12-21-2016

Influence of Mutation Frequency on Mutation Profile in Colon Cancer

Michael J. Gooch
goochmi@gmail.com

Influence of Mutation Frequency on Mutation Profile in Colon Cancer

Michael James Gooch, PhD

University of Connecticut, 2017

Abstract

Colorectal cancers display a vast range in the number of mutations per tumor. It is reasonable to assume that most mutations found in tumors are harmless passenger mutations and that only a small fraction of mutations found in these tumors are driver mutations that are responsible for initiation, progression and maintenance of the tumor. My research project was to compare types of mutations, genes targeted and specificity of gene targeting in high versus low mutation frequency tumors. **My hypothesis is that there are qualitative and quantitative differences in the mutation spectrum of colorectal cancers that can be distinguished by the overall number of mutations detected in the tumors.** To address this hypothesis, I analyzed whole-genome sequencing data from 223 colorectal cancers in the Cancer Genome Atlas. I compared cancers with >1000 mutations per tumor to those that had 0-999 mutations per tumor. I found that while the majority of genes mutated were found in both groups, distinct subsets of mutated genes did occur in the two sample sets that were mutated more than expected and more than in the other group. I found that those in the low mutation frequency set had a high specificity for mutations in known cancer genes while those in the high frequency set showed no significant clustering of mutations in known cancer genes. Altogether my data supported that there were qualitative and quantitative differences in the mutation spectrum of colorectal cancers based on the frequency of mutation in the individual tumors.

Influence of Mutation Frequency on Mutation Profile in Colon Cancer

Michael James Gooch

B.S. Rutgers University, **2009**

A Dissertation

Submitted in Partial Fulfillment of the

Requirements for the Degree of

Doctor of Philosophy

at the

University of Connecticut

2017

APPROVAL PAGE

Doctor of Philosophy Dissertation

Influence of Mutation Frequency on Mutation Profile in Colon Cancer

Presented by

Michael James Gooch, B.S.

Major Advisor _____
Dr. Marc Hansen

Associate Advisor _____
Dr. William Mohler

Associate Advisor _____
Dr. Asis Das

Associate Advisor _____
Dr. Christopher Heinen

Associate Advisor _____
Dr. Gordon Carmichael

University of Connecticut

2017

Acknowledgements

I have finally completed my PhD requirements, and would not have been able to do so without the aid of many people. Both major labs I have been a member of during the development of this research project contributed significantly to my experience and successful completion of my dissertation.

My first thesis advisor, Dr. Richard Everson, helped me to develop the idea of examining the spectrum of mutations in cancer, while other projects in his lab helped to increase my experience and exposure to bioinformatics methods and available analysis software. He and I share a vision of the significant role computational analysis of genetic data will play in the future of biological and medical research.

My current thesis advisor, Dr. Marc Hansen, also helped me bring the ideas behind this project into fruition. He helped me stay on course, and provided key guidance that enabled me to solve problems that came up during analysis, and identify efficient and accurate solutions.

Cynthia Alander, our lab technician, is a great resource, a good friend, and great moral support. It was always a pleasure to share lab space with her.

I would like to thank Statisticians Dr. James Grady, and Dr. Yu-Bo Wang for their significant assistance in choosing viable analytical methods which enabled a numerical approach that created interesting results.

I would like to thank Dr. Ion Moraru and Jeffrey Dutton of CCAM for all of their technical support and computational resources that I have used over the last several years. Their help was instrumental in my acquisition of bioinformatics experience, and having access to an environment capable of the computational demands of some projects.

I would like to thank Dr. Pramod Srivastiva and the Neag Cancer Center for continuing to support me through my research and dissertation preparation. Without this aid I could not have completed the program.

My committee members, Dr. Gordon Carmichael, Dr. Asis K. Das, Dr. Christopher D. Heinen, and Dr. William Mohler, have also been an enormous resource for both my dissertation work, and personal and professional development. Their questions and observations improved the quality of my research in ways that should not be taken for granted.

I also thank my friends and family for being supportive when times were difficult, and being a significant source of motivation and moral support.

I dedicate this dissertation to my wonderfully independent and cheerful daughter, Kaelyn Siqi Gooch.

Table of Contents

List of Figures

Chapter II

Chapter III

Chapter 1

Genome sequencing, Cancer, The Cancer Genome Atlas, Mutation Rates, and Mutation Profiles

Introduction

There were three technological breakthroughs that were critical to the success of my thesis project. The first of these was the successful completion of the human genome sequencing project that provided the template upon which my project could be built. Accompanying and enabling the human genome sequencing project was the development of technologies that enabled rapid and accurate sequencing of entire genomes. Finally, the development of highly curated databases storing both the normal genome sequence as well as the variation within the “normal” human genome together with the genomic sequences of human tumors was a critical preliminary step in my project. Parallel with this technological development, there was the conceptual advance that mutations in certain genes were critical in driving the process of tumorigenesis while mutations in other genes were simply passengers that were carried along as part of a stochastic process. Together these prior steps laid the foundation for my work.

History of Human Genome Sequencing Project

The Human Genome Project was initiated in 1990 with the goal of obtaining the full human genomic DNA sequence (1). The project did not seek to sequence heterochromatic regions such as centromeres or telomeres, but rather focused on euchromatic regions (1,2). When the project began, the NIH Genome program was headed by James Watson, who was succeeded by Francis Collins in 1993 (3).

There was also a privately funded quest launched by Craig Venter and the firm Celera Genomics in 1998 (4). It was able to proceed much faster and more cheaply than the publicly funded HGP by making use of data that was released by the HGP. This effort was a profit seeking one, and Celera attempted to obtain patents on a large number of

genes (5). Celera promised to publish their results but there was suspicion that they would not permit free redistribution or scientific use of the data (5–7). Their intentions compelled the publicly funded project to publish their results first (6,7). In March 2000, then president Bill Clinton announced that the human genome sequence should be made freely available to all researchers and that access should be unencumbered (8).

Initial drafts of the human genome became available in June 2000 and working drafts were completed by February 2001(1,6,7,9). The project was declared complete in April 2003 (10). The sequencing speeds during this 13-year time period increased dramatically as technologies improved, allowing the project to be completed 2 years earlier than initially planned. The genomic sequence continues to be updated and revised as directed by improvements in technology and accuracy of the underlying data (11).

Sequencing technologies

DNA sequencing technologies began being developed and used in the 1970s (12). Initially, labor intensive processes like Maxam & Gilbert sequencing resulted in maximum read lengths of about 100 bases at the time of its development (12). This was eventually overtaken by Sanger's enzyme-driven sequencing process, which was able to achieve significantly reduced manual labor requirements and increased read lengths as improvements in gel and dye technologies were developed (12,13). Successive improvements have led to current processes including SOLiD and Illumina next generation high throughput sequencing technologies (13–15), and very powerful variants of Sanger's original sequencing method that have been greatly improved by modern technological advances (13,16). Parallel improvements in multiple technologies, including electrophoresis gels, DNA base identification and detection, ranging from radiolabels, to

detecting pyrophosphate released during nucleobase-specific reactions, to various technology-specific fluorescent dyes, while machine driven automation technologies have brought DNA sequencing from a very labor intensive process resulting in very inefficient output, to a relatively easier and much cheaper process that enables the sequencing of enormous eukaryotic genomes within days, generating vast amounts of data that enable the detection of mutations throughout the genome. Various read selection techniques like exome sequencing (17,18), shotgun sequencing (19–21), cDNA sequencing of RNAs (19) and others have been developed that make specific types of experimental analysis possible.

The Sanger sequencing method can now make use of colored fluorescent dyes allowing sequencing to proceed in a single reaction, paired with arrays of capillary gels in re-usable capillaries that allow the use of more powerful electric fields resulting in faster sequencing than could be achieved in slab gels (12,13). This technology was achieving read lengths of about 1300 bp in about 2 hours in the year 2000 (12,13). The benefit of these very long read lengths and arrays of capillaries allowing multiple samples to run at once, analogous to parallel “lanes” in older slab gels, enabled the rapid completion of the human genome project (13,16).

The current Next Generation high throughput methods such as Illumina sequencing do not achieve the same long individual read lengths as Sanger sequencing (13,16), but what they lack in length, they more than make up for in read numbers, allowing sequencing runs to cover entire genomes with large numbers of reads per locus (13). The paired end nature of the sequencing (13) also allows software analysis to provide insights into phenomena such as alternative splicing, despite the shorter read

lengths. This technology can read 2 x 150bp in most current machines and 2 x 300bp on the MiSeq series (14,22). (2 x N refers to the first N bp on each “side” of the DNA strands). The machines range in output from 25 million reads per run to 6 billion reads per run, and run times range from 4-24 hours to more than 3 days, depending on the sequencer platform.

Implications for research in human disease

The completion of the human genome project’s main goal and the advent of these sequencing technologies has been a boon for all biological research, but in particular it enabled a much deeper probing of the genomic changes occurring in heritable diseases and various kinds of cancers than had previously been possible.

History of TCGA

The Cancer Genome Atlas (TCGA) is an ambitious project that seeks to map mutations and clinical characteristics of 33 types of human cancers (23). It was formed as a collaboration between the National Cancer Institute and the National Human Genome Research Institute. The TCGA dataset was generated by the TCGA Research Network, which is composed of a broad coalition of different research centers and laboratories. Much of the data has been made publicly available. Identifiable data, including a portion of the patient information and any germ-line mutations or relevant SNPs, as well as the raw sequencing data, require additional agreements and security procedures from researchers and their institutions in order to gain access.

Across all tumor types, 11091 samples have been sent to the TCGA project for analysis. Out of these, 11077 have data available in the database. These are not evenly

distributed amongst the tissue types, probably due to varying rates of incidence. 16 tumor types have between 300-600 samples with data, 9 tumor types have between 100-300 samples with data, 8 tumor types have less than 100 samples with data, and just one, breast invasive carcinoma has more than 600, with 1097 samples with data. I noticed that 34 rows were in the table of tumor types on the TCGA website. I believe that while colon and rectal cancers are sometimes listed separately, the project treats them as the same cancer type when they provide the value of 33 for the number of tumor types.

Concept of Drivers and passengers

One of the big questions in cancer research is how many mutations are necessary to cause a tumor and what types of genes are involved (24). Bert Vogelstein and others have long proposed that multiple steps are necessary during the tumorigenic process (25). Early attempts to correlate mutations with tumor stage particularly in colon cancer resulted in a linear multistep process now known as the Vogelstein model (26). It proposed that each of three steps, initiation, promotion and progression could be correlated with specific mutations (27). Although this model has now been shown to be overly simplistic (28), it did provide the foundation for subsequent work on the concept of driver mutations.

In two seminal papers by Sjöblom et al (29) and Wood et al (30) from Bert Vogelstein's group, they were some of the earliest to describe the use of genome sequencing approaches to determine what mutations contribute to cancer pathogenesis, and to classify them appropriately (31). Mutations that do not contribute to cancer pathogenesis are called passengers, while the mutations that do contribute to cancer pathogenesis are called drivers (32).

In addition to this concept they examined mutation patterns, and frequency of mutations at different loci within populations of tumors (29–31). One of the ways they visualize these mutations is a 3 dimensional histogram “map”. On one axis is the chromosome number and on the other axis is the position along the chromosome. The height of the peaks is determined by the number of mutations within the region. Using this technique, they are able to find that certain genes are “gene mountains”, with very tall peaks, some are “gene hills” which are mutated an intermediate amount, while other genes are barely a blip on such a graph. They then examine how mutations cluster within the protein structure of a gene product as well as where the mutated genes fit within different signaling pathways.

While these efforts were groundbreaking, there are some important details regarding how scientists might assign driver and passenger status to a mutation that need to be further explored. It is the combination of which gene is mutated, at which location, and the precise change, that actually determine whether the mutation will be pathogenic or not. It isn’t sufficient to consider a mutation a driver, simply because it happens within a gene known to have associations with cancer.

Additionally, it is possible that there are some mutations that may contribute conditionally, if paired with other mutations. One such theoretical example would be Myc mutations in Burkitt’s lymphoma described by Bauer et al (33). In this case, a driver mutation in Myc leads to increased proliferation as well as increased rates of apoptosis. The increased rate of apoptosis balances the increased rate of proliferation preventing the tumor from increasing in size. A subsequent mutation in another gene that reduces the rate of apoptosis would lead to an increased growth of the tumor. Thus the initial driver

mutation is dependent on the second driver mutation for there to be a benefit to tumorigenesis. Other more complex scenarios may exist in genes that are part of signaling pathways. For example, multiple mutations disrupting different parts of a pathway might or might not produce a greater tumorigenic effect together than they would by themselves.

Another example of conditional driver mutations would be a hemizygous mutation in a tumor suppressor gene. If one copy of a tumor suppressor gene was mutated somatically or was inherited with a defective sequence and did not produce a haplo-insufficiency effect, and then a subsequent deletion or disruptive mutation that disabled the remaining functional copy on the other chromosome would be required to produce the tumorigenic effect. However, if only 1 such mutation occurred and the other functional copy remained, this mutation would not contribute to tumorigenesis. Technically this kind of mutation would rightly be considered a driver, but it would be a conditional driver, since its driver effect is dependent on the absence of both functional alleles.

While silent mutations that do not affect the amino acid composition of a gene, or functionally synonymous amino acid changes, would likely be passenger mutations (since they do not affect the function of the gene), in some cases even mutations that destroy the function of a gene could also be considered passenger mutations. If the tumorigenic process requires a gain-of-function mutation in a gene to produce a tumorigenic effect, then mutations that inactivate the protein would actually be passenger mutations.

I wanted to examine the effect that significant differences in mutation frequency had on the patterns of mutations that were found in tumors. Heritable defects in genes related to DNA mismatch repair are found in the familial syndrome, Lynch Syndrome (34),

which is known to cause very high rates of colon cancer (34), as well as broadly raise the rates of cancers in other tissues as well (34). This same repair mechanism can also be damaged or disabled in a somatic way in a subset of tumors in individuals that do not have a heritable defect (34,35). For my thesis project, it seemed that colon cancer would be a good model cancer type in which to examine mutation patterns to see if the idea that variations in mutation frequency in each tumor would significantly affect which genes were mutated in that tumor was plausible and to examine some details of the phenomenon if the hypothesis was true.

Colon Cancer

Colon adenocarcinoma is the third most common cancer in the USA (36,37). According to Cancer.org, 93,090 new cases of colon cancer and 39,610 new cases of rectal cancer were predicted to occur in the U.S. in 2015 of which 49,700 deaths from colon and rectal combined were anticipated (38). From 2003-2007 men showed an incidence of 57.2/100,000 and mortality of 21.2/100,000 and women had an incidence of 42.5/100,000 and a mortality of 14.9/100,000. Understanding the defects that lead to these cancers may give us better tools to manage disease and potentially make available new avenues of attack to destroy tumors more effectively without serious harm to patients.

A fraction (25%) of colorectal cancers result from inherited mutations (36), while the rest (75%) are a result of somatically acquired mutations. Colorectal cancers arise through several different pathways. One commonly observed mechanism in colorectal cancer involves defects in the DNA mismatch repair (MMR) pathway (39). MMR defects account for 15% of colon cancers (36) while most other cases (~85%) are caused by other processes involving chromosomal instability. In either case, cells that lose their

ability to repair replication errors and/or DNA damage have an increased mutation rate many times greater than that of normal cells (40,41). Presumably, any condition that leads cells to have either increased rate of mutation, an inability to recognize damage and undergo apoptosis, or a reduced ability to repair DNA damage or replication errors would increase the risk of cancer by contributing to tumor initiation, progression and metastasis.

Even before the discovery of specific mechanisms that caused increased genetic instability and mutation rates, genetic instability had been considered to be highly important in human cancers (42). While mechanisms of DNA repair, their mutations, and the effects of mutations in these genes on colon cancers have been an area of intense research (43–46), the patterns of mutations that occur with different repair defects are still poorly understood (31). The specific types of mutations that occur could be significantly affected by the identity of the initially disrupted repair gene or the nature of the mutation, which could lead to distinct patterns of targeted mutations.

I was curious to determine whether the spectrum of genes in which somatic mutations occurred in colorectal cancers differed depending upon the overall number of mutations that occurred in each individual tumor. I wanted to know if separating tumors with high numbers of mutations from those with relatively few mutations altered the patterns of driver mutations in the tumors. To do this, I proposed to analyze whole genome sequencing data from colorectal cancer tumors to test whether the tumors have different patterns of mutated genes based on the overall number of mutations that occurred in the tumors. My reasoning is that if I can show that there is a difference in types of mutations or genes that are being mutated between tumors with a very large number of mutations and those with a lower mutation frequency, future studies may discover that these

differences contain patterns that may be associated with the probability of recurrence, effectiveness of treatments, and patient outcome.

Chapter 2

Broad analysis of TCGA COAD Sample Populations, Sample Grouping, and Differences Between Groups

Introduction

In order to study the effect of mutation frequency on tumor genetics I needed to choose a tumor mutation dataset to use as an example. The Cancer Genome Atlas project has been collecting and analyzing tumor samples from numerous kinds of cancers and collecting the data into a publicly accessible database (47). Of primary interest to me was the effect of mutation rate on mutation spectrum in cancer. I chose to use colon adenocarcinoma as the model system for my analysis due to its association with tumors having genomic instability and high mutation rates (35,48).

Colorectal cancer in general has high rates of genomic instability (49–51). Colon cancer genomic instability has previously been investigated in regard to diseases that produce microsatellite instability, such as hereditary nonpolyposis colorectal cancer (52–54), (Lynch Syndrome), and similar conditions associated with defects in DNA mismatch repair proteins or failure to express them (35,55).

Ongoing clonal adaption is a common and necessary trait of cancers leading to the concept that cancer is a Darwinian evolutionary process (50,56). As cancers and precancerous tissues undergo random selection during initiation and progression, their rates of mutation show variation (57,58). This variation in mutation rate drives the deterministic mechanism for both passenger and driver mutations. As a consequence, it is likely that those cancers that have lower mutation rates would likely show a retrospective bias towards mutations that actually contribute to the cancer phenotype while cancers that have a high rate of mutation would show a retrospective bias towards more stochastic mutations. This is because a low rate of mutation would provide greater opportunity for the developing tumor to undergo clonal selection before many mutations

accumulate. In my analysis, I lacked any physical access to any cells from the tumors that I proposed to examine. Therefore, I could not measure the mutation rate in these tumor cells directly. However, numbers of mutations occurring in each cancer could act as a surrogate for mutation rate. Therefore, by comparing the number of mutations that are found in individual tumors, it may be possible to correlate mutation frequency with its effects on clonal selection.

