Using these two verified sets of high accuracy loci, we analyzed them for possible mechanistic (ontology, pathway, function, drug-ability, etc.) relationships to illustrate their potential role in lung cancer.

### Analysis of whole exome sequencing data for cancer and control germline samples

To compare MST genotypic variation in lung cancer germline samples and non-cancer germline control samples, 266 LUAD and 222 LUSC germline cancer exome samples were downloaded from The Cancer Genome Atlas (TCGA) and 390 germline non-cancer control exome sequencing data were downloaded from the 1000 Genomes Project (1kGP). With our 95% accurate[14] MST allele calling method, a total of 1.8 million microsatellite loci were analyzed by comparing modal (most frequent genotype in control samples) and non-modal genotype distributions in the two lung cancer sub-types and the non-cancer control samples. Two sets (one each for LUAD and LUSC) of MST loci were identified having significantly different genotypic distributions compared to non-cancer controls. Of these two sets, 96 LUAD and 67 LUSC MST loci (Supplemental table 6) passed false discovery rate tests. A classification model was developed to assess how well each set of markers differentiates the disease samples from healthy controls[10]. The Receiver Operating Characteristic (ROC) demonstrates the predictive power of this classification scheme as well as the value of the underlying sets of loci: the area under the curve (AUC) is 0.94 (LUAD) and 0.92 (LUSC) (Supplemental figures 3 and 4). For each classification scheme an "at risk" score was established by plotting accuracy vs. cutoff, a sample with 39% or more of the 96 LUAD signature MST loci set with non-modal genotype will be classified as 'at-risk' for adenocarcinoma of the lung (Figure 2A) while a sample with 37% or more of the 67 LUSC signature MST loci set with non-modal genotype will be classified as ‘at-risk’ for squamous cell carcinoma (Figure 2B). The specificity and sensitivity of the LUAD classification scheme is 0.87 and 0.87, respectively. The specificity and sensitivity of the LUSC classification scheme is 0.82 and 0.88, respectively. The specificity and sensitivity of the classification power of the LUAD signature set was found to be 0.87 and 0.87 the same for the LUSC signature set was found to be 0.82 and 0.88.

### High-depth target sequencing of computationally harvested disease specific MST loci

To assess the differentiating power of the computationally harvested 119 (96 LUAD and 67 LUSC; of which 44 were in common) MST loci using high-depth enabled high accuracy genotyping to validate the computational findings, the 119 MST loci along with 144 MST loci computationally found to be specific for other cancers, and control loci were combined and enriched in 30 lung cancer samples (Supplemental table 3) and 89 non-cancer control samples (Supplemental table 4). Supplementary figures 1 and 2 show that more than 93% of the loci were called in all the lung cancer and non-cancer control samples. The average read depth per loci across all samples was 579x (Standard Deviation, 315x). The minimum read depth was 83x.

### Control loci genotyping

A set of 84 control loci were added to the enrichment kit to demonstrate the hypermutable nature of forensic and paternity test MST loci and the resilience of random MST loci in exonic regions (highly conserved) of the genome to MSI in both the lung cancer and non-cancer control sample groups. Of the 84 control loci, 79 were reliably called in both sample groups. About 70% of the 64 control loci found in exon regions were found to have the same predominant genotype (found in greater than 50% of the control or cancer samples) in both sample groups while 86% of the

15 hyper mutable loci were found to have a wide spectrum of genotypes. None of the 79 control loci were found to have consistent genotypes that differed between the sample groups (cancer and normal), that is none of the control loci were informative for differentiating the two groups, as expected.

### Validation of LUSC and LUAD MST loci sets

We successfully called the genotype for 105 out of 119 microsatellite markers that significantly differ in the lung cancer germline samples compared to non-cancer controls. Specifically, the predominant genotype for these 105 markers was calculated using high depth sequencing of 30 lung cancer samples (Supplemental table 3) and 89 non-cancer control samples (Supplemental table 4). A subset of 13 markers (from the 105) were found to have differing predominant genotypes between the high depth lung cancer and non-cancer control datasets.

### Genotyping MST loci from other diseases in the lung cancer samples

Recent findings from pan-cancer studies suggest that different cancer types share oncogenic signatures[18, 19]. We investigated this possibility by including 144 informative MST loci identified in studies of breast cancer, ovarian cancer, lower grade glioma, glioblastoma, melanoma, and medulloblastoma to the target enrichment kit. Of the 144 loci, 137 loci were reliably genotyped in both sample groups. Among these, 8 loci (Table 1) were found to have differing predominant genotypes in the high depth lung cancer and 1kGP non-cancer control datasets.

### Performance of the high-depth informative loci as a classifier

A binary classification model was developed to assess the power of the 13 validated MST loci set (Table 1) to differentiate lung cancer samples from non-cancer control samples. The 13 MST loci signature differentiated lung cancer samples from non-cancer control samples with a sensitivity of 0.90 and specificity of 0.94. The area under the ROC curve was 0.96 (Supplemental figure 5A). An optimal cutoff of 0.61 was identified by calculating the accuracy vs. cutoff (Supplemental figure 5B). This result has a simple interpretation: 8 or more predominant genotypes (out of 13) indicate an increased risk for non-small cell lung cancer. (Figure 3A).

A similar classification model was generated for all 21 validated MST loci set (Table 1). This model has higher classification power with sensitivity and specificity values of 0.93 and 0.97 respectively. The area under the ROC curve is 0.97 (Supplemental figure 6A). The accuracy vs. cutoff plot suggests a cutoff of 0.57. The 21 MST classifier (Supplemental figure 6B) shows that any sample with 57% or more of the 21 MST loci with predominant cancer genotype will be classified as 'at-risk' for non-small cell lung cancer (Figure 3B).

While the statistical power of the 21 MST loci to differentiate lung cancer from non-cancer control samples is significant, a leave one out cross validation was performed to estimate the performance of this model. The leave one out analysis (see methods) predicted 28 out of 30 lung cancer samples to be 'at-risk' and 88 out of 89 non-cancer control samples to be 'healthy'. The average sensitivity and specificity of this cross validation effort, corresponding to the 119 leave one out iterations (due to the 30 + 89 sample count), was found to be 0.93 and 0.97, respectively. This cross validation demonstrates the consistency of this prediction method.

### Potential roles of the genes that harbor these informative loci

Of the 13 MST loci that are found to differentiate lung cancer samples from control samples, all were in the intronic regions of genes. To understand the mechanistic roles of these genes, the occurrence of mutations in these 13 genes were examined in 5 TCGA lung cancer studies. On average 37% of the lung cancer samples in these 5 studies contained mutations in at least one of the 13 genes (Supplemental table

10). An LUSC study with 177 lung cancer samples had ~50% of the samples with mutations in at least one out of the 13 genes (Supplemental table 10). Nine gene pairs were found to co-occur significantly (Supplemental table 11). Of these gene pairs, the REL gene significantly co-occurred with 4 genes (PPP1R21, CCDC88A, ATG3, and PRPF18) and ARID1B co-occurred with 2 genes (IMPG1, FUBP3). Interestingly, when these 13 genes were inspected for possible association with cancer using the COSMIC Cancer Gene Census[20], only REL and ARID1B were found to be previously implicated in cancer[21-23]. When all 13 genes were examined for possible drug-ability, using DrugDB[24], REL and ARID1B were found to be clinically actionable. When clustering the 13 genes using the David ontology database we found alternative splicing (P value: 0.005) and splice variants (P value: 0.046) to be significant ontological characterizations (Supplemental table 12). It should be noted that all the 13 MST loci that are found to be lung cancer differentiating are found in the intron regions of genes (Table 1). It has been shown previously that alterations in the MST loci in the intronic regions of the genes can influence transcription, alternative splicing, or mRNA export to the cytoplasm[25]. Upon further investigation of the 13 genes using Reactome[26], we found that two pathways were statistically enriched: the cellular response to stress pathway and the chromatin organization pathway.

## DISCUSSION

Although substantial effort has been directed at identifying diagnostic actionable markers, there is a significant gap between the known genetic contributors to lung cancer and the number and power of known inherited and somatic variants. About 85% of all lung cancers are non-small cell lung cancer (NSCLC), and with a better understanding of the heterogeneity of NSCLC, more patient specific treatment options are on the rise[27]. The success of tailored lung cancer treatment arises from improvements in genetic and epigenetic biomarker discovery[28]; however, early detection still remains the most significant factor in cancer survival. Important to early discovery is the identification of genetic risk markers, inherited or spontaneous, that will aid in identifying high-risk patients for enhanced monitoring or preventative measures. Recent studies have found that 20% of newly diagnosed lung cancer patients are never-smokers, underscoring the need and potential for new genetic risk markers.[2, 5, 6] The markers found in this study will potentially fill the gap by enabling risk stratification, allowing clinicians to monitor high-risk patients more closely leading to earlier detection.

By analyzing the modal and non-modal genotypic distribution of about 1.8 million MSTs in the two lung cancer sub-types in comparison with the non-cancer control samples, we found 67 LUSC and 96 LUAD MST loci (Supplemental table 7) that can differentiate their corresponding lung cancer sub-type from the non-cancer controls at significant sensitivities and specificities (Supplemental figures 3 and 4). Although we have previously demonstrated our genotyping accuracy from exome datasets to be 95%, the modest (~15x) read coverage in publically downloaded TCGA and 1000 Genome Project exome datasets limits the accuracy and ability to call genotypes at all loci[13]. Low coverage, reduced sequence complexity and non-random variation, have hampered microsatellite based biomarker discovery[13, 29]. While we have addressed these limitations in our previous efforts to identify cancer associated MST loci and tuned our genotyping algorithms accordingly, here we endeavored to mitigate the main source of genotyping error by dramatically increasing the depth of coverage at informative loci[8-10, 12]. Hence we specifically enriched 347 (disease and control) MST loci (Supplemental table 1) in 119 multiplexed samples and attained an average per locus sample read depth of 579 (Supplementary figure 2) which is ~20 times the usual exome read depth[8-10, 12].

With high-depth enabled high accuracy genotyping, 13 MST loci were found to have one predominant genotype (a genotype found in more than 50% of group members) that differed in the lung cancer and the non-cancer control groups (Supplemental table 5). Eight MST loci previously found to be specific for other cancers were also able to differentiate lung cancer samples from the non-cancer control samples. The culling of uninformative loci in the validation study and incorporation of control loci significantly increased the overall reliability, sensitivity and specificity of the assay.

All MST loci were found in genes (Table 1), specifically within the introns of genes, which we have previously shown to influence alternative splicing[25]. Of these genes, all 13 are expressed in the lung, giving confidence that genes could play a role in lung carcinogenesis. TCGA analysis suggests REL, ARID1B and associated genes can drive lung carcinogenesis through DNA damage and chromatin remodeling induced genomic instability (Figure 4).

ARID1B is a transcriptional modulator of specific genes through chromatin remodeling. ARID1B is a part of the switch/sucrose non-fermenting (SWI/SNF) complex that is implicated in several cancers[30]. More recently a study reported that loss of function in the SWI/SNF complex leads to genomic instability in lung cancer[31]. Lung cancer subtypes showed no significant differences between histology, implying that the loss of SWI/SNF function caused genomic instability regardless of lung cancer subtype, consistent with our observation that ARID1B is a marker for both LUAD and LUSC [Table 1].

REL is a proto-oncogene and member of the NFκB transcription factor family. Rel/NFκB transcription factors are critically involved in innate and adaptive immune responses through the up-regulation of chemokines, cytokines, cell adhesion molecules and proteases. It has been shown that chronic inflammation increases the likelihood of tumorigenesis through increased proliferation and DNA damage[32]. The role of tobacco smoke as a carcinogen has been highly correlated with lung cancer and one explanation is the production of reactive oxygen species that is known to cause DNA damage and to activate NFκB[33]. It can be deduced that alterations in REL could predispose a smoker to increased risk of cancer compared to a non-smoker. Overexpression of REL has been associated to many lymphoid cancers such as Primary mediastinal B-cell lymphoma, Classical Hodgkin's Lymphoma, and solid tumors such as Breast cancer, Pancreatic cancer, and Head and Neck cancer but not lung cancer[21, 34-37]. Interestingly, our lung cancer risk classifying locus harbored by the REL gene lies only 68 base pairs downstream of Exon 6 and 17 base pairs upstream of Exon 7, both of which are included in the REL Homology Domain (RHD). The RHD is an N-terminal protein domain which is shared by REL genes, which mediates DNA binding, inhibitor binding, nuclear localization signal and dimerization[33]. It can be inferred that intronic mutations located in between two exons in close proximity of each other can affect protein structure in the RHD that can influence downstream effects of the NFκB pathway and consequently predispose individuals to cancer.

Further, REL, ARID1B and other genes in our set are members of pathways, the cellular response to stress pathway and the chromatin organization pathway, that could potentially play a role in carcinogenesis. Chromatin remodeling is a dynamic process which regulates DNA repair, recombination, and gene transcription, which if impaired can play a pivotal role in carcinogenesis[38, 39]. Few studies have reported the association of the chromatin-remodeling pathway to lung cancer risk. For example, one recent study identified polymorphisms in the chromatin-remodeling pathway as a lung cancer risk classifier in a Chinese population[40]. The cellular response to stress pathway is involved in damage control through protective or destructive cell response mechanisms that promote survival or initiate cell death[41]. Dysfunction in the cell response stress pathway can lead to inappropriate response to stress and accumulate mutations. For example, our previous studies revealed that a single exposure to carcinogens such as nicotine can cause mutations to accumulate in epithelial tissue, which can contribute to carcinogenesis[16].

Taken together, our results propose a 13 loci lung cancer risk classifier that may reveal insight into the mechanism of lung carcinogenesis. Dysfunctions in the two significantly enriched pathways can possibly encourage lung carcinogenesis through chromatin remodeling, inflammation and tumor microenvironment restructuring. The genes ARID1B and REL are of special interest because of their druggability, oncogenic implications, odds risk ratio scores (Supplemental table 11) and co-occurrence with other implicated loci. These findings may be of interest because of the clinical potential value of this lung cancer risk classifier for novel therapeutic target discovery, lung cancer prediction, and cancer risk assessment.

# FIGURES

## Figure 3-1: Flow chart of informative loci identification and validation

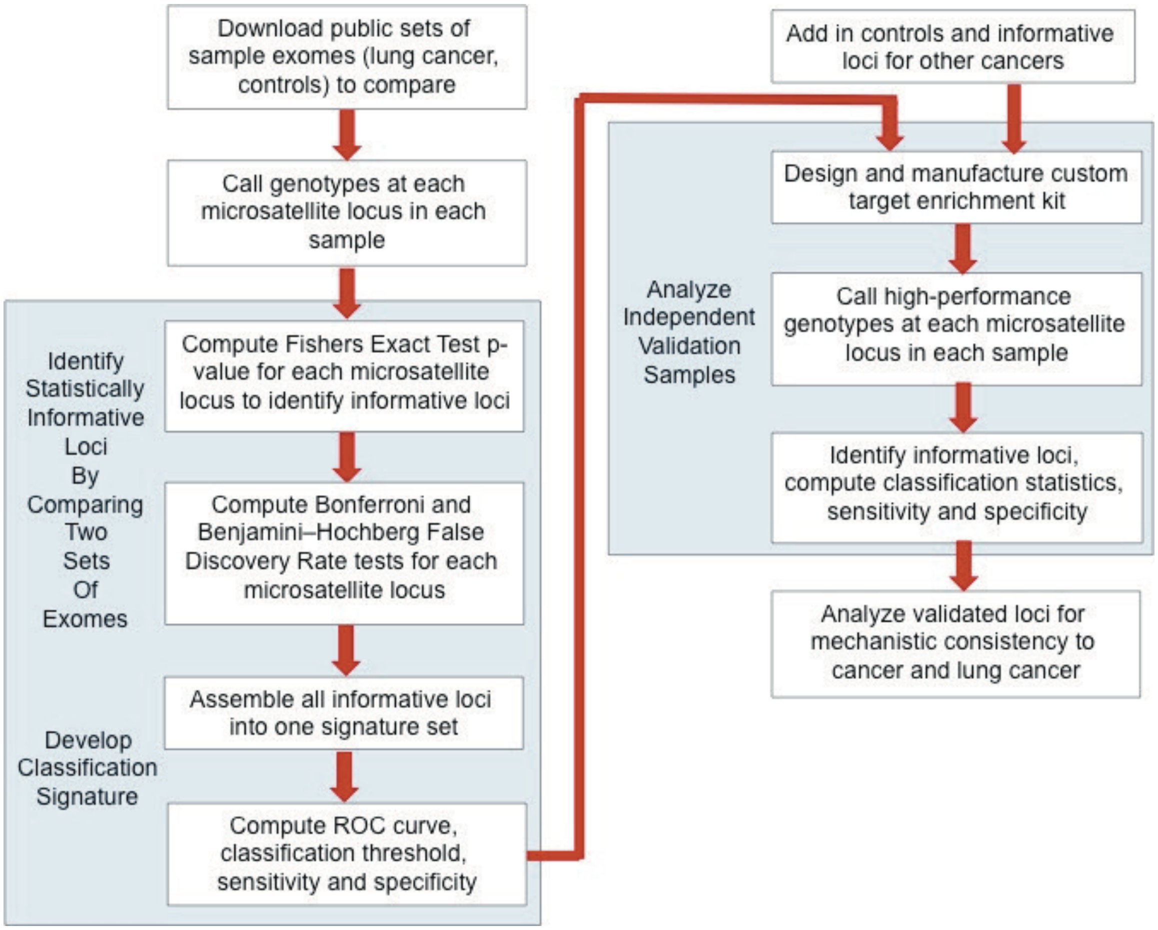

Download public sets of sample exomes (lung cancer, controls) to compare
Call genotypes at each microsatellite locus in each sample
Identify Statistically Informative Loci By Comparing Two Sets Of Exomes
Compute Fishers Exact Test p-value for each microsatellite locus to identify informative loci
Compute Bonferroni and Benjamini–Hochberg False Discovery Rate tests for each microsatellite locus
Develop Classification Signature
Assemble all informative loci into one signature set
Compute ROC curve, classification threshold, sensitivity and specificity
Add in controls and informative loci for other cancers
Analyze Independent Validation Samples
Design and manufacture custom target enrichment kit
Call high-performance genotypes at each microsatellite locus in each sample
Identify informative loci, compute classification statistics, sensitivity and specificity
Analyze validated loci for mechanistic consistency to cancer and lung cancer

**Figure 3-2: The computationally harvested LUAD and LUSC MST loci differentiate their corresponding cancer type from 1000 genomes non-cancer control samples with high sensitivity (LUAD: 0.87, LUSC: 0.88).** (A) A sample with 39% (vertical black line; identified via ROC analysis) or more of the 96 LUAD specific MST loci with cancer genotype will be called 'at-risk' for adenocarcinoma of the lung. (B) A sample with 37% or more of the 67 LUSC specific MST loci with cancer genotype will be called 'at-risk' for squamous cell carcinoma. The orange and blue bars represent cancer germline and control germline samples, respectively.

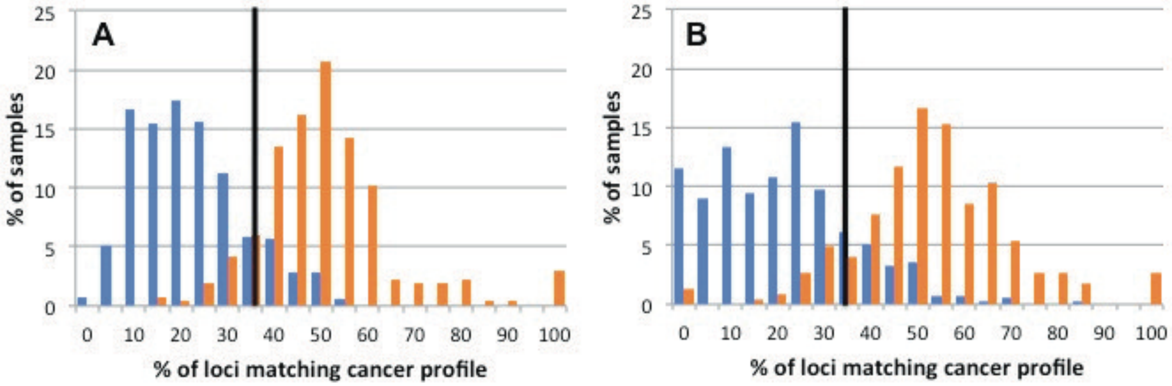


**Figure 3-3: The 13 lung cancer specific MST loci and 8 MST loci specific for other diseases can differentiate between the lung cancer and non-cancer control sample groups.** The blue and red bars represent the non-cancer control and lung cancer samples, respectively. (A) A sample with 61% or more of the 13 MST loci with cancer genotype will be termed 'at-risk' for lung cancer with sensitivity and specificity values of 0.90 and 0.94. (B) A sample with 57% or more of the 21 MST loci with cancer genotype will be termed 'at-risk' for lung cancer with sensitivity and specificity values of 0.93 and 0.97. The vertical black line corresponds to the optimum cut-off values found from the ROC analysis.

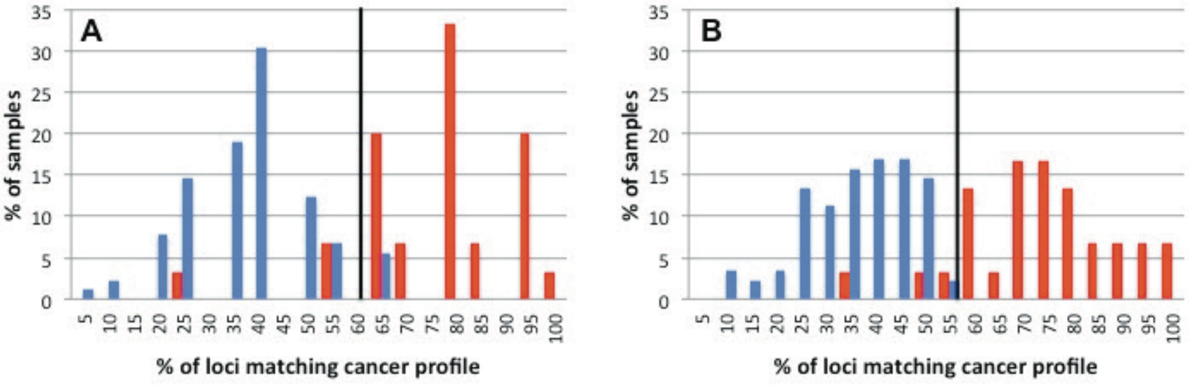


**Figure 3-4: Schematic describing potential mechanism underlying lung carcinogenesis**. Two genes out of 13 have significant oncogenic potential.

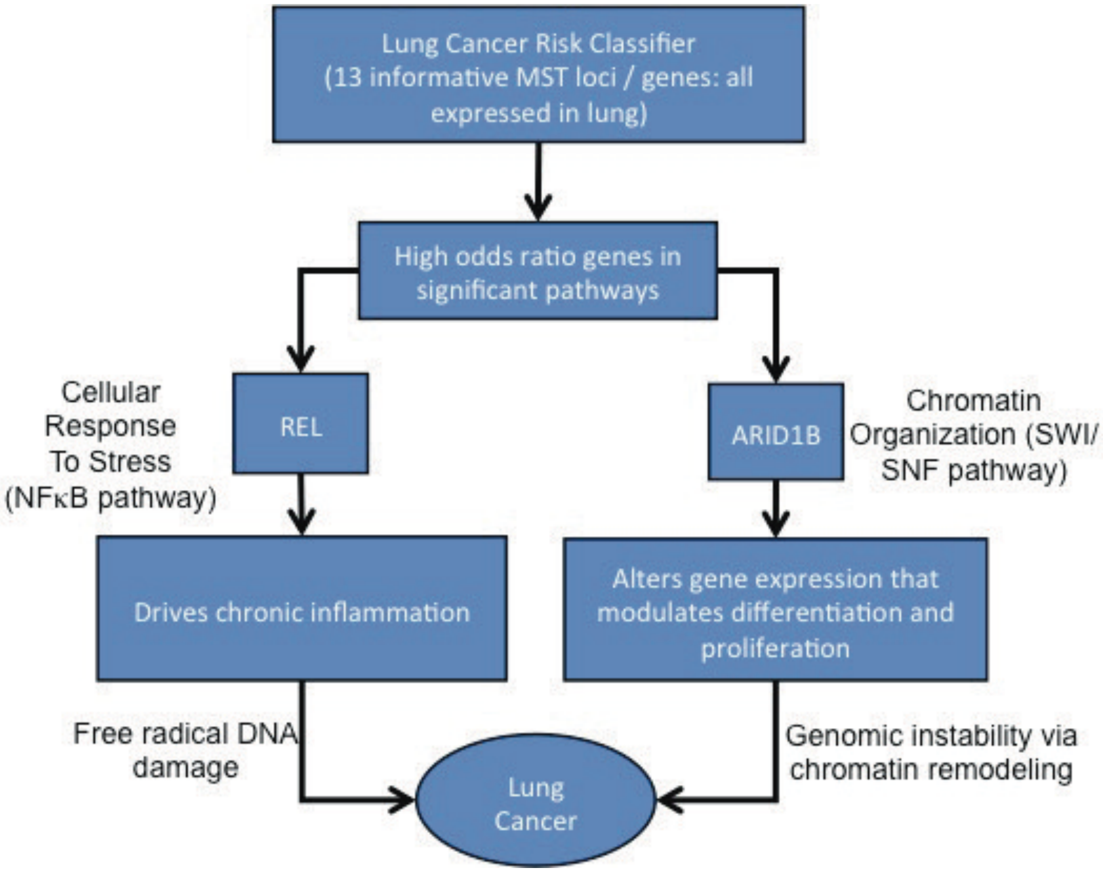

Lung Cancer Risk Classifier
(13 informative MST loci / genes: all expressed in lung)
High odds ratio genes in significant pathways
Cellular Response To Stress (NFκB pathway)
REL
ARID1B
Chromatin Organization (SWI/SNF pathway)
Drives chronic inflammation
Alters gene expression that modulates differentiation and proliferation
Free radical DNA damage
Genomic instability via chromatin remodeling
Lung Cancer

**Table 3-1: MST loci that can precisely differentiate between the lung cancer samples and non-tumor samples.**

| Genomic position | Repeat | Gene region | Gene | Entrez ID | Disease | Odds ratio |
|---|---|---|---|---|---|---|
| chr2:60918364-60918376 | T | Intron | REL | 5966 | LUAD, LUSC | 39.92 |
| chr6:157174818-157174831 | T | Intron | ARID1B | 57492 | LUAD, LUSC, MB, SKCM | 13.57 |
| chr6:76018867-76018880 | A | Intron | IMPG1 | 3617 | LUSC, OV | 12.28 |
| chr3:94035443-94035458 | T | Intron | ARL13B | 200894 | GBM, LUAD, SKCM | 11.20 |
| chr3:112534347-112534360 | A | Intron | ATG3 | 64422 | GBM, LGG, LUAD, LUSC | 10.29 |
| chr8:129862369-129862381 | A | Intron | FAM49B | 51571 | LUAD, LUSC | 7.01 |
| chr9:130622843-130622857 | A | Intron | FUBP3 | 8939 | GBM, LGG, LUAD, LUSC, MB, OV | 6.93 |
| chr7:135414296-135414309 | A | Intron | CNOT4 | 4850 | LUAD | 5.07 |
| chr2:48461120-48461133 | T | Intron | KLRAQ1 | 129285 | LGG, LUAD, LUSC, MB, SKCM | 4.43 |
| chr2:55332516-55332530 | A | Intron | CCDC88A | 55704 | LUAD, LUSC, SKCM | 3.90 |
| chr13:31148484-31148500 | A | Intron | HSPH1 | 3315 | LUAD, SKCM | 3.70 |
| chr15:20458509-20458521 | A | Intron | HERC2P3 | 283755 | LUAD | 3.00 |

| | | | | | | |
|---|---|---|---|---|---|---|
| chr10:13591929-13591943 | T | Intron | PRPF18 | 8559 | LUSC | 2.25 |
| chr2:202815832-202815844 | A | Intron | ICA1L | 130026 | BC, GBM, OV | 7.61 |
| chr13:114236623-114236635 | T | Intron | CDC16 | 8881 | LGG, SKCM | 6.19 |
| chr12:106106383-106106396 | A | Intron | NUAK1 | 9891 | OV | 5.74 |
| chr3:98580864-98580876 | A | Intron | CPOX | 1371 | BC, OV | 5.02 |
| chr16:70839964-70839978 | T | Intron | HYDIN | 54768 | GBM | 4.58 |
| chr2:233460070-233460083 | A | Intron | DGKD | 8527 | OV | 3.82 |
| chr5:87383860-87383873 | T | Intron | RASA1 | 5921 | OV | 2.93 |
| chr8:23852057-23852082 | TG | Intron | STC1 | 6781 | BC | 1.87 |

## REFERENCES

1 Society AC. Cancer Facts & Figures 2016, 2016.

2 Institute NC. SEER Stat Fact Sheets: Lung and Bronchus Cancer, 2016.

3 Society AC. What is non-small cell lung cancer?, 2016.

4 Society AC. Non-small cell lung cancer survival rates, by stage, 2016.

5 Li Y, Sheu CC, Ye Y, de Andrade M, Wang L, Chang SC *et al*. Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. The Lancet Oncology 2010; 11: 321-330.

6 Rivera GA, Wakelee H. Lung Cancer in Never Smokers. Advances in experimental medicine and biology 2016; 893: 43-57.

7 Budworth H, McMurray CT. A brief history of triplet repeat diseases. Methods in molecular biology 2013; 1010: 3-17.

8 Fonville NC, Vaksman Z, McIver LJ, Garner HR. Population analysis of microsatellite genotypes reveals a signature associated with ovarian cancer. Oncotarget 2015; 6: 11407-11420.

9 Karunasena E, McIver LJ, Rood BR, Wu X, Zhu H, Bavarva JH *et al*. Somatic intronic microsatellite loci differentiate glioblastoma from lower-grade gliomas. Oncotarget 2014; 5: 6003-6014.

10 McIver LJ, Fonville NC, Karunasena E, Garner HR. Microsatellite genotyping reveals a signature in breast cancer exomes. Breast cancer research and treatment 2014; 145: 791-798.

11 Hause RJ, Pritchard CC, Shendure J, Salipante SJ. Classification and characterization of microsatellite instability across 18 cancer types. Nature medicine 2016.

12 Karunasena E, McIver LJ, Bavarva JH, Wu X, Zhu H, Garner HR. 'Cut from the same cloth': Shared microsatellite variants among cancers link to ectodermal tissues-neural tube and crest cells. Oncotarget 2015; 6: 22038-22047.

13 McIver LJ, Fondon JW, 3rd, Skinner MA, Garner HR. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. Genomics 2011; 97: 193-199.

14 McIver LJ, McCormick JF, Martin A, Fondon JW, 3rd, Garner HR. Population-scale analysis of human microsatellites reveals novel sources of exonic variation. Gene 2013; 516: 328-334.

15 Vaksman Z, Garner HR. Somatic microsatellite variability as a predictive marker for colorectal cancer and liver cancer progression. Oncotarget 2015; 6: 5760-5771.

16 Bavarva JH, Tae H, McIver L, Garner HR. Nicotine and oxidative stress induced exomic variations are concordant and overrepresented in cancer-associated genes. Oncotarget 2014; 5: 4788-4798.

17 Bavarva JH, Tae H, McIver L, Karunasena E, Garner HR. The dynamic exome: acquired variants as individuals age. Aging 2014; 6: 511-521.

18 Ciriello G, Miller ML, Aksoy BA, Senbabaoglu Y, Schultz N, Sander C. Emerging landscape of oncogenic signatures across human cancers. Nature genetics 2013; 45: 1127-1133.

19 Zack TI, Schumacher SE, Carter SL, Cherniack AD, Saksena G, Tabak B *et al*. Pan-cancer patterns of somatic copy number alteration. Nature genetics 2013; 45: 1134-1140.

20 Forbes SA, Beare D, Gunasekaran P, Leung K, Bindal N, Boutselakis H *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. Nucleic Acids Res 2015; 43: D805-811.

21 Geismann C, Grohmann F, Sebens S, Wirths G, Dreher A, Hasler R *et al*. c-Rel is a critical mediator of NF-kappaB-dependent TRAIL resistance of pancreatic cancer cells. Cell death & disease 2014; 5: e1455.

22 Hunter JE, Leslie J, Perkins ND. c-Rel and its many roles in cancer: an old story with new twists. British journal of cancer 2016; 114: 1-6.

23 Khursheed M, Kolla JN, Kotapalli V, Gupta N, Gowrishankar S, Uppin SG *et al*. ARID1B, a member of the human SWI/SNF chromatin remodeling complex, exhibits tumour-suppressor activities in pancreatic cancer cell lines. British journal of cancer 2013; 108: 2056-2062.

24 Wagner AH, Coffman AC, Ainscough BJ, Spies NC, Skidmore ZL, Campbell KM *et al*. DGIdb 2.0: mining clinically relevant drug-gene interactions. Nucleic Acids Res 2016; 44: D1036-1044.

25 Lian Y, Garner HR. Evidence for the regulation of alternative splicing via complementary DNA sequence repeats. Bioinformatics 2005; 21: 1358-1364.

26 Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R *et al*. The Reactome pathway Knowledgebase. Nucleic Acids Res 2016; 44: D481-487.

27 Korpanty GJ, Graham DM, Vincent MD, Leighl NB. Biomarkers That Currently Affect Clinical Practice in Lung Cancer: EGFR, ALK, MET, ROS-1, and KRAS. Frontiers in oncology 2014; 4: 204.

28 Vargas AJ, Harris CC. Biomarker development in the precision medicine era: lung cancer as a case study. Nature reviews Cancer 2016; 16: 525-537.

29 Freimer NB, Slatkin M. Microsatellites: evolution and mutational processes. Ciba Foundation symposium 1996; 197: 51-67; discussion 67-72.

30 Biegel JA, Busse TM, Weissman BE. SWI/SNF chromatin remodeling complexes and cancer. American journal of medical genetics Part C, Seminars in medical genetics 2014; 166C: 350-366.

31 Huang HT, Chen SM, Pan LB, Yao J, Ma HT. Loss of function of SWI/SNF chromatin remodeling genes leads to genome instability of human lung cancer. Oncology reports 2015; 33: 283-291.

32 Meira LB, Bugni JM, Green SL, Lee CW, Pang B, Borenshtein D *et al*. DNA damage induced by chronic inflammation contributes to colon carcinogenesis in mice. The Journal of clinical investigation 2008; 118: 2516-2525.

33 Hasnis E, Bar-Shai M, Burbea Z, Reznick AZ. Mechanisms underlying cigarette smoke-induced NF-kappaB activation in human lymphocytes: the role of reactive nitrogen species. Journal of physiology and pharmacology : an official journal of the Polish Physiological Society 2007; 58 Suppl 5: 275-287.

34 Weniger MA, Gesk S, Ehrlich S, Martin-Subero JI, Dyer MJ, Siebert R *et al*. Gains of REL in primary mediastinal B-cell lymphoma coincide with nuclear accumulation of REL protein. Genes, chromosomes & cancer 2007; 46: 406-415.

35 Joos S, Menz Ck Fau - Wrobel G, Wrobel G Fau - Siebert R, Siebert R Fau - Gesk S, Gesk S Fau - Ohl S, Ohl S Fau - Mechtersheimer G *et al*. Classical Hodgkin lymphoma is characterized by recurrent copy number gains of the short arm of chromosome 2 2002.

36 Cogswell PC, Guttridge DC, Funkhouser WK, Baldwin AS, Jr. Selective activation of NF-kappa B subunits in human breast cancer: potential roles for NF-kappa B2/p52 and for Bcl-3. Oncogene 2000; 19: 1123-1131.

37 Lu H, Yang X Fau - Duggal P, Duggal P Fau - Allen CT, Allen Ct Fau - Yan B, Yan B Fau - Cohen J, Cohen J Fau - Nottingham L *et al*. TNF-alpha promotes c-REL/DeltaNp63alpha interaction and TAp73 dissociation from key genes that mediate growth arrest and apoptosis in head and neck cancer 2011.

38 Waterborg JH. Steady-state levels of histone acetylation in Saccharomyces cerevisiae. The Journal of biological chemistry 2000; 275: 13007-13011.

39 Celeste A, Petersen S, Romanienko PJ, Fernandez-Capetillo O, Chen HT, Sedelnikova OA *et al*. Genomic instability in mice lacking histone H2AX. Science 2002; 296: 922-927.

40 Geng L, Zhu M, Wang Y, Cheng Y, Liu J, Shen W *et al*. Genetic variants in chromatin-remodeling pathway associated with lung cancer risk in a Chinese population. Gene 2016; 587: 178-182.

41 Kultz D. Molecular and evolutionary basis of the cellular stress response. Annual review of physiology 2005; 67: 225-257.

42 Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics 2005; 21: 3940-3941.

43 Team RC. R: A Language and Environment for Statistical Computing, 2016.

44 Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014; 30: 2114-2120.

45 Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. Journal of clinical epidemiology 2003; 56: 1129-1135.

46 Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC *et al*. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol 2003; 4: P3.

47 Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO *et al*. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013; 6: pl1.

48 Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. Nature 2014; 511: 543-550.

49 Cancer Genome Atlas Research N. Comprehensive genomic characterization of squamous cell lung cancers. Nature 2012; 489: 519-525.

50 Campbell JD, Alexandrov A, Kim J, Wala J, Berger AH, Pedamallu CS *et al*. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. Nature genetics 2016; 48: 607-616.

## SUPPLEMENTARY MATERIAL

**Figure S3-1: Callable loci distribution in the lung cancer and non-tumor samples.** In 93% of non-tumor samples and 100% of the lung cancer samples, 90 to 100% of the 326 MST loci were called. The figure shows that in the 89 + 30 target enriched samples a very high percentage of the loci included in the enrichment kit can be called with confidence.

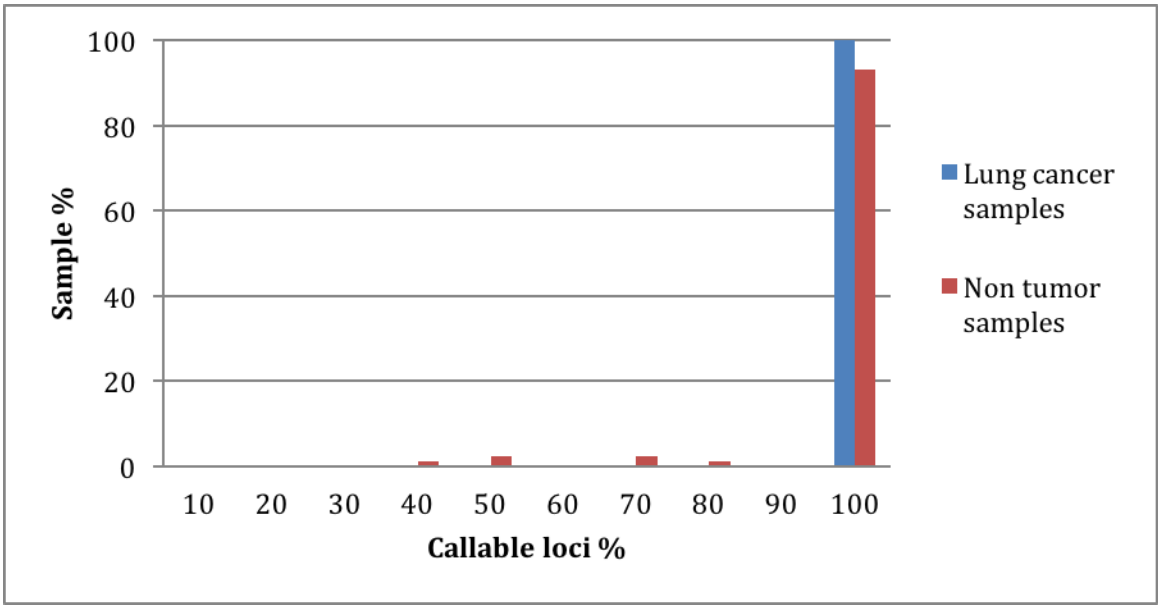


**Figure S3-2: Read depth distribution in the lung cancer and non-tumor validation samples.**

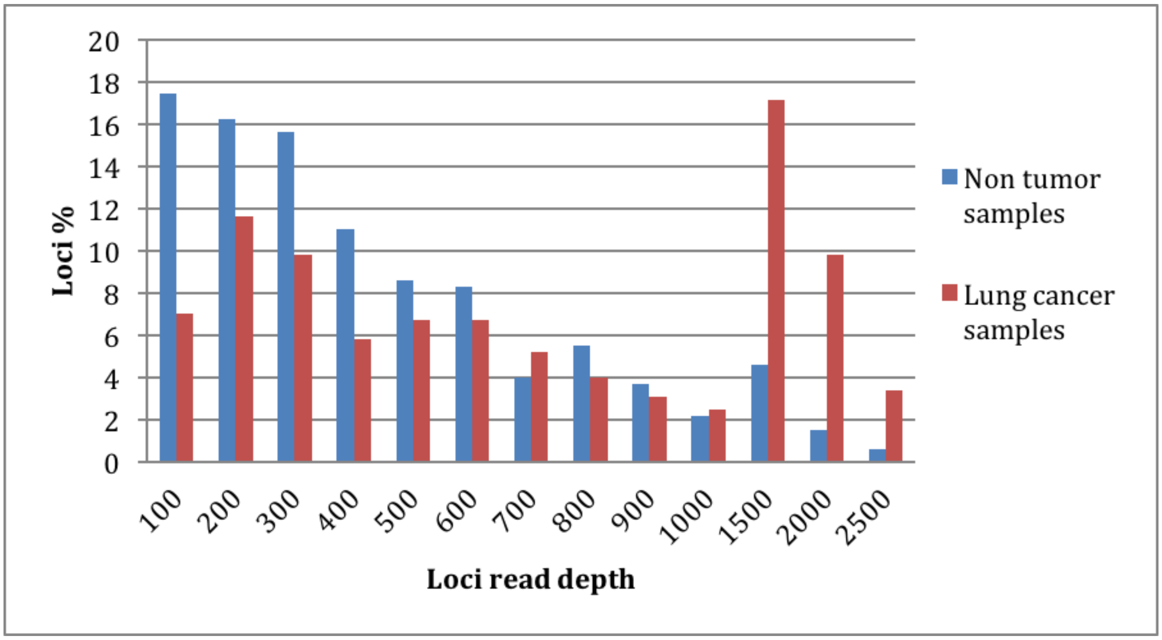


**Figure S3-3: The 96 computationally found LUAD specific MST loci set differentiated LUAD germline samples from non-cancer 1kGP control samples with a sensitivity of 0.87 and a specificity of 0.87.** The AUC value of this prediction was found to be 0.94.

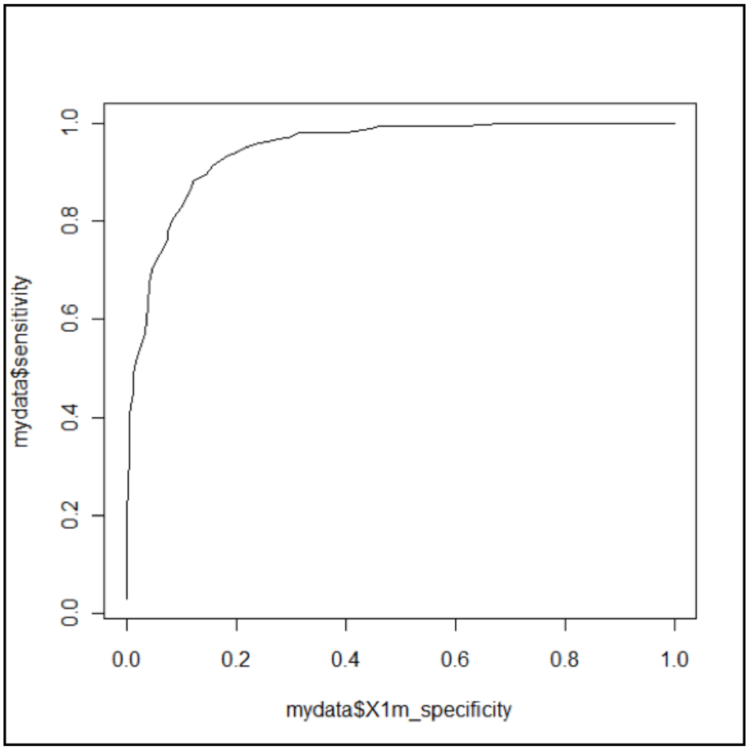


**Figure S3-4: The 67 computationally found LUSC specific MST loci set differentiated LUSC germline samples from non-cancer 1kGP control samples with a sensitivity of 0.88 and a specificity of 0.82.** The AUC value of this prediction was found to be 0.92.

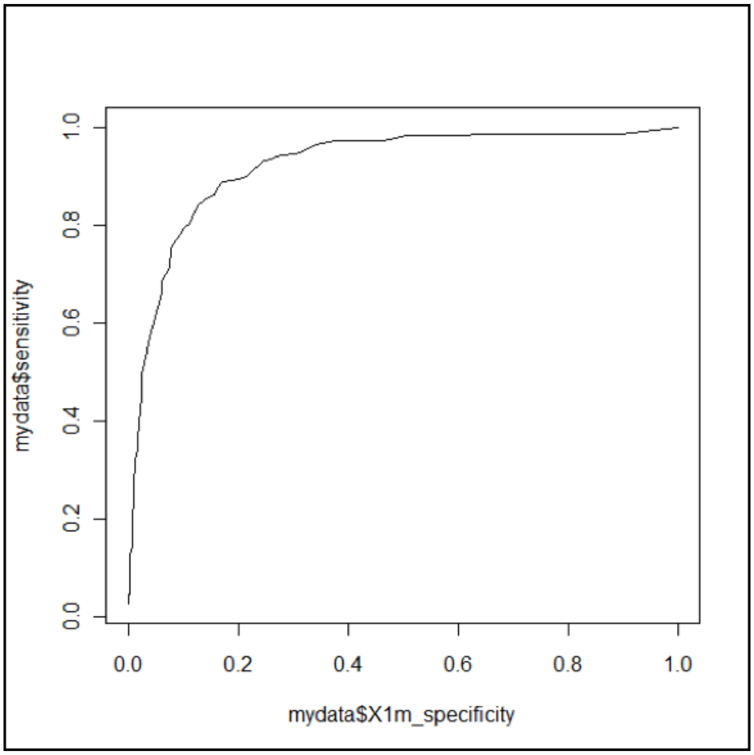


**Figure S3-5: A set of 13 MST loci that were computationally associated with lung cancer was validated by the high-depth enabled high accuracy genotyping to differentiate lung cancer samples from non-cancer 1kGP control samples with a sensitivity and specificity of 0.90 and 0.94, respectively.** (A) The AUC value for the classifier was found to be 0.96. (B) The accuracy versus cutoff plot gives the optimum point where the true positive value is high and the false positive value is low**.** According to the accuracy versus cutoff plot, a sample with 61% or more of the 13 MST loci set with predominant cancer genotype will be classified as 'at-risk' for NSCLC.

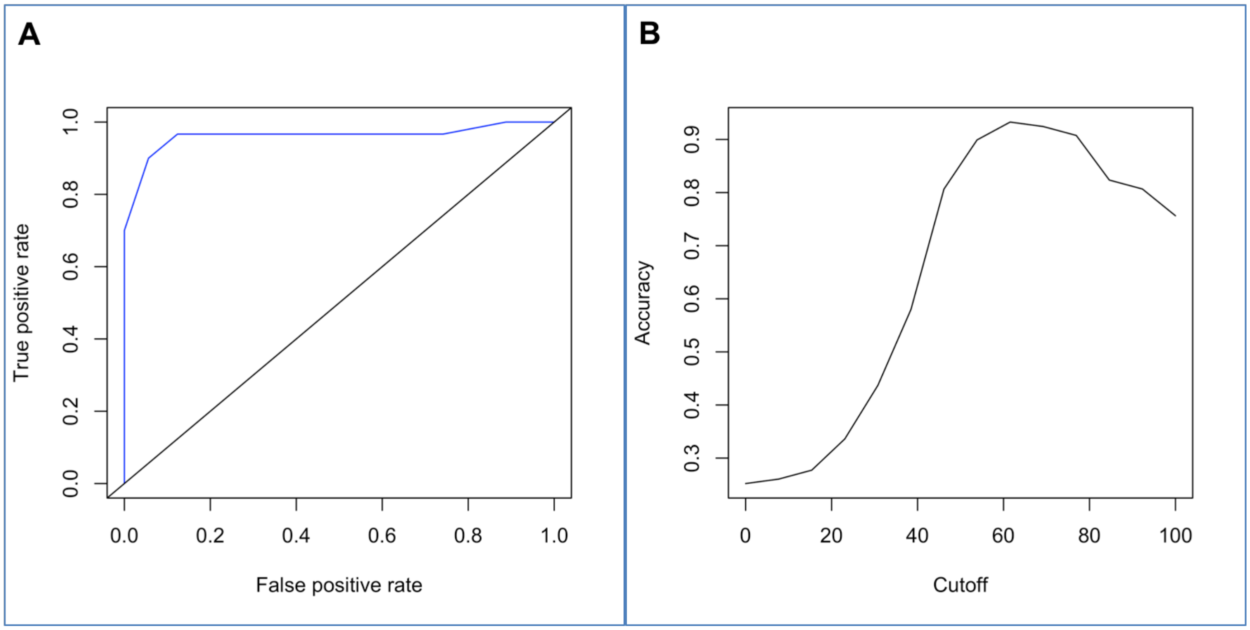


**Figure S3-6: A set of 21 (13 computationally known to be associated with lung cancer and 8 that were found to be associated for other diseases) MST loci was validated by the high-depth enabled high accuracy genotyping to differentiate lung cancer samples from non-cancer 1kGP control samples with a sensitivity and specificity of 0.93 and 0.97, respectively.** (A) The AUC value for the classifier was found to be 0.97. (B) The accuracy versus cutoff plot gives the optimum point where the true positive value is high and the false positive value is low**.** According to the accuracy versus cutoff plot, a sample with 57% or more of the 21 MST loci set with predominant cancer genotype will be classified as 'at-risk' for NSCLC.

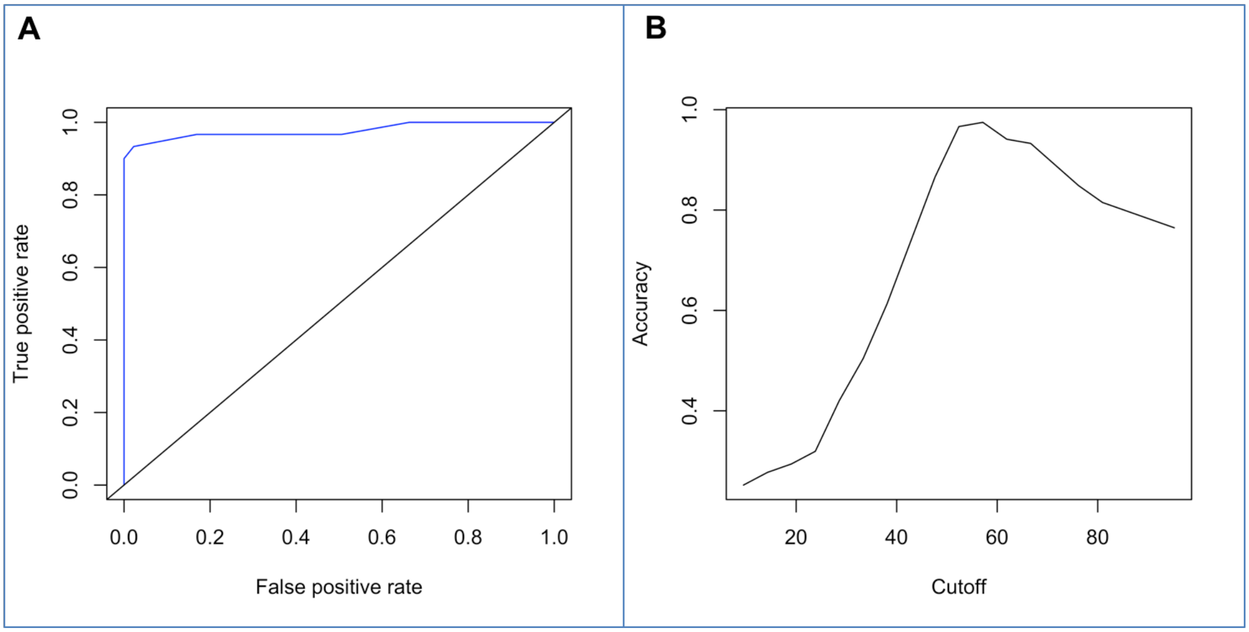


**Table S3-1: The Specific Microsatellite Target Enrichment Kit (SMTEK) consisted of 347 MST loci of which 322 were genotyped in at least 10 tumor and non-cancer control samples. All 322 successfully genotyped loci had at least 10 reads in all the samples they were called.**

| | | After sequencing | |
|---|---|---|---|
| **Loci Type** | **Total before sequencing** | **Passed** | **Failed** |
| Disease | 263 | 242 | 21 |
| Control | 84 | 79 | 5 |
| Total | 347 | 322 | 28 |

**Table S3-2: Genotyping all MST loci in several TCGA germline cancer sample types and comparing them to 1000 Genome Project sample genotypes yield disease specific sets of MSTs that can differentiate cancer and non-cancer control samples.** Note, some loci were found to be informative for several cancer disease types

| **Cancer disease type** | **# MST loci** |
|---|---|
| Lung squamous cell carcinoma | 67 |
| Lung adenocarcinoma | 96 |
| Skin cutaneous melanoma | 68 |
| Breast cancer | 55 |
| Ovarian cancer | 57 |
| Glioblastoma | 48 |
| Medulloblastoma | 25 |
| Lower grade glioma | 38 |

Lung squamous cell carcinoma = LUSC; Lung adenocarcinoma = LUAD; Skin cutaneous melanoma = SKCM; Breast cancer = BC; Ovarian cancer = OV; Glioblastoma = GBM; Medulloblastoma = MB; Lower grade glioma = LGG.