Other factors may also have an effect on the retrospective bias for mutations. There may be structural or chemical factors that could lead specific genes to be more likely to experience mutation than others depending on the mechanisms driving the change in rate of retained unrepaired mutation. It is also possible that the cancer staging may correlate to how severely genetically damaged the tumor cells have become although I did not actually expect later staging to correlate very strongly with mutations, since high mutation rate tumors seemed to have more favorable outcomes in general than normal tumors.

Together, these factors suggested to me that it was likely that there would be a qualitative difference in the spectrum of genes that underwent mutation during cancer that was dependent on the number of mutations in the individual tumors.

Hypothesis

As the number of mutations within a tumor increases, there will be a shift in which genes undergo mutations in tumors collected from cancer patients. Tumors with low numbers of mutations should show a bias towards mutations in cancer driver genes that should not be apparent in tumors with high numbers of mutations.

Methods

The TCGA dataset mutation data is stored in MAF format. A program was prepared to count mutation entries in the dataset according to sample ID, mutation type, and Gene Symbol. Individual counting functions were prepared for each of these with corresponding tab separated output. Originally there was a pair of files, with data resulting from SOLiD sequencing being kept separate from data that resulted from Illumina sequencing. The bulk of the data were in the Illumina dataset, so to avoid potential complication in the analysis, I chose not to use the SOLiD file.

As noted, I lacked physical access to any cells from these tumors with which to attempt to measure a mutation rate directly. The TCGA project itself did not include measurement of mutation rate as one of their analysis methods either. As a result, I had to use an indirect method to get a rough handle on the mutation rate. I decided to use the total count of somatic mutations within the TCGA dataset as a proxy for this. I chose not to analyze larger chromosomal structural changes as I felt that this would complicate the analysis unnecessarily.

I wrote a program in Python to count single nucleotide mutation types (substitutions, insertions, and deletions), separating the counts according to original base and resulting base. I also enabled this tool to count mutations by Sample ID and Gene symbol. This was accomplished using a counter object sub-classed into multiple other types defined to use different counting rules. My method was to use strings in a map data type that stored integers using strings as the key. Gene Symbol and Sample ID were used as keys, and I constructed a key for the mutation types using the original and changed bases, using

dash as a placeholder for indels. The script allowed any combination of the counters to be used simultaneously, each outputting their counts to separate files.

To determine the effect of the mutator phenotype on these counts, I also decided to divide the tumors into two groups, a high mutation group and a low mutation group, and ran the same counting process on the separated groups as I ran on the entire dataset in the previous analysis. To accomplish this, I wrote an additional program to use the output of the sample ID counting function in combination with integers provided by the user, to split an MAF file, entry by entry, into grouped outputs. These output entries, still in MAF format, were then run through the original counting script again and the results were examined.

Count boundaries used to split the data were selected based on the location in a sorted graph of total mutation counts where it seemed there was a significant difference in the number. I did check the MSI status of these samples at a later time, and the results of this split matched fairly well with this boundary. The MSI high samples mostly fell into the higher mutation group where the MSI low were mostly in the lower mutation group. Clinical staging was analyzed for correlation to mutation counts as well.

I requested clinical information alongside the somatic mutation data, when I first obtained the TCGA dataset. These files were examined for their structure and information content to determine where the relevant data were, and then were used to combine this information into a useful table containing the information of concern. The sample IDs were used to cross compare the various categorizations and mutation counts to determine if there were any statistically significant correlations.

Results

Population Level Mutation Information

The first analysis performed was to count the number of mutations in each sample and view this data. I chose to sort the results from highest to lowest and plot them in a bar graph. The values of the mutations per tumor showed a distribution of samples with a long trailing tail of lower value counts and a relatively shorter collection of samples with very high count values. There seemed to be at least two different trends in the population, which could be the result of shifts in mutation rate. If one was to draw trend lines for the low mutation side of this plot, and the high mutation side of this plot, separately, they would have very different slopes.

In order to compare the two potential trends in the populations, it was necessary to choose a breakpoint between the two populations. The choice of a breakpoint at 1000 mutations seemed like a reasonable boundary to split the samples into two groups since it was close to the long low count tail, and it seemed to be a point where there was a shift in the trend. I divided the tumors into two separate populations of ≥1000 mutations/tumor and <999 mutations/tumor. I then examined the identity of the genes that were mutated in these two separate groups by running the gene mutation counter on the split MAF files.

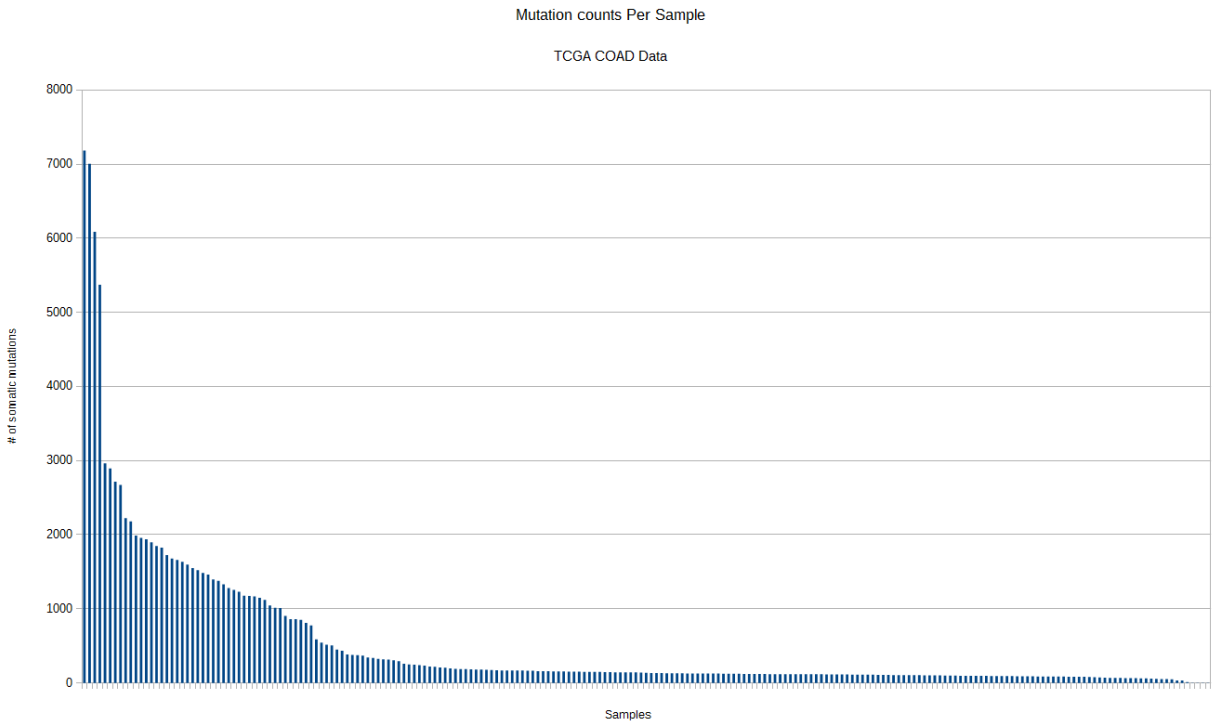

Figure 1. Number of mutations per tumor (TCGA database: COAD Illumina Dataset)

Each blue bar is a count of all mutation entries for an individual sample in the dataset

When looking at the results for all genes, since a very large portion of the human genome was included in the set of genes with mutations, there was an expected significant amount of overlap between the two groups of tumors. In total, both lists represented 17,046 genes, which is a majority of the predicted 20,000 to 25,000 human genes (2,59). However, even within this very large gene set, the two groups had distinct sets of genes. So I decided to look at overlaps within smaller, more significantly mutated subsets.

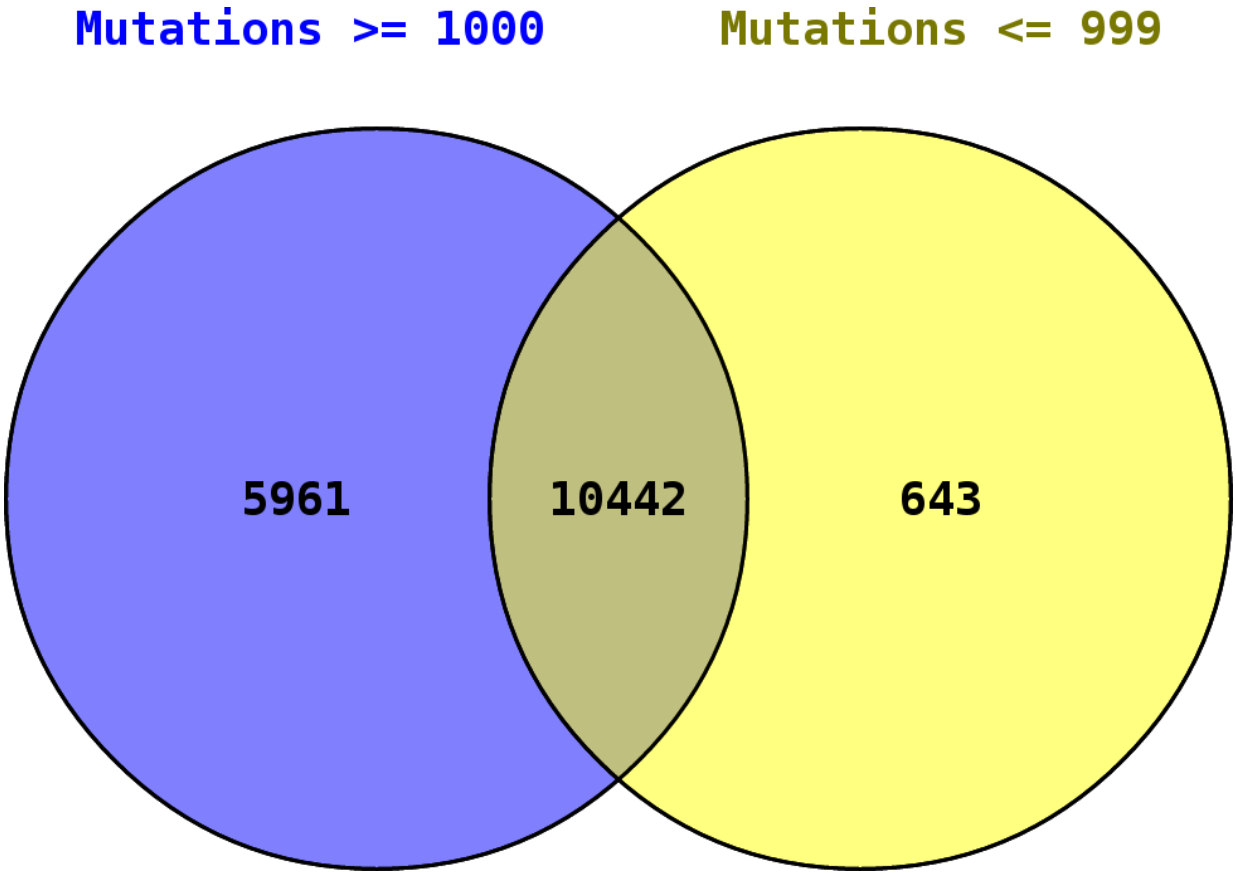

Figure 2. Venn diagram of all mutated genes

Lists were generated by counting mutations from the split groups, and binning according to gene symbol using the counting program described in methods.

The High mutation group is shown in blue while the low mutation group is yellow. The overlapping shared region is a darker yellow. A large number of mutated genes are shared between the groups, but there are unique genes in both groups.

Comparison of Most Mutated Genes between groups

This Venn diagram was created by selecting the genes with at least the same mutation counts as the 100th most mutated gene in each group. While there was again significant overlap, there were still potentially relevant differences. Some of these differences appeared to actually lie within the overlapping region of the original all-inclusive Venn diagram if the gene in the top 100 in either group was mutated within both populations, but was not mutated at the same frequency in the two groups. This, however would still be a potentially interesting qualitative difference.

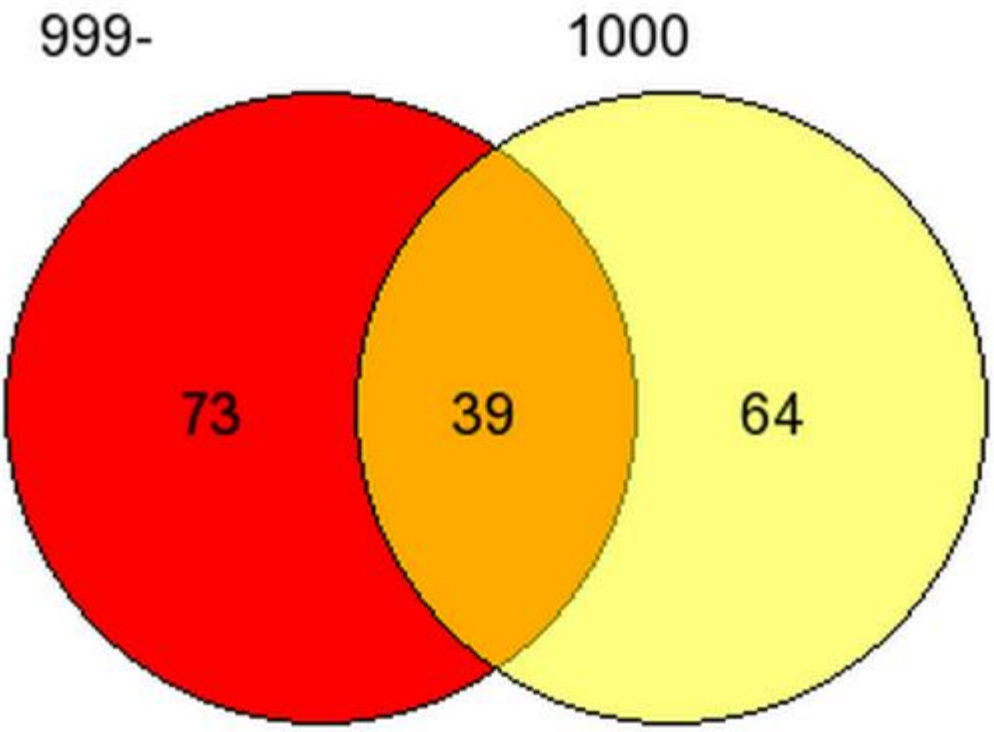

Figure 3. Venn Diagram of Mutated Genes, First ~100

Lists were generated by selecting all genes that had at least the same number of mutations as the 100th after sorting by mutation count. The list of gene based counts was generated using the counting program on the split data files. The High mutation group is yellow, the low mutation group is red, and the overlap is orange.

As in the previous diagram, this Venn diagram was created by selecting the genes with at least the same mutation counts as the 50th most mutated gene in each group. This smaller list again recapitulated the pattern observed in the larger previous lists. This same pattern of overlapping and non-matching genes remains visible even when filtering the list according to the most mutated genes. This resembles a chaotic equation plot, like a fractal (60–62), where the appearance remains similar regardless of magnification.

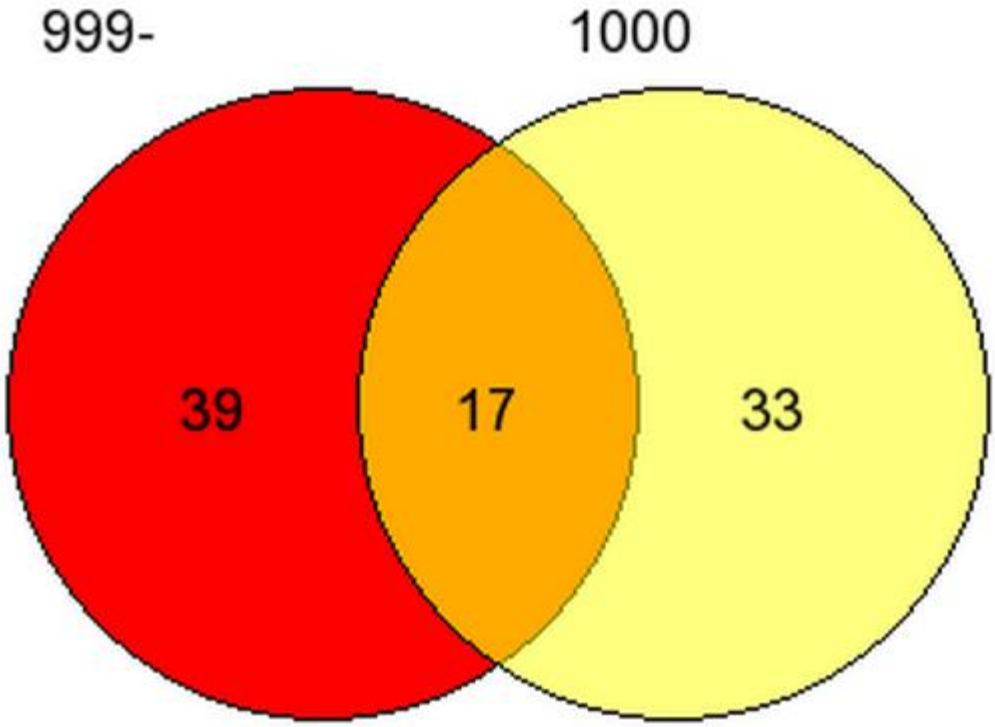

Figure 4. Venn Diagram of Mutated Genes, First ~50

Lists were generated by selecting all genes that had at least the same number of mutations as the 50th after sorting by mutation count. The list of gene based counts was generated using the counting program on the split data files. The High mutation group is yellow, the low mutation group is red, and the overlap is orange.

Most Mutated Genes Lists

When the list of genes from each of the tumor sets was compared using the DAVID functional annotation software (63,64), the list derived from the lower mutation group showed a much more targeted set of cancer related genes. While the list derived from the higher mutation group did still contain known cancer related genes, there were larger numbers of seemingly unrelated genes, as well. These unrelated genes were probably passenger mutations.

Functional clustering by gene ontology of the list of mutated genes from the lower mutation group using DAVID resulted in one functional cluster that listed many cancer-related sub-terms in addition to two other functional clusters which referred to colorectal cancer specifically. In contrast, results for the list of mutated genes from the higher mutation group only had one cluster specifically containing cancer terms, limited to “pathways in cancer” and “small cell lung cancer”.

Mutations <= 999			Mutations >= 1000	
#Samples	180		#Samples	39
Gene Symbol	Mutation Count		Gene Symbol	Mutation Count
APC	221		TTN	294
TTN	129		MUC16	129
TP53	117		SYNE1	97
NEFH	108		OBSCN	93
KRAS	84		CCDC168	78
DSPP	70		RYR2	66
OBSCN	66		SSPO	62
ERICH6B	60		MUC17	59
ZNF814	56		DNAH17	58
IRF5	54		NEB	58
MUC16	53		XIRP2	57
TBP	52		FAT4	57
ATXN1	52		CSMD3	56
PIK3CA	49		PLEC	56
KRTAP4-5	45		DNAH8	55
MUC4	43		RYR3	55
PRIC285	43		GPR98	54
SYNE1	41		DSPP	53
TMPRSS13	40		DNAH11	53
FAT4	40		DCHS2	51
PHF2	38		RYR1	50
GPRIN2	35		PCLO	49
CACNA1B	33		ZFHX4	48
ZFHX4	32		FAT1	47
RYR2	32		MUC5B	46
ZFPM1	30		LAMA1	45
MAML2	30		LAMA5	45
PLEC	29		LRP1B	45
FAT3	29		AHNAK2	45
KCNN3	28		CSMD1	44
SSPO	28		DNAH5	44
CACNA1H	28		USH2A	44
ABCA13	27		LRP2	44
RP1L1	27		MLL2	44
AR	26		DNAH3	43
SOX9	25		ABCA13	43
MUC5B	25		HMCN1	43
TPRX1	24		PCNT	43
CRIPAK	24		DNAH2	42
CROCC	24		SYNE2	42
DNAH5	24		TNXB	40
SMAD4	24		ANK3	40

Mutations <= 999			Mutations >= 1000	
#Samples	180		#Samples	39
Gene Symbol	Mutation Count		Gene Symbol	Mutation Count
CSMD3	24		NCOR2	40
PCDHA7	23		FREM2	40
COL18A1	23		FRAS1	39
TCHH	23		DST	39
LAMA5	23		WDR87	39
FLG	23		MACF1	39
KRT1	22		CMYA5	39
ANKLE1	22		PKHD1L1	39
IRF2BPL	22		DNAH14	38
KNDC1	22		DNAH10	38
SDK1	22		EYS	38
NCOR2	22		ASPM	38
PCLO	22		ALPK2	37
ZNF469	22		APC	37
OGFR	21		ZNF469	37
FBXW7	21		APOB	37
KIF26A	21		MYCBP2	37
COL6A3	21		ZFHX3	37
TCF15	20		FAT3	37
NUMBL	20		DNAH12	36
EP400	20		PDE4DIP	36
CSMD1	20		FBN3	36
RYR1	20		PKD1	36
APOB	20		NEFH	35
LRP1B	20		LRP1	35
KRTAP4-3	19		UBR4	35
MEFV	19		CACNA1H	35
PPM1E	19		TEX15	35
CSMD2	19		EP400	35
PTPRT	19		DNAH9	34
USH2A	19		RP1L1	34
DNAH11	19		MEGF8	34
CRCT1	18		SACS	34
PTPLA	18		LAMA2	34
OPRD1	18		AHNAK	34
ZNF837	18		WDFY4	34
ESPNL	18		MUC4	33
DSCAM	18		RNF43	33
TRPS1	18		KIAA1109	33
KRTAP4-11	17		PKD1L1	32
GRIN3B	17		HERC2	32
ABCA7	17		LRRK2	32

Mutations <= 999			Mutations >= 1000	
#Samples	180		#Samples	39
Gene Symbol	Mutation Count		Gene Symbol	Mutation Count
ADAMTS7	17		NBEA	32
WNK2	17		HSPG2	32
DNAH8	17		DNAH1	32
ANK2	17		MYO15A	32
LAMC3	17		COL6A5	31
SCN5A	17		ATM	31
HECW1	17		PCDH15	31
UNC13C	17		POLE	31
GLI3	17		CACNA1A	31
UNC80	17		MLL4	31
RYR3	17		RNF213	31
KRTAP4-8	16		ABCA7	30
NTSR2	16		NLRC5	30
MAP1S	16		CUBN	30
PCDHB8	16		CEP192	30
ADAMTS2	16		FMN2	30
POLRMT	16		VPS13B	30
ATP10A	16		MUC6	30
TNXB	16		FLNC	30
FAM123B	16			
NOTCH3	16			
ODZ4	16			
DCHS2	16			
PCDH17	16			
HSPG2	16			
WDR87	16			
ZNF831	16			
GPR98	16			

Figure 5. List of ~100 most mutated genes. Lists were generated by selecting all genes that had at least the same number of mutations as the 100th gene after sorting by mutation count. The list of counts grouped by gene symbol was generated using the counting program described in the methods section on the split data files.

Correlation Between Mutation Count Group and Cancer Staging

During one of my Research in Progress meetings, someone from the audience raised a question about correlation between tumor staging (65) and mutation rate. I chose to examine this by pulling the publicly available clinical data and matching it with the mutation counts. The high mutation count group has a higher proportion of Stage II and Stage IIA tumors and a lower proportion of Stage III and IV tumors than the low mutation count group (Figure 6). The proportion of Stage I tumors was very similar in the two groups, but was slightly higher in the high mutation group (Figure 6). The differences for Stage II (without a subtype) were statistically significant, as were the difference between the pooled counts for all subtypes of Stage II together (Figure 6). Differences for Stage III only showed statistical significance when pooled (Figure 6). Stage I and IV did not pass requirements for statistical significance via a two tailed Z test of proportions (Figure 6).

Group	1000-above		0-999					
# Samples	39		180					
Stage	Count	Proportion	Count	Proportion	Pooled_Prop	Std. Err	Z	P Value
[Not Available]	2	0.051	6	0.033	0.037	0.033	0.542	0.588
Stage I	8	0.205	23	0.128	0.142	0.062	1.256	0.209
Stage IA	0	0.000	1	0.006	0.005	0.012	-0.467	0.641
Stage IB	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
Stage IC	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
Stage II	7	0.179	8	0.044	0.068	0.045	3.027	0.002
Stage IIA	16	0.410	54	0.300	0.320	0.082	1.339	0.181
Stage IIB	0	0.000	2	0.011	0.009	0.017	-0.661	0.508
Stage IIC	0	0.000	1	0.006	0.005	0.012	-0.467	0.641
Stage III	1	0.026	6	0.033	0.032	0.031	-0.248	0.804
Stage IIIA	0	0.000	9	0.050	0.041	0.035	-1.426	0.154
Stage IIIB	2	0.051	30	0.167	0.146	0.062	-1.849	0.064
Stage IIIC	1	0.026	14	0.078	0.068	0.045	-1.169	0.243
Stage IV	1	0.026	13	0.072	0.064	0.043	-1.078	0.281
Stage IVA	1	0.026	12	0.067	0.059	0.042	-0.983	0.326
Stage IVB	0	0.000	1	0.006	0.005	0.012	-0.467	0.641
Stage IVC	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
Stage I (ALL)	8	0.205	24	0.133	0.146	0.062	1.151	0.250
Stage II (ALL)	23	0.590	65	0.361	0.402	0.087	2.640	0.008
Stage III (ALL)	4	0.103	59	0.328	0.288	0.080	-2.817	0.005
Stage IV (ALL)	2	0.051	26	0.144	0.128	0.059	-1.580	0.114

Figure 6. Tumor Staging

This data was pulled from text files containing clinical information on the cancer samples in the TCGA database. A two-tailed two sample Z test of proportions was performed on the counts of tumors grouped by stage.

Correlation Between Mutation Count Group and Mutations in Repair Genes

Similarly, during one of my committee meetings, the question was raised as to whether there was a correlation between mutation frequency and mutations within mismatch repair, and other DNA repair genes. I included the main genes known to be associated with Lynch Syndrome (66–68), as well as some genes involved in mismatch repair (69). PolG was included as well due to there being an association with colorectal cancer (70,71), and PolG being involved in mitochondrial DNA repair (70). I queried the mutation counts file for the specific genes of interest and compiled a list (Figure 7). There were more mutations within the DNA mismatch repair related genes in the higher mutation group. These could be functionally destructive mutations leading to a retention of unrepaired DNA damage, or they could be symptomatic of an already disrupted repair system, or an aberrantly regulated cell cycle checkpoint, or some other mechanism leading to an increased mutation rate. Out of these genes, PMS1, PMS2, RFC2, RFC5, and RFC6 did not meet statistical requirements for significance in the difference of the counts. However, given the low frequency of mutation in these genes, even with the two tailed Z test of proportions, some of these counts were hard to interpret. The one mutation in RFC3 for instance, might not actually be biologically significant or relevant, even if it was statistically significant, as it resulted from only 1 mutation in the smaller group.

	180 Samples		39 Samples					
	(Samples Mutations <= 999)		(Samples Mutations >= 1000)					
Hugo Symbol	#Mutations	Proportion	#Mutations	Proportion	Pooled_Prop	Std. Err	Z	P Value
MSH2	1	0.006	12	0.308	0.059	0.042	-7.239	0.000
MSH3	2	0.011	6	0.154	0.037	0.033	-4.308	0.000
MSH6	6	0.033	10	0.256	0.073	0.046	-4.853	0.000
MLH1	4	0.022	6	0.154	0.046	0.037	-3.570	0.000
PMS1	5	0.028	3	0.077	0.037	0.033	-1.483	0.138
PMS2	4	0.022	2	0.051	0.027	0.029	-1.008	0.314
MLH3	2	0.011	11	0.282	0.059	0.042	-6.492	0.000
EXO1	0	0.000	9	0.231	0.041	0.035	-6.582	0.000
PCNA	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
POLE	8	0.044	31	0.795	0.178	0.068	-11.105	0.000
POLE2	1	0.006	9	0.231	0.046	0.037	-6.108	0.000
POLE3	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
POLE4	0	0.000	1	0.026	0.005	0.012	-2.153	0.031
POLD1	6	0.033	18	0.462	0.110	0.055	-7.761	0.000
POLD2	1	0.006	2	0.051	0.014	0.021	-2.227	0.026
POLD3	0	0.000	8	0.205	0.037	0.033	-6.191	0.000
POLD4	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!
POLG	5	0.028	5	0.128	0.046	0.037	-2.724	0.006
RFC1	2	0.011	13	0.333	0.068	0.045	-7.222	0.000
RFC2	1	0.006	0	0.000	0.005	0.012	0.467	0.641
RFC3	0	0.000	1	0.026	0.005	0.012	-2.153	0.031
RFC4	3	0.017	1	0.026	0.018	0.024	-0.379	0.704
RFC5	2	0.011	2	0.051	0.018	0.024	-1.698	0.089
RFC6	0	0.000	0	0.000	0.000	0.000	#DIV/0!	#DIV/0!

Figure 7. Genes Related to Repair and Replication

Lists of counts grouped by genes were searched for the entries corresponding to genes of interest. A two-tailed two sample Z test of proportions was performed on the counts.

Following up the mismatch repair gene mutation list, I also looked into DNA microsatellite instability status, which is associated with the mismatch repair deficiency phenotype (46,54,72,73), as a mutation within these genes would not necessarily produce the well characterized phenotype. The MSI status was also available within the clinical data. The high mutation count population was enriched for tumors designated as MSI-H, while the population with lower mutation counts was enriched for MSI stable tumors and those categorized as MSI-L.

I plotted this breakdown of the tumors according to MSI categories (Figure 8) to show the total somatic mutation count in a manner similar to Figure1. A few of the MSI-H tumors had comparatively lower counts compared to the other MSI-H tumors while 1 MSI-H tumor showed a very high mutation count in comparison to the others. Most of the MSI-stable tumors were fairly low in mutation count, but a small number of them had between 300 and 600 mutations, which was significantly more than the rest (Figure 8).

A handful of MSI stable tumors had even more mutations than those in the low mutation population (Figure 8). These samples ranged from more than 7 thousand mutations to one with approximately 2000. Most of the MSI-H samples were between 1000 and 2000 mutations. A handful were between 2000 and 3000, and one was above 5000.

The difference in proportions between the high and low groups for all of these classification groups were highly statistically significant according to a two tailed Z test.

Group		# Samples	#MSS	#MSI-L	#MSI-H	#MSI(any)
0-999	Values	180	139	33	8	41
	Proportions	-	0.772	0.1833	0.0444	0.227778
1000-above	Values	39	6	0	33	33
	Proportions	-	0.154	0	0.8462	0.846154
	Pooled_Prop	-	0.662	0.151	0.187	0.338
	Std. Err	-	0.084	0.063	0.069	0.084
	Z	-	7.402	2.901	-11.636	-7.402
	P Value	-	0.000	0.004	0.000	0.000

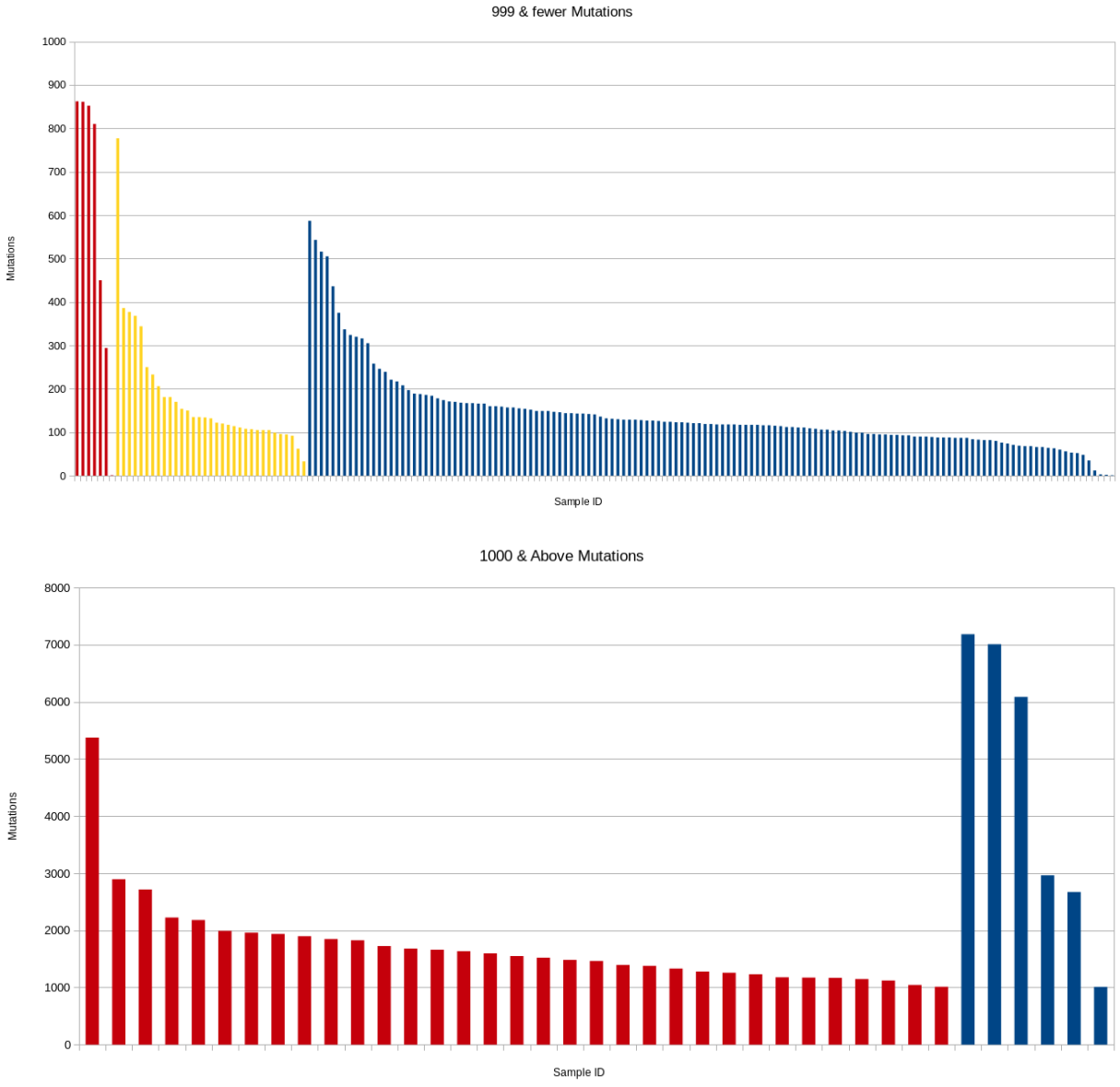

Figure 8. DNA Microsatellite status of the samples. Table shows grouped counts and the results of a two tailed z-test of proportions. Charts are the same numbers as in

Figure 1, but grouped by MSI status and split by mutation count group. The first bar graph is the low mutation count group, and the second is the high mutation count group. Red is MSI-H, Yellow is MSI-L, and Blue is MSS

Testing alternate split value for high and low groups

The value of 1000 for splitting the high and low mutation cancers into two groups was somewhat of an arbitrary choice. It happened to coincide nicely with the population of MSI-HIGH tumors as compared to the normal tumors and MSI-LOW tumors. In the interest of seeing how well the observed differences exist at other break points, I also split the samples into two groups at 300 mutations, and compared the lists of mutated genes as I had done for the 1000 breakpoint.

The overall Venn diagram resulting from this split is shown in figure 9. As before, there are differences in which genes appear with mutations, with some uniquely mutated genes in either group.

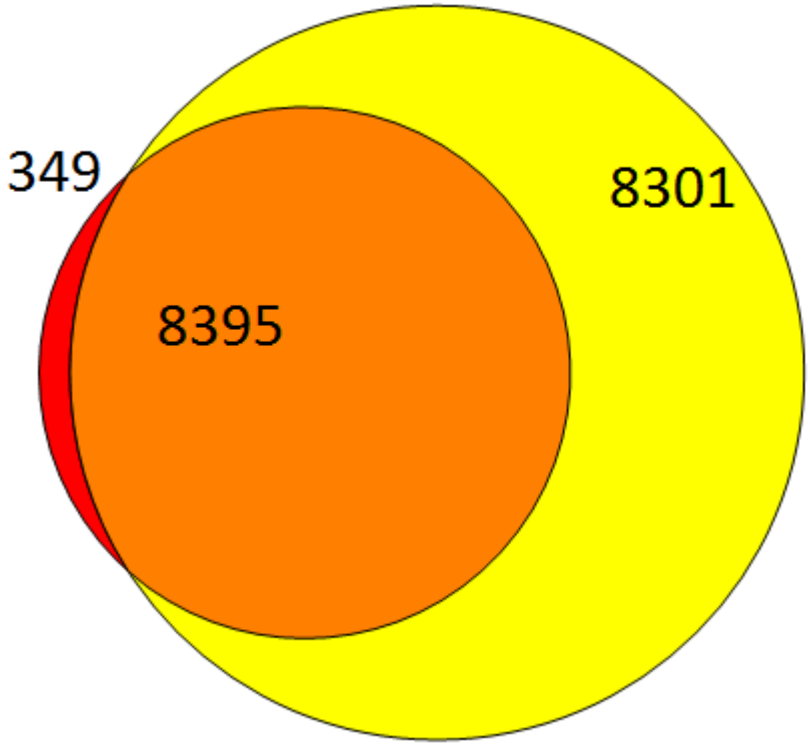

Figure 9. Venn Diagram of High and Low mutation groups split at 300 mutations.