**Table S3-3: Details of the 30 lung cancer samples used to validate the LUSC and LUAD markers via SMTEK nextgen sequencing.**

| Sample # | Tissue | Sample source |
|---|---|---|
| 1 | Lung cancer | Origene 3398 |
| 2 | Lung cancer | Origene 3538 |
| 3 | Lung cancer | Origene 3619 |
| 4 | Lung cancer | Origene 3803 |
| 5 | Lung cancer | Origene 3842 |
| 6 | Lung cancer | Origene 3844 |
| 7 | Lung cancer | Origene 3863 |
| 8 | Lung cancer | Origene 3931 |
| 9 | Lung cancer | Origene 3950 |
| 10 | Lung cancer | Origene 3989 |
| 11 | Lung cancer | Origene 4031 |
| 12 | Lung cancer | Origene 4033 |
| 13 | Lung cancer | Origene 4056 |
| 14 | Lung cancer | Origene 4119 |
| 15 | Lung cancer | Origene 4176 |
| 16 | Lung cancer | Origene CD07 |
| 17 | Lung cancer | Origene CD22 |
| 18 | Lung cancer | Origene CD26 |
| 19 | Lung cancer | Origene CD28 |
| 20 | Lung cancer | Origene CD36 |
| 21 | Lung cancer | Origene CD44 |
| 22 | Lung cancer | Origene CD56 |
| 23 | Lung cancer | Origene CD57 |
| 24 | Lung cancer | Origene CD59 |
| 25 | Lung cancer | Origene CD61 |
| 26 | Lung cancer | Origene CD62 |
| 27 | Lung cancer | Origene CD66 |
| 28 | Lung cancer | Origene CD71 |
| 29 | Lung cancer | Origene CD77 |
| 30 | Lung cancer | Origene CD97 |

**Table S3-4: Details of the 1000 genomes germline samples used as control samples, i.e. were normal healthy individuals.**

| Sample # | Cell type | Sample source |
|---|---|---|
| 1 | B-Lymphocyte | Coriell HG00313 |
| 2 | B-Lymphocyte | Coriell HG00180 |
| 3 | B-Lymphocyte | Coriell HG00181 |
| 4 | B-Lymphocyte | Coriell HG00138 |
| 5 | B-Lymphocyte | Coriell HG00113 |
| 6 | B-Lymphocyte | Coriell HG00100 |
| 7 | B-Lymphocyte | Coriell HG00116 |
| 8 | B-Lymphocyte | Coriell HG00107 |
| 9 | B-Lymphocyte | Coriell HG00278 |
| 10 | B-Lymphocyte | Coriell HG00269 |
| 11 | B-Lymphocyte | Coriell HG00125 |
| 12 | B-Lymphocyte | Coriell HG00182 |
| 13 | B-Lymphocyte | Coriell HG00272 |
| 14 | B-Lymphocyte | Coriell HG00273 |
| 15 | B-Lymphocyte | Coriell HG00121 |
| 16 | B-Lymphocyte | Coriell HG00186 |
| 17 | B-Lymphocyte | Coriell HG00134 |
| 18 | B-Lymphocyte | Coriell HG00131 |
| 19 | B-Lymphocyte | Coriell HG00190 |
| 20 | B-Lymphocyte | Coriell HG00102 |
| 21 | B-Lymphocyte | Coriell HG00127 |
| 22 | B-Lymphocyte | Coriell HG00135 |
| 23 | B-Lymphocyte | Coriell HG00110 |
| 24 | B-Lymphocyte | Coriell HG00274 |
| 25 | B-Lymphocyte | Coriell HG00118 |
| 26 | B-Lymphocyte | Coriell HG00115 |
| 27 | B-Lymphocyte | Coriell HG00187 |
| 28 | B-Lymphocyte | Coriell HG00104 |
| 29 | B-Lymphocyte | Coriell HG00111 |
| 30 | B-Lymphocyte | Coriell HG00122 |
| 31 | B-Lymphocyte | Coriell HG00277 |
| 32 | B-Lymphocyte | Coriell HG00139 |

| | | |
|---|---|---|
| 33 | B-Lymphocyte | Coriell HG00282 |
| 34 | B-Lymphocyte | Coriell HG00183 |
| 35 | B-Lymphocyte | Coriell HG00309 |
| 36 | B-Lymphocyte | Coriell HG00119 |
| 37 | B-Lymphocyte | Coriell HG00268 |
| 38 | B-Lymphocyte | Coriell HG00310 |
| 39 | B-Lymphocyte | Coriell HG00097 |
| 40 | B-Lymphocyte | Coriell HG00312 |
| 41 | B-Lymphocyte | Coriell HG00108 |
| 42 | B-Lymphocyte | Coriell HG00308 |
| 43 | B-Lymphocyte | Coriell HG00178 |
| 44 | B-Lymphocyte | Coriell HG00132 |
| 45 | B-Lymphocyte | Coriell HG00266 |
| 46 | B-Lymphocyte | Coriell HG00129 |
| 47 | B-Lymphocyte | Coriell HG00117 |
| 48 | B-Lymphocyte | Coriell HG00099 |
| 49 | B-Lymphocyte | Coriell HG00136 |
| 50 | B-Lymphocyte | Coriell HG00133 |
| 51 | B-Lymphocyte | Coriell HG00171 |
| 52 | B-Lymphocyte | Coriell HG00188 |
| 53 | B-Lymphocyte | Coriell HG00275 |
| 54 | B-Lymphocyte | Coriell HG00176 |
| 55 | B-Lymphocyte | Coriell HG00306 |
| 56 | B-Lymphocyte | Coriell HG00103 |
| 57 | B-Lymphocyte | Coriell HG00140 |
| 58 | B-Lymphocyte | Coriell HG00098 |
| 59 | B-Lymphocyte | Coriell HG00281 |
| 60 | B-Lymphocyte | Coriell HG00177 |
| 61 | B-Lymphocyte | Coriell HG00109 |
| 62 | B-Lymphocyte | Coriell HG00271 |
| 63 | B-Lymphocyte | Coriell HG00106 |
| 64 | B-Lymphocyte | Coriell HG00105 |
| 65 | B-Lymphocyte | Coriell HG00137 |
| 66 | B-Lymphocyte | Coriell HG00128 |
| 67 | B-Lymphocyte | Coriell HG00124 |
| 68 | B-Lymphocyte | Coriell HG00096 |

|  |  |  |
|---|---|---|
| 69 | B-Lymphocyte | Coriell HG00142 |
| 70 | B-Lymphocyte | Coriell HG00284 |
| 71 | B-Lymphocyte | Coriell HG00120 |
| 72 | B-Lymphocyte | Coriell HG00285 |
| 73 | B-Lymphocyte | Coriell HG00276 |
| 74 | B-Lymphocyte | Coriell HG00123 |
| 75 | B-Lymphocyte | Coriell HG00173 |
| 76 | B-Lymphocyte | Coriell HG00280 |
| 77 | B-Lymphocyte | Coriell HG00112 |
| 78 | B-Lymphocyte | Coriell HG00174 |
| 79 | B-Lymphocyte | Coriell HG00101 |
| 80 | B-Lymphocyte | Coriell HG00311 |
| 81 | B-Lymphocyte | Coriell HG00179 |
| 82 | B-Lymphocyte | Coriell HG00114 |
| 83 | B-Lymphocyte | Coriell HG00267 |
| 84 | B-Lymphocyte | Coriell HG00130 |
| 85 | B-Lymphocyte | Coriell HG00126 |
| 86 | B-Lymphocyte | Coriell HG00189 |
| 87 | B-Lymphocyte | Coriell HG00141 |
| 88 | B-Lymphocyte | Coriell HG00270 |
| 89 | B-Lymphocyte | Coriell HG00185 |

**Table S3-5: A set of 13 lung cancer specific loci and 8 other loci that were found to be specific for other diseases were found to have differing genotypes in both the sample groups.** All the 21 loci have predominant genotypes (larger than the sum of all the other genotypes, i.e. more than 50%) in both the groups. The genomic coordinates furnished correspond to the HG38 genome reference build.

| Genomic position | Modal control GT | Modal control GT sample % | Predominant cancer GT | Predominant cancer GT sample % | Odds ratio |
|---|---|---|---|---|---|
| chr2:60918364-60918376 | 13_13 | 49/89 (55%) | 13_12 | 29/30 (97%) | 39.92 |
| chr6:157174818-157174831 | 14_14 | 51/89 (57%) | 14_13 | 27/30 (90%) | 13.57 |
| chr6:76018867-76018880 | 14_14 | 60/85 (71%) | 14_13 | 24/30 (80%) | 12.28 |
| chr3:94035443-94035458 | 16_16 | 46/88 (52%) | 16_15 | 27/30 (90%) | 11.20 |
| chr3:112534347-112534360 | 15_15 | 46/83 (55%) | 15_14 | 26/30 (87%) | 10.29 |
| chr8:129862369-129862381 | 13_13 | 57/86 (66%) | 13_12 | 22/30 (73%) | 7.01 |
| chr9:130622843-130622857 | 15_15 | 55/84 (65%) | 15_14 | 22/30 (73%) | 6.93 |
| chr7:135414296-135414309 | 14_14 | 47/87 (54%) | 14_13 | 23/30 (77%) | 5.07 |
| chr2:48461120-48461133 | 14_14 | 69/89 (78%) | 14_13 | 16/30 (53%) | 4.43 |
| chr2:55332516-55332530 | 15_15 | 53/85 (62%) | 15_14 | 20/30 (67%) | 3.90 |

| | | | | | |
|---|---|---|---|---|---|
| chr13:31148484-31148500 | 17_16 | 43/83 (52%) | 16_15 | 18/30 (60%) | 3.70 |
| chr15:20458509-20458521 | 13_12 | 48/89 (54%) | 12_11 | 20/30 (67%) | 3.00 |
| chr10:13591929-13591943 | 15_15 | 56/89 (63%) | 15_14 | 16/30 (53%) | 2.25 |
| chr2:202815832-202815844 | 13_13 | 69/85 (81%) | 13_12 | 18/30 (60%) | 7.61 |
| chr13:114236623-114236635 | 13_13 | 51/89 (57%) | 13_12 | 24/30 (80%) | 6.19 |
| chr12:106106383-106106396 | 14_14 | 40/61 (66%) | 14_13 | 14/22 (64%) | 5.74 |
| chr3:98580864-98580876 | 13_13 | 50/83 (60%) | 13_12 | 22/30 (73%) | 5.02 |
| chr16:70839964-70839978 | 15_12 | 59/89 (66%) | 12_12 | 19/30 (63%) | 4.58 |
| chr2:233460070-233460083 | 14_14 | 44/88 (50%) | 14_13 | 23/30 (77%) | 3.82 |
| chr5:87383860-87383873 | 14_14 | 45/83 (54%) | 14_13 | 20/30 (67%) | 2.93 |
| chr8:23852057-23852082 | 26_26 | 44/84 (52%) | 26_24 | 16/30 (53%) | 1.87 |

**Table S3-6: Out of 119 MST loci computationally found to be specific for LUAD and/or LUSC cancer types, 105 produced data in both sample groups.** All loci produced at least 10 reads per loci in at least 10 samples in both sample groups. The genomic coordinates furnished attribute to the HG38 build.

| Genomic position | Repeat | Gene region | Gene | Disease |
|---|---|---|---|---|
| chr6:36484827-36484842 | A | intron | KCTD20 | BC, GBM, LGG, LUAD, LUSC, SKCM |
| chr13:44943348-44943377 | AC | intron | NUFIP1 | BC, GBM, LUAD |
| chrX:13757634-13757649 | T | intron | OFD1 | BC, GBM, LUAD |
| chr1:10297149-10297165 | T | intron | KIF1B | BC, LGG, LUAD, SKCM |
| chr17:16070104-16070120 | T | intron | NCOR1 | BC, LGG, LUSC, SKCM |
| chr14:50881564-50881580 | T | intron | ABHD12B | BC, LUSC |
| chr3:112534347-112534360 | A | intron | ATG3 | GBM, LGG, LUAD, LUSC |
| chr9:130622843-130622857 | A | intron | FUBP3 | GBM, LGG, LUAD, LUSC, MB, OV |
| chr17:42834438-42834469 | GA | intron | PSME3 | GBM, LGG, LUAD, LUSC, MB, SKCM |
| chr13:113310584-113310595 | T | intron | LAMP1 | GBM, LGG, LUAD, LUSC, MB, SKCM |
| chr4:5745180-5745201 | TTC | intron | EVC | GBM, LGG, LUAD, LUSC, MB, SKCM |

| | | | | |
|---|---|---|---|---|
| chrX:132097403-132097440 | AC | intron | FRMD7 | GBM, LUAD |
| chr10:33182834-33182862 | CA | intron | NRP1 | GBM, LUAD |
| chr6:31864580-31864594 | A | intron | SLC44A4 | GBM, LUAD, LUSC |
| chr2:86894983-86894997 | T | intergenic | - | GBM, LUAD, LUSC, MB |
| chr1:16564320-16564331 | A | intron | NBPF1 | GBM, LUAD, LUSC, MB |
| chr15:84512873-84512887 | A | 3utr | FLJ40113 | GBM, LUAD, MB |
| chr1:111762785-111762800 | A | intron | DDX20 | GBM, LUAD, SKCM |
| chr3:94035443-94035458 | T | intron | ARL13B | GBM, LUAD, SKCM |
| chr3:196361939-196361954 | A | intron | UBXN7 | LGG, LUAD |
| chr4:127699990-127700002 | T | intron | INTU | LGG, LUAD |
| chr3:50118451-50118476 | GA | exon | RBM5 | LGG, LUAD, LUSC |
| chr2:110963566-110963604 | TG | intron | ACOXL | LGG, LUAD, LUSC, MB, SKCM |
| chr13:27559820-27559834 | A | intron | LNX2 | LGG, LUAD, LUSC, MB, SKCM |
| chr7:65961068-65961081 | A | intron | GUSB | LGG, LUAD, LUSC, MB, SKCM |

| | | | | |
|---|---|---|---|---|
| chr2:48461120-48461133 | T | intron | KLRAQ1 | LGG, LUAD, LUSC, MB, SKCM |
| chr16:66912992-66913023 | GT | intron | CDH16 | LGG, LUAD, LUSC, SKCM |
| chr12:95094564-95094577 | A | intron | FGD6 | LGG, LUAD, LUSC, SKCM |
| chr4:112186674-112186688 | T | intron | C4orf32 | LGG, LUAD, LUSC, SKCM |
| chr13:77217965-77217977 | A | intron | MYCBP2 | LGG, LUAD, LUSC, SKCM |
| chr15:73126401-73126414 | T | intron | NEO1 | LGG, LUAD, LUSC, SKCM |
| chr5:137677662-137677675 | A | intron | KLHL3 | LGG, LUAD, LUSC, SKCM |
| chr15:43710473-43710501 | TG | intergenic | - | LGG, LUAD, LUSC, SKCM |
| chr5:72889765-72889779 | T | intron | TNPO1 | LGG, LUAD, LUSC, SKCM |
| chr4:22442629-22442643 | A | intron | GPR125 | LGG, LUAD, OV, SKCM |
| chr9:115402097-115402108 | T | intron | DEC1' | LGG, LUAD, SKCM |
| chr21:43068646-43068659 | A | intron | CBS | LGG, LUSC |
| chr8:7359489-7359500 | T | intergenic | - | LUAD |
| chr3:46709584-46709610 | AAG | exon | TMIE | LUAD |
| chr16:19608281-19608294 | T | intron | C16orf62 | LUAD |

|  |  |  |  |  |
|---|---|---|---|---|
| chr12:80845943-80845957 | A | intron | LIN7A | LUAD |
| chr12:118383996-118384007 | T | intron | SUDS3 | LUAD |
| chrX:133216975-133216989 | A | exon | TFDP3 | LUAD |
| chr4:84635240-84635255 | T | intron | CDS1 | LUAD |
| chr14:24102652-24102664 | A | intron | PCK2 | LUAD |
| chr14:58208106-58208119 | A | intron | ACTR10 | LUAD |
| chr14:62950568-62950583 | A | intron | KCNH5 | LUAD |
| chr7:135414296-135414309 | A | intron | CNOT4 | LUAD |
| chr6:13316549-13316562 | A | intron | TBC1D7 | LUAD |
| chr6:95586994-95587006 | TA | intron | MANEA | LUAD |
| chr15:43635426-43635437 | A | intron | CATSPER2 | LUAD |
| chr15:20458509-20458521 | A | intergenic | - | LUAD |
| chr1:186361564-186361576 | A | intron | TPR | LUAD |
| chr1:1988996-1989009 | A | intron | KIAA1751 | LUAD |
| chr18:74129655-74129666 | A | intron | FBXO15 | LUAD |

|  |  |  |  |  |
|---|---|---|---|---|
| chr5:172994758-172994772 | T | intron | ATP6V0E1 | LUAD |
| chr5:157098910-157098931 | AG | intron | HAVCR2 | LUAD |
| chr6:7595009-7595021 | T | intron | SNRNP48 | LUAD |
| chr2:190673068-190673082 | T | intron | NAB1 | LUAD |
| chr3:108505839-108505853 | A | intron | MYH15 | LUAD |
| chr3:132502233-132502248 | T | intron | DNAJC13 | LUAD |
| chr11:89914202-89914215 | A | intron | LOC729384 | LUAD |
| chr16:74893610-74893622 | A | intron | WDR59 | LUAD |
| chr2:60918364-60918376 | T | intron | REL | LUAD, LUSC |
| chr6:157901855-157901868 | T | intron | SNX9 | LUAD, LUSC |
| chr16:12051827-12051841 | T | upstream | SNX29 | LUAD, LUSC |
| chr12:96912826-96912839 | T | 5utr | NEDD1 | LUAD, LUSC |
| chr14:81108485-81108507 | T | intron | TSHR | LUAD, LUSC |
| chr4:109347460-109347474 | T | intergenic | - | LUAD, LUSC |
| chr4:44689779-44689792 | A | intron | GUF1 | LUAD, LUSC |

| | | | | |
|---|---|---|---|---|
| chr10:104038216-104038281 | AC | intron | COL17A1 | LUAD, LUSC |
| chr20:33367014-33367027 | A | intron | CDK5RAP1 | LUAD, LUSC |
| chr8:129862369-129862381 | A | intron | FAM49B | LUAD, LUSC |
| chr8:109523216-109523230 | T | intron | PKHD1L1 | LUAD, LUSC |
| chr11:108188043-108188057 | T | intron | NPAT | LUAD, LUSC |
| chr11:124754810-124754824 | A | intron | ESAM | LUAD, LUSC |
| chr3:161238095-161238109 | T | intron | NMD3 | LUAD, LUSC, MB, SKCM |
| chr6:157174818-157174831 | T | intron | ARID1B | LUAD, LUSC, MB, SKCM |
| chr11:108271229-108271243 | T | intron | ATM | LUAD, LUSC, MB, SKCM |
| chr4:185267220-185267233 | A | intron | SNX25 | LUAD, LUSC, OV |
| chrX:52753170-52753183 | T | intron | SSX2 | LUAD, LUSC, SKCM |
| chr6:136389532-136389546 | A | intron | MAP7 | LUAD, LUSC, SKCM |
| chr22:35323386-35323400 | A | intron | TOM1 | LUAD, LUSC, SKCM |
| chr2:55332516-55332530 | A | intron | CCDC88A | LUAD, LUSC, SKCM |

|  |  |  |  |  |
|---|---|---|---|---|
| chr1:62578975-62578989 | A | intron | DOCK7 | LUAD, LUSC, SKCM |
| chr4:73144839-73144853 | A | intron | ANKRD17 | LUAD, MB |
| chrX:52895580-52895606 (HG19) | GT | intron | XAGE3 | LUAD, MB |
| chr16:10689232-10689244 | A | intron | TEKT5 | LUAD, OV |
| chr2:24327999-24328012 | A | intron | ITSN2 | LUAD, SKCM |
| chr13:31148484-31148500 | A | intron | HSPH1 | LUAD, SKCM |
| chr7:5199689-5199704 | A | intron | WIPI2 | LUAD, SKCM |
| chr1:172608725-172608738 | T | intron | C1orf9 | LUAD, SKCM |
| chr1:100077147-100077161 | T | intron | HIAT1 | LUAD, SKCM |
| chr4:165467674-165467685 | T | intron | CPE | LUAD, SKCM |
| chrX:48354751-48354764 | A | intron | SSX3 | LUAD, SKCM |
| chr12:21638477-21638491 | A | intron | LDHB | LUSC |
| chr10:13591929-13591943 | T | intron | PRPF18 | LUSC |
| chr3:186804505-186804518 | A | intron | RFC4 | LUSC |
| chr19:6833152-6833167 | T | intron | VAV1 | LUSC |

| | | | | |
|---|---|---|---|---|
| chr1:153645035-153645049 | T | intron | C1orf77 | LUSC, MB, SKCM |
| chr6:76018867-76018880 | A | intron | IMPG1 | LUSC, OV |
| chr4:145110085-145110100 | T | intron | ABCE1 | LUSC, SKCM |
| chr1:94498584-94498598 | T | intron | ABCD3 | LUSC, SKCM |
| chr1:52481679-52481693 | A | intron | ZCCHC11 | LUSC, SKCM |
| chr12:7368968-7368991 | GA | intron | CD163L1 | LUSC, SKCM |

Genomic coordinates that have ‘HG19’ mentioned on the side were partially deleted in HG38 hence the coordinates from the previous build are given.

**Table S3-7: Out of 144 MST loci computationally found to be specific for other cancer types, 137 produced data in both sample groups.** All loci produced at least 10 reads per loci in at least 10 samples in both sample groups. The genomic coordinates furnished correspond to the HG38 genome reference build.

| Genomic position | Repeat | Gene region | Gene | Disease |
|---|---|---|---|---|
| chr8:39749565-39749600 | GT | intron | ADAM2 | BC |
| chr8:23852057-23852082 | TG | intron | STC1 | BC |
| chr2:202765380-202765400 | T | intron | FAM117B | BC |
| chr3:155116591-155116607 | TA | intron | MME | BC |
| chr3:113360927-113360938 | A | intron | WDR52 | BC |
| chr16:20944777-20944802 | AC | intron | DNAH3 | BC |
| chr12:110396226-110396243 | A | intron | ANAPC7 | BC |
| chr16:56684104-56684123 | T | exon | MT1X | BC |
| chr4:76144324-76144338 | A | intron | NUP54 | BC |
| chr14:102083733-102083750 | A | intron | HSP90AA1 | BC |
| chr17:59586236-59586253 | A | intron | DHX40 | BC |
| chr20:20038239-20038260 | A | intron | CRNKL1 | BC |
| chr7:148797703-148797719 | T | intron | CUL1 | BC |
| chrX:10141619-10141634 | A | exon | WWC3 | BC |
| chr1:113829711-113829722 | A | intron | PTPN22 | BC |
| chr10:45073089-45073105 | T | intergenic | - | BC |
| chr15:81345017-81345037 | GA | intron | TMC3 | BC |
| chr4:54264835-54264851 | A | intron | PDGFRA | BC |
| chr22:37912036-37912064 | TG | intron | MICALL1 | BC |
| chr18:46812342-46812357 | A | exon | PIAS2 | BC |
| chr5:87383679-87383696 | A | intron | RASA1 | BC |
| chr3:33836009-33836020 | T | intron | PDCD6IP | BC |
| chr2:197469873-197469884 | A | intron | COQ10B | BC |

| | | | | |
|---|---|---|---|---|
| chr2:75692147-75692171 | AT | intron | C2orf3 | BC |
| chr3:198153260-198153301 | GCA | exon | FAM157A | BC |
| chr3:196257948-196257959 | A | intron | PCYT1A | BC |
| chr11:118482323-118482338 | T | intron | MLL | BC |
| chr15:83804574-83804590 | T | intron | ADAMTSL3 | BC |
| chr1:23082431-23082446 | T | intron | AOF2 | BC |
| chrX:71592599-71592613 | T | intron | ACRC | BC, GBM |
| chr4:47744586-47744598 | A | intron | CORIN | BC, GBM, MB |
| chr8:106692713-106692726 | A | intron | OXR1 | BC, GBM, OV |
| chr2:202815832-202815844 | A | intron | ICA1L | BC, GBM, OV |
| chr7:38242530-38242549 | GT | intron | TRG | BC, MB |
| chr9:5798652-5798666 | A | intron | ERMP1 | BC, OV |
| chr17:65750900-65750913 | A | intron | CCDC46 | BC, OV |
| chr20:5186510-5186522 | T | intron | CDS2 | BC, OV |
| chr7:123117666-123117678 | A | intron | SLC13A1 | BC, OV |
| chr11:110258201-110258215 | A | intron | RDX | BC, OV |
| chr6:170572302-170572314 | T | exon | TBP | BC, OV |
| chr8:31076301-31076312 | T | intron | WRN | BC, OV |
| chr3:98580864-98580876 | A | intron | CPOX | BC, OV |
| chr11:62798437-62798472 | AAAAGA | intron | NXF1 | BC, OV |
| chr15:89268652-89268664 | T | intron | FANCI | BC, OV |
| chr5:134608354-134608369 | T | intron | SAR1B | BC, OV |
| chr19:29615224-29615240 | T | intron | POP4 | BC, OV |
| chr15:62748318-62748333 | A | intron | TLN2 | BC, SKCM |
| chr6:70240579-70240595 | AT | intron | COL9A1 | GBM |
| chr3:171126228-171126241 | A | intron | TNIK | GBM |
| chr16:70839964-70839978 | T | intron | HYDIN | GBM |
| chr4:168275913-168275928 | A | intron | DDX60 | GBM |
| chr14:95099732-95099772 | AC | intron | DICER1 | GBM |

| | | | | |
|---|---|---|---|---|
| chr17:56904211-56904226 | A | intron | TRIM25 | GBM |
| chr4:188142208-188142243 | GT | intron | TRIML1 | GBM |
| chr7:103185541-103185553 | A | 3utr | DPY19L2P2 | GBM |
| chr7:73307734-73307743 | CAA | exon | NSUN5 | GBM |
| chr21:10516457-10516469 | A | intergenic | - | GBM |
| chr7:83392484-83392501 | A | intron | SEMA3E | GBM |
| chr10:87057822-87057837 | A | intron | GLUD1 | GBM |
| chr15:43618669-43618701 | CAG | exon | STRC | GBM |
| chr14:35865700-35865714 | T | intron | BRMS1L | GBM |
| chr10:121496816-121496831 | T | intron | FGFR2 | GBM |
| chr3:121483587-121483611 | A | intron | POLQ | GBM |
| chr2:138550814-138550849 | TC | intron | SPOPL | GBM |
| chr3:113000945-113000960 | A | exon | GTPBP8 | GBM |
| chr3:154284569-154284580 | T | intron | DHX36 | GBM |
| chr11:119274082-119274098 | T | intron | CBL | GBM |
| chr1:225519570-225519585 | A | intron | ENAH | GBM |
| chr1:117062509-117062522 | T | intron | TTF2 | GBM |
| chr12:33426063-33426109 | CA | intron | SYT10 | GBM |
| chr2:91698005-91698016 | A | intergenic | - | GBM, OV |
| chr9:52626-52640 | A | intergenic | - | GBM, OV, SKCM |
| chr14:50595519-50595543 | TC | intron | ATL1 | LGG |
| chr14:21468604-21468616 | A | intron | RAB2B | LGG |
| chr17:15613747-15613758 | A | intron | CDRT1 | LGG |
| chr7:96146537-96146550 | A | intron | SLC25A13 | LGG |
| chrX:18164978-18164992 | A | exon | BEND2 | LGG |
| chr1:145456733-145456746 (HG19) | A | intron | POLR3GL | LGG |
| chr3:132447305-132447317 | T | intron | DNAJC13 | LGG |

| | | | | |
|---|---|---|---|---|
| chr10:100505295-100505339 | CA | intron | SEC31B | LGG, MB |
| chr19:21375214-21375230 | TG | intergenic | - | LGG, OV |
| chr16:70142419-70142432 | T | intron | PDPR | LGG, SKCM |
| chr12:50660091-50660105 | T | intron | DIP2B | LGG, SKCM |
| chr13:114236623-114236635 | T | intron | CDC16 | LGG, SKCM |
| chr12:129081740-129081756 | T | intron | TMEM132D | MB |
| chr5:36629700-36629712 | A | intron | SLC1A3 | MB |
| chr11:17089651-17089679 | GT | exon | PIK3C2A | MB |
| chr10:119037061-119037072 | C | intron | EIF3A | MB, SKCM |
| chr2:233460070-233460083 | A | intron | DGKD | OV |
| chr6:49848161-49848174 | T | intron | CRISP1 | OV |
| chr3:50057664-50057685 | T | intron | RBM6 | OV |
| chr17:68045756-68045769 | T | intron | KPNA2 | OV |
| chr17:49821919-49821932 | A | intron | MYST2 | OV |
| chr19:20646413-20646427 | AC | intron | ZNF626 | OV |
| chr14:91462502-91462516 | T | intron | SMEK1 | OV |
| chr12:106106383-106106396 | A | intron | NUAK1 | OV |
| chr13:49376888-49376921 | ATAG | intron | CAB39L | OV |
| chr4:71022616-71022630 | T | intron | DCK | OV |
| chr7:31092622-31092634 | T | intron | ADCYAP1R1 | OV |
| chr7:82066527-82066542 | A | intron | CACNA2D1 | OV |
| chr7:36425998-36426012 | T | intron | ANLN | OV |
| chrX:11169774-11169785 | T | intron | ARHGAP6 | OV |
| chr10:67939722-67939740 | AT | intron | HERC4 | OV |
| chr10:92506574-92506588 | T | intron | IDE | OV |
| chr10:22226073-22226095 | A | intergenic | - | OV |
| chr15:64680562-64680589 | TG | intron | ZNF609 | OV |
| chr1:236558153-236558165 | A | intron | HEATR1 | OV |
| chr1:149929094-149929109 | A | intron | MTMR11 | OV |

|  |  |  |  |  |
|---|---|---|---|---|
| chr10:91819355-91819375 | T | intron | TNKS2 | OV |
| chr18:23540418-23540433 | A | intron | NPC1 | OV |
| chr8:120506629-120506642 | T | intron | MTBP | OV |
| chr2:222474811-222474831 | T | intron | SGPP2 | OV |
| chr11:89800992-89801004 | A | intron | TRIM49 | OV |
| chr11:30417412-30417426 | T | intron | MPPED2 | OV |
| chr1:169586130-169586142 | A | intron | F5 | OV |
| chr5:87383860-87383873 | T | intron | RASA1 | OV |
| chr5:159084572-159084586 | A | intron | EBF1 | OV |
| chr5:123378441-123378458 | A | intron | CEP120 | OV |
| chr18:2960515-2960527 | A | intron | LPIN2 | OV |
| chr12:75508182-75508196 | A | intron | KRR1 | OV |
| chr4:140527442-140527455 | T | intron | ELMOD2 | OV, SKCM |
| chr6:88929270-88929284 | A | intron | RNGTT | OV, SKCM |
| chr8:98042609-98042620 | A | intron | RPL30 | SKCM |
| chr6:125928610-125928624 | T | intron | NCOA7 | SKCM |
| chr16:3508263-3508275 | T | intron | CLUAP1 | SKCM |
| chr20:59922427-59922441 | A | exon | SYCP2 | SKCM |
| chr7:138749322-138749335 | A | intron | ATP6V0A4 | SKCM |
| chr6:100540391-100540403 | A | intron | ASCC3 | SKCM |
| chr15:32424717-32424732 | A | 3utr | FAM7A1 | SKCM |
| chr11:115209592-115209633 | TGG | exon | CADM1 | SKCM |
| chr18:46844689-46844703 | T | intron | PIAS2 | SKCM |
| chr3:180961448-180961461 | T | intron | FXR1 | SKCM |
| chr1:243572909-243572923 | T | intron | AKT3 | SKCM |
| chr5:138177681-138177692 | A | intron | BRD8 | SKCM |
| chr19:21167718-21167729 | A | intron | ZNF431 | SKCM |
| chr19:4947051-4947063 | T | intron | UHRF1 | SKCM |
| chr12:95992869-95992882 | CCCT | intron | HAL | SKCM |