Figure 10 corresponds to Figure 3, but for the 300 mutation split point. As before, when selecting the 100 most mutated genes from either group, there were shared genes as well as significant numbers of unique genes. This pattern supports the idea that there really are differences between the high and low mutation groups, and that the split point of 1000 is not the only breakpoint that reveals these differences in mutation. There were slightly more shared genes and fewer unique genes, although it was hard to determine how significant this difference was, as the numbers are still very similar.

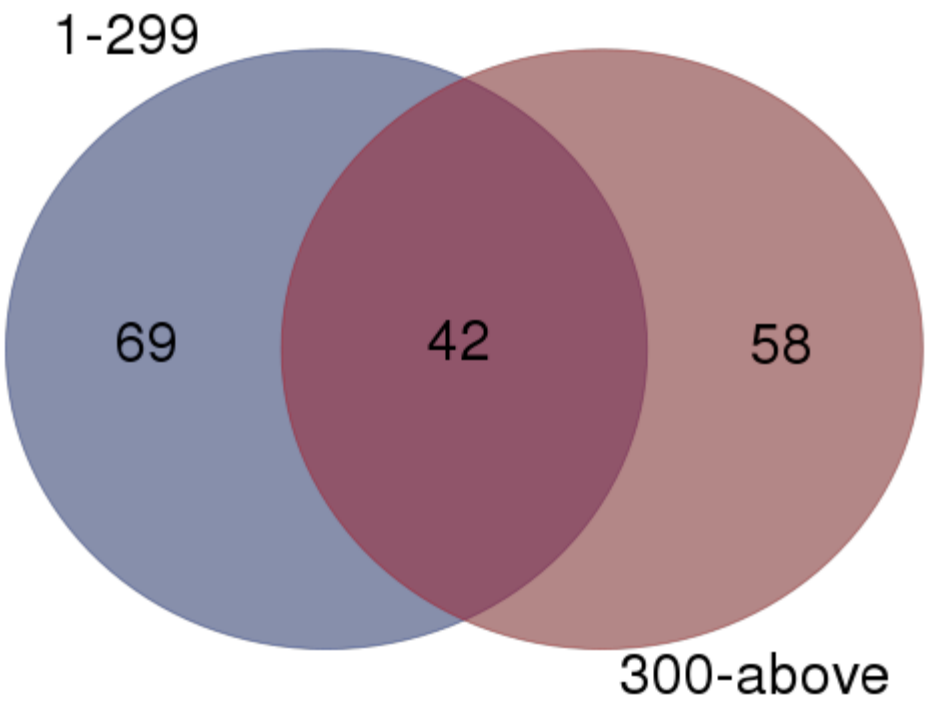

Figure 10. Venn Diagram of ~100 most mutated genes in High and Low mutation groups split at 300 mutations.

Figure 11 corresponds to Figure 4, but for the 300 mutation split point. As with the 1000 split point, this more restricted list maintained a similar pattern of unique genes and shared genes to the ~100 gene list. With this split point however, there were more shared genes and fewer unique genes within the high mutation group. This was likely due to the fact that including the samples with 300-999 mutations in the higher mutation group shifted the mutation patterns of the group slightly, but the overall fact that there were still differences remains.

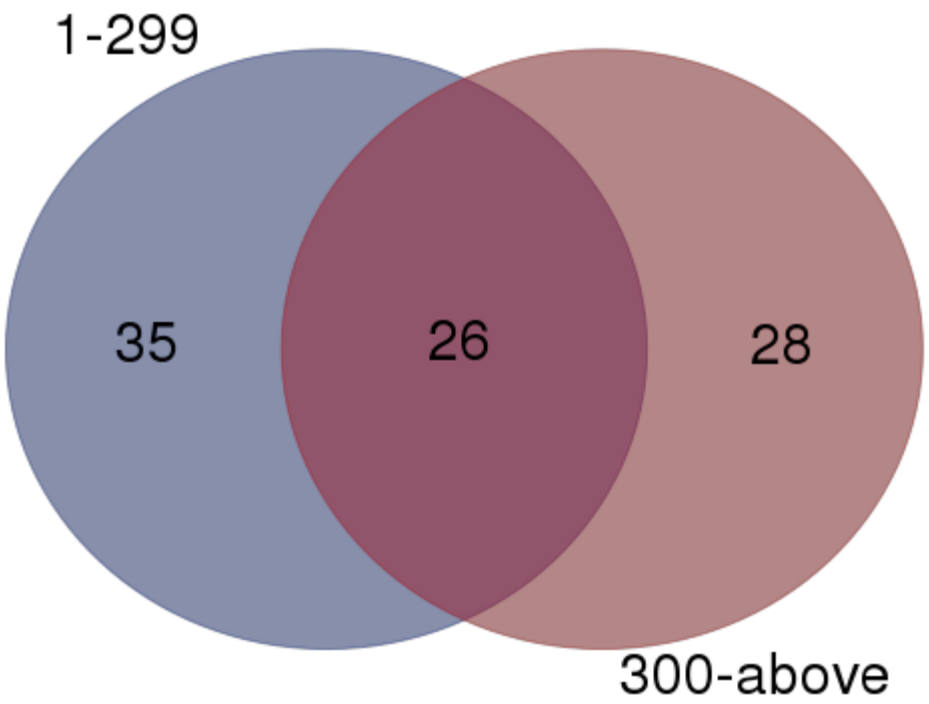

Figure 11. Venn Diagram of ~50 most mutated genes in High and Low mutation groups split at 300 mutations.

Discussion

These tumors exhibited an interesting divide between the high and low mutation frequency groups that held at both the large scale and when focusing on the most mutated genes. There did appear to be a relationship to DNA mismatch repair genes in the high mutation frequency group, and curiously, lower staging seems to correlate with the high mutation frequency group tumors instead of higher staging as one might be inclined to expect.

When looking broadly at the most mutated genes between these low and high mutation groups, it was apparent that the lower mutation group exhibits a bias towards mutations in cancer focused genes, while the higher mutation group had some cancer genes in its list but also many other kinds of genes. It was apparent that there was indeed a difference between these populations. Thus a more detailed examination of those differences became of interest.

Chapter 3

Mutation Rate Group Differences in Mutation Types and Gene Mutation Counts; Positively Deviating Outliers in Mutation Count & Gene Length Trend

Introduction

In the previous chapter, I had noticed when examining the counts of mutations by Official Gene Symbol, that there was a general trend of mutation count with gene size. As an example, Titin (TTN; OMIM 188840) was generally near the top of both lists when sorted by mutation count. TTN is encoded by an 82kb mRNA making it one of the largest genes in the human genome (74). This suggested that its mutation frequency might be a consequence of its size. However, after examining other genes from annotation databases I noticed that there were some other genes that appeared more frequently than their size would dictate.

The concept of mutation spectrum, also called genome or mutation landscapes, refers to the total of all mutations contained in the genome of tumor cells, usually classified as passenger or driver mutations (28–32,75). It is sometimes visualized in a histogram-like manner using mutation counts as the height variable, across the chromosomes on one axis and the chromosome number on the other axis.

There are many types of mutations and each can have varying causes. Small scale mutations include substitutions, insertions, and deletions each of which involves a single nucleotide or multiple nucleotides, and small scale inversions, which would necessarily involve more than one nucleotide. Large scale mutations, involving very large changes to chromosomes including copy number changes, duplications, deletions, and movement of large segments of DNA within or between chromosomes, or even gain or loss of entire chromosomes, would require a more complex analysis and were not included in my project.

Mutations within genes can have complications in their assignment. Due to splicing of transcripts there are sometimes multiple transcript variants that use different combinations of exons. A mutation may fall within one of the optional exons and only affect some transcript variants.

I used an approach of examining at the gene scale rather than differentiating between transcripts. There could perhaps be something of interest involving alternative transcripts, but it would introduce complexities into my analysis that were not directly relevant to answering the question I was seeking to answer.

Hypothesis

I hypothesized that the collection of genes mutated within the low mutation count population would tend to be more selective and have a greater degree of cancer specificity than the higher mutation count group. Further, in the higher mutation count group, I expected the influence of gene length on mutation count to increase significantly and the influence of selection, related to clonal evolution within the tumors, to be less significant in the high mutation population compared to the low mutation population.

Methods

In an idealized mutation scenario without the influence of clonal selection, one would expect a linear relationship between gene size and the number of mutations that would appear in a population of cells over time. If the mutations found in the tumors were to become more random and less selected, then there ought to be an increase in the association between gene size and mutation count within a gene. I wanted to be able to determine which genes were outliers from this trend. The impact of selection would place

some genes outside of this trend, in a positive direction if the mutations provided selective benefit or negative direction if they were deleterious for the cells in some way. I am primarily interested in those genes which had positive deviation from the trend.

I wrote a program to count mutation types both for the group as a whole and with a breakdown by sample. I also wrote a program to perform two tailed t-tests on the resulting tables of values, as well as their proportions against the total number of mutations within each tumor using a matching table structure, and to perform a linear regression on count data binned by gene symbol, and use that regression to obtain studentized residuals for the mutation counts of each gene. These were used to assess differences between the high mutation group and the low mutation group. The residuals and linear regression aimed to reduce the impact of gene length on the ranking and to identify genes that were positively deviated outliers in the roughly linear gene length and mutation count relationship.

The results of mutation count grouped by gene symbol were used to plot gene size against total mutation count. I used Excel's trend-line feature to place a trend line on this graph. Note that this trend line was not produced by the same software that performed the linear regression, and so may not be exactly the same as the line that was used to produce the residuals.

Files containing gene lengths were obtained from the genome browser table viewer. These were matched up with the entries in the MAF files in order to normalize against the effect of gene length. Genes with multiple transcript variants had isoform lengths averaged. The counts were divided by the lengths and the number of samples in their respective group.

I added these lengths to the selection of ~100 genes with the highest mutation counts from the previous chapter to show the variation of the gene lengths at the high end of the mutation counts. For these, I examined the 100th gene in the count-sorted list, and included any others beyond that point which had the same number of counts, in each group. I also repeated this using the studentized residual values instead of the counts, to show the genes with the highest residuals.

Results

Differences in Mutation Types

Figures 12,13,14, and 15 show the mutation type counts for the low and high mutation count groups respectively. Figures 14 and 15 are calculated values based on categories of mutation type, and mutations that are technically chemically indistinguishable due to DNA's base pairing. Figures 16, 17, and 18 show these values converted into percentages based on the total sum of mutations within each sample. Figure 18 corresponds to both Figures 14 and 15, which were kept separate due to size. These mutations are labeled using a "reference_mutation" pattern, with a dash "-" standing in for missing bases in insertions or deletions. MNC stands for multi-nucleotide change. Due to difficulty and complexity in the analysis of mutations involving more than one base, I opted to bin these types of mutations into one category. These types of mutations were less frequent than single base mutations in general, but are still prevalent enough to be potentially important.

The raw data and some calculated values are being shown in Figures 12-18 to facilitate explanation of some details about the mutation types, and because it was

feasible to fit these tables into this document, albeit in a very dense and compact form. The results of the t-test are shown in Figure 19.

Figure 12 depicts the result of the uncategorized mutation counts within the low mutation count group. There was quite a lot of variation in mutation count within this group, and the tumors with low numbers of total mutations frequently had zeroes for certain values. The values of corresponding mutations (such as C_T and G_A) usually matched quite well within each row. C_T and G_A were the most common mutation type by far, which was probably frequently the result of 5-methylcytosine deamination.

Figure 13 is the same type of table as Figure 12, but for the high mutation group. The counts were broadly raised, but there was a subset of these tumors that had very large counts for some types and not others. For instance, the 2nd to last row was a sample that had 799 C_T mutations, but only 1 C_G mutation and very low numbers of insertion mutations.

sample_ID	ALL	A_C	A_T	A_G	-_A	A_-	C_A	C_T	C_G	-_C	C_-	T_A	T_C	T_G	-_T	T_-	G_A	G_C	G_T	-_G	G_-	MNC
TCGA-CM-6169-01A-11D-1650-10	129	1	3	3	0	0	8	39	3	1	2	2	3	3	0	1	36	1	14	0	0	9
TCGA-D5-5541-01A-01D-1650-10	101	0	1	7	1	0	4	33	2	2	0	3	3	3	0	1	28	1	4	2	1	5
TCGA-G4-6314-01A-11D-1719-10	143	2	1	6	0	0	10	37	9	0	0	6	4	3	1	0	44	8	11	0	0	1
TCGA-G4-6297-01A-11D-1719-10	305	5	4	21	0	3	22	93	15	1	0	4	28	3	2	4	57	11	13	0	1	18
TCGA-F4-6703-01A-11D-1835-10	450	5	7	25	3	8	25	125	5	0	5	5	27	7	1	11	126	1	27	2	10	25
TCGA-DM-A1DA-01A-11D-A152-10	344	8	12	20	0	1	36	75	13	0	1	4	17	9	0	1	83	16	36	0	1	11
TCGA-F4-6806-01A-11D-1835-10	90	1	2	3	1	0	4	29	1	2	0	0	0	0	0	0	31	3	5	1	0	7
TCGA-A6-6650-01A-11D-1771-10	149	2	4	8	0	0	11	40	6	0	0	1	2	5	1	1	39	1	14	0	1	13
TCGA-F4-6805-01A-11D-1835-10	53	0	0	2	0	0	2	11	1	0	1	3	5	2	0	0	14	3	4	0	1	4
TCGA-D5-6898-01A-11D-1924-10	76	1	3	2	2	0	4	26	2	0	0	2	2	0	1	0	18	3	9	0	0	1
TCGA-CA-5797-01A-01D-1650-10	56	0	3	1	0	1	2	14	3	0	0	1	1	1	2	1	18	2	5	0	0	1
TCGA-G4-6321-01A-11D-1719-10	377	9	7	36	0	0	23	79	12	0	0	3	43	7	0	0	96	19	18	0	2	23
TCGA-CM-6679-01A-11D-1835-10	48	0	0	1	0	0	1	15	1	1	0	4	0	0	0	0	22	0	1	1	0	1
TCGA-F4-6809-01A-11D-1835-10	124	0	4	4	3	1	10	29	4	0	0	7	8	2	0	1	19	6	13	0	1	12
TCGA-CK-5915-01A-11D-1650-10	143	1	0	6	2	0	2	57	5	2	0	4	8	2	0	1	38	4	8	0	0	3
TCGA-D5-6533-01A-11D-1719-10	171	4	9	11	1	1	13	36	6	0	0	9	10	3	0	0	42	6	11	0	1	8
TCGA-AU-3779-01A-01D-1719-10	104	3	2	3	0	0	8	27	1	0	0	5	6	2	0	0	35	2	7	0	0	3
TCGA-AD-6901-01A-11D-1924-10	250	2	4	4	0	0	19	74	12	0	1	7	12	6	0	0	65	10	22	3	2	7
TCGA-D5-6529-01A-11D-1771-10	117	0	0	5	0	0	6	46	4	0	0	3	3	1	0	1	35	3	4	0	2	4
TCGA-CM-6164-01A-11D-1650-10	60	1	2	5	0	1	8	18	0	0	0	3	5	0	0	0	10	2	2	1	0	2
TCGA-DM-A28F-01A-11D-A16V-10	239	11	8	10	1	2	16	51	10	4	1	7	8	11	1	1	62	10	16	2	0	7
TCGA-G4-6311-01A-11D-1719-10	337	10	10	35	1	1	26	70	14	0	2	3	18	8	0	2	82	11	22	0	0	22
TCGA-D5-6532-01A-11D-1719-10	111	5	2	3	2	1	4	30	8	0	2	3	3	3	0	0	31	4	7	0	0	3
TCGA-AZ-4682-01B-01D-1408-10	96	3	3	3	0	0	5	38	0	0	1	1	0	1	0	0	34	2	3	0	0	2
TCGA-CM-5863-01A-21D-1835-10	84	1	0	4	0	2	4	24	3	0	0	1	4	1	0	2	20	2	7	0	0	9
TCGA-D5-6536-01A-11D-1719-10	103	1	3	4	0	0	5	22	6	2	0	3	3	1	0	0	36	2	10	0	1	4
TCGA-DM-A28H-01A-11D-A16V-10	120	3	2	2	0	2	4	27	4	1	0	4	5	3	2	1	36	8	5	1	1	9
TCGA-F4-6460-01A-11D-1771-10	127	2	2	3	0	0	11	40	5	0	0	8	6	2	1	0	27	5	13	0	1	1
TCGA-D5-6531-01A-11D-1719-10	197	3	11	5	0	0	12	52	7	0	1	5	9	5	1	0	58	5	13	0	1	9
TCGA-G4-6304-01A-11D-1924-10	294	1	7	12	0	7	18	107	1	0	0	6	16	6	1	3	87	5	13	1	3	0
TCGA-D5-6926-01A-11D-1924-10	131	0	2	4	0	0	4	36	5	0	0	2	8	3	1	0	43	4	11	0	0	8
TCGA-CM-5868-01A-01D-1650-10	189	5	7	16	0	2	10	44	1	1	0	5	14	4	0	2	40	5	17	2	0	14
TCGA-A6-2675-01A-02D-1719-10	83	2	4	5	1	0	3	31	1	0	0	0	3	0	0	0	26	3	1	0	0	3
TCGA-DM-A1DB-01A-11D-A152-10	80	1	1	4	0	1	4	26	3	0	0	0	4	0	0	0	20	6	6	0	0	4
TCGA-CM-4750-01A-01D-1408-10	99	3	1	3	1	0	6	30	0	0	0	3	3	1	0	0	29	0	11	0	1	7
TCGA-AY-5543-01A-01D-1650-10	126	2	0	6	0	0	10	38	6	0	1	3	6	3	0	0	39	4	7	1	0	0
TCGA-AD-6548-01A-11D-1835-10	105	1	0	1	0	2	4	34	0	0	0	3	7	2	0	0	36	3	7	1	0	4
TCGA-AZ-4323-01A-21D-1835-10	33	0	2	2	0	0	2	14	1	0	0	0	0	1	0	0	8	1	1	0	0	1
TCGA-CM-5864-01A-01D-1650-10	221	6	7	9	0	2	21	54	11	1	2	4	11	7	0	0	55	8	19	0	0	4
TCGA-AA-3489-01A-21D-1835-10	82	2	3	2	1	0	7	27	0	1	0	4	2	0	0	0	21	2	5	0	0	5
TCGA-A6-6140-01A-11D-1771-10	505	10	6	30	1	0	30	157	14	0	0	12	21	8	1	0	169	22	21	0	0	3
TCGA-A6-5664-01A-21D-1835-10	52	1	0	2	0	0	6	18	0	0	0	0	3	1	0	0	16	0	0	0	0	5
TCGA-CM-5860-01A-01D-1650-10	122	0	1	5	0	0	11	41	2	1	0	0	4	1	0	0	41	3	8	0	0	4
TCGA-DM-A28K-01A-21D-A16V-10	115	1	2	2	2	0	7	37	4	1	0	1	3	1	0	0	40	1	2	2	1	8
TCGA-F4-6807-01A-11D-1835-10	160	1	4	4	0	1	23	30	5	0	0	5	5	1	0	0	39	2	27	1	0	12
TCGA-AD-6899-01A-11D-1924-10	157	0	4	5	0	0	7	55	4	0	1	1	2	1	0	0	59	6	8	1	0	3
TCGA-CM-5862-01A-01D-1650-10	94	1	0	5	0	0	3	31	2	2	1	2	6	2	0	0	24	1	8	0	0	6
TCGA-AY-6196-01A-11D-1719-10	386	15	10	32	0	0	19	65	22	1	1	12	33	9	1	1	86	30	34	0	1	14
TCGA-CK-4948-01B-11D-1650-10	108	3	3	4	1	0	8	34	3	0	0	2	4	2	0	0	31	5	3	0	0	5
TCGA-D5-5539-01A-01D-1650-10	106	0	0	3	0	0	4	35	1	3	0	4	7	2	0	0	35	7	4	0	0	1
TCGA-G4-6626-01A-11D-1771-10	157	0	2	6	0	0	10	55	4	0	0	6	5	0	0	0	48	3	13	0	0	5
TCGA-D5-6537-01A-11D-1719-10	375	10	3	28	0	0	23	91	8	0	0	8	37	8	1	0	87	14	27	0	2	28
TCGA-DM-A28G-01A-11D-A16V-10	114	1	1	4	0	0	11	36	1	1	0	4	2	3	1	0	35	2	7	0	0	5
TCGA-AA-3662-01A-01D-1719-10	516	16	7	75	0	3	23	109	26	1	0	5	51	12	0	1	103	25	33	2	3	21
TCGA-DM-A1D8-01A-11D-A152-10	154	8	4	4	1	0	17	36	4	0	0	8	6	3	2	0	36	6	15	1	0	3
TCGA-CM-6163-01A-11D-1650-10	99	1	0	5	0	0	9	36	1	0	0	1	4	0	0	0	29	4	7	0	0	2
TCGA-D5-7000-01A-11D-1924-10	167	0	1	7	1	1	7	63	5	0	0	3	6	0	0	2	33	3	9	2	2	22
TCGA-CA-6719-01A-11D-1835-10	122	0	1	4	3	0	7	42	1	0	1	0	6	1	0	0	43	0	6	0	0	7

sample_ID	ALL	A_C	A_T	A_G	-_A	A_-	C_A	C_T	C_G	-_C	C_-	T_A	T_C	T_G	-_T	T_-	G_A	G_C	G_T	-_G	G_-	MNC
TCGA-DM-A282-01A-12D-A16V-10	90	1	1	2	0	1	8	28	1	0	0	3	2	0	2	1	29	4	5	0	0	2
TCGA-AD-6888-01A-11D-1924-10	117	4	2	3	0	0	6	33	3	1	0	1	2	1	1	1	46	1	6	0	0	6
TCGA-CA-6716-01A-11D-1835-10	208	1	1	10	4	1	16	74	4	2	0	1	7	5	2	0	62	4	8	1	0	5
TCGA-CK-5912-01A-11D-1650-10	63	0	1	1	0	0	0	29	3	0	1	1	1	1	1	1	13	2	5	0	0	3
TCGA-CK-4950-01A-01D-1719-10	436	11	5	24	1	1	38	98	18	0	2	6	28	14	0	3	112	20	31	0	2	22
TCGA-A6-5660-01A-01D-1650-10	116	3	1	6	1	0	9	34	3	1	0	1	6	1	0	1	36	4	3	1	0	5
TCGA-AZ-6605-01A-11D-1835-10	136	2	1	6	0	0	8	45	4	0	1	0	10	0	1	0	40	5	7	1	0	5
TCGA-DM-A1D6-01A-21D-A152-10	128	2	2	5	0	1	3	47	4	0	0	3	4	2	1	1	38	5	6	0	0	4
TCGA-CM-6165-01A-11D-1650-10	112	0	1	9	0	0	14	26	0	0	1	2	1	0	3	1	26	5	17	0	1	5
TCGA-D5-6541-01A-11D-1719-10	95	1	1	3	1	1	4	31	4	0	1	1	4	1	1	0	29	0	2	1	1	8
TCGA-CM-6677-01A-11D-1835-10	130	1	1	3	0	1	8	45	1	0	0	2	6	2	0	0	45	1	7	2	0	5
TCGA-CM-4744-01A-01D-1408-10	166	4	1	3	0	0	14	47	1	3	1	4	4	3	2	0	48	4	12	0	1	14
TCGA-DM-A1D7-01A-11D-A152-10	141	4	7	3	1	0	20	37	1	0	0	3	4	2	0	0	41	8	9	0	0	1
TCGA-AA-3510-01A-01W-1461-10	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
TCGA-A6-4105-01A-02D-1771-10	149	0	4	7	0	0	11	48	4	1	0	4	5	3	1	0	39	7	11	0	0	4
TCGA-G4-6306-01A-11D-1771-10	124	0	3	7	1	0	11	46	2	0	2	1	4	1	0	0	32	0	4	1	0	9
TCGA-D5-6535-01A-11D-1719-10	160	5	1	7	1	0	14	43	4	0	1	4	6	5	0	0	32	11	21	0	0	5
TCGA-CM-5349-01A-21D-1719-10	104	3	2	7	0	0	4	40	1	0	1	4	2	2	2	0	25	1	3	0	1	6
TCGA-A6-6138-01A-11D-1771-10	117	1	4	4	1	0	3	41	2	0	2	0	7	3	1	0	35	2	9	0	1	1
TCGA-A6-5656-01A-21D-1835-10	106	1	0	1	0	1	6	44	2	0	0	2	5	0	1	1	31	2	4	0	0	5
TCGA-CM-4747-01A-01D-1408-10	96	2	0	2	0	0	6	34	1	0	0	1	6	1	2	1	35	1	2	0	0	2
TCGA-D5-5540-01A-01D-1650-10	132	4	4	1	0	0	8	45	1	2	0	6	6	2	1	0	35	3	11	0	0	3
TCGA-CA-5796-01A-01D-1650-10	66	0	2	5	1	0	6	22	0	1	1	1	1	2	0	0	17	1	5	0	0	1
TCGA-D5-6539-01A-11D-1719-10	66	2	0	6	0	0	6	20	1	0	0	2	0	3	0	0	16	3	3	0	0	4
TCGA-G4-6294-01A-11D-1806-10	184	3	10	5	1	1	11	48	14	0	0	8	8	4	0	1	28	10	19	0	1	12
TCGA-F4-6808-01A-11D-1835-10	123	1	3	3	0	0	10	36	4	0	0	1	3	1	0	3	39	2	11	0	0	6
TCGA-F4-6704-01A-11D-1835-10	35	0	2	0	0	0	1	7	1	0	0	2	0	0	0	0	14	1	2	0	0	5
TCGA-G4-6310-01A-11D-1719-10	117	1	2	1	0	0	5	42	3	0	0	2	3	1	1	0	38	4	10	0	1	3
TCGA-DM-A28M-01A-12D-A16V-10	108	1	2	4	1	0	7	32	0	0	0	2	4	1	2	1	42	1	5	0	1	2
TCGA-A6-6654-01A-21D-1835-10	168	0	1	7	1	1	6	71	3	0	2	4	0	1	1	0	63	0	3	0	1	3
TCGA-D5-6927-01A-21D-1924-10	810	8	10	51	10	30	39	195	5	5	35	3	61	6	11	19	206	5	47	15	23	26
TCGA-A6-2671-01A-01D-1408-10	88	3	1	2	0	0	7	20	1	0	1	1	2	1	1	2	26	2	9	0	0	9
TCGA-CM-4748-01A-01D-1408-10	68	0	0	3	0	0	7	19	2	0	0	1	2	3	1	0	16	2	6	0	1	5
TCGA-AA-3660-01A-01D-1719-10	118	1	0	4	0	0	8	38	1	1	1	2	5	5	0	0	38	2	3	0	0	9
TCGA-D5-6534-01A-21D-1924-10	155	5	5	2	1	1	17	44	3	0	0	5	2	8	0	0	38	5	12	0	0	7
TCGA-F4-6854-01A-11D-1924-10	123	5	3	5	0	0	11	35	2	0	0	6	5	1	2	0	32	1	7	0	1	7
TCGA-CK-4952-01A-01D-1719-10	324	7	5	22	1	1	14	85	8	1	1	3	27	8	0	4	97	6	15	2	0	17
TCGA-G4-6317-01A-11D-1719-10	114	2	0	1	1	0	8	42	2	0	1	5	3	0	0	0	37	2	4	0	0	6
TCGA-CK-4947-01B-11D-1650-10	93	2	3	5	0	1	6	29	5	0	0	3	5	2	0	2	25	0	4	0	0	1
TCGA-CM-6166-01A-11D-1650-10	96	2	0	0	2	0	13	28	5	2	0	2	2	0	1	0	28	0	5	0	0	6
TCGA-DM-A1D4-01A-21D-A152-10	181	2	4	7	1	1	20	59	4	1	0	3	4	3	1	0	53	3	12	1	0	2
TCGA-CM-5341-01A-01D-1408-10	167	2	4	4	0	0	6	53	3	0	3	6	6	3	0	0	47	10	11	1	0	8
TCGA-CM-6170-01A-11D-1650-10	99	2	1	2	0	0	2	39	4	0	0	2	3	1	1	1	32	1	3	1	0	4
TCGA-AZ-6603-01A-11D-1835-10	119	1	1	9	0	3	5	27	3	0	0	1	4	4	1	0	40	2	4	0	0	14
TCGA-A6-5662-01A-01D-1650-10	68	1	1	1	0	1	4	20	1	1	0	3	2	0	1	1	22	1	7	0	0	1
TCGA-AA-3655-01A-02D-1719-10	87	0	0	5	1	0	6	27	1	0	0	1	8	0	0	0	27	4	3	1	0	3
TCGA-CM-4746-01A-01D-1408-10	862	3	4	12	13	33	32	323	4	10	23	2	27	3	11	17	260	6	29	8	17	25
TCGA-A6-5666-01A-01D-1650-10	105	3	1	5	2	0	8	31	3	1	0	1	4	3	2	0	26	1	11	0	1	2
TCGA-CM-6161-01A-11D-1650-10	107	0	5	7	0	1	5	36	2	0	0	2	4	1	1	0	29	3	7	0	0	4
TCGA-AD-6890-01A-11D-1924-10	118	1	0	2	1	1	10	38	1	1	0	2	0	1	1	2	38	1	9	0	1	8
TCGA-D5-6920-01A-11D-1924-10	119	1	2	5	1	1	7	38	5	0	0	3	3	0	0	1	37	1	4	0	2	8
TCGA-AZ-6607-01A-11D-1835-10	69	4	3	2	1	0	5	20	1	0	0	2	1	0	0	0	25	0	2	0	0	3
TCGA-CM-6676-01A-11D-1835-10	92	1	3	2	0	0	4	43	1	0	0	2	0	1	1	0	29	0	4	0	0	1
TCGA-DM-A1D0-01A-11D-A152-10	117	3	6	4	1	0	4	36	7	0	0	7	7	1	0	0	19	4	10	0	1	7
TCGA-AD-6965-01A-11D-1924-10	147	1	3	4	1	0	9	38	4	0	0	3	7	1	2	0	51	4	10	0	1	8
TCGA-A6-6142-01A-11D-1771-10	71	3	2	7	0	3	4	23	0	0	0	3	3	1	0	1	15	0	2	0	0	4
TCGA-D5-6932-01A-11D-1924-10	121	2	5	5	0	1	14	28	6	0	2	7	3	5	0	0	27	4	9	0	0	3
TCGA-F4-6461-01A-11D-1771-10	127	1	3	7	1	0	6	36	2	0	0	1	7	2	0	0	40	5	9	1	0	6

sample_ID	ALL	A_C	A_T	A_G	-_A	A_-	C_A	C_T	C_G	-_C	C_-	T_A	T_C	T_G	-_T	T_-	G_A	G_C	G_T	-_G	G_-	MNC
TCGA-A6-5657-01A-01D-1650-10	88	2	1	0	1	0	12	35	3	1	1	0	1	0	0	0	23	1	7	0	0	0
TCGA-DM-A1HA-01A-11D-A152-10	170	0	6	2	2	0	11	45	3	0	0	5	8	1	1	1	50	12	17	0	2	4
TCGA-CM-6680-01A-11D-1835-10	178	3	0	11	2	1	7	48	4	0	1	5	8	4	0	1	46	4	10	0	0	23
TCGA-A6-6651-01A-21D-1835-10	111	1	2	6	0	2	5	22	6	1	0	3	6	1	0	2	25	3	6	0	1	19
TCGA-D5-6922-01A-11D-1924-10	116	3	1	3	1	0	2	39	7	0	0	1	2	1	1	0	40	5	7	0	0	3
TCGA-F4-6459-01A-11D-1771-10	112	4	5	3	0	2	10	29	7	2	1	3	4	0	0	1	26	7	3	0	2	3
TCGA-AA-3697-01A-01D-1719-10	777	21	5	80	1	1	47	148	48	2	1	12	85	22	1	2	166	40	58	2	5	30
TCGA-AD-6963-01A-11D-1924-10	89	1	3	2	0	0	2	31	6	0	0	5	3	1	0	0	19	6	4	0	0	6
TCGA-CM-5348-01A-21D-1719-10	159	3	4	4	0	0	5	57	3	0	0	3	8	2	0	0	48	8	7	1	1	5
TCGA-DM-A0XD-01A-12D-A152-10	181	2	1	2	1	0	11	62	6	1	1	3	8	2	0	1	60	6	8	0	0	6
TCGA-AA-3511-01A-21D-1835-10	118	3	6	1	0	0	6	35	5	0	0	6	7	7	1	0	22	5	5	1	0	8
TCGA-AZ-6599-01A-11D-1771-10	258	2	1	2	0	0	3	121	3	0	1	2	3	0	0	0	101	2	8	0	2	7
TCGA-CK-5914-01A-11D-1650-10	142	2	4	8	1	0	8	42	3	0	1	4	6	2	0	1	38	6	8	0	0	8
TCGA-CM-6675-01A-11D-1835-10	87	1	3	5	0	0	4	23	2	0	0	1	4	3	3	1	21	1	10	0	0	5
TCGA-AZ-6600-01A-11D-1771-10	149	2	4	4	2	0	10	43	5	0	0	6	13	3	2	0	33	6	10	0	0	6
TCGA-A6-6652-01A-11D-1771-10	90	0	1	5	0	0	9	24	3	0	3	3	2	1	0	1	26	2	5	0	1	4
TCGA-G4-6293-01A-11D-1719-10	233	8	4	13	1	1	15	56	6	1	1	5	15	2	1	1	63	9	9	1	1	20
TCGA-F4-6569-01A-11D-1771-10	109	6	0	5	0	1	4	31	4	0	1	3	3	4	0	1	26	7	9	0	1	3
TCGA-A6-6649-01A-11D-1771-10	129	1	2	0	1	0	4	47	5	0	0	4	7	1	0	0	45	1	6	0	0	5
TCGA-G4-6295-01A-11D-1719-10	150	2	7	7	0	0	9	21	15	1	0	8	11	1	0	0	38	8	16	0	1	5
TCGA-CK-5913-01A-11D-1650-10	852	3	11	54	15	25	54	210	2	8	32	7	41	7	12	23	230	5	48	18	30	17
TCGA-CM-4752-01A-01D-1408-10	152	3	3	8	0	3	5	45	2	0	0	0	8	0	0	0	56	3	10	0	0	6
TCGA-AA-3502-01A-01D-1408-10	170	3	4	2	2	1	7	56	5	0	0	2	3	2	0	0	60	7	14	0	0	2
TCGA-D5-5538-01A-01D-1650-10	111	1	5	7	0	0	8	30	2	1	1	3	6	1	0	0	28	4	13	0	1	0
TCGA-CM-6678-01A-11D-1835-10	132	4	2	2	0	0	6	59	0	0	0	1	2	0	2	0	43	1	4	0	1	5
TCGA-AZ-4681-01A-01D-1408-10	2	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TCGA-CM-6674-01A-11D-1835-10	905	3	7	27	8	25	29	292	2	12	24	8	40	6	15	23	284	3	28	9	26	34
TCGA-AZ-5403-01A-01D-1650-10	87	1	4	4	0	0	5	26	3	0	0	3	1	2	0	1	26	3	4	0	0	4
TCGA-AZ-4315-01A-01W-1461-10	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
TCGA-A6-5667-01A-21D-1719-10	217	3	3	19	2	2	16	46	7	1	0	6	18	5	0	0	54	10	9	0	2	14
TCGA-DM-A28E-01A-11D-A16V-10	12	0	0	0	0	4	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	6
TCGA-G4-6625-01A-21D-1771-10	88	1	2	1	2	0	7	18	4	1	0	3	5	0	1	0	29	5	7	0	0	2
TCGA-G4-6299-01A-11D-1771-10	146	5	7	4	3	0	11	38	6	0	1	3	3	2	0	0	36	8	8	0	3	8
TCGA-DM-A0X9-01A-11D-A152-10	186	4	6	5	0	0	8	65	5	0	0	7	7	3	3	0	42	3	20	0	1	7
TCGA-A6-5659-01A-01D-1650-10	188	3	3	5	1	1	24	78	2	0	0	3	5	2	0	0	42	2	12	0	1	4
TCGA-DM-A28C-01A-11D-A16V-10	95	1	1	3	0	1	9	32	4	0	1	0	3	0	1	0	21	6	3	1	2	6
TCGA-AA-3712-01A-21D-1719-10	587	16	4	61	1	1	32	112	54	0	1	9	55	12	1	0	106	46	52	1	1	22
TCGA-AZ-4616-01A-21D-1835-10	62	2	0	2	0	0	3	20	2	0	0	2	2	4	1	0	18	3	2	0	0	1
TCGA-D5-6931-01A-11D-1924-10	320	25	8	20	0	1	16	82	3	1	0	8	18	30	2	0	78	2	16	1	0	9
TCGA-G4-6309-01A-21D-1835-10	861	6	5	26	14	29	20	267	3	11	25	5	26	5	17	22	269	3	28	7	28	45
TCGA-CM-4746-01A-01W-1461-10	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TCGA-A6-6648-01A-11D-1771-10	93	1	1	4	0	1	4	32	3	0	0	5	3	1	0	0	29	1	3	0	1	4
TCGA-D5-6929-01A-31D-1924-10	134	3	2	7	0	1	16	32	5	0	0	3	6	2	1	2	29	3	13	0	0	9
TCGA-G4-6322-01A-11D-1719-10	95	1	0	3	2	0	7	26	2	0	1	2	3	0	0	0	32	2	8	0	0	6
TCGA-D5-6538-01A-11D-1719-10	135	4	3	5	1	0	4	47	0	0	0	4	5	2	1	1	46	3	5	1	0	3
TCGA-AZ-5407-01A-01D-1719-10	316	12	9	31	1	4	15	76	15	0	1	4	27	6	0	0	62	18	12	0	1	22
TCGA-G4-6323-01A-11D-1719-10	82	1	0	4	0	0	5	25	3	0	0	2	3	0	0	1	26	3	7	0	0	2
TCGA-DM-A0XF-01A-11D-A152-10	135	1	5	1	0	2	10	38	5	0	0	6	7	2	0	0	28	5	15	0	2	8
TCGA-G4-6307-01A-11D-1719-10	74	1	0	1	0	0	12	20	1	0	0	3	3	1	1	0	15	5	6	0	2	3
TCGA-G4-6298-01A-11D-1719-10	368	6	5	35	2	0	20	85	11	0	0	5	35	8	0	2	85	16	28	1	1	23
TCGA-D5-6924-01A-11D-1924-10	174	2	5	1	2	0	11	47	9	0	0	10	7	7	0	0	43	6	13	0	0	11
TCGA-F4-6463-01A-11D-1719-10	246	2	4	25	0	2	11	54	13	1	1	5	22	6	1	0	57	12	9	1	1	19
TCGA-DM-A285-01A-11D-A16V-10	121	4	3	7	0	0	14	21	3	0	1	8	2	0	0	2	35	7	8	1	0	5
TCGA-D5-5537-01A-21D-1924-10	129	1	2	3	2	0	8	43	2	0	0	0	4	2	1	0	39	2	12	1	1	6
TCGA-AY-6386-01A-21D-1719-10	543	16	9	54	2	0	26	124	22	0	0	8	72	13	2	2	110	29	25	0	0	29
TCGA-CM-6168-01A-11D-1650-10	206	0	0	9	0	2	15	73	9	1	0	2	13	1	0	1	49	6	21	0	1	3
TCGA-DM-A1D9-01A-11D-A152-10	144	1	9	4	2	1	9	38	6	0	0	8	2	3	1	2	30	9	10	0	0	9
TCGA-CM-5344-01A-21D-1719-10	64	0	2	2	0	0	4	17	3	0	0	0	0	1	1	0	22	1	7	0	1	3