**Table 3S-8: Out of 84 MST loci computationally found to be specific for other cancer types, 79 produced data in both sample groups.** All loci produced at least 10 reads per loci in at least 10 samples in both sample groups. The genomic coordinates furnished attribute to the HG38 build.

| Genomic posiiton | Repeat | Gene region | Gene/ Repeat ID |
|---|---|---|---|
| chr8:19957980-19958041 | AAAT | intron | LPL |
| chr8:25503396-25503417 | GAAAG | exon | CDCA2 |
| chr1:209432292-209432332 | AGC | exon | LOC642587 |
| chr2:48515217-48515259 | TC | exon | KLRAQ1 |
| chr2:1489620-1489686 | AATG | intron | TPOX |
| chr6:110895439-110895479 | GTTTT | exon | AMD1 |
| chr7:149239672-149239691 | CGG | exon | ZNF212 |
| chr6:109632842-109632919 | ATAG | exon | AKD1 |
| chr3:49115555-49115580 | TCTTCC | exon | USP19 |
| chr3:18349641-18349687 | CTG | exon | SATB1 |
| chr16:69693627-69693682 | CAG | exon | NFAT5 |
| chr12:49033867-49033913 | TGC | exon | MLL2 |
| chr12:114355457-114355501 | TC | exon | TBX5 |
| chr9:127506124-127506152 | CTCA | exon | FAM129B |
| chr9:132896597-132896621 | GCT | exon | TSC1 |
| chr9:71768574-71768616 | CCTCCG | exon | TMEM2 |
| chr16:67980293-67980324 | GCA | exon | DPEP3 |
| chr17:58756094-58756155 | CCGAAC | exon | PPM1E |
| chr17:12992120-12992153 | TGAT | exon | ELAC2 |

| | | | |
|---|---|---|---|
| chr17:50835991-50836015 | GCT | exon | WFIKKN2 |
| chr17:17136248-17136283 | CAG | exon | MPRIP |
| chr4:90127564-90127590 | CT | exon | FAM190A |
| chr4:139889485-139889540 | GCT | exon | MAML3 |
| chr4:68349770-68349795 | CCG | exon | YTHDC1 |
| chr19:47742125-47742165 | TTCC | exon | EHD2 |
| chr14:94842043-94842108 | [CTGT]n[CTAT]n | | D14S1434 |
| chr14:61321907-61321932 | GGGA | exon | PRKCH |
| chr14:103385444-103385464 | GCG | exon | MARK3 |
| chr14:22883256-22883286 | GCACAC | exon | REM2 |
| chr12:76031129-76031189 | GCT | exon | PHLDA1 |
| chr12:4914286-4914330 | ACAA | exon | KCNA1 |
| chr12:5983962-5984074 | AGAT | intron | VWA |
| chr17:42699813-42699840 | GCTGT | exon | CNTNAP1 |
| chr17:81996788-81996845 | AGCAGG | exon | ASPSCR1 |
| chr17:52158102-52158125 | GCG | exon | CA10 |
| chr4:54740022-54740046 | AAAAC | exon | KIT |
| chr4:93829017-93829037 | CCG | exon | ATOH1 |
| chr20:49905305-49905338 | AAAAC | exon | SPATA2 |
| chrX:53077999-53078047 | TC | exon | GPR173 |
| chrX:134481487-134481569 | AGAT | intron | HPRTB |
| chr1:965304-965347 | TTTA | exon | KLHL17 |
| chr1:15746939-15746954 | TGC | exon | TMEM82 |
| chr6:6321353-6321413 | AAAG | | |
| chr6:45422678-45422748 | GCA | exon | RUNX2 |
| chr10:129294237-129294302 | GGAA | | D10S1248 |
| chr10:75028875-75028954 | GAGGAA | exon | MYST4 |

| | | | |
|---|---|---|---|
| chr15:40036336-40036424 | CTG | exon | SRP14 |
| chr1:247876957-247876991 | CT | exon | TRIM58 |
| chr1:2556549-2556574 | TTCTCT | exon | TNFRSF14 |
| chr1:26429657-26429681 | CTTCC | exon | LIN28 |
| chrY:18639710-18639764 | GAAA | | DYS385_a/b |
| chr11:3855884-3855906 | TCTCT | exon | STIM1 |
| chr3:45540732-45540809 | AGAT | | D3S1358 |
| chr10:117548492-117548574 | GA | exon | EMX2 |
| chr10:49016119-49016146 | TTTG | exon | C10orf72 |
| chr3:126988622-126988658 | TGC | exon | PLXNA1 |
| chr18:63281662-63281742 | GAAA | | D18S51 |
| chr22:37140281-37140335 | ATT | | D22S1045 |
| chr18:69867388-69867414 | TCTCT | exon | CD226 |
| chr8:144530000-144530021 | CAGGA | exon | KIAA1688 |
| chr5:113017812-113017852 | ATATCT | exon | DCP2 |
| chr5:88684765-88684804 | TC | exon | LOC645323 |
| chr5:153490748-153490778 | AAGG | exon | GRIA1 |
| chr11:2171078-2171122 | AATG | intron | TH01 |
| chr16:67842859-67842951 | GCA | exon | THAP11 |
| chr1:226993391-226993405 | TGC | exon | CDC42BPA |
| chr5:150076298-150076384 | AGAT | intron | CSF1PO |
| chr5:146233608-146233675 | CCAGGC | exon | RBM27 |
| chr5:148982979-148983006 | TGACAT | exon | SH3TC2 |
| chr5:150651587-150651611 | CTGGG | exon | SYNPO |
| chr19:11164698-11164760 | CCAT | exon | KANK2 |
| chr19:11466768-11466792 | GGGCC | exon | ELAVL3 |
| chr19:29926209-29926306 | (AAGG)(AAAG)(AAGG)(TAGG)[AAGG]n | | D19S433 |
| chr19:2514952-2515006 | GA | exon | GNG7 |

|  |  |  |  |
|---|---|---|---|
| chr18:37243909-37243929 | CGG | exon | BRUNOL4 |
| chr11:6390698-6390746 | GGCGCT | exon | SMPD1 |
| chr12:12297001-12297100 | [AGAT]8-17[AGAC]6-10[AGAT]0-1 |  | D12S391 |
| chr12:114406047-114406082 | TC | exon | TBX5 |
| chr12:12643661-12643693 | TCTAG | exon | CREBL2 |

**Table S3-9: The leave one out cross validation confirms that the predictive power of the model sustains when applied to a test set.** The model consistently predicted 28 out of 30 lung cancer germlines samples to be 'at-risk' and 88 out of 89 thousand genome non-cancer control samples to be 'healthy'. The explanations of the sample types are as follows: LC – Lung cancer, 1kGPC – 1000 genome project control.

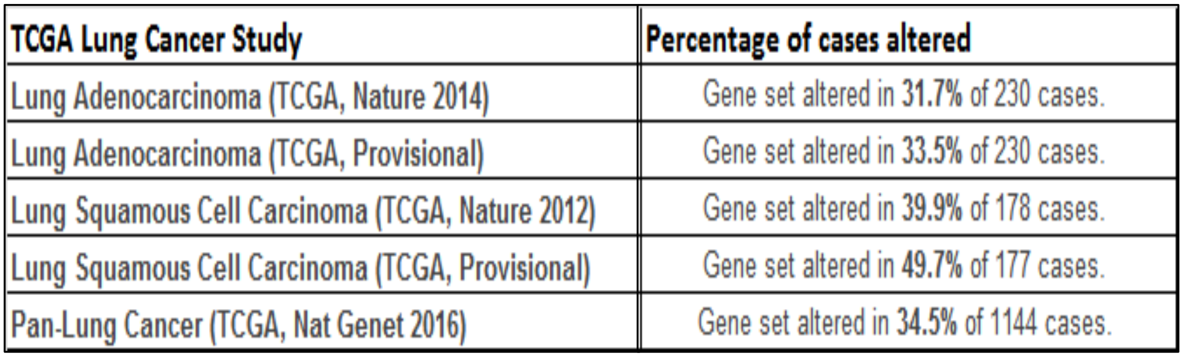
| Sample # | Sample Type | % of loci with cancer GT | Leave one out at-risk % cut off | Leave one out prediction |
|---|---|---|---|---|
| 1 | LC | 95.2 | 52.4 | At risk |
| 2 | LC | 63.6 | 50.0 | At risk |
| 3 | LC | 81.0 | 52.4 | At risk |
| 4 | LC | 54.5 | 50.0 | At risk |
| 5 | LC | 68.2 | 50.0 | At risk |
| 6 | LC | 57.1 | 52.4 | At risk |
| 7 | LC | 72.7 | 50.0 | At risk |
| 8 | LC | 81.8 | 50.0 | At risk |
| 9 | LC | 95.2 | 52.4 | At risk |
| 10 | LC | 77.3 | 50.0 | At risk |
| 11 | LC | 76.2 | 52.4 | At risk |
| 12 | LC | 47.6 | 52.4 | Healthy |
| 13 | LC | 57.1 | 52.4 | At risk |
| 14 | LC | 71.4 | 52.4 | At risk |
| 15 | LC | 90.5 | 52.4 | At risk |
| 16 | LC | 71.4 | 52.4 | At risk |
| 17 | LC | 72.7 | 50.0 | At risk |
| 18 | LC | 63.6 | 50.0 | At risk |
| 19 | LC | 66.7 | 52.4 | At risk |
| 20 | LC | 68.2 | 50.0 | At risk |
| 21 | LC | 72.7 | 50.0 | At risk |
| 22 | LC | 59.1 | 50.0 | At risk |
| 23 | LC | 57.1 | 52.4 | At risk |
| 24 | LC | 81.8 | 50.0 | At risk |
| 25 | LC | 31.8 | 50.0 | Healthy |
| 26 | LC | 66.7 | 52.4 | At risk |
| 27 | LC | 63.6 | 50.0 | At risk |

|  |  |  |  |  |
|---|---|---|---|---|
| 28 | LC | 71.4 | 52.4 | At risk |
| 29 | LC | 50.0 | 50.0 | At risk |
| 30 | LC | 90.5 | 52.4 | At risk |
| 31 | 1KGPC | 25.0 | 55.0 | Healthy |
| 32 | 1KGPC | 45.0 | 55.0 | Healthy |
| 33 | 1KGPC | 47.6 | 52.4 | Healthy |
| 34 | 1KGPC | 40.0 | 55.0 | Healthy |
| 35 | 1KGPC | 40.0 | 55.0 | Healthy |
| 36 | 1KGPC | 52.4 | 52.4 | Healthy |
| 37 | 1KGPC | 42.9 | 52.4 | Healthy |
| 38 | 1KGPC | 30.0 | 55.0 | Healthy |
| 39 | 1KGPC | 25.0 | 55.0 | Healthy |
| 40 | 1KGPC | 50.0 | 59.1 | Healthy |
| 41 | 1KGPC | 57.1 | 52.4 | At risk |
| 42 | 1KGPC | 47.6 | 52.4 | Healthy |
| 43 | 1KGPC | 40.0 | 55.0 | Healthy |
| 44 | 1KGPC | 38.1 | 52.4 | Healthy |
| 45 | 1KGPC | 38.1 | 52.4 | Healthy |
| 46 | 1KGPC | 25.0 | 55.0 | Healthy |
| 47 | 1KGPC | 47.6 | 52.4 | Healthy |
| 48 | 1KGPC | 47.6 | 52.4 | Healthy |
| 49 | 1KGPC | 31.8 | 59.1 | Healthy |
| 50 | 1KGPC | 40.0 | 55.0 | Healthy |
| 51 | 1KGPC | 33.3 | 52.4 | Healthy |
| 52 | 1KGPC | 52.4 | 52.4 | Healthy |
| 53 | 1KGPC | 52.4 | 52.4 | Healthy |
| 54 | 1KGPC | 40.0 | 55.0 | Healthy |
| 55 | 1KGPC | 38.1 | 52.4 | Healthy |
| 56 | 1KGPC | 35.0 | 55.0 | Healthy |
| 57 | 1KGPC | 52.4 | 52.4 | Healthy |
| 58 | 1KGPC | 38.1 | 52.4 | Healthy |
| 59 | 1KGPC | 45.5 | 59.1 | Healthy |
| 60 | 1KGPC | 33.3 | 52.4 | Healthy |
| 61 | 1KGPC | 45.5 | 59.1 | Healthy |
| 62 | 1KGPC | 33.3 | 52.4 | Healthy |
| 63 | 1KGPC | 38.1 | 52.4 | Healthy |
| 64 | 1KGPC | 14.3 | 52.4 | Healthy |
| 65 | 1KGPC | 31.8 | 59.1 | Healthy |
| 66 | 1KGPC | 28.6 | 52.4 | Healthy |
| 67 | 1KGPC | 50.0 | 59.1 | Healthy |

| | | | | |
|---|---|---|---|---|
| 68 | 1KGPC | 42.9 | 52.4 | Healthy |
| 69 | 1KGPC | 25.0 | 55.0 | Healthy |
| 70 | 1KGPC | 36.4 | 59.1 | Healthy |
| 71 | 1KGPC | 50.0 | 55.0 | Healthy |
| 72 | 1KGPC | 40.0 | 55.0 | Healthy |
| 73 | 1KGPC | 40.9 | 59.1 | Healthy |
| 74 | 1KGPC | 38.1 | 52.4 | Healthy |
| 75 | 1KGPC | 45.5 | 59.1 | Healthy |
| 76 | 1KGPC | 18.2 | 59.1 | Healthy |
| 77 | 1KGPC | 25.0 | 55.0 | Healthy |
| 78 | 1KGPC | 33.3 | 52.4 | Healthy |
| 79 | 1KGPC | 42.9 | 52.4 | Healthy |
| 80 | 1KGPC | 9.5 | 52.4 | Healthy |
| 81 | 1KGPC | 36.4 | 59.1 | Healthy |
| 82 | 1KGPC | 23.8 | 52.4 | Healthy |
| 83 | 1KGPC | 23.8 | 52.4 | Healthy |
| 84 | 1KGPC | 35.0 | 55.0 | Healthy |
| 85 | 1KGPC | 28.6 | 52.4 | Healthy |
| 86 | 1KGPC | 47.6 | 52.4 | Healthy |
| 87 | 1KGPC | 36.4 | 59.1 | Healthy |
| 88 | 1KGPC | 36.4 | 59.1 | Healthy |
| 89 | 1KGPC | 20.0 | 55.0 | Healthy |
| 90 | 1KGPC | 23.8 | 52.4 | Healthy |
| 91 | 1KGPC | 45.5 | 59.1 | Healthy |
| 92 | 1KGPC | 40.0 | 55.0 | Healthy |
| 93 | 1KGPC | 42.9 | 52.4 | Healthy |
| 94 | 1KGPC | 50.0 | 59.1 | Healthy |
| 95 | 1KGPC | 50.0 | 59.1 | Healthy |
| 96 | 1KGPC | 42.9 | 52.4 | Healthy |
| 97 | 1KGPC | 31.8 | 59.1 | Healthy |
| 98 | 1KGPC | 25.0 | 55.0 | Healthy |
| 99 | 1KGPC | 40.0 | 55.0 | Healthy |
| 100 | 1KGPC | 50.0 | 59.1 | Healthy |
| 101 | 1KGPC | 38.1 | 52.4 | Healthy |
| 102 | 1KGPC | 35.0 | 55.0 | Healthy |
| 103 | 1KGPC | 15.0 | 55.0 | Healthy |
| 104 | 1KGPC | 40.0 | 55.0 | Healthy |
| 105 | 1KGPC | 31.8 | 59.1 | Healthy |
| 106 | 1KGPC | 10.0 | 55.0 | Healthy |
| 107 | 1KGPC | 25.0 | 55.0 | Healthy |

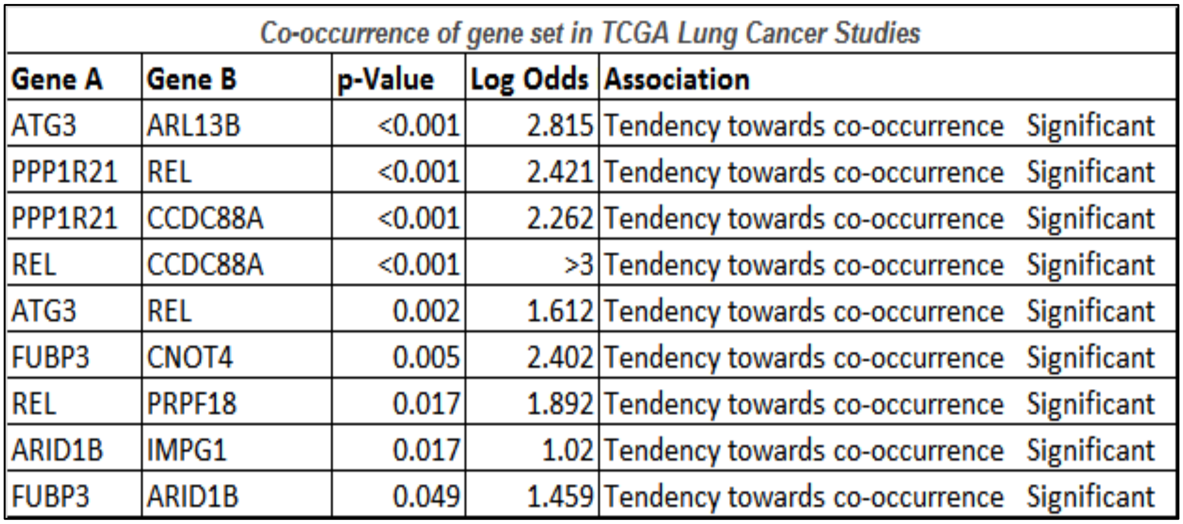
| | | | | |
|---|---|---|---|---|
| 108 | 1KGPC | 54.5 | 59.1 | Healthy |
| 109 | 1KGPC | 23.8 | 52.4 | Healthy |
| 110 | 1KGPC | 40.9 | 59.1 | Healthy |
| 111 | 1KGPC | 20.0 | 55.0 | Healthy |
| 112 | 1KGPC | 42.9 | 52.4 | Healthy |
| 113 | 1KGPC | 27.3 | 59.1 | Healthy |
| 114 | 1KGPC | 33.3 | 52.4 | Healthy |
| 115 | 1KGPC | 27.3 | 59.1 | Healthy |
| 116 | 1KGPC | 47.6 | 52.4 | Healthy |
| 117 | 1KGPC | 36.4 | 59.1 | Healthy |
| 118 | 1KGPC | 42.9 | 52.4 | Healthy |
| 119 | 1KGPC | 45.5 | 59.1 | Healthy |

**Table 3S-10: Investigation of TCGA lung cancer data reveal that alterations in genes in lung cancer risk classifiers occur on average in 37% of studied individuals.**



| TCGA Lung Cancer Study | Percentage of cases altered |
|---|---|
| Lung Adenocarcinoma (TCGA, Nature 2014) | Gene set altered in **31.7%** of 230 cases. |
| Lung Adenocarcinoma (TCGA, Provisional) | Gene set altered in **33.5%** of 230 cases. |
| Lung Squamous Cell Carcinoma (TCGA, Nature 2012) | Gene set altered in **39.9%** of 178 cases. |
| Lung Squamous Cell Carcinoma (TCGA, Provisional) | Gene set altered in **49.7%** of 177 cases. |
| Pan-Lung Cancer (TCGA, Nat Genet 2016) | Gene set altered in **34.5%** of 1144 cases. |

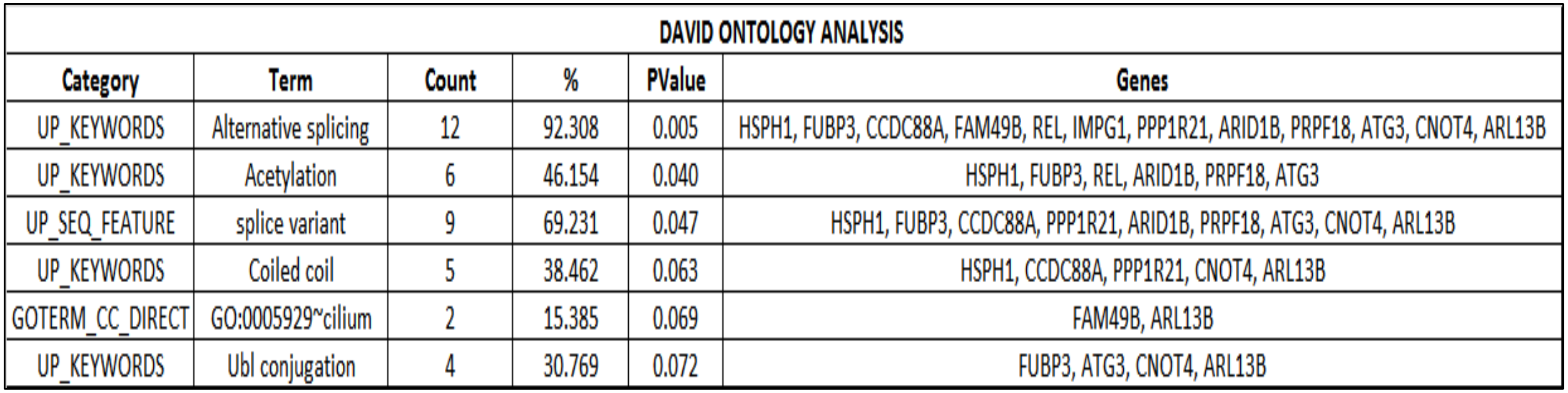
**Table S3-11: Analysis of TCGA Pan Cancer lung cancer studies reveal indicate that 9 gene pairs tend to significantly co-occur within lung cancer risk classifiers.**



*Co-occurrence of gene set in TCGA Lung Cancer Studies*

| Gene A | Gene B | p-Value | Log Odds | Association | |
|---|---|---|---|---|---|
| ATG3 | ARL13B | <0.001 | 2.815 | Tendency towards co-occurrence | Significant |
| PPP1R21 | REL | <0.001 | 2.421 | Tendency towards co-occurrence | Significant |
| PPP1R21 | CCDC88A | <0.001 | 2.262 | Tendency towards co-occurrence | Significant |
| REL | CCDC88A | <0.001 | >3 | Tendency towards co-occurrence | Significant |
| ATG3 | REL | 0.002 | 1.612 | Tendency towards co-occurrence | Significant |
| FUBP3 | CNOT4 | 0.005 | 2.402 | Tendency towards co-occurrence | Significant |
| REL | PRPF18 | 0.017 | 1.892 | Tendency towards co-occurrence | Significant |
| ARID1B | IMPG1 | 0.017 | 1.02 | Tendency towards co-occurrence | Significant |
| FUBP3 | ARID1B | 0.049 | 1.459 | Tendency towards co-occurrence | Significant |

**Table S3-12: Analysis of ontologies indicates alternative splicing, acetylation, and splice variants to be significant terms among the 12 genes in lung cancer risk classifier.**



| DAVID ONTOLOGY ANALYSIS | | | | | |
|---|---|---|---|---|---|
| Category | Term | Count | % | PValue | Genes |
| UP_KEYWORDS | Alternative splicing | 12 | 92.308 | 0.005 | HSPH1, FUBP3, CCDC88A, FAM49B, REL, IMPG1, PPP1R21, ARID1B, PRPF18, ATG3, CNOT4, ARL13B |
| UP_KEYWORDS | Acetylation | 6 | 46.154 | 0.040 | HSPH1, FUBP3, REL, ARID1B, PRPF18, ATG3 |
| UP_SEQ_FEATURE | splice variant | 9 | 69.231 | 0.047 | HSPH1, FUBP3, CCDC88A, PPP1R21, ARID1B, PRPF18, ATG3, CNOT4, ARL13B |
| UP_KEYWORDS | Coiled coil | 5 | 38.462 | 0.063 | HSPH1, CCDC88A, PPP1R21, CNOT4, ARL13B |
| GOTERM_CC_DIRECT | GO:0005929~cilium | 2 | 15.385 | 0.069 | FAM49B, ARL13B |
| UP_KEYWORDS | Ubl conjugation | 4 | 30.769 | 0.072 | FUBP3, ATG3, CNOT4, ARL13B |

# Chapter 4: Dysfunctional DNA repair pathway via defective FANCD2 gene engenders multifarious exomic and transcriptomic effects

# ABSTRACT

Fanconi Anemia (FA) affects only 1 in 130,000 births, but has severe and diverse clinical consequences. It has been theorized that defects in the FA DNA repair complex lead to a spectrum of variants that are responsible for those diverse clinical phenotypes. Here we show using nextgen sequencing that a clinically-derived FA cell line relative to its retrovirally corrected line continued to accumulate variants, especially INDELs. Over 200 SNV and 25 INDEL robust impact variants (introduction of start/stop codons, nonsense and missense mutations) distinguished the FA and corrected cell lines, including many high-impact variants. Further, sequences from biological replicates indicated that new mutations accumulated during a single 30-hour culture. These genetic modifications had a devastating effect on the transcriptome, with statistically significant changes in the expression of 270 genes. These genetic and transcriptomic variants significantly impacted many pathways and functional ontologies, spanning those associated with many diverse disease phenotypes/symptoms. The downstream continuously accumulating variant diversity is consistent with the disease diversity seen in FA patients. These results underscore the consequences of defects in DNA repair mechanisms, and indicate that accumulating diverse mutations from individual parent cells make it difficult to anticipate the longitudinal clinical behavior of emerging disease states in an individual with FA.

## INTRODUCTION:

Fanconi anemia (FA) is a congenital disorder that is phenotypically expressed in the daughter generation usually due to an autosomal recessive genetic condition in the parent generation. FA is characterized by bone marrow failure, infection susceptibility and a predisposition to cancer, specifically acute myeloid leukemia (AML)[1]. FA cells have been found to be enriched with chromosomal aberrations that occur due to unrepaired DNA crosslinking that are left uncorrected by the dysfunctional DNA repair mechanism; the FA pathway [2-4]. The presence of increased chromosomal aberrations is used as a preliminary diagnostic test that can confirm FA[5]. DNA crosslinking agents have been used as cytotoxic drugs to disrupt cell division in cancer but the cell's response to this method has not been entirely understood[6]. Recruiting DNA repair proteins to the interstrand crosslinks (ICLs) has been known to be the primary function of the FA pathway, and recent studies show the involvement of the FA pathway in maintaining general genome stability[7]. About 16 genes have been identified to be coding for the FA pathway core protein complex. These genes have been classified into complementation groups: A, B, C, D1, D2, E, F, G, I. J. L, M and N [8-15].