sample_ID	ALL	A_C	A_T	A_G	-_A	A_-	C_A	C_T	C_G	-_C	C_-	T_A	T_C	T_G	-_T	T_-	G_A	G_C	G_T	-_G	G_-	MNC
TCGA-CM-6172-01A-11D-1650-10	94	3	1	1	0	0	3	37	3	0	0	1	1	2	0	0	37	0	5	0	0	0
TCGA-A6-6782-01A-11D-1835-10	144	5	6	4	1	0	16	39	3	0	1	5	5	1	0	0	38	7	7	0	0	6
TCGA-G4-6303-01A-11D-1771-10	105	3	1	1	2	1	6	31	6	0	0	1	2	2	1	0	38	0	6	0	1	3
TCGA-A6-6137-01A-11D-1771-10	118	2	1	1	0	0	5	39	4	0	0	3	2	3	3	0	43	2	5	0	1	4
TCGA-F4-6855-01A-11D-1924-10	154	2	2	4	0	0	18	45	4	0	0	2	8	1	2	0	43	3	10	0	0	10
TCGA-G4-6315-01A-11D-1719-10	166	3	2	4	1	2	11	52	2	0	0	6	5	2	0	0	45	8	15	1	1	6

Figure 12. Low Mutation Count Group Mutation Type Counts.

ALL is the total sum of mutations for a given sample. MNC stands for multi-nucleotide change. All other columns were composed of a pattern of REFERENCE-BASE_MUTATED-BASE, with a dash standing in for missing bases in the case of insertions or deletions. (Insertions were of the form -_INSERTED-BASE and deletions were of the form REFERENCE-BASE_-).

sample_ID	ALL	A_C	A_T	A_G	-_A	A_-	C_A	C_T	C_G	-_C	C_-	T_A	T_C	T_G	-_T	T_-	G_A	G_C	G_T	-_G	G_-	MNC
TCGA-AA-3713-01A-21D-1719-10	1596	16	20	83	25	50	59	403	25	17	83	14	75	24	22	39	392	15	87	23	68	56
TCGA-AD-6889-01A-11D-1924-10	1990	16	23	112	28	81	120	466	8	24	78	9	108	12	30	59	520	9	104	27	92	64
TCGA-CM-6162-01A-11D-1650-10	1168	4	3	24	14	18	29	426	0	7	20	3	32	4	9	16	477	6	27	9	23	17
TCGA-A6-6141-01A-11D-1771-10	1010	49	5	26	0	0	189	247	2	0	0	1	25	49	1	1	253	3	156	0	0	3
TCGA-AZ-4615-01A-01D-1408-10	1847	26	16	157	19	95	85	434	27	16	41	15	129	20	12	50	444	36	103	7	46	69
TCGA-A6-6781-01A-22D-1924-10	1897	3	9	49	20	62	55	670	4	14	70	4	32	5	29	48	637	9	56	9	66	46
TCGA-G4-6302-01A-11D-1719-10	1044	16	7	77	0	13	80	295	18	0	3	11	82	15	0	2	319	21	65	1	2	17
TCGA-AD-6964-01A-11D-1924-10	1395	14	36	110	17	30	61	390	0	13	27	35	103	3	17	25	386	4	63	3	37	21
TCGA-F4-6570-01A-11D-1771-10	1550	19	24	72	17	55	92	423	9	16	57	19	49	11	19	39	434	8	84	6	54	43
TCGA-G4-6586-01A-11D-1771-10	1279	16	10	75	11	50	52	356	6	10	45	6	75	15	9	48	318	9	59	15	63	31
TCGA-CA-6717-01A-11D-1835-10	7007	594	46	404	9	1	1094	1398	10	1	0	61	394	582	7	4	1352	15	1030	0	0	5
TCGA-D5-6928-01A-11D-1924-10	1680	14	14	54	12	24	101	524	4	10	34	22	63	13	8	21	566	6	128	5	26	31
TCGA-AD-5900-01A-11D-1650-10	1521	12	12	68	16	43	89	413	2	13	57	15	71	12	23	32	432	8	70	16	66	51
TCGA-AA-3663-01A-01D-1719-10	2894	57	20	281	30	74	168	578	35	19	109	32	309	60	18	90	575	41	192	24	96	86
TCGA-AZ-4315-01A-01D-1408-10	6086	326	18	320	13	2	678	1690	10	3	1	24	355	324	9	5	1641	12	654	0	0	1
TCGA-A6-6653-01A-11D-1771-10	1148	7	3	55	15	28	48	330	6	15	46	10	48	10	12	38	346	5	39	9	40	38
TCGA-AZ-6598-01A-11D-1771-10	2715	15	25	126	34	107	121	706	7	29	127	21	110	21	29	98	712	13	119	39	149	107
TCGA-A6-5665-01A-01D-1650-10	1937	27	20	156	40	65	85	431	9	51	68	22	147	18	24	45	442	8	93	41	69	76
TCGA-CM-6171-01A-11D-1650-10	1231	11	6	60	9	40	58	348	7	20	49	9	52	8	15	36	335	9	53	13	45	48
TCGA-AU-6004-01A-11D-1719-10	1483	15	8	97	20	64	54	368	22	19	59	8	99	18	14	38	364	20	63	12	62	59
TCGA-CK-5916-01A-11D-1650-10	1660	12	11	89	16	55	70	433	7	15	70	12	98	8	23	45	467	12	65	18	78	56
TCGA-AM-5820-01A-01D-1650-10	7183	142	131	851	16	33	278	1731	280	15	29	104	800	187	11	30	1842	282	269	10	39	103
TCGA-D5-6540-01A-11D-1719-10	1378	20	17	35	17	51	90	370	11	15	73	10	35	15	17	33	377	3	85	14	59	31
TCGA-G4-6320-01A-11D-1719-10	1013	5	7	23	6	19	61	350	3	8	35	8	26	3	5	13	327	7	61	5	25	16
TCGA-CM-4743-01A-01D-1719-10	1122	13	9	77	9	43	60	248	5	11	52	8	54	9	19	24	297	10	67	14	58	35
TCGA-D5-6930-01A-11D-1924-10	1331	7	8	24	10	37	61	433	7	8	50	15	37	8	15	35	443	3	57	6	39	28
TCGA-CM-5861-01A-01D-1650-10	1463	20	13	161	23	30	92	330	11	18	49	16	134	26	16	30	322	3	73	14	54	28
TCGA-AD-6895-01A-11D-1924-10	1725	11	74	66	19	35	52	552	18	13	33	86	71	10	13	29	502	13	57	6	25	40
TCGA-F4-6856-01A-11D-1924-10	1173	6	16	51	17	31	38	338	3	13	41	11	43	13	17	31	363	6	34	11	44	46
TCGA-AZ-6601-01A-11D-1771-10	1826	8	9	47	2	8	112	708	5	7	7	11	55	13	5	7	717	3	79	3	12	8
TCGA-G4-6588-01A-11D-1771-10	2180	21	19	153	25	52	112	531	13	35	101	24	161	15	20	63	559	8	113	14	90	51
TCGA-A6-6780-01A-11D-1835-10	1257	19	17	85	20	71	71	257	11	18	61	4	72	21	20	47	296	7	55	14	42	49
TCGA-CA-6718-01A-11D-1835-10	2671	216	12	108	4	0	497	513	4	1	1	16	93	226	4	0	513	6	453	0	0	4
TCGA-AA-3492-01A-01D-1408-10	2224	14	34	117	22	97	97	618	14	18	61	28	112	24	19	50	621	13	101	18	80	66
TCGA-A6-5661-01A-01D-1650-10	1178	6	11	40	32	37	79	339	5	13	32	12	46	5	21	23	340	7	55	12	30	33
TCGA-AM-5821-01A-01D-1650-10	5373	105	103	661	3	4	181	1418	190	0	3	79	648	108	1	2	1374	221	182	2	5	83
TCGA-AY-6197-01A-11D-1719-10	1634	15	17	164	17	66	80	365	22	7	55	22	162	24	18	35	341	23	87	18	44	52
TCGA-AA-3510-01A-01D-1408-10	2963	136	6	74	5	0	458	799	1	0	0	17	72	128	0	0	804	2	455	1	1	4
TCGA-G4-6628-01A-11D-1835-10	1959	7	15	60	46	65	113	540	15	14	62	16	68	9	52	60	603	11	99	14	49	41

Figure 13. High Mutation Count Group Mutation Type Counts

Columns are as described for Figure 12.

I created additional data columns based on potentially relevant categories by adding together the values present in the above table. These values included insertions, a sum of all single-base insertions, deletions, a sum of all single base deletions, indels, a sum of both of those, transitions, and transversions, sums of the mutations that matched these definitions, and a set of mutations that would be considered indistinguishable due to double-strand base-pairing ({A-T and T-A}, {C-G and G-C}, {C-T and G-A}, {A-C and T-G}, {A-G and T-C}, {C-A and G-T}). The values for these sums and their proportions were also included in the t tests.

Figure 14 depicts these values for the low mutation group. Insertions and deletions showed quite a bit of variability. In some tumors the insertions and deletions roughly matched, while in others the numbers were very different. Indels in general were less frequent than other single-base mutations. With the chemically matching mutations added together, the dominance of the C_T mutation became even more apparent.

In Figure 15, the data for the high mutation group mutation type categories drive home again how many more mutations these samples had on average than the others. C-T and G_A were once again the most common, with C_A and G_T generally having slightly more mutations than A_G and T_C. It was interesting that the other 3 mutation types were still fairly low. There were quite a lot of single base indels in these samples as well.

sample_ID	ALL	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-CM-6169-01A-11D-1650-10	129	1	3	4	81	35	5	4	75	4	6	22
TCGA-D5-5541-01A-01D-1650-10	101	5	2	7	71	18	4	3	61	3	10	8
TCGA-G4-6314-01A-11D-1719-10	143	1	0	1	91	50	7	17	81	5	10	21
TCGA-G4-6297-01A-11D-1719-10	305	3	8	11	199	77	8	26	150	8	49	35
TCGA-F4-6703-01A-11D-1835-10	450	6	34	40	303	82	12	6	251	12	52	52
TCGA-DM-A1DA-01A-11D-A152-10	344	0	4	4	195	134	16	29	158	17	37	72
TCGA-F4-6806-01A-11D-1835-10	90	4	0	4	63	16	2	4	60	1	3	9
TCGA-A6-6650-01A-11D-1771-10	149	1	2	3	89	44	5	7	79	7	10	25
TCGA-F4-6805-01A-11D-1835-10	53	0	2	2	32	15	3	4	25	2	7	6
TCGA-D5-6898-01A-11D-1924-10	76	3	0	3	48	24	5	5	44	1	4	13
TCGA-CA-5797-01A-01D-1650-10	56	2	2	4	34	17	4	5	32	1	2	7
TCGA-G4-6321-01A-11D-1719-10	377	0	2	2	254	98	10	31	175	16	79	41
TCGA-CM-6679-01A-11D-1835-10	48	2	0	2	38	7	4	1	37	0	1	2
TCGA-F4-6809-01A-11D-1835-10	124	3	3	6	60	46	11	10	48	2	12	23
TCGA-CK-5915-01A-11D-1650-10	143	4	1	5	109	26	4	9	95	3	14	10
TCGA-D5-6533-01A-11D-1719-10	171	1	2	3	99	61	18	12	78	7	21	24
TCGA-AU-3779-01A-01D-1719-10	104	0	0	0	71	30	7	3	62	5	9	15
TCGA-AD-6901-01A-11D-1924-10	250	3	3	6	155	82	11	22	139	8	16	41
TCGA-D5-6529-01A-11D-1771-10	117	0	3	3	89	21	3	7	81	1	8	10
TCGA-CM-6164-01A-11D-1650-10	60	1	1	2	38	18	5	2	28	1	10	10
TCGA-DM-A28F-01A-11D-A16V-10	239	8	4	12	131	89	15	20	113	22	18	32
TCGA-G4-6311-01A-11D-1719-10	337	1	5	6	205	104	13	25	152	18	53	48
TCGA-D5-6532-01A-11D-1719-10	111	2	3	5	67	36	5	12	61	8	6	11
TCGA-AZ-4682-01B-01D-1408-10	96	0	1	1	75	18	4	2	72	4	3	8
TCGA-CM-5863-01A-21D-1835-10	84	0	4	4	52	19	1	5	44	2	8	11
TCGA-D5-6536-01A-11D-1719-10	103	2	1	3	65	31	6	8	58	2	7	15
TCGA-DM-A28H-01A-11D-A16V-10	120	4	4	8	70	33	6	12	63	6	7	9
TCGA-F4-6460-01A-11D-1771-10	127	1	1	2	76	48	10	10	67	4	9	24
TCGA-D5-6531-01A-11D-1719-10	197	1	2	3	124	61	16	12	110	8	14	25
TCGA-G4-6304-01A-11D-1924-10	294	2	13	15	222	57	13	6	194	7	28	31
TCGA-D5-6926-01A-11D-1924-10	131	1	0	1	91	31	4	9	79	3	12	15
TCGA-CM-5868-01A-01D-1650-10	189	3	4	7	114	54	12	6	84	9	30	27
TCGA-A6-2675-01A-02D-1719-10	83	1	0	1	65	14	4	4	57	2	8	4
TCGA-DM-A1DB-01A-11D-A152-10	80	0	1	1	54	21	1	9	46	1	8	10
TCGA-CM-4750-01A-01D-1408-10	99	1	1	2	65	25	4	0	59	4	6	17
TCGA-AY-5543-01A-01D-1650-10	126	1	1	2	89	35	3	10	77	5	12	17
TCGA-AD-6548-01A-11D-1835-10	105	1	2	3	78	20	3	3	70	3	8	11
TCGA-AZ-4323-01A-21D-1835-10	33	0	0	0	24	8	2	2	22	1	2	3
TCGA-CM-5864-01A-01D-1650-10	221	1	4	5	129	83	11	19	109	13	20	40
TCGA-AA-3489-01A-21D-1835-10	82	2	0	2	52	23	7	2	48	2	4	12
TCGA-A6-6140-01A-11D-1771-10	505	2	0	2	377	123	18	36	326	18	51	51
TCGA-A6-5664-01A-21D-1835-10	52	0	0	0	39	8	0	0	34	2	5	6
TCGA-CM-5860-01A-01D-1650-10	122	1	0	1	91	26	1	5	82	1	9	19
TCGA-DM-A28K-01A-21D-A16V-10	115	5	1	6	82	19	3	5	77	2	5	9
TCGA-F4-6807-01A-11D-1835-10	160	1	1	2	78	68	9	7	69	2	9	50
TCGA-AD-6899-01A-11D-1924-10	157	1	1	2	121	31	5	10	114	1	7	15
TCGA-CM-5862-01A-01D-1650-10	94	2	1	3	66	19	2	3	55	3	11	11
TCGA-AY-6196-01A-11D-1719-10	386	2	3	5	216	151	22	52	151	24	65	53
TCGA-CK-4948-01B-11D-1650-10	108	1	0	1	73	29	5	8	65	5	8	11
TCGA-D5-5539-01A-01D-1650-10	106	3	0	3	80	22	4	8	70	2	10	8
TCGA-G4-6626-01A-11D-1771-10	157	0	0	0	114	38	8	7	103	0	11	23
TCGA-D5-6537-01A-11D-1719-10	375	1	2	3	243	101	11	22	178	18	65	50
TCGA-DM-A28G-01A-11D-A16V-10	114	2	0	2	77	30	5	3	71	4	6	18
TCGA-AA-3662-01A-01D-1719-10	516	3	7	10	338	147	12	51	212	28	126	56
TCGA-DM-A1D8-01A-11D-A152-10	154	4	0	4	82	65	12	10	72	11	10	32
TCGA-CM-6163-01A-11D-1650-10	99	0	0	0	74	23	1	5	65	1	9	16
TCGA-D5-7000-01A-11D-1924-10	167	3	5	8	109	28	4	8	96	0	13	16