A major portion of previous research that has been conducted has tried to understand the causes of these DNA lesions and have extensively discussed the complementation groups of genes that have mutations that lead to a dysfunctional FA pathway[1,8-12,16-18]. The complementation studies show that the dysfunction of anyone of the 16 genes can cause FA. In this work, we undertake an effort to understand the genome-wide and transcriptome-wide downstream effects of dysfunctional FANCD2 gene in PD20 cells. We hypothesize that one of the downstream consequences of FA caused ICL is disrupted transcription, which can lead to altered transcriptional products. It has been shown that transcriptional products can be altered during cancer[19]. A significant level of transcriptome instability has been shown to always be present in cells and several stabilizing mechanisms have been known to be in place to respond to them[20]. However, in the absence of a functional FA pathway, we propose that an increased level of unrepaired ICLs can overwhelm the transcriptome stability maintaining mechanisms and lead to altered transcriptional products, in which case consequent impairments of cellular function is imminent. While the ICL repair function of FA pathway is well known, there are studies that suggest that the FA pathway may be involved in the general upkeep of genomic stability[7]. Microsatellite instability has been linked to dysfunctional mismatch repair but recent studies show the contribution of the crosslink repair (FA) pathway to microsatellite instability[21-23]. Hence we also aim to understand possible effects of a dysfunctional FA pathway on the microsatellite instability.

## METHODS:

**Cell lines and DNA/RNA sample preparation:** PD20 cell lines containing a defective FANCD2 gene and retrovirally corrected FANCD2 were obtained from the FA cell repository at Oregon Health and Science University (http://www.ohsu.edu/research/fanconi-anemia/celllines.cfm/forum/index.cfm). The cells were seeded onto six well plates at 1.5 x $10^4$ cells/well. After ~24 hours of incubation, the medium (5 ml of DMEM-10% FBS) was changed. After 4 hours of incubation, the cells were tripsin-ized and collected for DNA (Qiagen Allprep kit) and RNA (Qiagen RNA prep kit) extraction. DNA and RNA from two PD20 samples and two PD20 RV corrected for FANCD2 gene were sent for exomic sequencing and RNASeq analysis.

**DNA variant detection:** Paired-end sequencing reads of the 4 Fanconi Anemia DNA samples were obtained in the form of fastq files. The samples had a mean of 30 million reads with a standard deviation of 3 million reads. The samples were then checked for quality using the Trimmomatic tool. The Trimmomatic tool utilizes a window based system to check for the sequencing quality of bases. The tool requires user input for three parameters: sliding window length, quality score threshold and minimum length of sequencing read. The following values were used: 10, 20 and 70. Only sequences that contained both the mate pairs reads were used for further processing. After quality trimming, the samples on overage had 27 million reads with a standard deviation of 3 million. The 'mem' function of the BWA package was used to map the sequencing reads to GRCh38 reference genome. On average 99% of reads in all the samples were mapped to the reference genome. The SAMTOOLS package was used to convert the SAM file into a sorted and indexed BAM file. The AddOrReplaceReadGroups function in the Picard package was used to add read groups to the BAM files. All the four BAM files were pooled together to create INDEL realignment targets using the RealignerTargetCreator function in GATK package. The GATK IndelRealigner function was then used on individual samples to create INDEL realigned BAM files. The GATK HaplotypeCaller was used to create VCF files with INDEL and SNP calls and the SelectVariants function was used to separate SNPs and INDELs.

Python and shell scripts were written to calculate the number of commonly present SNPs and INDELs in the FA and FA_RV samples (Table 2). The BEDTOOLS coverage function was used to calculate the percentage of the exome that is covered by the reads and the SAMTOOLS depth function (samtools depth sample.bam | awk '{sum+=$3} END { print "Average = ",sum/NR}') was used to calculate the coverage of the read covered exome (Table 3).

The VCF files were fed to the SNPEff program to annotate the SNP and INDEL calls. The SNPEff annotation output was used to classify the detected variants according to the type and genomic region (Supplemental table 2). The SNPEff program outputs the possible effects of a given variant and classifies the effect into 4 categories: high, low, moderate and modifier.

Custom written python scripts were used perform the pairwise comparison of DNA variants (tables) that were predicted as high impact by SNPEff. A variant is considered to be specific to a sample, during the pairwise comparison, if the variant position is covered in the both samples and does not have the same variant call.

The calculations describing the pairwise comparisons in Table 5 and 6 are described in detail in the Supplemental table 3 and 4.

**RNA-Seq data processing and analysis:** The single end sequencing reads of the RNA samples were obtained in the form of fastq files. On average each sample contained 36 million reads with a standard deviation of 1 million. They were quality trimmed using Trimmomatic tool with the same parameters used for the DNA samples. After quality trimming, the samples retained on average 35 million reads (standard deviation 1 million). The quality trimmed RNA sequence reads were then mapped to the GRCh38 reference genome by the Tophat2 program with a prebuilt transcriptome index. On average 90% (Standard deviation: 3%) of the reads were uniquely mapped. The mapped BAM file was passed through the Cufflinks and Cuffmerge programs and were finally passed through the Cuffdiff program as 4 separate sample groups. The Cuffdiff program generates a pairwise comparison of the FPKM values for each gene in all the 4 samples, providing 6 pairwise comparisons. Genes found to have statistically significant (P value < 0.05) differential expression in comparisons across the FA and FA_RV sample groups and no significant differential expression within the FA and FA_RV sample groups are provided in the supplemental table 4 as differentially expressed between the two sample groups.

**Microsatellite genotyping:** Microsatellite (MST) genotyping of the Fanconi Anemia samples was performed using the Repeatseq program[24]. The program requires three user inputs: mapped sequencing reads in the form of a BAM file, a list of MST genomic coordinates and the reference genome. Repeatseq outputs a variant call file listing all possible alleles detected for each microsatellite. Custom written perl scripts were written to generate a list of MSTs with their primary, secondary and minor allele information. An MST is considered to have a minor allele only if the minor allele is covered by at least 3 reads.

**Generating the query list of microsatellites:** A list of microsatellites in version 19 of the human reference genome was generated with a custom perl script 'searchTandemRepeats.pl' using default parameters. This script has been used in previous microsatellite studies and is available online at http://genotan.sourceforge.net/#_Toc324410847. The initial list generated with this script included 1,671,121 microsatellites. To mitigate the likelihood of improper read mapping between microsatellites we removed all subsets of microsatellites possessing the same motif between identical 3' and 5' flanking regions. For example, the microsatellites 'GCTGC(A)$^{34}$CTTAG' and 'GCTGC(A)$^{15}$CTTAG' were preemptively removed from our initial list of microsatellites. The fact that there were many of these potentially ambiguous regions is not surprising considering microsatellites are often embedded in larger repetitive motifs such as LINES and SINES. Our filtered list included 611,515 microsatellites.

**Mechanistic analysis:** The DAVID online bioinformatics tool was used to perform the ontological gene enrichment analysis[25]. The DAVID tool was used on the 270 genes that were found to be differentially expressed between the FA and FA_RV sample groups. The REACTOME pathway online tool was used to find the pathways in which the two sets of genomic variant associated genes are involved[26]. The pathway hits that were found to have significant (P value < 0.05) involvement are considered for analysis and are furnished in supplemental tables 8 and 9.

## RESULTS AND DISCUSSION

### Verification of FANCD2 retroviral correction:

To assess the extent of genomic damage that can be caused by the dysfunction of only one of the DNA repair genes, in this case FANCD2, it is first necessary to establish the differential expression of FANCD2 in the two sample groups i.e. 2 biological replicates of the original clinical FA sample and 2 FA samples that were corrected for FANCD2 expression using a retroviral vector. Table 1 shows the statistically significant differential expression of FANCD2 gene in these two sample groups by comparing the FPKM values from RNA sequencing. All the four comparisons (FA1 vs FA_RV1, FA2 vs FA_RV1, FA1 vs FA_RV2, FA2 vs FA_RV2) of FPKM (Fragments Per Kilobase of transcript per Million mapped reads) values across the two sample groups show significant difference while both the comparison within groups (FA1 vs FA2, FA_RV1 vs FA_RV2) do not show any such difference. This confirms that expression of FANCD2 is quenched in the FA samples, and restored in the corrected samples.

The RNA sequencing derived differential expression data in Table 1 demonstrates that the retrovirus mediated correction of the FANCD2 gene in the FA_RV sample group is restoring proper expression of the gene and that the FA_RV sample group can be utilized as a control to study the propagation of genomic damage in the FA sample group. The gene expression fold change of all 17 Fanconi Anemia related genes show that only FANCD2 had a significant gene expression change between FA and FA_RV sample groups and the other 16 genes did not show any such change in gene expression (Supplemental Table 1). Hence, the genomic damage and associated transcriptional profile change in non-FA genes found to be specific to the

FA sample group can be attributed to the dysfunctional DNA repair mechanism caused by the incorrect expression of the FANCD2 gene.

**Genomic variant analysis:**

The four DNA samples, 2 biological replicates for FA cell line and 2 biological replicates for the FA_RV cell line were sequenced and mapped to the human genome version 38 reference to identify genomic variants, such as SNPs and INDELs, and assess their potential impact. About 55% of the INDELs were found in all four samples while 65% of the SNPs were commonly found in all the four samples (Table 2), thus confirming a common origin, i.e. that these had existed in the clinical sample prior to creating the cell lines. About 70% of INDELs and 75% of SNPs were found to be shared by samples within groups (Table 2). The remaining fraction of un-shared variants indicate that the biological replicates continue to develop different variants that go uncorrected, illustrating a mechanism for enhanced heterogeneity.

A significant increase in the fraction of INDELs in all the four samples was observed, in comparison to the 1000 Genome Project samples (1kGP) (Table 3). Chromosomal aberrations such as DNA lesions are signature variants of Fanconi Anemia[5]. The increased fraction of INDEL variants observed in all 4 Fanconi Anemia samples is consistent with the specific correction mechanism defect in FA. It should be noted that such a difference in SNP-INDEL ratio is found while the fraction of the exome being sequenced and the depth of coverage are approximately equal in the 4 Fanconi Anemia samples and the 1kGP samples. The constancy of the SNP: INDEL ratio in three different 1kGP samples (9.2:0.8) serves as a reference which shows all the FA samples have a higher fraction of INDELs (8.8:1.2) than what is seen in a normal population.

Analyzing the high impact genomic variants in all the 4 Fanconi anemia samples, a set of 82 genes was found to be associated with variants that were found only in the FA samples and a set of 618 genes was found to be associated with variants that were found commonly in all the 4 samples (FA and FA_RV sample groups).

Microsatellite genotyping was done to understand the effect of crosslink repair mechanism on MST instability. It has been suggested that MST instability is not only caused by the mismatch repair mechanism but could also be caused by nucleotide excision repair and crosslink repair mechanisms[21-23]. The fraction (5% - 256 MSTs) of heterozygous MST in the Fanconi anemia samples is found to be higher than the fraction (3.5% - 138 MSTs) in the 1kGP samples which is consistent with the increase in the percentage of MST with minor alleles.  The fraction of callable MSTs with minor alleles is 5% (200 MSTs) in the 1kGP samples while the fraction of callable MSTs with minor alleles is 8.5% (437 MSTs) in the Fanconi anemia samples.

To explore in detail the heterogeneity in these 4 samples, pairwise comparisons of SNP and INDEL occurrences were made. SNP and INDEL-containing loci were compared only if the corresponding genomic location in the other sample was sequenced and called. On average, 24% of SNPs was found to sample specific, demonstrating, again, the high heterogeneity between these samples (Table 5 and Supplemental Table 3). Around 14% of INDELs were found to be sample specific (Table 6 and supplemental table 4). It should be noted that such high level of variance in genome is consistent with the fact that Fanconi Anemia patients are highly predisposed to cancer[27].  ICLs can directly affect DNA replication by phenomenally increasing DNA errors which can lead to cell death or uncontrolled cell growth[27].

**RNA sequencing to measure expression changes:**

Having realized the extent of genomic damage caused by a dysfunctional DNA repair gene, it is pertinent to examine the downstream effect of this DNA damage on the transcriptome, i.e. on gene expression. Along with FANCD2, a set of 270 genes (Supplemental Table 5) were found to be differentially expressed in the biological replicates between the FA and the FA_RV samples groups. Significant changes in expression ranged from 1.1 to 11.5. The top most 8 genes with altered expression are shown in Table 7.

The pairwise comparisons of the number of differentially expressed genes in all 4 samples showed that on average all 4 pairwise comparisons across sample groups had 375 differentially expressed genes (FA1 - FA_RV1 = 376; FA1 - FA_RV2 = 384; FA2 - FA_RV1 = 373; FA2 - FA_RV2 = 369) while the two comparisons within sample groups (FA1-FA2 = 1, FA_RV1-FA_RV2 = 1) had only 1 differentially expressed gene each. This confirms that although there was some divergence in the genomes of the biological replicates, that the effect was very minor on the transcriptome. Also, as a control, and to put these numbers into context, three 1kGP RNA-Seq samples were downloaded and pairwise analyzed for differentially expressed genes. The number of differentially expressed genes in these 3 samples ranged from 105 to 236. It should be noted that the expression measurements in the 1kGP samples were for different individuals under differing conditions, while the FA samples were from the same individual under controlled culture conditions.

In addition to the raw expression level changes, there are indications that these genes are alternatively spliced. Pairwise comparisons of exon count in genes within and across the FA and FA_RV sample groups show that on average 29% of expressed genes have varying exon counts across sample groups while 24% of expressed genes have varying exon counts within sample groups. An increase of 5% of genes (795 genes) with varying exon counts between FA and FA_RV sample groups shows the effects of the dysfunctional FANCD2 gene on alternative splicing. These findings might suggest that a dysfunctional DNA repair mechanism leads to DNA damage which in turn affects gene expression through truncated RNA transcripts rather than directly affecting the number of genes that are expressed.

### Mechanistic analysis of genes with differential expression and genomic variants:

In order to understand the mechanistic role of genes that were affected by a DNA repair gene (FANCD2) dysfunction, three sets of genes where analyzed for their gene ontology term enrichment and their overrepresentation in pathways: 1) set of genes (270) were found to be differentially expressed (RNA-Seq) between the FA and FA_RV sample groups; 2) set of genes (82) that were associated with DNA variants that were specific to the FA sample group and were not found in the FA_RV sample group and; 3) set of genes (618) that were found to associated with genomic variants (SNPs and INDELs) that were occur in both sample groups.

The gene ontology enrichment analysis of differentially expressed genes and pathway analysis of the 82 genes that were found to be specific for the FA sample group are consistent. Both sets showed the common involvement of genes in immune functions such as antigen presentation and signaling pathways (Supplemental Table 7 and 8). These genes were also found to be involved in immune related signaling pathways (Supplemental Table 8).

Pathway analysis of the FA specific gene set, in comparison to the gene variants that are commonly found in all samples, show involvement in a wider variety of functionalities such as apoptosis and transcriptional regulation, along with known immune related FA functions such as antigen presentation and immune related signaling. The common gene variant set, on the other hand, is only involved in immune related pathways. This could suggest that earlier mutations target the immune system while further addition of mutations that are caused by the unrepaired FA pathway may lead to a wider variety of genomic effects that can initiate apoptosis.

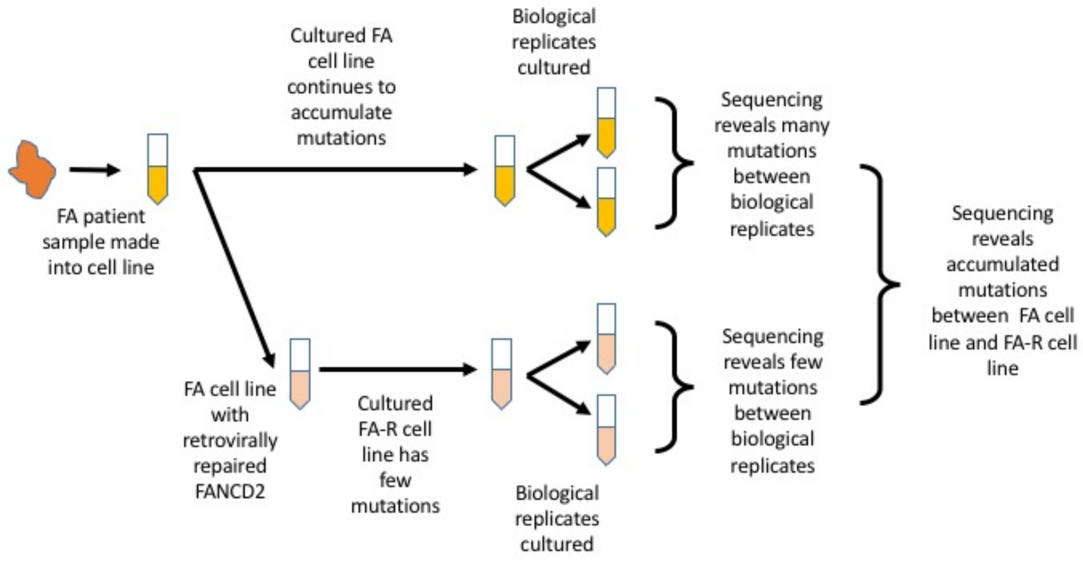
Pathway analysis of the differentially expressed genes show that signaling genes are specifically involved in immunological processes such as T cell receptor signaling. One of the highly overrepresented pathways as indicated by the differentially expressed genes was found to be the endosomal pathway. The Cathespin L gene (CTSL) that is mainly involved in the lysosomal degradation of proteins was not only differentially expressed between the FA and FA_RV sample group but it was up-regulated in the FA samples and down-regulated in the FA_RV samples. It has already been suggested that dysfunction of FANCD2 affects the transcripome by the production of broken transcripts which can be translated into malformed proteins that can require overactive lysosomal activity. Other overrepresented pathways are the PD-1 (programmed cell death) signaling pathway and ER-phagosome pathway. For example, one gene involved in the phagosome pathway is MYD88 which again was found to be more highly expressed in the FA sample group and not in the FA_RV sample group. We observed the up-regulation of 12 collagen genes (COL4A2, COL5A2, COL1A2, COL6A2, COL6A1, COL8A1, COL11A1, COL4A1, COL3A1, COL16A1, COL5A3, COL6A3) in the FA sample group. Cell viability and the regulation of the extra cellular matrix has been convincingly linked to cancer progression and the up-regulation of a collection of collagen coding proteins is consistent with up-regulation of cell death signaling and phagosomal activity[28]. It should be noted that an up-regulation of cell death and collagen genes can lead to cancer and, recalling, Fanconi Anemia patient are predisposed to acute myeloid leukemia[28].

## CONCLUSION:

These findings are unexpected and surprising. We find it surprising that so many variants, so many high-impact variants, accumulate so quickly as a consequence of uncorrected genomic mistakes. We find it unexpected that so many genes can be altered and/or disrupted, yet tolerated, at least such that the cells survive. Further, we find it surprising that although there were many coding variants, these and other non-coding variants resulted in fewer genes with altered expression. One must consider cause and effect. How much does the genomic variants modify the expression level, and how much has the expression level of various genes shifted to compensate and ensure cell survival. One must also wonder how many cellular/genomic trajectories are possible, and how cells compensate accordingly. In this experiment, we quantified the variants from but one of the many (currently estimated at 16) genes that are members of this DNA repair complex. It begs the question as to how each of these genes contribute to the repair process, and how defects in each manifest themselves as to the number and spectrum of variants that persist.

## FIGURES:

**Figure 4-1: Two types of Fanconi anemia – PD20 cell lines were included in this experiment.** The cell lines represented in yellow are FA cell lines with a dysfunctional FANCD2 gene while the cell lines represented in pink are FA cell lines that were corrected with a functional FANCD2 gene using retroviral transduction. Variants with respect to the human genome reference found in common among all samples are indicative of variants accumulated prior to extraction from the patient. Variants found in common in each of the pairs of biological replicates are indicative of mutations that diverged when the original cell line was repaired. Differences between the biological replicates indicates the divergence of each of the cell line cultures as different uncorrected errors are made in each sample.



FA patient sample made into cell line
Cultured FA cell line continues to accumulate mutations
Biological replicates cultured
Sequencing reveals many mutations between biological replicates
FA cell line with retrovirally repaired FANCD2
Cultured FA-R cell line has few mutations
Biological replicates cultured
Sequencing reveals few mutations between biological replicates
Sequencing reveals accumulated mutations between FA cell line and FA-R cell line

# TABLES

**Table 4-1: Pairwise comparison of FANCD2 gene expression in 2 FA samples and 2 FANCD2 RV corrected FA samples.** Gene expression calculated by FPKM confirms that compared to FA samples, the expression of FANCD2 in the FA_RV samples is significant. This verifies the retroviral correction of the FANCD2 gene in the FA_RV samples.

| Sample 1 | Sample 2 | Sample 1 FPKM | Sample2 FPKM | log2(fold_ change) | P value | Significance |
|---|---|---|---|---|---|---|
| FA1 | FA2 | 4.6 | 3.4 | -0.4 | 0.4 | No |
| FA1 | FA_RV1 | 4.6 | 108.6 | 4.5 | 0.0 | Yes |
| FA2 | FA_RV1 | 3.4 | 108.6 | 5.0 | 0.0 | Yes |
| FA1 | FA_RV2 | 4.6 | 124.2 | 4.7 | 0.0 | Yes |
| FA2 | FA_RV2 | 3.4 | 124.2 | 5.1 | 0.0 | Yes |
| FA_RV1 | FA_RV2 | 108.6 | 124.2 | 0.1 | 0.5 | No |

FPKM is the Fragments Per Kilobase of exon per Million fragments mapped.

**Table 4-2: Variant calls for the 2 FA and 2 FANCD2 RV corrected FA sample, with respect to the human reference genome.** While a significantly large number of SNPs and INDELs were found to be repeated in all the 4 samples, the difference in the number of variants within biological replicates show evidence of heterogeneity in these replicate cell line cultures.

| **Variant Type** | **Sample** | **Sample Type** | **Variant count** | | |
|---|---|---|---|---|---|
| | | | **#** | **Repeated in replicates** | **Repeated in all samples** |
| INDEL | GRL1398 | FA | 28053 | 19556 | 15023 |
| | GRL1399 | FA | 27218 | | |
| | GRL1400 | FA_RV | 27386 | 18836 | |
| | GRL1401 | FA_RV | 25966 | | |
| SNP | GRL1398 | FA | 201940 | 153043 | 128179 |
| | GRL1399 | FA | 198816 | | |
| | GRL1400 | FA_RV | 194954 | 144266 | |
| | GRL1401 | FA_RV | 191593 | | |

**Table 4-3: Distribution of SNPs and INDELs in the FA, FA_RV and 1kGP samples.** While the ratio of SNP vs INDELs in the FA and FA_RV samples are constant, an increase in the INDEL events is seen in the FA and FA_RV samples, compared to the 1kGP samples, which we include as another type of control.

| Sample | Sample Type | Total variants | SNP | INDEL | SNP-INDEL ratio | Exome % covered | Coverage |
|---|---|---|---|---|---|---|---|
| GRL1398 | FA1 | 234244 | 205,613 | 28631 | 8.8 : 1.2 | 92 | 13 |
| GRL1399 | FA2 | 229972 | 202,186 | 27786 | 8.8 : 1.2 | 92 | 10 |
| GRL1400 | FA_RV_1 | 226132 | 198,180 | 27952 | 8.8 : 1.2 | 92 | 11 |
| GRL1401 | FA_RV_2 | 221086 | 194,671 | 26415 | 8.8 : 1.2 | 91 | 11 |
| HG02003 | 1KG | 281236 | 258,715 | 22521 | 9.2 : 0.8 | 86 | 12 |
| HG02008 | 1KG | 271595 | 250,147 | 21448 | 9.2 : 0.8 | 87 | 10 |
| HG02009 | 1KG | 252543 | 232,261 | 20282 | 9.2 : 0.8 | 86 | 11 |
| HG02010 | 1KG | 278559 | 255,466 | 23093 | 9.2 : 0.8 | 86 | 11 |

**Table 4-4: Microsatellite genotyping of the 4 Fanconi Anemia samples and healthy controls from the 1kGP.** On average, 8.4% of the callable MST loci have minor allele while only 5% of the callable MST in 1kGP samples have minor alleles which indicates a higher rate of mutations in MSTs in the Fanconi Anemia samples.

| Sample | Callable MST | Homozygous | Heterozygous | Minor alleles |
|---|---|---|---|---|
| FA1 | 5395 | 95.4 | 4.6 | 7.6 |
| FA2 | 5011 | 94.9 | 5.1 | 7.7 |
| FA_RV1 | 5818 | 94.9 | 5.1 | 9.2 |
| FA_RV2 | 4519 | 95.0 | 5.0 | 9.3 |
| HG02003 | 4485 | 96.6 | 3.4 | 5.5 |
| HG02008 | 3834 | 97.1 | 2.9 | 4.5 |
| HG02009 | 3211 | 96.6 | 3.4 | 4.0 |
| HG02010 | 4158 | 95.6 | 4.4 | 6.1 |

**Table 4-5: Pairwise comparison of high impact SNPs in FA and FA_RV samples show the extent of SNP variability within biological replicates.** To ensure accurate comparison, only SNPs that were sequenced in both samples in a sample pair were considered.

| | **FA1** | **FA2** | **FA_RV1** | **FA_RV2** |
|---|---|---|---|---|
| **FA1** | * | 142 | 167 | 252 |
| **FA2** | | * | 169 | 255 |
| **FA_RV1** | | | * | 234 |
| **FA_RV2** | | | | * |

**Table 4-6: Pairwise comparison of high impact INDELs in FA and FA_RV samples show the extent of INDEL variability within biological replicates.** To ensure accurate comparison, only genomic positions that were sequenced in both samples were used.

| | **FA1** | **FA2** | **FA_RV1** | **FA_RV2** |
|---|---|---|---|---|
| **FA1** | * | 27 | 25 | 36 |
| **FA2** | | * | 23 | 30 |
| **FA_RV1** | | | * | 29 |
| **FA_RV2** | | | | * |

**Table 4-7: Of 270 differentially expressed genes, the most significant are illustrated here. Eight had a high gene expression fold change and was significantly divergent from the exponential distribution of the full set of genes, when comparing FA and FA_RV sample groups.** A “+” indicates higher gene expression in FA_RV sample group and a “-” sign indicates higher gene expression in FA sample group. See supplement table 5 for the full list.

| # | Gene ID | Gene Symbol | Genomic position | GE fold change | High expression in FA_RV |
|---|---|---|---|---|---|
| 1 | XLOC_003069 | - | chr1:239266341-239270392 | 11.5 | + |
| 2 | XLOC_014162 | COLEC12 | chr18:318126-500729 | 10.5 | - |
| 3 | XLOC_015674 | ZNF626 | chr19:20619938-20661596 | 10.0 | - |
| 4 | XLOC_033910 | FGF13 | chrX:138631570-139222889 | 9.8 | + |
| 5 | XLOC_011076 | MT1E | chr16:56625653-56627112 | 9.5 | - |
| 6 | XLOC_002609 | GLUL | chr1:182381703-182392206 | 8.8 | - |
| 7 | XLOC_018786 | COL6A3 | chr2:237324011-237422190 | 8.8 | - |
| 8 | XLOC_002401 | C1orf85 | chr1:156292686-156295689 | 8.8 | - |

## REFERENCES:

1 Kutler, D. I. *et al.* A 20-year perspective on the International Fanconi Anemia Registry (IFAR). *Blood* **101**, 1249-1256 (2003).

2 Donahue, S. L. & Campbell, C. A Rad50-dependent pathway of DNA repair is deficient in Fanconi anemia fibroblasts. *Nucleic Acids Research* **32**, 3248-3257, doi:10.1093/nar/gkh649 (2004).

3 Mace-Aime, G., Couve, S., Khassenov, B., Rosselli, F. & Saparbaev, M. K. The Fanconi anemia pathway promotes DNA glycosylase-dependent excision of interstrand DNA crosslinks. *Environmental and molecular mutagenesis* **51**, 508-519, doi:10.1002/em.20548 (2010).

4 Deans, A. J. & West, S. C. DNA interstrand crosslink repair and cancer. *Nature reviews. Cancer* **11**, 467-480, doi:10.1038/nrc3088 (2011).

5 Meyer, S., Neitzel, H. & Tönnies, H. Chromosomal Aberrations Associated with Clonal Evolution and Leukemic Transformation in Fanconi Anemia: Clinical and Biological Implications. *Anemia* **2012**, 349837, doi:10.1155/2012/349837 (2012).

6 Brulikova, L., Hlavac, J. & Hradil, P. DNA interstrand cross-linking agents and their chemotherapeutic potential. *Current medicinal chemistry* **19**, 364-385 (2012).

7 Michl, J., Zimmer, J. & Tarsounas, M. Interplay between Fanconi anemia and homologous recombination pathways in genome integrity. *The EMBO Journal* **35**, 909-923 (2016).

8 de Winter, J. P. *et al.* Isolation of a cDNA representing the Fanconi anemia complementation group E gene. *Am J Hum Genet* **67**, 1306-1308, doi:10.1016/S0002-9297(07)62959-0 (2000).

9 de Winter, J. P. *et al.* The Fanconi anaemia group G gene FANCG is identical with XRCC9. *Nat Genet* **20**, 281-283, doi:10.1038/3093 (1998).

10 Dorsman, J. C. *et al.* Identification of the Fanconi anemia complementation group I gene, FANCI. *Cell Oncol* **29**, 211-218 (2007).

11 Howlett, N. G. *et al.* Biallelic inactivation of BRCA2 in Fanconi anemia. *Science* **297**, 606-609, doi:10.1126/science.1073834 (2002).

12 Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. *Nat Genet* **37**, 934-935, doi:10.1038/ng1625 (2005).

13 Reid, S. *et al.* Biallelic mutations in PALB2 cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat Genet* **39**, 162-164, doi:10.1038/ng1947 (2007).

14 Strathdee, C. A., Gavish, H., Shannon, W. R. & Buchwald, M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. *Nature* **356**, 763-767, doi:10.1038/356763a0 (1992).

15 Timmers, C. *et al.* Positional cloning of a novel Fanconi anemia gene, FANCD2. *Mol Cell* **7**, 241-248 (2001).

16 Meetei, A. R. *et al.* X-linked inheritance of Fanconi anemia complementation group B. *Nat Genet* **36**, 1219-1224, doi:10.1038/ng1458 (2004).

17 Meetei, A. R. *et al.* A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat Genet* **37**, 958-963, doi:10.1038/ng1626 (2005).

18 Smogorzewska, A. *et al.* Identification of the FANCI protein, a monoubiquitinated FANCD2 paralog required for DNA repair. *Cell* **129**, 289-301, doi:10.1016/j.cell.2007.03.009 (2007).

19 Silva, J. M. *et al.* Identification of Long stress-induced non-coding transcripts that have altered expression in cancer. *Genomics* **95**, 355-362, doi:http://dx.doi.org/10.1016/j.ygeno.2010.02.009 (2010).

20 Radhakrishnan, A. & Green, R. Connections Underlying Translation and mRNA Stability. *Journal of Molecular Biology*, doi:http://dx.doi.org/10.1016/j.jmb.2016.05.025.

21 Hubert, L., Jr., Lin, Y., Dion, V. & Wilson, J. H. Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. *Hum Mol Genet* **20**, 4822-4830, doi:10.1093/hmg/ddr421 (2011).

22 Lin, Y., Hubert, L., Jr. & Wilson, J. H. Transcription destabilizes triplet repeats. *Mol Carcinog* **48**, 350-361, doi:10.1002/mc.20488 (2009).

23 Concannon, C. & Lahue, R. S. Nucleotide excision repair and the 26S proteasome function together to promote trinucleotide repeat expansions. *DNA Repair (Amst)* **13**, 42-49, doi:10.1016/j.dnarep.2013.11.004 (2014).

24 Highnam, G. *et al.* Accurate human microsatellite genotypes from high-throughput resequencing data using informed error profiles. *Nucleic Acids Res* **41**, e32, doi:10.1093/nar/gks981 (2013).

25 Huang, D. W. *et al.* DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res* **35**, W169-175, doi:10.1093/nar/gkm415 (2007).

26 Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).

27 Medicine, U. S. N. L. o. *Fanconi Anemia*, <https://ghr.nlm.nih.gov/condition/fanconi-anemia - genes> (

28 Takamoto, N., Leppert, P. C. & Yu, S. Y. Cell death and proliferation and its relation to collagen degradation in uterine involution of rat. *Connect Tissue Res* **37**, 163-175 (1998).

## SUPPLEMNTAL INFORMATION

## FIGURES

**Figure S1-1: Of 270 genes that were found to have significant difference in gene expression between the FA and FA_RV sample groups, 8 were found to have a high gene expression fold change that diverged significantly from the exponential fit.** These genes, with the most significant gene expression difference, were chosen to illustrate these significant changes. The full list of genes is found below.

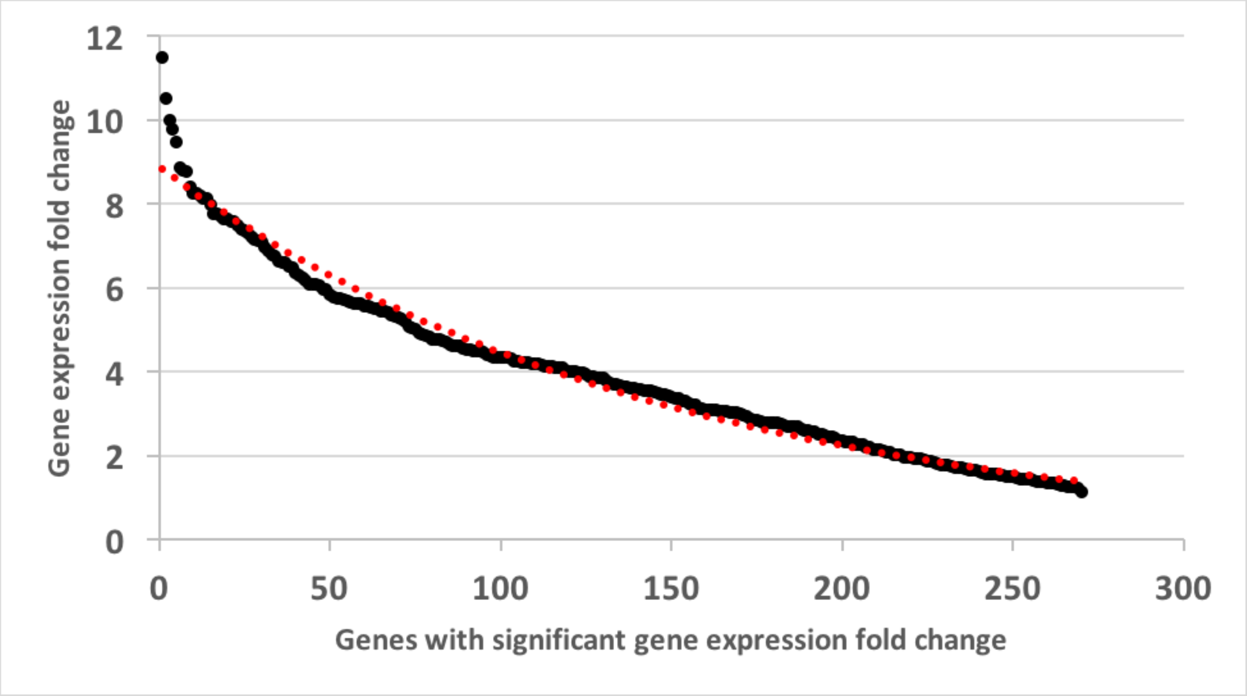


**Table S4-1: The log2(fold change) in gene expression for all the genes (16) that are known to be related to Fanconi Anemia for all 6 pairwise comparisons of 2 FA cell lines and FANCD2 RV corrected FA cell lines.** Only FANCD2 gene shows any discernible fold change in gene expression. This confirms that the FANCD2 gene was disrupted in the FA cell lines and restored in the FA_RV cell lines.

| Gene | Gene expression fold change | | | | | |
|---|---|---|---|---|---|---|
| | FA1/FA2 | FA1/FA_RV1 | FA2/FA_RV1 | FA1/FA_RV2 | FA2/FA_RV2 | FA_RV1/FA_RV2 |
| BRCA1 | 0.11 | 0.10 | -0.02 | 0.03 | -0.08 | -0.06 |
| BRCA2 | 0.12 | -0.18 | -0.30 | -0.06 | -0.18 | 0.12 |
| FANCD2 | -0.45 | 4.54 | 5.00 | 4.73 | 5.19 | 0.19 |
| PALB2 | 0.16 | 0.29 | 0.13 | 0.32 | 0.16 | 0.02 |
| FANCG | -0.08 | 0.15 | 0.23 | 0.11 | 0.19 | -0.04 |
| FANCF | 0.24 | 0.77 | 0.53 | 0.68 | 0.44 | -0.09 |
| FANCE | -0.03 | 0.59 | 0.62 | 0.64 | 0.66 | 0.04 |
| FANCC | 0.14 | -0.12 | -0.26 | -0.21 | -0.35 | -0.09 |
| FANCB | 0.23 | -0.38 | -0.61 | -0.34 | -0.57 | 0.04 |
| FANCA | -0.17 | 0.71 | 0.88 | 0.59 | 0.76 | -0.12 |
| BRIP1 | 0.10 | -0.81 | -0.91 | -1.03 | -1.14 | -0.22 |
| FANCM | 0.13 | 0.16 | 0.04 | 0.20 | 0.07 | 0.04 |
| FANCL | -0.07 | -0.05 | 0.01 | -0.22 | -0.16 | -0.17 |
| RAD51C | 0.14 | -0.56 | -0.70 | -0.59 | -0.72 | -0.02 |
| FANCI | 0.12 | 0.22 | 0.10 | 0.24 | 0.11 | 0.02 |
| SLX4 | 0.16 | -0.02 | -0.18 | 0.11 | -0.05 | 0.13 |

**Table S4-2: Distribution of the DNA variants with respect to the human genome reference (build - Grch38) in all the four Fanconi Anemia samples.** There were a significant number of variants found, many that are high impact. The classification of variants was done using the SNPEff program that annotates GATK VCF output. This also illustrates that many variants had already been acquired when the primary cells were taken from the patient to establish the defective and corrected cell lines.

| # | DNA variant event | FA1 | FA2 | FA_RV1 | FA_RV2 |
|---|---|---|---|---|---|
| 1 | intron_variant | 664039 | 650794 | 644703 | 631624 |
| 2 | downstream_gene_variant | 289232 | 283457 | 278135 | 276387 |
| 3 | upstream_gene_variant | 200121 | 194620 | 190464 | 188488 |
| 4 | 3_prime_UTR_variant | 59656 | 59324 | 57570 | 57931 |
| 5 | non_coding_transcript_exon_variant | 59060 | 58437 | 56577 | 56950 |
| 6 | synonymous_variant | 33460 | 33362 | 32162 | 32662 |
| 7 | missense_variant | 30275 | 30518 | 29686 | 31696 |
| 8 | intergenic_region | 30699 | 30131 | 29501 | 27426 |
| 9 | splice_region_variant | 14149 | 14279 | 13791 | 14389 |
| 10 | 5_prime_UTR_variant | 13877 | 13618 | 13058 | 13405 |
| 11 | sequence_feature | 5716 | 5569 | 5509 | 5455 |
| 12 | structural_interaction_variant | 2965 | 3044 | 2998 | 2987 |
| 13 | 5_prime_UTR_premature_start_codon_gain_variant | 1855 | 1815 | 1746 | 1831 |
| 14 | TF_binding_site_variant | 1175 | 1074 | 1120 | 1130 |
| 15 | frameshift_variant | 643 | 650 | 647 | 727 |
| 16 | stop_gained | 354 | 410 | 368 | 767 |
| 17 | splice_acceptor_variant | 391 | 424 | 343 | 428 |
| 18 | splice_donor_variant | 361 | 308 | 343 | 367 |
| 19 | disruptive_inframe_deletion | 317 | 307 | 311 | 312 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| 20 | disruptive_inframe_insertion | 227 | 229 | 219 | 210 |
| 21 | conservative_inframe_insertion | 170 | 149 | 167 | 176 |
| 22 | protein_protein_contact | 154 | 154 | 154 | 153 |
| 23 | conservative_inframe_deletion | 72 | 101 | 89 | 66 |
| 24 | start_lost | 70 | 76 | 61 | 68 |
| 25 | stop_lost | 52 | 52 | 56 | 55 |
| 26 | stop_retained_variant | 46 | 47 | 39 | 41 |
| 27 | non_coding_transcript_variant | 37 | 27 | 30 | 23 |
| 28 | intragenic_variant | 24 | 20 | 19 | 19 |
| 29 | TFBS_ablation | 10 | 15 | 11 | 5 |
| 30 | initiator_codon_variant | 4 | 4 | 4 | 4 |
| 31 | bidirectional_gene_fusion | 2 | 2 | 3 | 3 |
| 32 | gene_fusion | 1 | 1 | 1 | 1 |

**Table S4-3: Variability in high impact SNPs.** This table presents the calculation of pairwise SNP comparisons in and across the FA and RA_RV sample groups. The number of unique high impact SNPs that are found specific in each sample shows the heterogeneity of the Fanconi anemia samples.

| | # high impact SNPs | | SNPs specific to sample | | | | |
|---|---|---|---|---|---|---|---|
| Sample pair | S1 | S2 | S1 | S2 | Mean high impact | Mean specific SNPs | % of specific SNPs |
| FA1 - FA2 | 810 | 808 | 139 | 145 | 809 | 142 | 18 |
| FA1 - FA_RV1 | 810 | 786 | 178 | 155 | 798 | 167 | 21 |
| FA1 - FA_RV2 | 810 | 960 | 173 | 331 | 885 | 252 | 28 |
| FA2 - FA_RV1 | 808 | 786 | 182 | 155 | 797 | 169 | 21 |
| FA2 - FA_RV2 | 808 | 960 | 179 | 331 | 884 | 255 | 29 |
| FA_RV1 - FA_RV2 | 786 | 960 | 147 | 321 | 873 | 234 | 27 |

S1 – First sample of the samples in the sample pair column. S2 – Second sample of the samples in the sample pair column.

**Table S4-4: Variability in high impact INDELs.** This table presents the calculation of pairwise INDEL comparisons in and across the FA and RA_RV sample groups. The number of unique high impact INDELs that are found specific in each sample confirms the heterogeneity of the Fanconi anemia samples.

|  | # high impact INDELs |  | INDELs specific to sample |  |  |  |  |
|---|---|---|---|---|---|---|---|
| **Sample pair** | **S1** | **S2** | **S1** | **S2** | **Mean high impact** | **Mean specific INDELs** | **% of specific INDELs** |
| FA1 - FA2 | 205 | 209 | 25 | 28 | 207 | 26 | 13 |
| FA1 - FA_RV1 | 205 | 204 | 25 | 25 | 204 | 25 | 12 |
| FA1 - FA_RV2 | 205 | 209 | 34 | 38 | 207 | 36 | 17 |
| FA2 - FA_RV1 | 209 | 204 | 24 | 21 | 206 | 22 | 11 |
| FA2 - FA_RV2 | 209 | 209 | 30 | 29 | 209 | 29 | 14 |
| FA_RV1 - FA_RV2 | 204 | 209 | 27 | 31 | 206 | 29 | 14 |

S1 – First sample of the samples in the sample pair column. S2 – Second sample of the samples in the sample pair column.

**Table S4-5: A total of 270 genes were found to be statistically significantly differentially expressed between the FA and FA_RV sample groups.** The genes are sorted according to the gene expression fold change. This table furnishes the entire list of 270 genes that were differentially expressed. The very large number of genes, with very large expression changes illustrate the significant impact that the many variants have on the cell lines.

| Gene Id | Gene symbol | Genomic position | GE fold change | High expression in FA_RV |
|---|---|---|---|---|
| XLOC_003069 | - | chr1:239266341-239270392 | 11.48 | + |
| XLOC_014162 | COLEC12 | chr18:318126-500729 | 10.50 | - |
| XLOC_015674 | ZNF626 | chr19:20619938-20661596 | 10.00 | - |
| XLOC_033910 | FGF13 | chrX:138631570-139222889 | 9.79 | + |
| XLOC_011076 | MT1E | chr16:56625653-56627112 | 9.46 | - |
| XLOC_002609 | GLUL | chr1:182381703-182392206 | 8.85 | - |
| XLOC_018786 | COL6A3 | chr2:237324011-237422190 | 8.80 | - |
| XLOC_002401 | C1orf85 | chr1:156292686-156295689 | 8.78 | - |
| XLOC_022046 | CDCP1 | chr3:45082273-45146422 | 8.39 | - |
| XLOC_019959 | MX1 | chr21:41420557-41459214 | 8.25 | - |
| XLOC_006994 | CLEC2B | chr12:9852368-9869859 | 8.25 | - |
| XLOC_017545 | GALNT5 | chr2:157257547-157314078 | 8.18 | - |
| XLOC_008950 | LOC440157 | chr14:19298728-19303582 | 8.14 | + |
| XLOC_022006 | SUSD5 | chr3:33149927-33219215 | 8.14 | - |
| XLOC_021235 | MYD88 | chr3:38138477-38143022 | 7.97 | - |
| XLOC_011492 | XYLT1 | chr16:17102323-17470881 | 7.77 | - |
| XLOC_008904 | LOC102723726,TNFAIP2 | chr14:103121351-103137439 | 7.76 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_008934 | CRIP1 | chr14:105481517-105488789 | 7.70 | - |
| XLOC_031419 | - | chr8:122168573-122171056 | 7.65 | + |
| XLOC_029150 | PEG10 | chr7:94656324-94669695 | 7.65 | - |
| XLOC_007989 | ZIC2 | chr13:99981771-99986765 | 7.57 | + |
| XLOC_008412 | FLJ39632 | chr14:19076243-19096796 | 7.57 | + |
| XLOC_015095 | RCN3 | chr19:49527617-49543634 | 7.48 | - |
| XLOC_015837 | ATP1A3 | chr19:41966475-41994276 | 7.40 | + |
| XLOC_032446 | SLC35D2 | chr9:96313436-96383710 | 7.35 | - |
| XLOC_030357 | C8orf48 | chr8:13566842-13568288 | 7.31 | - |
| XLOC_017040 | EMILIN1 | chr2:27078566-27086403 | 7.22 | - |
| XLOC_030680 | - | chr8:122139981-122168546 | 7.16 | + |
| XLOC_017096 | QPCT | chr2:37344609-37373322 | 7.13 | - |
| XLOC_024738 | ZNF354C | chr5:179060406-179083771 | 7.09 | - |
| XLOC_003948 | PCBD1 | chr10:70882279-70888784 | 6.97 | - |
| XLOC_008753 | IFI27 | chr14:94110732-94116699 | 6.87 | - |
| XLOC_002724 | C1orf116 | chr1:207018520-207032761 | 6.78 | - |
| XLOC_031978 | PHYHD1 | chr9:128920894-128942041 | 6.76 | - |
| XLOC_010362 | CYP1A1 | chr15:74719541-74725610 | 6.63 | + |
| XLOC_006805 | HSPB8 | chr12:119178789-119194746 | 6.59 | - |
| XLOC_011239 | PLCG2 | chr16:81779293-81958294 | 6.59 | - |
| XLOC_026014 | CUL9 | chr6:43182174-43224587 | 6.50 | - |
| XLOC_027033 | LAMA4 | chr6:112107930-112306683 | 6.47 | - |
| XLOC_000273 | THEMIS2 | chr1:27872542-27886685 | 6.37 | - |
| XLOC_015133 | CD33 | chr19:51225078-51240019 | 6.29 | + |
| XLOC_002809 | ITPKB | chr1:226631689-226739327 | 6.25 | - |
| XLOC_032358 | - | chr9:78848554-78863649 | 6.17 | + |
| XLOC_012078 | ZFP3 | chr17:5078458-5096374 | 6.09 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_023003 | LIMCH1 | chr4:41359606-41700044 | 6.08 | + |
| XLOC_012093 | XAF1 | chr17:6755836-6775647 | 6.07 | - |
| XLOC_022879 | STK32B | chr4:5051545-5501001 | 6.06 | - |
| XLOC_000296 | SERINC2 | chr1:31409564-31434680 | 5.97 | - |
| XLOC_028763 | FAM20C | chr7:192958-260774 | 5.96 | - |
| XLOC_007833 | FREM2 | chr13:38687035-38887131 | 5.85 | + |
| XLOC_020937 | - | chr22:15854177-15855201 | 5.77 | + |
| XLOC_009291 | FBLN5 | chr14:91869409-91947702 | 5.75 | - |
| XLOC_014387 | - | chr18:14858969-14863974 | 5.75 | + |
| XLOC_009235 | MLH3 | chr14:75013763-75051532 | 5.71 | - |
| XLOC_006696 | - | chr12:98289603-98305388 | 5.69 | + |
| XLOC_034013 | CLIC2 | chrX:155276206-155334681 | 5.66 | - |
| XLOC_005225 | H19,MIR675 | chr11:1995175-1997835 | 5.62 | + |
| XLOC_007091 | TMTC1 | chr12:29500812-29784759 | 5.62 | + |
| XLOC_013665 | - | chr17:41110001-41112904 | 5.62 | + |
| XLOC_021327 | KLHDC8B | chr3:49171565-49176486 | 5.56 | - |
| XLOC_004657 | SERPING1 | chr11:57597553-57614853 | 5.55 | - |
| XLOC_022019 | SCN5A | chr3:38548061-38649675 | 5.54 | + |
| XLOC_007107 | KIF21A | chr12:39293227-39443390 | 5.51 | + |
| XLOC_030536 | TRIM55 | chr8:66127042-66175485 | 5.50 | - |
| XLOC_005657 | CD248 | chr11:66314486-66317044 | 5.44 | - |
| XLOC_010530 | - | chr15:30488358-30490284 | 5.43 | - |
| XLOC_017645 | COL3A1 | chr2:188974372-189012746 | 5.42 | - |
| XLOC_012724 | C1QTNF1 | chr17:79019208-79049788 | 5.34 | - |
| XLOC_011438 | PPL | chr16:4882506-4937135 | 5.33 | - |
| XLOC_006569 | DTX3 | chr12:57604326-57609804 | 5.28 | - |
| XLOC_012992 | USP32P2 | chr17:18511261-18531380 | 5.25 | - |
| XLOC_019537 | TSPY26P | chr20:32186497-32190526 | 5.21 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_023332 | GUCY1B3 | chr4:155758973-155807631 | 5.07 | + |
| XLOC_005429 | PAMR1 | chr11:35431826-35530300 | 5.05 | - |
| XLOC_012186 | TRPV2 | chr17:16415541-16437003 | 5.00 | - |
| XLOC_019717 | LAMA5 | chr20:62309059-62367312 | 4.93 | - |
| XLOC_023129 | ANXA3 | chr4:78551587-78610451 | 4.89 | - |
| XLOC_021138 | FANCD2 | chr3:10026383-10108291 | 4.87 | + |
| XLOC_015509 | COL5A3 | chr19:9959560-10010471 | 4.84 | - |
| XLOC_029774 | NSUN5 | chr7:73302515-73308867 | 4.77 | - |
| XLOC_000113 | TNFRSF1B | chr1:12166942-12209220 | 4.76 | - |
| XLOC_011062 | MMP2 | chr16:55479168-55506674 | 4.76 | - |
| XLOC_025963 | MAPK13 | chr6:36130483-36144524 | 4.74 | - |
| XLOC_025635 | DSP | chr6:7540450-7586713 | 4.70 | - |
| XLOC_017001 | KCNS3 | chr2:17877846-17932985 | 4.65 | + |
| XLOC_004032 | PPP1R3C | chr10:91628439-91633101 | 4.63 | - |
| XLOC_030912 | NEFL | chr8:24950954-24956869 | 4.62 | - |
| XLOC_024221 | CCDC152 | chr5:42756805-42811922 | 4.61 | - |
| XLOC_004154 | ABLIM1 | chr10:114431109-114779903 | 4.56 | - |
| XLOC_022647 | BDH1 | chr3:197509782-197573323 | 4.53 | - |
| XLOC_015498 | ZNF560 | chr19:9466354-9498603 | 4.52 | - |
| XLOC_000682 | GSTM1 | chr1:109687795-109693745 | 4.49 | - |
| XLOC_002107 | GSTM3 | chr1:109733931-109741038 | 4.49 | - |
| XLOC_007460 | TMEM119 | chr12:108589845-108598118 | 4.48 | - |
| XLOC_019438 | FERMT1 | chr20:6074844-6123544 | 4.47 | - |
| XLOC_005789 | ME3 | chr11:86441022-86672636 | 4.41 | - |
| XLOC_032395 | SEMA4D | chr9:89360790-89498014 | 4.36 | - |
| XLOC_017690 | CDK15 | chr2:201790453-201895550 | 4.35 | - |
| XLOC_023402 | SLC25A4 | chr4:185143262-185150384 | 4.35 | - |