sample_ID	ALL	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-CA-6719-01A-11D-1835-10	122	3	1	4	95	16	1	1	85	1	10	13
TCGA-DM-A282-01A-12D-A16V-10	90	2	2	4	61	23	4	5	57	1	4	13
TCGA-AD-6888-01A-11D-1924-10	117	2	1	3	84	24	3	4	79	5	5	12
TCGA-CA-6716-01A-11D-1835-10	208	9	1	10	153	40	2	8	136	6	17	24
TCGA-CK-5912-01A-11D-1650-10	63	1	2	3	44	13	2	5	42	1	2	5
TCGA-CK-4950-01A-01D-1719-10	436	1	8	9	262	143	11	38	210	25	52	69
TCGA-A6-5660-01A-01D-1650-10	116	3	1	4	82	25	2	7	70	4	12	12
TCGA-AZ-6605-01A-11D-1835-10	136	2	1	3	101	27	1	9	85	2	16	15
TCGA-DM-A1D6-01A-21D-A152-10	128	1	2	3	94	27	5	9	85	4	9	9
TCGA-CM-6165-01A-11D-1650-10	112	3	3	6	62	39	3	5	52	0	10	31
TCGA-D5-6541-01A-11D-1719-10	95	3	3	6	67	14	2	4	60	2	7	6
TCGA-CM-6677-01A-11D-1835-10	130	2	1	3	99	23	3	2	90	3	9	15
TCGA-CM-4744-01A-01D-1408-10	166	5	2	7	102	43	5	5	95	7	7	26
TCGA-DM-A1D7-01A-11D-A152-10	141	1	0	1	85	54	10	9	78	6	7	29
TCGA-AA-3510-01A-01W-1461-10	1	0	0	0	1	0	0	0	1	0	0	0
TCGA-A6-4105-01A-02D-1771-10	149	2	0	2	99	44	8	11	87	3	12	22
TCGA-G4-6306-01A-11D-1771-10	124	2	2	4	89	22	4	2	78	1	11	15
TCGA-D5-6535-01A-11D-1719-10	160	1	1	2	88	65	5	15	75	10	13	35
TCGA-CM-5349-01A-21D-1719-10	104	2	2	4	74	20	6	2	65	5	9	7
TCGA-A6-6138-01A-11D-1771-10	117	2	3	5	87	24	4	4	76	4	11	12
TCGA-A6-5656-01A-21D-1835-10	106	1	2	3	81	17	2	4	75	1	6	10
TCGA-CM-4747-01A-01D-1408-10	96	2	1	3	77	14	1	2	69	3	8	8
TCGA-D5-5540-01A-01D-1650-10	132	3	0	3	87	39	10	4	80	6	7	19
TCGA-CA-5796-01A-01D-1650-10	66	2	1	3	45	17	3	1	39	2	6	11
TCGA-D5-6539-01A-11D-1719-10	66	0	0	0	42	20	2	4	36	5	6	9
TCGA-G4-6294-01A-11D-1806-10	184	1	3	4	89	79	18	24	76	7	13	30
TCGA-F4-6808-01A-11D-1835-10	123	0	3	3	81	33	4	6	75	2	6	21
TCGA-F4-6704-01A-11D-1835-10	35	0	0	0	21	9	4	2	21	0	0	3
TCGA-G4-6310-01A-11D-1719-10	117	1	1	2	84	28	4	7	80	2	4	15
TCGA-DM-A28M-01A-12D-A16V-10	108	3	2	5	82	19	4	1	74	2	8	12
TCGA-A6-6654-01A-21D-1835-10	168	2	4	6	141	18	5	3	134	1	7	9
TCGA-D5-6927-01A-21D-1924-10	810	41	107	148	513	123	13	10	401	14	112	86
TCGA-A6-2671-01A-01D-1408-10	88	1	3	4	50	25	2	3	46	4	4	16
TCGA-CM-4748-01A-01D-1408-10	68	1	1	2	40	21	1	4	35	3	5	13
TCGA-AA-3660-01A-01D-1719-10	118	1	1	2	85	22	2	3	76	6	9	11
TCGA-D5-6534-01A-21D-1924-10	155	1	1	2	86	60	10	8	82	13	4	29
TCGA-F4-6854-01A-11D-1924-10	123	2	1	3	77	36	9	3	67	6	10	18
TCGA-CK-4952-01A-01D-1719-10	324	4	6	10	231	66	8	14	182	15	49	29
TCGA-G4-6317-01A-11D-1719-10	114	1	1	2	83	23	5	4	79	2	4	12
TCGA-CK-4947-01B-11D-1650-10	93	0	3	3	64	25	6	5	54	4	10	10
TCGA-CM-6166-01A-11D-1650-10	96	5	0	5	58	27	2	5	56	2	2	18
TCGA-DM-A1D4-01A-21D-A152-10	181	4	1	5	123	51	7	7	112	5	11	32
TCGA-CM-5341-01A-01D-1408-10	167	1	3	4	110	45	10	13	100	5	10	17
TCGA-CM-6170-01A-11D-1650-10	99	2	1	3	76	16	3	5	71	3	5	5
TCGA-AZ-6603-01A-11D-1835-10	119	1	3	4	80	21	2	5	67	5	13	9
TCGA-A6-5662-01A-01D-1650-10	68	2	2	4	45	18	4	2	42	1	3	11
TCGA-AA-3655-01A-02D-1719-10	87	2	0	2	67	15	1	5	54	0	13	9
TCGA-CM-4746-01A-01D-1408-10	862	42	90	132	622	83	6	10	583	6	39	61
TCGA-A6-5666-01A-01D-1650-10	105	5	1	6	66	31	2	4	57	6	9	19
TCGA-CM-6161-01A-11D-1650-10	107	1	1	2	76	25	7	5	65	1	11	12
TCGA-AD-6890-01A-11D-1924-10	118	3	4	7	78	25	2	2	76	2	2	19
TCGA-D5-6920-01A-11D-1924-10	119	1	4	5	83	23	5	6	75	1	8	11
TCGA-AZ-6607-01A-11D-1835-10	69	1	0	1	48	17	5	1	45	4	3	7
TCGA-CM-6676-01A-11D-1835-10	92	1	0	1	74	16	5	1	72	2	2	8
TCGA-DM-A1D0-01A-11D-A152-10	117	1	1	2	66	42	13	11	55	4	11	14
TCGA-AD-6965-01A-11D-1924-10	147	3	1	4	100	35	6	8	89	2	11	19
TCGA-A6-6142-01A-11D-1771-10	71	0	4	4	48	15	5	0	38	4	10	6
TCGA-D5-6932-01A-11D-1924-10	121	0	3	3	63	52	12	10	55	7	8	23

sample_ID	ALL	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-F4-6461-01A-11D-1771-10	127	2	0	2	90	29	4	7	76	3	14	15
TCGA-A6-5657-01A-01D-1650-10	88	2	1	3	59	26	1	4	58	2	1	19
TCGA-DM-A1HA-01A-11D-A152-10	170	3	3	6	105	55	11	15	95	1	10	28
TCGA-CM-6680-01A-11D-1835-10	178	2	3	5	113	37	5	8	94	7	19	17
TCGA-A6-6651-01A-21D-1835-10	111	1	5	6	59	27	5	9	47	2	12	11
TCGA-D5-6922-01A-11D-1924-10	116	2	0	2	84	27	2	12	79	4	5	9
TCGA-F4-6459-01A-11D-1771-10	112	2	6	8	62	39	8	14	55	4	7	13
TCGA-AA-3697-01A-01D-1719-10	777	6	9	15	479	253	17	88	314	43	165	105
TCGA-AD-6963-01A-11D-1924-10	89	0	0	0	55	28	8	12	50	2	5	6
TCGA-CM-5348-01A-21D-1719-10	159	1	1	2	117	35	7	11	105	5	12	12
TCGA-DM-A0XD-01A-12D-A152-10	181	2	2	4	132	39	4	12	122	4	10	19
TCGA-AA-3511-01A-21D-1835-10	118	2	0	2	65	43	12	10	57	10	8	11
TCGA-AZ-6599-01A-11D-1771-10	258	0	3	3	227	21	3	5	222	2	5	11
TCGA-CK-5914-01A-11D-1650-10	142	1	2	3	94	37	8	9	80	4	14	16
TCGA-CM-6675-01A-11D-1835-10	87	3	1	4	53	25	4	3	44	4	9	14
TCGA-AZ-6600-01A-11D-1771-10	149	4	0	4	93	46	10	11	76	5	17	20
TCGA-A6-6652-01A-11D-1771-10	90	0	5	5	57	24	4	5	50	1	7	14
TCGA-G4-6293-01A-11D-1719-10	233	4	4	8	147	58	9	15	119	10	28	24
TCGA-F4-6569-01A-11D-1771-10	109	0	4	4	65	37	3	11	57	10	8	13
TCGA-A6-6649-01A-11D-1771-10	129	1	0	1	99	24	6	6	92	2	7	10
TCGA-G4-6295-01A-11D-1719-10	150	1	1	2	77	66	15	23	59	3	18	25
TCGA-CK-5913-01A-11D-1650-10	852	53	110	163	535	137	18	7	440	10	95	102
TCGA-CM-4752-01A-01D-1408-10	152	0	3	3	117	26	3	5	101	3	16	15
TCGA-AA-3502-01A-01D-1408-10	170	2	1	3	121	44	6	12	116	5	5	21
TCGA-D5-5538-01A-01D-1650-10	111	1	2	3	71	37	8	6	58	2	13	21
TCGA-CM-6678-01A-11D-1835-10	132	2	1	3	106	18	3	1	102	4	4	10
TCGA-AZ-4681-01A-01D-1408-10	2	0	0	0	2	0	0	0	2	0	0	0
TCGA-CM-6674-01A-11D-1835-10	905	44	98	142	643	86	15	5	576	9	67	57
TCGA-AZ-5403-01A-01D-1650-10	87	0	1	1	57	25	7	6	52	3	5	9
TCGA-AZ-4315-01A-01W-1461-10	3	0	0	0	2	1	0	0	2	0	0	1
TCGA-A6-5667-01A-21D-1719-10	217	3	4	7	137	59	9	17	100	8	37	25
TCGA-DM-A28E-01A-11D-A16V-10	12	1	5	6	0	0	0	0	0	0	0	0
TCGA-G4-6625-01A-21D-1771-10	88	4	0	4	53	29	5	9	47	1	6	14
TCGA-G4-6299-01A-11D-1771-10	146	3	4	7	81	50	10	14	74	7	7	19
TCGA-DM-A0X9-01A-11D-A152-10	186	3	1	4	119	56	13	8	107	7	12	28
TCGA-A6-5659-01A-01D-1650-10	188	1	2	3	130	51	6	4	120	5	10	36
TCGA-DM-A28C-01A-11D-A16V-10	95	2	4	6	59	24	1	10	53	1	6	12
TCGA-AA-3712-01A-21D-1719-10	587	3	3	6	334	225	13	100	218	28	116	84
TCGA-AZ-4616-01A-21D-1835-10	62	1	0	1	42	18	2	5	38	6	4	5
TCGA-D5-6931-01A-11D-1924-10	320	4	1	5	198	108	16	5	160	55	38	32
TCGA-G4-6309-01A-21D-1835-10	861	49	104	153	588	75	10	6	536	11	52	48
TCGA-CM-4746-01A-01W-1461-10	1	0	0	0	1	0	0	0	1	0	0	0
TCGA-A6-6648-01A-11D-1771-10	93	0	2	2	68	19	6	4	61	2	7	7
TCGA-D5-6929-01A-31D-1924-10	134	1	3	4	74	47	5	8	61	5	13	29
TCGA-G4-6322-01A-11D-1719-10	95	2	1	3	64	22	2	4	58	1	6	15
TCGA-D5-6538-01A-11D-1719-10	135	3	1	4	103	25	7	3	93	6	10	9
TCGA-AZ-5407-01A-01D-1719-10	316	1	6	7	196	91	13	33	138	18	58	27
TCGA-G4-6323-01A-11D-1719-10	82	0	1	1	58	21	2	6	51	1	7	12
TCGA-DM-A0XF-01A-11D-A152-10	135	0	4	4	74	49	11	10	66	3	8	25
TCGA-G4-6307-01A-11D-1719-10	74	1	2	3	39	29	3	6	35	2	4	18
TCGA-G4-6298-01A-11D-1719-10	368	3	3	6	240	99	10	27	170	14	70	48
TCGA-D5-6924-01A-11D-1924-10	174	2	0	2	98	63	15	15	90	9	8	24
TCGA-F4-6463-01A-11D-1719-10	246	3	4	7	158	62	9	25	111	8	47	20
TCGA-DM-A285-01A-11D-A16V-10	121	1	3	4	65	47	11	10	56	4	9	22
TCGA-D5-5537-01A-21D-1924-10	129	4	1	5	89	29	2	4	82	3	7	20
TCGA-AY-6386-01A-21D-1719-10	543	4	2	6	360	148	17	51	234	29	126	51
TCGA-CM-6168-01A-11D-1650-10	206	1	4	5	144	54	2	15	122	1	22	36
TCGA-DM-A1D9-01A-11D-A152-10	144	3	3	6	74	55	17	15	68	4	6	19

sample_ID	ALL	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-CM-5344-01A-21D-1719-10	64	1	1	2	41	18	2	4	39	1	2	11
TCGA-CM-6172-01A-11D-1650-10	94	0	0	0	76	18	2	3	74	5	2	8
TCGA-A6-6782-01A-11D-1835-10	144	1	1	2	86	50	11	10	77	6	9	23
TCGA-G4-6303-01A-11D-1771-10	105	3	2	5	72	25	2	6	69	5	3	12
TCGA-A6-6137-01A-11D-1771-10	118	3	1	4	85	25	4	6	82	5	3	10
TCGA-F4-6855-01A-11D-1924-10	154	2	0	2	100	42	4	7	88	3	12	28
TCGA-G4-6315-01A-11D-1719-10	166	2	3	5	106	49	8	10	97	5	9	26

Figure 14. Low mutation Group Mutation Type Count Categories

ALL is the same as Figures 12 and 13 and was repeated for visual comparative purposes.

INS was a sum of all insertions. DEL was a sum of all deletions. INDEL was a sum of all insertions and deletions. Transition was a sum of all transition mutations, and transversion was a sum of all transversion mutations.

The remaining columns were sums of chemically equivalent mutation types. These were represented using a pattern of base1-base2 with a ‘|’ character separating the pair of mutation representations. Thus an A to T mutation, which could be considered biochemically equivalent to a T to A mutation due to base pairing was represented as “A-T|T-A”, and had a count formed by the sum of the A_T and T_A columns from the previous table.

sample_ID	ALL	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-AA-3713-01A-21D-1719-10	1596	87	240	327	953	260	34	40	795	40	158	146
TCGA-AD-6889-01A-11D-1924-10	1990	109	310	419	1206	301	32	17	986	28	220	224
TCGA-CM-6162-01A-11D-1650-10	1168	39	77	116	959	76	6	6	903	8	56	56
TCGA-A6-6141-01A-11D-1771-10	1010	1	1	2	551	454	6	5	500	98	51	345
TCGA-AZ-4615-01A-01D-1408-10	1847	54	232	286	1164	328	31	63	878	46	286	188
TCGA-A6-6781-01A-22D-1924-10	1897	72	246	318	1388	145	13	13	1307	8	81	111
TCGA-G4-6302-01A-11D-1719-10	1044	1	20	21	773	233	18	39	614	31	159	145
TCGA-AD-6964-01A-11D-1924-10	1395	50	119	169	989	216	71	4	776	17	213	124
TCGA-F4-6570-01A-11D-1771-10	1550	58	205	263	978	266	43	17	857	30	121	176
TCGA-G4-6586-01A-11D-1771-10	1279	45	206	251	824	173	16	15	674	31	150	111
TCGA-CA-6717-01A-11D-1835-10	7007	17	5	22	3548	3432	107	25	2750	1176	798	2124
TCGA-D5-6928-01A-11D-1924-10	1680	35	105	140	1207	302	36	10	1090	27	117	229
TCGA-AD-5900-01A-11D-1650-10	1521	68	198	266	984	220	27	10	845	24	139	159
TCGA-AA-3663-01A-01D-1719-10	2894	91	369	460	1743	605	52	76	1153	117	590	360
TCGA-AZ-4315-01A-01D-1408-10	6086	25	8	33	4006	2046	42	22	3331	650	675	1332
TCGA-A6-6653-01A-11D-1771-10	1148	51	152	203	779	128	13	11	676	17	103	87
TCGA-AZ-6598-01A-11D-1771-10	2715	131	481	612	1654	342	46	20	1418	36	236	240
TCGA-A6-5665-01A-01D-1650-10	1937	156	247	403	1176	282	42	17	873	45	303	178
TCGA-CM-6171-01A-11D-1650-10	1231	57	170	227	795	161	15	16	683	19	112	111
TCGA-AU-6004-01A-11D-1719-10	1483	65	223	288	928	208	16	42	732	33	196	117
TCGA-CK-5916-01A-11D-1650-10	1660	72	248	320	1087	197	23	19	900	20	187	135
TCGA-AM-5820-01A-01D-1650-10	7183	52	131	183	5224	1673	235	562	3573	329	1651	547
TCGA-D5-6540-01A-11D-1719-10	1378	63	216	279	817	251	27	14	747	35	70	175
TCGA-G4-6320-01A-11D-1719-10	1013	24	92	116	726	155	15	10	677	8	49	122
TCGA-CM-4743-01A-01D-1719-10	1122	53	177	230	676	181	17	15	545	22	131	127
TCGA-D5-6930-01A-11D-1924-10	1331	39	161	200	937	166	23	10	876	15	61	118
TCGA-CM-5861-01A-01D-1650-10	1463	71	163	234	947	254	29	14	652	46	295	165
TCGA-AD-6895-01A-11D-1924-10	1725	51	122	173	1191	321	160	31	1054	21	137	109
TCGA-F4-6856-01A-11D-1924-10	1173	58	147	205	795	127	27	9	701	19	94	72
TCGA-AZ-6601-01A-11D-1771-10	1826	17	34	51	1527	240	20	8	1425	21	102	191
TCGA-G4-6588-01A-11D-1771-10	2180	94	306	400	1404	325	43	21	1090	36	314	225
TCGA-A6-6780-01A-11D-1835-10	1257	72	221	293	710	205	21	18	553	40	157	126
TCGA-CA-6718-01A-11D-1835-10	2671	9	1	10	1227	1430	28	10	1026	442	201	950
TCGA-AA-3492-01A-01D-1408-10	2224	77	288	365	1468	325	62	27	1239	38	229	198
TCGA-A6-5661-01A-01D-1650-10	1178	78	122	200	765	180	23	12	679	11	86	134
TCGA-AM-5821-01A-01D-1650-10	5373	6	14	20	4101	1169	182	411	2792	213	1309	363
TCGA-AY-6197-01A-11D-1719-10	1634	60	200	260	1032	290	39	45	706	39	326	167
TCGA-AA-3510-01A-01D-1408-10	2963	6	1	7	1749	1203	23	3	1603	264	146	913
TCGA-G4-6628-01A-11D-1835-10	1959	126	236	362	1271	285	31	26	1143	16	128	212

Figure 15. High mutation Group Mutation Type Count Categories

Columns are as in Figure 14.