|  |  |  |  |  |
|---|---|---|---|---|
| XLOC_014991 | APOE | chr19:44905781-44909393 | 4.34 | - |
| XLOC_001078 | PRRX1 | chr1:170662727-170739400 | 4.34 | - |
| XLOC_033198 | FAM133A | chrX:93674012-93712274 | 4.33 | + |
| XLOC_020061 | TMPRSS15 | chr21:18268866-18477284 | 4.30 | + |
| XLOC_003939 | AIFM2 | chr10:70052600-70132934 | 4.26 | - |
| XLOC_009515 | SNRPN,SNURF | chr15:24823646-24978582 | 4.25 | - |
| XLOC_001304 | MARK1 | chr1:220528182-220664457 | 4.23 | + |
| XLOC_007768 | TNFRSF19 | chr13:23570369-23676105 | 4.22 | - |
| XLOC_019508 | NINL | chr20:25452696-25585531 | 4.21 | - |
| XLOC_008016 | TEX29 | chr13:111320667-111344247 | 4.20 | - |
| XLOC_029826 | SEMA3A | chr7:83957817-84492768 | 4.20 | - |
| XLOC_018505 | DPP4 | chr2:161992240-162074542 | 4.18 | - |
| XLOC_008768 | BDKRB2 | chr14:96204797-96244329 | 4.15 | - |
| XLOC_017714 | ADAM23 | chr2:206443543-206621130 | 4.14 | + |
| XLOC_020507 | APOBEC3G | chr22:39077004-39087743 | 4.14 | - |
| XLOC_025666 | GMPR | chr6:16238579-16295549 | 4.13 | - |
| XLOC_013997 | ANKRD30B | chr18:14748239-14854702 | 4.11 | + |
| XLOC_001631 | MFAP2 | chr1:16974501-16981586 | 4.09 | - |
| XLOC_014195 | PIEZO2 | chr18:10670189-11149534 | 4.09 | - |
| XLOC_019417 | ADAM33 | chr20:3667972-3682131 | 4.03 | - |
| XLOC_009672 | GCHFR | chr15:40764086-40767713 | 4.00 | - |
| XLOC_002292 | CTSK | chr1:150796207-150808441 | 4.00 | - |
| XLOC_011431 | CDIP1 | chr16:4510674-4538815 | 4.00 | - |
| XLOC_019053 | SPTLC3 | chr20:13008953-13169001 | 3.99 | - |
| XLOC_020436 | TCN2 | chr22:30607082-30627060 | 3.97 | - |
| XLOC_021152 | PPARG | chr3:12287849-12434356 | 3.95 | - |
| XLOC_009076 | FOXA1 | chr14:37589551-37595120 | 3.89 | + |
| XLOC_019744 | STMN3 | chr20:63639704-63654977 | 3.88 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_001759 | COL16A1 | chr1:31652246-31704242 | 3.86 | - |
| XLOC_002087 | COL11A1 | chr1:102876466-103108496 | 3.86 | - |
| XLOC_032002 | AIF1L | chr9:131096475-131123152 | 3.84 | + |
| XLOC_011603 | STX1B | chr16:30989255-31010508 | 3.79 | - |
| XLOC_018582 | FRZB | chr2:182833274-182866770 | 3.74 | - |
| XLOC_024507 | TGFBI | chr5:136028894-136063818 | 3.70 | - |
| XLOC_024998 | F2RL2 | chr5:76403254-76708132 | 3.70 | - |
| XLOC_007942 | SLAIN1 | chr13:77697736-77764242 | 3.66 | + |
| XLOC_029710 | GRB10 | chr7:50590062-50793462 | 3.65 | - |
| XLOC_032192 | ELAVL2 | chr9:23690098-23826344 | 3.64 | + |
| XLOC_016959 | RSAD2 | chr2:6877664-6898232 | 3.61 | - |
| XLOC_017085 | LTBP1 | chr2:32947153-33399509 | 3.60 | - |
| XLOC_032494 | ABCA1 | chr9:104781001-104928246 | 3.57 | - |
| XLOC_000275 | XKR8 | chr1:27959992-27968093 | 3.57 | - |
| XLOC_020790 | FOXRED2 | chr22:36487185-36507101 | 3.56 | + |
| XLOC_021231 | CTDSPL | chr3:37862152-37984469 | 3.54 | - |
| XLOC_013620 | SECTM1 | chr17:82321023-82334045 | 3.54 | - |
| XLOC_007808 | MEDAG | chr13:30882561-30932608 | 3.52 | - |
| XLOC_020014 | COL6A1 | chr21:45981748-46005049 | 3.50 | - |
| XLOC_019958 | MX2 | chr21:41362022-41408943 | 3.47 | - |
| XLOC_007542 | OASL | chr12:121020291-121039242 | 3.46 | - |
| XLOC_007614 | CHFR | chr12:132840351-132887618 | 3.42 | - |
| XLOC_002410 | BCAN | chr1:156640499-156661441 | 3.39 | - |
| XLOC_012316 | CCL2 | chr17:34255276-34257203 | 3.37 | - |
| XLOC_028862 | GPNMB | chr7:23246685-23275110 | 3.35 | - |
| XLOC_031933 | OLFML2A | chr9:124777137-124814891 | 3.30 | - |
| XLOC_022011 | TRANK1 | chr3:36826816-36945057 | 3.29 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_031238 | MTSS1 | chr8:124550769-124728507 | 3.25 | + |
| XLOC_006782 | OAS1 | chr12:112906933-112919907 | 3.22 | - |
| XLOC_002412 | CRABP2 | chr1:156699605-156713174 | 3.22 | - |
| XLOC_014114 | SERPINB2 | chr18:63887704-63903890 | 3.13 | - |
| XLOC_024819 | CMBL | chr5:10277594-10308056 | 3.13 | - |
| XLOC_020015 | COL6A2 | chr21:46098118-46132849 | 3.10 | - |
| XLOC_003369 | PLAU | chr10:73909968-73922777 | 3.10 | - |
| XLOC_008041 | LINC00452 | chr13:113883636-113926238 | 3.10 | - |
| XLOC_018250 | CAPG | chr2:85394747-85414074 | 3.08 | - |
| XLOC_032636 | CRAT | chr9:129094793-129110791 | 3.07 | - |
| XLOC_026169 | POU3F2 | chr6:98834703-98838790 | 3.06 | + |
| XLOC_022447 | HLTF | chr3:149029382-149102823 | 3.05 | - |
| XLOC_018202 | PAIP2B | chr2:71182663-71227103 | 3.04 | + |
| XLOC_021488 | COL8A1 | chr3:99638364-99799220 | 3.02 | - |
| XLOC_022698 | - | chr3:75483604-75489296 | 3.02 | + |
| XLOC_002345 | S100A4 | chr1:153543495-153545806 | 2.99 | - |
| XLOC_006784 | OAS2 | chr12:112978346-113011723 | 2.96 | - |
| XLOC_033765 | TMSB15A | chrX:102513681-102516771 | 2.94 | + |
| XLOC_018606 | SDPR | chr2:191834304-191847280 | 2.88 | + |
| XLOC_004976 | FAT3 | chr11:92224640-92896533 | 2.85 | + |
| XLOC_026946 | ELOVL4 | chr6:79914811-79947598 | 2.85 | - |
| XLOC_027008 | CD24 | chr6:106969830-106975454 | 2.83 | + |
| XLOC_015120 | EMC10 | chr19:50466787-50505802 | 2.80 | - |
| XLOC_010197 | FBN1 | chr15:48408305-48645788 | 2.80 | - |
| XLOC_000565 | IFI44 | chr1:78649791-78664078 | 2.80 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_031709 | PGM5 | chr9:68355188-68531061 | 2.80 | - |
| XLOC_015457 | C3 | chr19:6677834-6720651 | 2.77 | - |
| XLOC_004350 | EPS8L2 | chr11:706116-727727 | 2.74 | - |
| XLOC_003177 | KIAA1217 | chr10:23694745-24557525 | 2.71 | - |
| XLOC_018507 | FAP | chr2:162114440-162243535 | 2.70 | - |
| XLOC_000564 | IFI44L | chr1:78620381-78646255 | 2.70 | - |
| XLOC_004340 | IFITM1 | chr11:313990-315272 | 2.68 | - |
| XLOC_008248 | KCTD12 | chr13:76880168-76886405 | 2.68 | + |
| XLOC_026181 | AIM1 | chr6:106360807-106570460 | 2.64 | - |
| XLOC_002167 | TBX15 | chr1:118882758-118989556 | 2.61 | - |
| XLOC_004418 | TRIM22 | chr11:5689586-5710863 | 2.59 | - |
| XLOC_018593 | COL5A2 | chr2:189031914-189179879 | 2.58 | - |
| XLOC_000213 | EPHB2 | chr1:22710769-22915330 | 2.56 | - |
| XLOC_020308 | USP18 | chr22:18149953-18177397 | 2.52 | - |
| XLOC_029590 | IGF2BP3 | chr7:23310208-23470491 | 2.51 | - |
| XLOC_001732 | IFI6 | chr1:27666060-27672213 | 2.49 | - |
| XLOC_020801 | RAC2 | chr22:37225260-37244299 | 2.46 | - |
| XLOC_013434 | CYB561 | chr17:63432303-63446363 | 2.45 | - |
| XLOC_032599 | ANGPTL2 | chr9:126914773-127223166 | 2.43 | - |
| XLOC_031215 | ENPP2 | chr8:119557076-119673404 | 2.37 | - |
| XLOC_006963 | C1R | chr12:7080208-7092570 | 2.36 | - |
| XLOC_008336 | COL4A1 | chr13:110148962-110307149 | 2.34 | - |
| XLOC_017455 | INHBB | chr2:120346142-120351807 | 2.34 | + |
| XLOC_031995 | ASS1 | chr9:130444706-130501274 | 2.33 | - |
| XLOC_023921 | DDX60 | chr4:168216290-168318807 | 2.28 | - |
| XLOC_004926 | DGAT2 | chr11:75768732-75801536 | 2.27 | + |
| XLOC_004725 | FADS2 | chr11:61799624-61867354 | 2.27 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_005840 | MMP1 | chr11:102783675-102843611 | 2.20 | - |
| XLOC_012791 | NXN | chr17:799312-979775 | 2.19 | - |
| XLOC_023355 | CPE | chr4:165378944-165498330 | 2.15 | - |
| XLOC_017529 | KIF5C | chr2:148875222-149026759 | 2.14 | + |
| XLOC_019324 | CDH4 | chr20:61252425-61940617 | 2.14 | + |
| XLOC_009361 | AHNAK2 | chr14:104924849-104978357 | 2.12 | - |
| XLOC_006972 | SLC2A3 | chr12:7919227-7936296 | 2.07 | - |
| XLOC_000012 | ISG15 | chr1:1013466-1014540 | 2.07 | - |
| XLOC_026239 | GJA1 | chr6:121435598-121449727 | 2.03 | - |
| XLOC_022561 | CAMK2N2 | chr3:184249656-184293031 | 2.02 | + |
| XLOC_023165 | HERC6 | chr4:88378685-88443097 | 2.02 | - |
| XLOC_014654 | CD97 | chr19:14380590-14408725 | 1.96 | - |
| XLOC_026329 | SASH1 | chr6:148212113-148552049 | 1.96 | - |
| XLOC_030733 | LY6K | chr8:142700110-142726973 | 1.95 | + |
| XLOC_006783 | OAS3 | chr12:112938443-112973251 | 1.94 | - |
| XLOC_031767 | CTSL | chr9:87726058-87731469 | 1.93 | - |
| XLOC_027083 | MOXD1 | chr6:132296054-132401525 | 1.93 | - |
| XLOC_026400 | MLLT4 | chr6:167826916-167972023 | 1.91 | - |
| XLOC_002032 | GBP1 | chr1:89052303-89065360 | 1.88 | - |
| XLOC_005937 | MCAM | chr11:119308523-119318377 | 1.86 | + |
| XLOC_026130 | TPBG | chr6:82363205-82367422 | 1.85 | - |
| XLOC_018508 | IFIH1 | chr2:162267078-162318708 | 1.81 | - |
| XLOC_002314 | S100A10 | chr1:151982909-151994238 | 1.79 | - |
| XLOC_023670 | ANTXR2 | chr4:79901217-80073472 | 1.78 | - |
| XLOC_029195 | PCOLCE | chr7:100586332-100608175 | 1.77 | - |

| | | | | |
|---|---|---|---|---|
| XLOC_026113 | CD109 | chr6:73694235-73828317 | 1.76 | - |
| XLOC_008010 | COL4A2 | chr13:110307283-110513027 | 1.72 | - |
| XLOC_003446 | IFIT1 | chr10:89392545-89406487 | 1.72 | - |
| XLOC_005802 | CHORDC1 | chr11:90200428-90223364 | 1.71 | + |
| XLOC_000513 | PGM1 | chr1:63593275-63660245 | 1.69 | - |
| XLOC_030901 | TNFRSF10D | chr8:23135587-23164030 | 1.66 | + |
| XLOC_002230 | TXNIP | chr1:145992434-145996631 | 1.66 | - |
| XLOC_008358 | GAS6 | chr13:113815609-113864073 | 1.65 | - |
| XLOC_005644 | RNASEH2C | chr11:65711995-65720938 | 1.62 | + |
| XLOC_013563 | LGALS3BP | chr17:78971252-78979979 | 1.60 | - |
| XLOC_029851 | SAMD9 | chr7:93099512-93118023 | 1.57 | - |
| XLOC_008580 | LGALS3 | chr14:55129216-55145430 | 1.57 | - |
| XLOC_005191 | IFITM3 | chr11:319672-320914 | 1.56 | - |
| XLOC_012598 | MRC2 | chr17:62627400-62693601 | 1.55 | - |
| XLOC_005821 | MRE11A | chr11:94417299-94493874 | 1.52 | + |
| XLOC_001666 | HSPG2 | chr1:21812264-21937257 | 1.52 | - |
| XLOC_006321 | EMP1 | chr12:13196667-13216774 | 1.51 | - |
| XLOC_003133 | OPTN | chr10:13100074-13138291 | 1.50 | - |
| XLOC_030737 | LY6E | chr8:143018484-143022410 | 1.50 | - |
| XLOC_007246 | ITGA5 | chr12:54395260-54419266 | 1.48 | - |
| XLOC_025847 | HLA-A | chr6:29887759-29945884 | 1.45 | - |
| XLOC_003443 | IFIT2 | chr10:89301948-89309276 | 1.44 | - |
| XLOC_032210 | DDX58 | chr9:32455301-32526324 | 1.44 | - |
| XLOC_004997 | CEP57 | chr11:95790460-95832693 | 1.44 | + |
| XLOC_026696 | HLA-B | chr6:31353871-31357212 | 1.42 | - |
| XLOC_030925 | CLU | chr8:27596916-27615031 | 1.39 | - |
| XLOC_019496 | CST3 | chr20:23627896-23638048 | 1.38 | - |

|  |  |  |  |  |
|---|---|---|---|---|
| XLOC_025239 | DPYSL3 | chr5:147390807-147510056 | 1.37 | - |
| XLOC_002343 | S100A6 | chr1:153534596-153536241 | 1.36 | - |
| XLOC_029148 | COL1A2 | chr7:94394560-94431232 | 1.35 | - |
| XLOC_013255 | VAT1 | chr17:43014566-43022442 | 1.35 | - |
| XLOC_021728 | MME | chr3:155079646-155183729 | 1.32 | - |
| XLOC_005507 | UBE2L6 | chr11:57551654-57568330 | 1.30 | - |
| XLOC_005851 | CASP4 | chr11:104942866-104968598 | 1.28 | + |
| XLOC_005005 | YAP1 | chr11:102110419-102233423 | 1.27 | + |
| XLOC_009729 | EID1 | chr15:48823678-48963507 | 1.26 | - |
| XLOC_019547 | E2F1 | chr20:33675485-33686404 | 1.25 | + |
| XLOC_002315 | S100A11 | chr1:152032505-152037035 | 1.22 | - |
| XLOC_004759 | STIP1 | chr11:64185271-64204548 | 1.13 | + |

**Table S4-6: Pairwise comparison of the number of exons in the most expressed transcript of genes in 4 Fanconi Anemia samples.** The % of expressed genes with varying exon counts when comparing within sample groups is 23.5 but % of expressed genes with varying exon counts when comparing samples across sample groups is about 28.5%. This table explains the effect of a dysfunctional FANCD2 gene on the transcriptome.

| Sample pair | Transcripts with equal exons | Expressed genes | Transcripts with varying exon count |
|---|---|---|---|
| FA1 - FA2 | 12221 | 15903 | 23 |
| FA_RV1 - FA_RV2 | 12016 | 15903 | 24 |
| FA1 - FA_RV1 | 11311 | 15903 | 29 |
| FA1 - FA_RV2 | 11351 | 15903 | 29 |
| FA2 - FA_RV1 | 11401 | 15903 | 28 |
| FA2 - FA_RV2 | 11432 | 15903 | 28 |

**Table S4-7: Ontological gene enrichment analysis of the 270 genes that were found to be differentially expressed the FA and FA_RV sample groups.** The exceptional p-values seen for these GO ontology terms, especially for terms involved in infectious disease, development and cancer illustrate the potential connections with the observed symptoms and diseases seen in FA patients. The DAVID online bioinformatics tool was used to this GO enrichment analysis.

| Term type | Term | Count | List Total | PValue | Genes |
|---|---|---|---|---|---|
| BP | GO:0060337~type I interferon signaling pathway | 17 | 241 | 4.71E-16 | IFITM1, IFITM3, OAS3, HLA-A, RSAD2, OAS1, HLA-B, OAS2, IFIT2, OASL, IFIT1, IFI27, ISG15, XAF1, MX1, MX2, IFI6 |
| BP | GO:0030574~collagen catabolic process | 16 | 241 | 1.17E-14 | COL4A2, COL4A1, MRC2, COL3A1, COL5A3, COL5A2, MMP2, MMP1, CTSL, CTSK, COL6A3, COL6A2, COL1A2, COL6A1, COL8A1, COL11A1 |
| BP | GO:0009615~response to virus | 19 | 241 | 1.73E-14 | IFIH1, CYP1A1, IFITM1, IFITM3, CLU, OAS3, RSAD2, OAS1, IFI44, OAS2, TRIM22, DDX58, IFIT2, OASL, IFIT1, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | MYD88, DDX60, MX1, MX2 |
| BP | GO:0030198~extracellular matrix organization | 23 | 241 | 7.75E-14 | COL4A2, COL4A1, COL3A1, FBN1, HSPG2, BCAN, OLFML2A, COL5A3, COL16A1, COL5A2, EMILIN1, LAMA4, ITGA5, LAMA5, FBLN5, COL6A3, TGFBI, COL1A2, COL6A2, COL6A1, MFAP2, COL8A1, COL11A1 |
| BP | GO:0007155~cell adhesion | 28 | 241 | 5.26E-10 | MTSS1, CCL2, FERMT1, BCAN, CDH4, LGALS3BP, FAP, TGFBI, COL6A3, COL6A2, COL6A1, CD24, GPNMB, COL8A1, ADAM23, COL16A1, MCAM, TPBG, GAS6, EMILIN1, LAMA4, PGM5, FREM2, ITGA5, CD33, SUSD5, SEMA4D, THEMIS2 |
| BP | GO:0051607~defense response to virus | 17 | 241 | 1.90E-09 | IFITM1, IFITM3, OAS3, RSAD2, APOBEC3G, IFI44L, OAS1, OAS2, TRIM22, IFIT2, OASL, |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
|  |  |  |  |  | IFIT1, ISG15, DDX60, MX1, MX2, GBP1 |
| BP | GO:0045071~negative regulation of viral genome replication | 10 | 241 | 3.99E-09 | IFIT1, OASL, ISG15, IFITM1, IFITM3, OAS3, RSAD2, APOBEC3G, OAS1, MX1 |
| BP | GO:0071230~cellular response to amino acid stimulus | 8 | 241 | 7.58E-06 | COL4A1, ASS1, COL3A1, COL1A2, COL6A1, COL16A1, MMP2, COL5A2 |
| BP | GO:0035987~endodermal cell differentiation | 6 | 241 | 3.58E-05 | COL4A2, ITGA5, COL6A1, COL8A1, COL11A1, MMP2 |
| BP | GO:0060333~interferon-gamma-mediated signaling pathway | 8 | 241 | 6.83E-05 | OASL, OAS3, HLA-A, OAS1, OAS2, HLA-B, TRIM22, GBP1 |
| BP | GO:0001525~angiogenesis | 13 | 241 | 9.88E-05 | COL4A2, CCL2, LAMA5, ITGA5, FAP, TGFBI, HSPG2, MCAM, COL8A1, TNFAIP2, MMP2, PLAU, EPHB2 |
| BP | GO:0022617~extracellular matrix disassembly | 8 | 241 | 1.06E-04 | CTSL, CTSK, CAPG, FBN1, HSPG2, BCAN, MMP2, MMP1 |
| BP | GO:0006955~immune response | 18 | 241 | 1.19E-04 | SECTM1, CRIP1, CCL2, ENPP2, C3, IFITM3, OAS3, HLA-A, OAS1, C1R, OAS2, HLA-B, TRIM22, TNFRSF1B, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | TNFRSF10D, CD24, SEMA4D, IFI6 |
| BP | GO:0009612~response to mechanical stimulus | 7 | 241 | 3.01E-04 | TXNIP, INHBB, CCL2, COL3A1, PPARG, BDKRB2, PIEZO2 |
| BP | GO:0070208~protein heterotrimerization | 4 | 241 | 0.0012 | C1QTNF1, COL1A2, COL6A2, COL6A1 |
| BP | GO:0010716~negative regulation of extracellular matrix disassembly | 3 | 241 | 0.0012 | FAP, CST3, DPP4 |
| BP | GO:0019941~modification-dependent protein catabolic process | 3 | 241 | 0.0020 | ISG15, UBE2L6, CHFR |
| BP | GO:0030199~collagen fibril organization | 5 | 241 | 0.0023 | COL3A1, COL1A2, COL5A3, COL11A1, COL5A2 |
| BP | GO:0006952~defense response | 6 | 241 | 0.0027 | INHBB, CST3, COLEC12, HLA-B, MX1, MX2 |
| BP | GO:0008637~apoptotic mitochondrial changes | 4 | 241 | 0.0027 | IFIT2, AIFM2, SLC25A4, CD24 |
| BP | GO:0045087~innate immune response | 15 | 241 | 0.0039 | DDX58, IFIH1, CASP4, MYD88, LGALS3, DDX60, CLU, PPARG, APOBEC3G, SERPING1, COLEC12, C1R, HLA-B, MX1, MX2 |
| BP | GO:0006508~proteolysis | 16 | 241 | 0.0041 | C3, ADAM23, MME, C1R, MMP2, PCOLCE, MMP1, CTSL, CTSK, CASP4, BACE2, FAP, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | ADAM33, DPP4, PLAU, TMPRSS15 |
| BP | GO:0035456~response to interferon-beta | 3 | 241 | 0.0069 | IFITM1, IFITM3, XAF1 |
| BP | GO:0050776~regulation of immune response | 8 | 241 | 0.0080 | IFITM1, C3, CD33, CLEC2B, COL3A1, HLA-A, COL1A2, HLA-B |
| BP | GO:0035455~response to interferon-alpha | 3 | 241 | 0.0085 | IFITM1, IFITM3, MX2 |
| BP | GO:0035457~cellular response to interferon-alpha | 3 | 241 | 0.0085 | IFIT2, OAS1, GAS6 |
| BP | GO:0032480~negative regulation of type I interferon production | 4 | 241 | 0.0088 | DDX58, IFIH1, ISG15, UBE2L6 |
| BP | GO:0007229~integrin-mediated signaling pathway | 6 | 241 | 0.0138 | ADAM23, LAMA5, ITGA5, COL3A1, ADAM33, COL16A1 |
| BP | GO:0051091~positive regulation of sequence-specific DNA binding transcription factor activity | 6 | 241 | 0.0174 | DDX58, FOXA1, PPARG, TRIM22, ANXA3, ZIC2 |
| BP | GO:0033627~cell adhesion mediated by integrin | 3 | 241 | 0.0189 | ITGA5, FBN1, COL16A1 |
| BP | GO:0006024~glycosaminoglycan biosynthetic process | 4 | 241 | 0.0220 | XYLT1, GALNT5, SLC35D2, HSPG2 |
| BP | GO:0090280~positive regulation of calcium ion import | 3 | 241 | 0.0240 | CCL2, LGALS3, TRPV2 |
| BP | GO:0030334~regulation of cell migration | 5 | 241 | 0.0255 | DDX58, LAMA4, LAMA5, ENPP2, DPYSL3 |

| | | | | | |
|---|---|---|---|---|---|
| BP | GO:0030168~platelet activation | 6 | 241 | 0.0256 | F2RL2, RAC2, PLCG2, COL3A1, COL1A2, GAS6 |
| BP | GO:0043691~reverse cholesterol transport | 3 | 241 | 0.0268 | APOE, CLU, ABCA1 |
| BP | GO:0044267~cellular protein metabolic process | 6 | 241 | 0.0273 | TGFBI, CST3, HSPG2, UBE2L6, MMP2, MMP1 |
| BP | GO:1902998~positive regulation of neurofibrillary tangle assembly | 2 | 241 | 0.0284 | APOE, CLU |
| BP | GO:0034344~regulation of type III interferon production | 2 | 241 | 0.0284 | DDX58, IFIH1 |
| BP | GO:0048285~organelle fission | 2 | 241 | 0.0284 | MX1, MX2 |
| BP | GO:0060700~regulation of ribonuclease activity | 2 | 241 | 0.0284 | OAS3, OAS1 |
| BP | GO:1900221~regulation of beta-amyloid clearance | 2 | 241 | 0.0284 | APOE, CLU |
| BP | GO:1902622~regulation of neutrophil migration | 2 | 241 | 0.0284 | MYD88, RAC2 |
| BP | GO:0032355~response to estradiol | 6 | 241 | 0.0328 | TXNIP, ASS1, C3, FOXA1, CST3, BDH1 |
| BP | GO:0008544~epidermis development | 5 | 241 | 0.0348 | CRABP2, IFT172, DSP, POU3F2, EMP1 |
| BP | GO:0042730~fibrinolysis | 3 | 241 | 0.0358 | SERPINB2, SERPING1, PLAU |
| BP | GO:0060279~positive regulation of ovulation | 2 | 241 | 0.0423 | INHBB, PLAU |
| BP | GO:0071400~cellular response to oleic acid | 2 | 241 | 0.0423 | ASS1, DGAT2 |
| BP | GO:0002486~antigen processing and presentation of endogenous | 2 | 241 | 0.0423 | HLA-A, HLA-B |

| | | | | | |
|---|---|---|---|---|---|
| | peptide antigen via MHC class I via ER pathway, TAP-independent | | | | |
| BP | GO:0035583~sequestering of TGFbeta in extracellular matrix | 2 | 241 | 0.0423 | LTBP1, FBN1 |
| BP | GO:0060741~prostate gland stromal morphogenesis | 2 | 241 | 0.0423 | CRIP1, FOXA1 |
| BP | GO:0070458~cellular detoxification of nitrogen compound | 2 | 241 | 0.0423 | GSTM1, GSTM3 |
| BP | GO:0039528~cytoplasmic pattern recognition receptor signaling pathway in response to virus | 2 | 241 | 0.0423 | DDX58, IFIH1 |
| BP | GO:0051260~protein homooligomerization | 7 | 241 | 0.0439 | KCNS3, GLUL, CEP57, C1QTNF1, DPYSL3, COLEC12, KCTD12 |
| BP | GO:0045880~positive regulation of smoothened signaling pathway | 3 | 241 | 0.0457 | FOXA1, IFT172, PRRX1 |
| BP | GO:0016032~viral process | 10 | 241 | 0.0476 | DDX58, IFIH1, ISG15, SLC25A4, HLA-A, RSAD2, APOBEC3G, HLA-B, TRIM22, MMP1 |
| BP | GO:0001501~skeletal system development | 6 | 241 | 0.0482 | COL3A1, FBN1, COL1A2, BCAN, FRZB, COL5A2 |
| MF | GO:0005201~extracellular matrix structural constituent | 10 | 232 | 4.10E-07 | COL4A2, LAMA4, COL4A1, COL3A1, FBN1, COL1A2, BCAN, COL5A3, COL11A1, COL5A2 |

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| MF | GO:0001730~2'-5'-oligoadenylate synthetase activity | 4 | 232 | 1.11E-05 | OASL, OAS3, OAS1, OAS2 |
| MF | GO:0005178~integrin binding | 10 | 232 | 1.81E-05 | ADAM23, LAMA5, ITGA5, FAP, FBLN5, TGFBI, COL3A1, FBN1, GPNMB, COL16A1 |
| MF | GO:0005518~collagen binding | 7 | 232 | 2.00E-04 | CTSL, CTSK, C1QTNF1, TGFBI, MRC2, COL5A3, PCOLCE |
| MF | GO:0003725~double-stranded RNA binding | 7 | 232 | 2.19E-04 | DDX58, IFIH1, OASL, DDX60, OAS3, OAS1, OAS2 |
| MF | GO:0048407~platelet-derived growth factor binding | 4 | 232 | 4.25E-04 | COL4A1, COL3A1, COL1A2, COL6A1 |
| MF | GO:0016740~transferase activity | 8 | 232 | 5.47E-04 | GSTM1, OASL, SPTLC3, XYLT1, GALNT5, OAS3, OAS1, OAS2 |
| MF | GO:0042802~identical protein binding | 20 | 232 | 0.0012 | S100A4, MTSS1, ASS1, PCBD1, PPARG, CST3, OPTN, EPHB2, DDX58, TRIM55, GLUL, GSTM3, MYD88, APOE, HSPB8, COL1A2, POU3F2, NEFL, DPP4, GBP1 |
| MF | GO:0005509~calcium ion binding | 22 | 232 | 0.0014 | S100A4, S100A6, ME3, LTBP1, ENPP2, CD248, PAMR1, FAM20C, FBN1, HSPG2, S100A11, S100A10, C1R, CDH4, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | MMP1, ANXA3, GAS6, FAT3, FBLN5, AIF1L, NINL, RCN3 |
| MF | GO:0005102~receptor binding | 14 | 232 | 0.0016 | MTSS1, CCL2, C3, HLA-A, GJA1, HLA-B, ABCA1, CRAT, GAS6, EPHB2, LAMA4, SEMA4D, ANGPTL2, DPP4 |
| MF | GO:0004252~serine-type endopeptidase activity | 11 | 232 | 0.0019 | CTSL, CTSK, C3, FAP, PAMR1, C1R, MMP2, DPP4, PLAU, MMP1, TMPRSS15 |
| MF | GO:0030023~extracellular matrix constituent conferring elasticity | 3 | 232 | 0.0029 | COL4A1, FBN1, EMILIN1 |
| MF | GO:0008270~zinc ion binding | 30 | 232 | 0.0030 | ABLIM1, S100A6, IFIH1, ENPP2, PPARG, MME, APOBEC3G, OAS1, OAS2, MMP2, MMP1, PEG10, CUL9, MT1E, ADAM33, XAF1, CRIP1, ADAM23, CA12, DTX3, HLTF, TRIM22, VAT1, QPCT, DDX58, TRIM55, CPE, LIMCH1, CHORDC1, CHFR |
| MF | GO:0030674~protein binding, bridging | 6 | 232 | 0.0047 | COL1A2, DSP, OPTN, TRIM22, COL11A1, NEFL |

| | | | | | |
|---|---|---|---|---|---|
| MF | GO:0050840~extracellular matrix binding | 4 | 232 | 0.0057 | CD248, TGFBI, OLFML2A, COL11A1 |
| MF | GO:0030246~carbohydrate binding | 9 | 232 | 0.0075 | LGALS3, CD33, CLEC2B, CD248, GALNT5, MRC2, BCAN, CD24, AIM1 |
| MF | GO:0043394~proteoglycan binding | 3 | 232 | 0.0101 | CTSL, CTSK, COL5A3 |
| MF | GO:0015075~ion transmembrane transporter activity | 3 | 232 | 0.0120 | S100A6, TRPV2, GJA1 |
| MF | GO:0005515~protein binding | 105 | 232 | 0.0139 | S100A4, S100A6, LTBP1, CRABP2, FAM20C, FGF13, APOBEC3G, ITPKB, MYD88, ISG15, APOE, ELOVL4, C8ORF48, FAP, CUL9, TGFBI, MX1, TMPRSS15, KIF5C, HLA-A, CST3, SERPING1, OPTN, MOXD1, GRB10, HSPB8, CD33, COL1A2, DSP, CHORDC1, EMP1, GBP1, EID1, IFIH1, IFITM1, CLU, MME, OAS1, IGF2BP3, BDKRB2, ABCA1, PPP1R3C, PEG10, PPL, AHNAK2, LGALS3, ADAM23, FOXRED2, MRC2, UBE2L6, |

| | | | | | |
|---|---|---|---|---|---|
| | | | | | S100A10, COL16A1, GAS6, EMILIN1, TRIM55, DDX58, C1ORF116, ITGA5, PLCG2, ANTXR2, IFI6, PLAU, E2F1, PPARG, GJA1, CDCP1, MLH3, FAM133A, USP18, DDX60, YAP1, DPP4, SLC25A4, CLIC2, HLTF, INHBB, CTSL, CTSK, FANCD2, SEMA4D, CHFR, C3, COL3A1, RSAD2, PCOLCE, PAIP2B, TNFRSF1B, CEP57, C1QTNF1, CD24, COL8A1, GPNMB, SCN5A, NEFL, GCHFR, TXNIP, FBN1, DTX3, HSPG2, ELAVL2, TRIM22, IFIT2, IFIT1, FBLN5, CAPG |
| MF | GO:0004222~metalloendopeptidase activity | 6 | 232 | 0.0238 | ADAM23, FAP, MME, ADAM33, MMP2, MMP1 |
| MF | GO:0005044~scavenger receptor activity | 4 | 232 | 0.0288 | LGALS3BP, ENPP2, COLEC12, TMPRSS15 |
| MF | GO:0046977~TAP binding | 2 | 232 | 0.0419 | HLA-A, HLA-B |
| MF | GO:0005031~tumor necrosis factor-activated receptor activity | 3 | 232 | 0.0449 | TNFRSF1B, TNFRSF10D, TNFRSF19 |

| | | | | | |
|---|---|---|---|---|---|
| MF | GO:0008022~protein C-terminus binding | 7 | 232 | 0.0462 | SASH1, FBLN5, HSPG2, STIP1, YAP1, OPTN, NEFL |

Count: The number of genes in the input list that was found to be associated with the given GO term. List total: The total number of genes in the input list that are associated with GO terms in the DAVID database.

**Table S4-8: Pathway analysis of 82 genes that were found to be associated with genomic variants that were specific to the FA sample group.** Pathway analysis also confirms the association of the differentially expressed genes with the observed signs, symptoms and diseases typically seen in FA patients. The pathway analysis was performed using the REACTOME pathway online tool.

| Pathway name | #Entities found | #Entities total | Entities pValue |
|---|---|---|---|
| Translocation of ZAP-70 to Immunological synapse | 4 | 42 | 0.0002 |
| Phosphorylation of CD3 and TCR zeta chains | 4 | 45 | 0.0003 |
| PD-1 signaling | 4 | 45 | 0.0003 |
| Generation of second messenger molecules | 4 | 58 | 0.0007 |
| Downstream TCR signaling | 5 | 124 | 0.0016 |
| MHC class II antigen presentation | 5 | 142 | 0.0028 |
| TCR signaling | 5 | 146 | 0.0032 |
| Costimulation by the CD28 family | 4 | 96 | 0.0042 |
| Synthesis of PIPs at the late endosome membrane | 2 | 21 | 0.0091 |
| TFAP2 (AP-2) family regulates transcription of growth factors and their receptors | 2 | 21 | 0.0091 |
| Metabolism of Angiotensinogen to Angiotensins | 2 | 26 | 0.0136 |
| Synthesis of PIPs at the early endosome membrane | 2 | 29 | 0.0167 |
| Activation of Matrix Metalloproteinases | 2 | 35 | 0.0237 |
| Defective SLC22A18 causes lung cancer (LNCR) and embryonal rhabdomyosarcoma 1 (RMSE1) | 1 | 4 | 0.0266 |
| Vpr-mediated induction of apoptosis by mitochondrial outer membrane permeabilization | 1 | 4 | 0.0266 |
| Adaptive Immune System | 13 | 1076 | 0.0284 |
| Interferon gamma signaling | 4 | 176 | 0.0317 |
| Transport of the SLBP independent Mature mRNA | 1 | 5 | 0.0332 |

| | | | |
|---|---|---|---|
| Defective CYP19A1 causes Aromatase excess syndrome (AEXS) | 1 | 5 | 0.0332 |
| Formation of apoptosome | 1 | 5 | 0.0332 |
| SMAC-mediated apoptotic response | 1 | 5 | 0.0332 |
| SMAC-mediated dissociation of IAP:caspase complexes | 1 | 5 | 0.0332 |
| SMAC binds to IAPs | 1 | 5 | 0.0332 |
| Lysosome Vesicle Biogenesis | 2 | 43 | 0.0345 |
| Transport of the SLBP Dependant Mature mRNA | 1 | 6 | 0.0397 |
| PLCG1 events in ERBB2 signaling | 1 | 6 | 0.0397 |
| Signaling by Overexpressed Wild-Type EGFR in Cancer | 1 | 6 | 0.0397 |
| Inhibition of Signaling by Overexpressed EGFR | 1 | 6 | 0.0397 |
| Activation of caspases through apoptosome-mediated cleavage | 1 | 7 | 0.0462 |
| Cytochrome c-mediated apoptotic response | 1 | 7 | 0.0462 |
| Transcriptional regulation by the AP-2 (TFAP2) family of transcription factors | 2 | 52 | 0.0486 |

Entities found: Number of genes in the input list that were found to be associated with a pathway. Entities total:

Total number of genes that are associated with a given pathway in the reactome database.

**Table S4-9: Pathway analysis of 618 genes that were found to be associated with variants that were commonly found in all 4 Fanconi Anemia samples.** The genomic variants are found in pathways that were highly correlated with the significant pathways identified through expression changes. The pathway analysis was performed using the REACTOME pathway online tool.

| Pathway name | #Entities found | #Entities total | Entities pValue |
|---|---|---|---|
| Interferon gamma signaling | 19 | 176 | 0.0016 |
| Antigen Presentation: Folding, assembly and peptide loading of class I MHC | 11 | 102 | 0.0142 |
| Defective GALNT3 causes familial hyperphosphatemic tumoral calcinosis (HFTC) | 4 | 20 | 0.0182 |
| Defective GALNT12 causes colorectal cancer 1 (CRCS1) | 4 | 20 | 0.0182 |
| Defective C1GALT1C1 causes Tn polyagglutination syndrome (TNPS) | 4 | 20 | 0.0182 |
| Translocation of ZAP-70 to Immunological synapse | 6 | 42 | 0.0193 |
| Endosomal/Vacuolar pathway | 9 | 82 | 0.0227 |
| Interferon alpha/beta signaling | 13 | 140 | 0.0244 |
| PD-1 signaling | 6 | 45 | 0.0259 |
| TRKA activation by NGF | 2 | 5 | 0.0260 |
| Olfactory Signaling Pathway | 31 | 432 | 0.0272 |
| Interferon Signaling | 22 | 291 | 0.0354 |
| APC truncation mutants are not K63 polyubiquitinated | 1 | 1 | 0.0483 |

Entities found: Number of genes in the input list that were found to be associated with a pathway. Entities total: Total number of genes that are associated with a given pathway in the reactome database.

# Chapter 5: Concluding remarks and future directions

# RESEARCH OVERVIEW AND CONCLUSION

The thesis presented here consists of three comprehensive studies that shed light on different research paradigms of repetitive DNA, microsatellites (MST) in particular, which have been relatively understudied, despite being linked to several disorders such as Huntington disease and Fragile X syndrome[1-3]. This thesis represents a much-needed improvement in building computational tools, a novel utilization of target specific sequencing techniques to facilitate MST genomics research, and a possible validation of the involvement of MST instability (MSI) in cancer progression based on MSI signatures for congenital inheritance of cancer[4]. A part of this work also investigates the role of molecular level cellular processes that may give rise to MST instability, other than the well-studied mismatch repair process[5-8].

The results of this work can be useful in three areas of research: 1. The utilization of global MST enrichment to further genomic research, 2. The power of target specific MST enrichment and its utility in the development of high-accuracy companion diagnostics in cancer (and other diseases) and 3. Recommendations for further MST instability research using DNA repair deficient cellular systems.

## Global MST enrichment:

Since the advent of genomic research, the focus of biologists on the coding regions of the genome and the development of tools for studying these protein coding genomic sequences have resulted in a large gap in our understanding of repeat regions that form about 47% of the human genome[4]. This has not only directly contributed to the existence of the incomplete knowledge of the euchromatic DNA, but has also contributed to the failure to address the possibility of uncovering important functional elements that could be hidden among the repetitive regions of the human genome. Since the completion of the human genome project, at which point 94% of the euchromatic DNA was known, recent efforts have advanced the euchromatic DNA framework to 99%[9,10]. Using our newly developed molecular techniques and computational methods we successfully sequenced novel MST sequences in the human genome, demonstrating the potential for detecting novel MST sequences that can help in the possible identification of the 1% of unknown euchromatic DNA. Specifically, our assembly technique can detect longer MSTs, which is not possible through currently available genomic enrichment and computational methods[4]. This work also identifies novel functional elements that could potentially shed light on biological issues such as the existence of multiple isoforms of rRNA.

The results, while showing potential for aiding in the completion of the euchromatic human DNA, also holds the promise of advancing the completion of other genomes that have even higher content of repetitive DNA sequences. Figure S2-4 shows the distribution of the length of the novel contiguous sequences that were assembled using our global MST enrichment technique. While sequencing of MSTs by whole genome or exome enrichments usually allow the detection of 100 to 200 bp MST sequences, the global MST enrichment, coupled with appropriate assembling tools, enables the detection of longer MST sequence regions (250 to 500 bp) (Figure S2-4).

**This technique shows promise in 3 ways:**

1. With the advent of better sequencing technologies, in the future, that can provide longer sequencing reads without compromising the read depth, longer novel contiguous sequences (contigs) can be assembled. This will aid in the sequencing of more functional elements, directly adding to the knowledge of human genetics and

disease. For example, 37 novel contigs assembled (Table 2-3) using the global MST enrichment technique were found in 1kGP RNA-Seq samples. The contig with the highest number of hits in the 10 RNA-Seq samples was found in large clone DNA sequences that were sequenced as a part of the original (pre-nextgen) human genome project but was, conspicuously, not found in the reference genome. This contig is 370 bp and a portion of which was found to match with high sequence identity (99%) to multiple rRNA subunits. Recent studies in parasite models suggest the diverse evolution of rRNA subunits to accommodate specific functions that are, specific to subnuclear compartments[11]. The study by Deveau *et al* shows the diversification of two rRNA subunits in a protozoan parasite. By the targeted sequencing of MST, novel functional elements identification is possible and their discoveries, like the rRNA sequence, have the potential to contribute to answering numerous biomolecular questions.

2. This technique may contribute even more when used for genomes of plants and amphibians. The repetitive DNA content of amphibians, plants and insects have been shown to be higher than that of the human genome[12]. The probability of finding functional elements and genes in the refractory regions of these repeat heavy genomes is high.

3. One of the most interesting findings of this study was the abundance of two pentamer MST repeats. While the telomere regions are known to have a high percentage of the hexamer repeat GGGTTA, the finding of these pentamer repeats, AATGG and GTGGA, and their comparative abundance sheds new light onto the repeat rich telomere and centromere regions of the chromosomes. Telomere shortening has been linked to aging, cancer and other diseases[13-16]. Access to the functional elements of these repeat heavy regions can possibly help understand the

reason behind the connection between telomere shortening and associated disorders[16].

**Target specific MST enrichment:**

Coupled with the inefficiency (especially low depth coverage, that limits accuracy) of normal exome enrichment methods for sequencing MSTs, the high sequence (40X) coverage requirement of MST genotyping programs present a highly challenging problem[4,17,18]. The MST genotyping methodology involves specific statistical steps that allow the accommodation of low MST sequence coverage from exome enrichments to be efficiently utilized[19-22]. While the Garner lab has shown convincingly the statistical significance of the MST instability in breast, ovarian and brain cancers, the validation of these findings remained to be completed[19-22] By using Illumina's target specific enrichment, this validation was made possible in a lung cancer scenario that involved lung cancer and normal sample sets apart from the publically downloaded LUSC and LUAD germline samples from the TCGA and 1kGP (Tables S3-3 and S3-4). This work resulted in a signature set of 21 MST loci (all located in gene regions), 57% or more of which were found in a cancer genotype with a sensitivity ratio of 0.93 (Figure 3-3).

**Applications and future directions:**

This work, while providing the first validation for a possible link between specific microsatellite mutations and cancer, is also a technique that can be adapted to be used as a companion diagnostic in oncological health care facilities. Next to the wide spread usage of the Bethesda markers for forensic identifications, this technique presents an opportunity for introducing a MST based marker kit for disease diagnostics.

While MSI based markers for colon cancer is well known, the involvement of MSI in cancer was not confirmed until recently [23-25]. This target specific MST enrichment will show the possibility of developing a next generation sequencing (ultra-high resolution) based diagnostic kit for all MSI positive disorders. More importantly, as demonstrated in Chapter 3, it is known that ultra-high-depth sequencing and normal exome sequencing can indicate different genotypes. This method, hence, not only presents an accurate way of MST genotyping but also introduces the opportunity to develop a genome wide MST specific enrichment sequencing process. Also, this enrichment method will allow us to revise observations from previous genome wide MST genotyping studies that may be incorrect, and will also provide a strong platform to understand in high resolution the heterozygosity-homozygosity ratio of cancer tissues and MST[17,18].

Perhaps one of the most appealing direct applications of this work, specific to lung cancer, is the early diagnosis of lung cancer in never smokers. About 20% of lung cancer cases with non-small cell lung cancer are reported to be non-smokers[26]. The theory that these cases could have inherited lung cancer signature from their parents is highly probable. As this work is based on genotyping MST from the germline cells of patients, this diagnostic kit could be applied to detect any future onsets of lung cancer in never smokers with a lung cancer familial trait[27,28].

## Fanconi anemia

### Fanconi anemia, as a DNA repair deficiency disorder:

Fanconi anemia is an autosomal recessive disorder that occurs mainly in Jewish populations[29]. When one of the 16 genes, that code for the DNA inter-strand crosslink repair complex of the Fanconi anemia (FA) pathway, harbor a protein modifying mutation, FA is likely to occur[30,31]. Much information about the FA genes that are involved in crosslink repair have been reported, thanks to the availability of a variety patient samples, which has divided the FA genes into complementation groups[32-37]. Although the FA core complex has been studied extensively, the details of this pathway are not entirely understood. It is well established that the FA pathway recruits DNA repair proteins to crosslink sites, however, there are also studies that suggest that the FA pathway might also be involved in the maintenance of general genomic stability[38]. Also, recent studies reveal the involvement of the FA pathway in MST instability[39-41].

The hypothesis of this project is twofold: 1. By sequencing PD20-FA cell lines with defective FANCD2 gene and PD20-FA cell lines with retrovirus corrected FANCD2 gene and understanding their genomic variants and difference in gene expression patterns, we will be able to understand the effects of a dysfunctional FANCD2 gene on the genome and the transcriptome and by that come to a first-hand understanding of the other roles of a functioning FA pathway. 2. By performing MST genotyping on FA samples we can estimate any possible effects of a dysfunctional FA pathway on MST instability.

**Over all conclusion:**

Genomic variant analysis on the ratio of SNPs vs. INDELs show an increased percentage of INDELs in the FA samples, compared to 1kGP samples. This suggests an increased amount of DNA lesions which can be the result of uncorrected DNA crosslinks (Table 4-3). The difference in SNP and INDEL variations within biological replicates (tables 4-5 and 4-6) suggest a high heterogeneity and a significant effect of a dysfunctional FA pathway on the transcriptome. Unlike non-existent gene expression difference between biological replicates, the exon count (transcript length) of the most expressed transcript of genes show that a large number of genes (23.5%) within biological replicates have varying exon counts (Table S4-6) which is consistent with the observed genomic damage. This, again, demonstrates the pronounced effect of a dysfunctional FA pathway on the transcriptome. Also, the significant increase in the percentage of MSTs with minor alleles in the FA samples (Table 4-4), compared to the 1kGP samples, suggest that the FA pathway, when not functional, could also be another source of MST instability, along with the DNA mismatch repair mechanisms.

**Future directions:**

Owing to the significant increase in the MST instability (MST with minor alleles) in the FA samples, the hypothesized possibility of the FA pathway contributing to MST instability becomes more evident. While this confirms the FA cell line system to be a prospective platform for studying the FA pathway and MSI, a few factors should be included in future experimental designs. A time-lapse longitudinal study that involves DNA extraction from FA and FA_RV samples at different time points or at different passage numbers during cell culturing is needed to understand the progression rate of DNA damage in FA cells and the progression rate of DNA damage in FA_RV cells.

Quantifying DNA breaks that occur in the form of chromosomal translocations in FA genomes are important to better understand inter-strand crosslink occurrences. While SNPs and INDELs form an essential variant determination paradigm, the quantification of DNA translocations is a critical way of determining the effect of a corrected or dysfunctional DNA repair gene on the genome. Recent advances in translocation sequencing present an opportunity to quantify double strand break caused chromosomal translocations[42-44]. DNA samples from cells that are treated with crosslink inducing agents such as mitomycin C were previously thought to be unsuitable for sequencing, but with the advent of translocation sequencing, the genome-wide effect of mitomycin and the counter effect of a functional FA pathway can be studied with high precision.

# REFERENCES

1 Science, U. S. N. L. o. *Huntington Disease*, <https://ghr.nlm.nih.gov/condition/huntington-disease>

2 Ellegren, H. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet* **5**, 435-445, doi:10.1038/nrg1348 (2004).

3 Budworth, H. & McMurray, C. T. A brief history of triplet repeat diseases. *Methods Mol Biol* **1010**, 3-17, doi:10.1007/978-1-62703-411-1_1 (2013).

4 Fonville, N. C. *et al.* Genomic leftovers: identifying novel microsatellites, over-represented motifs and functional elements in the human genome. *Sci Rep* **6**, 27722, doi:10.1038/srep27722 (2016).

5 Hite, J. M., Eckert, K. A. & Cheng, K. C. 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 (1996).

6 McMurray, C. T. Mechanisms of trinucleotide repeat instability during human development. *Nat Rev Genet* **11**, 786-799, doi:10.1038/nrg2828 (2010).

7 Huang, J. *et al.* MSH6 and MSH3 are rarely involved in genetic predisposition to nonpolypotic colon cancer. *Cancer Res* **61**, 1619-1623 (2001).

8 Iaccarino, I., Marra, G., Palombo, F. & Jiricny, J. hMSH2 and hMSH6 play distinct roles in mismatch binding and contribute differently to the ATPase activity of hMutSalpha. *EMBO J* **17**, 2677-2686, doi:10.1093/emboj/17.9.2677 (1998).

9 *The Human Genome Project*, <https://www.genome.gov/12011238/an-overview-of-the-human-genome-project/>

10 International Human Genome Sequencing, C. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931-945, doi:10.1038/nature03001 (2004).

11 Devaux, S. *et al.* Diversification of function by different isoforms of conventionally shared RNA polymerase subunits. *Mol Biol Cell* **18**, 1293-1301, doi:10.1091/mbc.E06-09-0841 (2007).

12 RepeatMasker. *RepeatMasker Genomic Datasets*, <http://www.repeatmasker.org/genomicDatasets/RMGenomicDatasets.html>

13 Chang, A. C. *et al.* Telomere shortening and metabolic compromise underlie dystrophic cardiomyopathy. *Proc Natl Acad Sci U S A* **113**, 13120-13125, doi:10.1073/pnas.1615340113 (2016).

14 Epel, E. S. *et al.* Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci U S A* **101**, 17312-17315, doi:10.1073/pnas.0407162101 (2004).

15 Mourkioti, F. *et al.* Role of telomere dysfunction in cardiac failure in Duchenne muscular dystrophy. *Nat Cell Biol* **15**, 895-904, doi:10.1038/ncb2790 (2013).

16 Raval, A. *et al.* Reversibility of Defective Hematopoiesis Caused by Telomere Shortening in Telomerase Knockout Mice. *PLoS One* **10**, e0131722, doi:10.1371/journal.pone.0131722 (2015).

17 McIver, L. J., Fondon, J. W., 3rd, Skinner, M. A. & Garner, H. R. Evaluation of microsatellite variation in the 1000 Genomes Project pilot studies is indicative of the quality and utility of the raw data and alignments. *Genomics* **97**, 193-199, doi:10.1016/j.ygeno.2011.01.001 (2011).

18 McIver, L. J., McCormick, J. F., Martin, A., Fondon, J. W., 3rd & Garner, H. R. Population-scale analysis of human microsatellites reveals novel sources of exonic variation. *Gene* **516**, 328-334, doi:10.1016/j.gene.2012.12.068 (2013).

19 Fonville, N. C., Vaksman, Z., McIver, L. J. & Garner, H. R. Population analysis of microsatellite genotypes reveals a signature associated with ovarian cancer. *Oncotarget* **6**, 11407-11420, doi:10.18632/oncotarget.2933 (2015).

20 Karunasena, E. *et al.* 'Cut from the same cloth': Shared microsatellite variants among cancers link to ectodermal tissues-neural tube and crest cells. *Oncotarget* **6**, 22038-22047, doi:10.18632/oncotarget.4194 (2015).

21 Karunasena, E. *et al.* Somatic intronic microsatellite loci differentiate glioblastoma from lower-grade gliomas. *Oncotarget* **5**, 6003-6014, doi:10.18632/oncotarget.2076 (2014).

22 McIver, L. J., Fonville, N. C., Karunasena, E. & Garner, H. R. Microsatellite genotyping reveals a signature in breast cancer exomes. *Breast Cancer Res Treat* **145**, 791-798, doi:10.1007/s10549-014-2908-8 (2014).

23 Gologan, A. *et al.* Performance of the revised Bethesda guidelines for identification of colorectal carcinomas with a high level of microsatellite instability. *Arch Pathol Lab Med* **129**, 1390-1397, doi:10.1043/1543-2165(2005)129[1390:POTRBG]2.0.CO;2 (2005).

24 Rodriguez-Moranta, F. *et al.* Clinical performance of original and revised Bethesda guidelines for the identification of MSH2/MLH1 gene carriers in patients with newly diagnosed colorectal cancer: proposal of a new and simpler set of recommendations. *Am J Gastroenterol* **101**, 1104-1111, doi:10.1111/j.1572-0241.2006.00522.x (2006).

25 Hause, R. J., Pritchard, C. C., Shendure, J. & Salipante, S. J. Classification and characterization of microsatellite instability across 18 cancer types. *Nat Med* **22**, 1342-1350, doi:10.1038/nm.4191 (2016).

26 Institute, N. C. *SEER Stat Fact Sheets: Lung and Bronchus Cancer*, <http://seer.cancer.gov/statfacts/html/lungb.html> (2016).

27 Li, Y. *et al.* Genetic variants and risk of lung cancer in never smokers: a genome-wide association study. *The Lancet. Oncology* **11**, 321-330, doi:10.1016/S1470-2045(10)70042-5 (2010).

28 Rivera, G. A. & Wakelee, H. Lung Cancer in Never Smokers. *Advances in experimental medicine and biology* **893**, 43-57, doi:10.1007/978-3-319-24223-1_3 (2016).

29 Kutler, D. I. *et al.* A 20-year perspective on the International Fanconi Anemia Registry (IFAR). *Blood* **101**, 1249-1256, doi:10.1182/blood-2002-07-2170 (2003).

30 Deans, A. J. & West, S. C. DNA interstrand crosslink repair and cancer. *Nat Rev Cancer* **11**, 467-480, doi:10.1038/nrc3088 (2011).

31 Walden, H. & Deans, A. J. The Fanconi anemia DNA repair pathway: structural and functional insights into a complex disorder. *Annu Rev Biophys* **43**, 257-278, doi:10.1146/annurev-biophys-051013-022737 (2014).

32 de Winter, J. P. *et al.* Isolation of a cDNA representing the Fanconi anemia complementation group E gene. *Am J Hum Genet* **67**, 1306-1308, doi:10.1016/S0002-9297(07)62959-0 (2000).

33 Dorsman, J. C. *et al.* Identification of the Fanconi anemia complementation group I gene, FANCI. *Cell Oncol* **29**, 211-218 (2007).

34 Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group J. *Nat Genet* **37**, 934-935, doi:10.1038/ng1625 (2005).

35 Meetei, A. R. *et al.* X-linked inheritance of Fanconi anemia complementation group B. *Nat Genet* **36**, 1219-1224, doi:10.1038/ng1458 (2004).

36 Meetei, A. R. *et al.* A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. *Nat Genet* **37**, 958-963, doi:10.1038/ng1626 (2005).

37 Strathdee, C. A., Gavish, H., Shannon, W. R. & Buchwald, M. Cloning of cDNAs for Fanconi's anaemia by functional complementation. *Nature* **356**, 763-767, doi:10.1038/356763a0 (1992).

38 Michl, J., Zimmer, J. & Tarsounas, M. Interplay between Fanconi anemia and homologous recombination pathways in genome integrity. *EMBO J* **35**, 909-923, doi:10.15252/embj.201693860 (2016).

39 Hubert, L., Jr., Lin, Y., Dion, V. & Wilson, J. H. Xpa deficiency reduces CAG trinucleotide repeat instability in neuronal tissues in a mouse model of SCA1. *Hum Mol Genet* **20**, 4822-4830, doi:10.1093/hmg/ddr421 (2011).

40 Lin, Y., Hubert, L., Jr. & Wilson, J. H. Transcription destabilizes triplet repeats. *Mol Carcinog* **48**, 350-361, doi:10.1002/mc.20488 (2009).

41 Concannon, C. & Lahue, R. S. Nucleotide excision repair and the 26S proteasome function together to promote trinucleotide repeat expansions. *DNA Repair (Amst)* **13**, 42-49, doi:10.1016/j.dnarep.2013.11.004 (2014).

42 Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107-119, doi:10.1016/j.cell.2011.07.049 (2011).

43 Hu, J. *et al.* Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat Protoc* **11**, 853-871, doi:10.1038/nprot.2016.043 (2016).

44 Klein, I. A. *et al.* Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell* **147**, 95-106, doi:10.1016/j.cell.2011.07.048 (2011).

## APPENDIX: Copyrights

The chapter 2 is published in Scientific Reports. Below is the copyright policy of the Scientific Reports journal.

"*Scientific Reports* does not require authors to assign copyright of their published original research papers to the journal. Articles are published under a CC BY license (Creative Commons Attribution 4.0 International License). The CC BY license allows for maximum dissemination and re-use of open access materials and is preferred by many research funding bodies. Under this license users are free to share (copy, distribute and transmit) and remix (adapt) the contribution including for commercial purposes, providing they attribute the contribution in the manner specified by the author or licensor"