It was apparent that due to the difference in magnitude of the numbers between the groups that most if not all of these mutation categories would have statistically significant differences. I created equivalently structured tables where the values for each sample were divided by the total number of mutations in that sample, to obtain percentage proportions.

Figure 16 shows the proportions for the low mutation population. After controlling for total mutation count by converting these numbers to percentages, the ratios were much less variant for the C_T and G_A mutations. There were some variations among other mutation types that might be more significant. MNC mutations were fairly consistently in the single digit percentages, although were more frequent in some tumors.

sample_ID	ALL	A_C	A_T	A_G	C_A	C_T	C_G	T_A	T_C	T_G	G_A	G_C	G_T	-_T	T_-	-_C	C_-	-_A	A_-	-_G	G_-	MNC
TCGA-CM-6169-01A-11D-1650-10	100.0%	0.8%	2.3%	2.3%	6.2%	30.2%	2.3%	1.6%	2.3%	2.3%	27.9%	0.8%	10.9%	0.0%	0.8%	0.8%	1.6%	0.0%	0.0%	0.0%	0.0%	7.0%
TCGA-D5-5541-01A-01D-1650-10	100.0%	0.0%	1.0%	6.9%	4.0%	32.7%	2.0%	3.0%	3.0%	3.0%	27.7%	1.0%	4.0%	0.0%	1.0%	2.0%	0.0%	1.0%	0.0%	2.0%	1.0%	5.0%
TCGA-G4-6314-01A-11D-1719-10	100.0%	1.4%	0.7%	4.2%	7.0%	25.9%	6.3%	4.2%	2.8%	2.1%	30.8%	5.6%	7.7%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%
TCGA-G4-6297-01A-11D-1719-10	100.0%	1.6%	1.3%	6.9%	7.2%	30.5%	4.9%	1.3%	9.2%	1.0%	18.7%	3.6%	4.3%	0.7%	1.3%	0.3%	0.0%	0.0%	1.0%	0.0%	0.3%	5.9%
TCGA-F4-6703-01A-11D-1835-10	100.0%	1.1%	1.6%	5.6%	5.6%	27.8%	1.1%	1.1%	6.0%	1.6%	28.0%	0.2%	6.0%	0.2%	2.4%	0.0%	1.1%	0.7%	1.8%	0.4%	2.2%	5.6%
TCGA-DM-A1DA-01A-11D-A152-10	100.0%	2.3%	3.5%	5.8%	10.5%	21.8%	3.8%	1.2%	4.9%	2.6%	24.1%	4.7%	10.5%	0.0%	0.3%	0.0%	0.3%	0.0%	0.3%	0.0%	0.3%	3.2%
TCGA-F4-6806-01A-11D-1835-10	100.0%	1.1%	2.2%	3.3%	4.4%	32.2%	1.1%	0.0%	0.0%	0.0%	34.4%	3.3%	5.6%	0.0%	0.0%	2.2%	0.0%	1.1%	0.0%	1.1%	0.0%	7.8%
TCGA-A6-6650-01A-11D-1771-10	100.0%	1.3%	2.7%	5.4%	7.4%	26.8%	4.0%	0.7%	1.3%	3.4%	26.2%	0.7%	9.4%	0.7%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.7%	8.7%
TCGA-F4-6805-01A-11D-1835-10	100.0%	0.0%	0.0%	3.8%	3.8%	20.8%	1.9%	5.7%	9.4%	3.8%	26.4%	5.7%	7.5%	0.0%	0.0%	0.0%	1.9%	0.0%	0.0%	0.0%	1.9%	7.5%
TCGA-D5-6898-01A-11D-1924-10	100.0%	1.3%	3.9%	2.6%	5.3%	34.2%	2.6%	2.6%	2.6%	0.0%	23.7%	3.9%	11.8%	1.3%	0.0%	0.0%	0.0%	2.6%	0.0%	0.0%	0.0%	1.3%
TCGA-CA-5797-01A-01D-1650-10	100.0%	0.0%	5.4%	1.8%	3.6%	25.0%	5.4%	1.8%	1.8%	1.8%	32.1%	3.6%	8.9%	3.6%	1.8%	0.0%	0.0%	0.0%	1.8%	0.0%	0.0%	1.8%
TCGA-G4-6321-01A-11D-1719-10	100.0%	2.4%	1.9%	9.5%	6.1%	21.0%	3.2%	0.8%	11.4%	1.9%	25.5%	5.0%	4.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	6.1%
TCGA-CM-6679-01A-11D-1835-10	100.0%	0.0%	0.0%	2.1%	2.1%	31.3%	2.1%	8.3%	0.0%	0.0%	45.8%	0.0%	2.1%	0.0%	0.0%	2.1%	0.0%	0.0%	0.0%	2.1%	0.0%	2.1%
TCGA-F4-6809-01A-11D-1835-10	100.0%	0.0%	3.2%	3.2%	8.1%	23.4%	3.2%	5.6%	6.5%	1.6%	15.3%	4.8%	10.5%	0.0%	0.8%	0.0%	0.0%	2.4%	0.8%	0.0%	0.8%	9.7%
TCGA-CK-5915-01A-11D-1650-10	100.0%	0.7%	0.0%	4.2%	1.4%	39.9%	3.5%	2.8%	5.6%	1.4%	26.6%	2.8%	5.6%	0.0%	0.7%	1.4%	0.0%	1.4%	0.0%	0.0%	0.0%	2.1%
TCGA-D5-6533-01A-11D-1719-10	100.0%	2.3%	5.3%	6.4%	7.6%	21.1%	3.5%	5.3%	5.8%	1.8%	24.6%	3.5%	6.4%	0.0%	0.0%	0.0%	0.0%	0.6%	0.6%	0.0%	0.6%	4.7%
TCGA-AU-3779-01A-01D-1719-10	100.0%	2.9%	1.9%	2.9%	7.7%	26.0%	1.0%	4.8%	5.8%	1.9%	33.7%	1.9%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.9%
TCGA-AD-6901-01A-11D-1924-10	100.0%	0.8%	1.6%	1.6%	7.6%	29.6%	4.8%	2.8%	4.8%	2.4%	26.0%	4.0%	8.8%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	1.2%	0.8%	2.8%
TCGA-D5-6529-01A-11D-1771-10	100.0%	0.0%	0.0%	4.3%	5.1%	39.3%	3.4%	2.6%	2.6%	0.9%	29.9%	2.6%	3.4%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	1.7%	3.4%
TCGA-CM-6164-01A-11D-1650-10	100.0%	1.7%	3.3%	8.3%	13.3%	30.0%	0.0%	5.0%	8.3%	0.0%	16.7%	3.3%	3.3%	0.0%	0.0%	0.0%	0.0%	0.0%	1.7%	1.7%	0.0%	3.3%
TCGA-DM-A28F-01A-11D-A16V-10	100.0%	4.6%	3.3%	4.2%	6.7%	21.3%	4.2%	2.9%	3.3%	4.6%	25.9%	4.2%	6.7%	0.4%	0.4%	1.7%	0.4%	0.4%	0.8%	0.8%	0.0%	2.9%
TCGA-G4-6311-01A-11D-1719-10	100.0%	3.0%	3.0%	10.4%	7.7%	20.8%	4.2%	0.9%	5.3%	2.4%	24.3%	3.3%	6.5%	0.0%	0.6%	0.0%	0.6%	0.3%	0.3%	0.0%	0.0%	6.5%
TCGA-D5-6532-01A-11D-1719-10	100.0%	4.5%	1.8%	2.7%	3.6%	27.0%	7.2%	2.7%	2.7%	2.7%	27.9%	3.6%	6.3%	0.0%	0.0%	0.0%	1.8%	1.8%	0.9%	0.0%	0.0%	2.7%
TCGA-AZ-4682-01B-01D-1408-10	100.0%	3.1%	3.1%	3.1%	5.2%	39.6%	0.0%	1.0%	0.0%	1.0%	35.4%	2.1%	3.1%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	0.0%	2.1%
TCGA-CM-5863-01A-21D-1835-10	100.0%	1.2%	0.0%	4.8%	4.8%	28.6%	3.6%	1.2%	4.8%	1.2%	23.8%	2.4%	8.3%	0.0%	2.4%	0.0%	0.0%	0.0%	2.4%	0.0%	0.0%	10.7%
TCGA-D5-6536-01A-11D-1719-10	100.0%	1.0%	2.9%	3.9%	4.9%	21.4%	5.8%	2.9%	2.9%	1.0%	35.0%	1.9%	9.7%	0.0%	0.0%	1.9%	0.0%	0.0%	0.0%	0.0%	1.0%	3.9%
TCGA-DM-A28H-01A-11D-A16V-10	100.0%	2.5%	1.7%	1.7%	3.3%	22.5%	3.3%	3.3%	4.2%	2.5%	30.0%	6.7%	4.2%	1.7%	0.8%	0.8%	0.0%	0.0%	1.7%	0.8%	0.8%	7.5%
TCGA-F4-6460-01A-11D-1771-10	100.0%	1.6%	1.6%	2.4%	8.7%	31.5%	3.9%	6.3%	4.7%	1.6%	21.3%	3.9%	10.2%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	0.8%
TCGA-D5-6531-01A-11D-1719-10	100.0%	1.5%	5.6%	2.5%	6.1%	26.4%	3.6%	2.5%	4.6%	2.5%	29.4%	2.5%	6.6%	0.5%	0.0%	0.0%	0.5%	0.0%	0.0%	0.0%	0.5%	4.6%
TCGA-G4-6304-01A-11D-1924-10	100.0%	0.3%	2.4%	4.1%	6.1%	36.4%	0.3%	2.0%	5.4%	2.0%	29.6%	1.7%	4.4%	0.3%	1.0%	0.0%	0.0%	0.0%	2.4%	0.3%	1.0%	0.0%
TCGA-D5-6926-01A-11D-1924-10	100.0%	0.0%	1.5%	3.1%	3.1%	27.5%	3.8%	1.5%	6.1%	2.3%	32.8%	3.1%	8.4%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.1%
TCGA-CM-5868-01A-01D-1650-10	100.0%	2.6%	3.7%	8.5%	5.3%	23.3%	0.5%	2.6%	7.4%	2.1%	21.2%	2.6%	9.0%	0.0%	1.1%	0.5%	0.0%	0.0%	1.1%	1.1%	0.0%	7.4%
TCGA-A6-2675-01A-02D-1719-10	100.0%	2.4%	4.8%	6.0%	3.6%	37.3%	1.2%	0.0%	3.6%	0.0%	31.3%	3.6%	1.2%	0.0%	0.0%	0.0%	0.0%	1.2%	0.0%	0.0%	0.0%	3.6%
TCGA-DM-A1DB-01A-11D-A152-10	100.0%	1.3%	1.3%	5.0%	5.0%	32.5%	3.8%	0.0%	5.0%	0.0%	25.0%	7.5%	7.5%	0.0%	0.0%	0.0%	0.0%	0.0%	1.3%	0.0%	0.0%	5.0%
TCGA-CM-4750-01A-01D-1408-10	100.0%	3.0%	1.0%	3.0%	6.1%	30.3%	0.0%	3.0%	3.0%	1.0%	29.3%	0.0%	11.1%	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	0.0%	1.0%	7.1%
TCGA-AY-5543-01A-01D-1650-10	100.0%	1.6%	0.0%	4.8%	7.9%	30.2%	4.8%	2.4%	4.8%	2.4%	31.0%	3.2%	5.6%	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%	0.8%	0.0%	0.0%
TCGA-AD-6548-01A-11D-1835-10	100.0%	1.0%	0.0%	1.0%	3.8%	32.4%	0.0%	2.9%	6.7%	1.9%	34.3%	2.9%	6.7%	0.0%	0.0%	0.0%	0.0%	0.0%	1.9%	1.0%	0.0%	3.8%
TCGA-AZ-4323-01A-21D-1835-10	100.0%	0.0%	6.1%	6.1%	6.1%	42.4%	3.0%	0.0%	0.0%	3.0%	24.2%	3.0%	3.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.0%
TCGA-CM-5864-01A-01D-1650-10	100.0%	2.7%	3.2%	4.1%	9.5%	24.4%	5.0%	1.8%	5.0%	3.2%	24.9%	3.6%	8.6%	0.0%	0.0%	0.5%	0.9%	0.0%	0.9%	0.0%	0.0%	1.8%
TCGA-AA-3489-01A-21D-1835-10	100.0%	2.4%	3.7%	2.4%	8.5%	32.9%	0.0%	4.9%	2.4%	0.0%	25.6%	2.4%	6.1%	0.0%	0.0%	1.2%	0.0%	1.2%	0.0%	0.0%	0.0%	6.1%
TCGA-A6-6140-01A-11D-1771-10	100.0%	2.0%	1.2%	5.9%	5.9%	31.1%	2.8%	2.4%	4.2%	1.6%	33.5%	4.4%	4.2%	0.2%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.6%
TCGA-A6-5664-01A-21D-1835-10	100.0%	1.9%	0.0%	3.8%	11.5%	34.6%	0.0%	0.0%	5.8%	1.9%	30.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	9.6%
TCGA-CM-5860-01A-01D-1650-10	100.0%	0.0%	0.8%	4.1%	9.0%	33.6%	1.6%	0.0%	3.3%	0.8%	33.6%	2.5%	6.6%	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	3.3%
TCGA-DM-A28K-01A-21D-A16V-10	100.0%	0.9%	1.7%	1.7%	6.1%	32.2%	3.5%	0.9%	2.6%	0.9%	34.8%	0.9%	1.7%	0.0%	0.0%	0.9%	0.0%	1.7%	0.0%	1.7%	0.9%	7.0%
TCGA-F4-6807-01A-11D-1835-10	100.0%	0.6%	2.5%	2.5%	14.4%	18.8%	3.1%	3.1%	3.1%	0.6%	24.4%	1.3%	16.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.6%	0.0%	7.5%
TCGA-AD-6899-01A-11D-1924-10	100.0%	0.0%	2.5%	3.2%	4.5%	35.0%	2.5%	0.6%	1.3%	0.6%	37.6%	3.8%	5.1%	0.0%	0.0%	0.0%	0.6%	0.0%	0.0%	0.6%	0.0%	1.9%
TCGA-CM-5862-01A-01D-1650-10	100.0%	1.1%	0.0%	5.3%	3.2%	33.0%	2.1%	2.1%	6.4%	2.1%	25.5%	1.1%	8.5%	0.0%	0.0%	2.1%	1.1%	0.0%	0.0%	0.0%	0.0%	6.4%
TCGA-AY-6196-01A-11D-1719-10	100.0%	3.9%	2.6%	8.3%	4.9%	16.8%	5.7%	3.1%	8.5%	2.3%	22.3%	7.8%	8.8%	0.3%	0.3%	0.3%	0.3%	0.0%	0.0%	0.0%	0.3%	3.6%
TCGA-CK-4948-01B-11D-1650-10	100.0%	2.8%	2.8%	3.7%	7.4%	31.5%	2.8%	1.9%	3.7%	1.9%	28.7%	4.6%	2.8%	0.0%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	4.6%
TCGA-D5-5539-01A-01D-1650-10	100.0%	0.0%	0.0%	2.8%	3.8%	33.0%	0.9%	3.8%	6.6%	1.9%	33.0%	6.6%	3.8%	0.0%	0.0%	2.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%
TCGA-G4-6626-01A-11D-1771-10	100.0%	0.0%	1.3%	3.8%	6.4%	35.0%	2.5%	3.8%	3.2%	0.0%	30.6%	1.9%	8.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	3.2%
TCGA-D5-6537-01A-11D-1719-10	100.0%	2.7%	0.8%	7.5%	6.1%	24.3%	2.1%	2.1%	9.9%	2.1%	23.2%	3.7%	7.2%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	7.5%
TCGA-DM-A28G-01A-11D-A16V-10	100.0%	0.9%	0.9%	3.5%	9.6%	31.6%	0.9%	3.5%	1.8%	2.6%	30.7%	1.8%	6.1%	0.9%	0.0%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	4.4%
TCGA-AA-3662-01A-01D-1719-10	100.0%	3.1%	1.4%	14.5%	4.5%	21.1%	5.0%	1.0%	9.9%	2.3%	20.0%	4.8%	6.4%	0.0%	0.2%	0.2%	0.0%	0.0%	0.6%	0.4%	0.6%	4.1%
TCGA-DM-A1D8-01A-11D-A152-10	100.0%	5.2%	2.6%	2.6%	11.0%	23.4%	2.6%	5.2%	3.9%	1.9%	23.4%	3.9%	9.7%	1.3%	0.0%	0.0%	0.0%	0.6%	0.0%	0.6%	0.0%	1.9%
TCGA-CM-6163-01A-11D-1650-10	100.0%	1.0%	0.0%	5.1%	9.1%	36.4%	1.0%	1.0%	4.0%	0.0%	29.3%	4.0%	7.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.0%
TCGA-D5-7000-01A-11D-1924-10	100.0%	0.0%	0.6%	4.2%	4.2%	37.7%	3.0%	1.8%	3.6%	0.0%	19.8%	1.8%	5.4%	0.0%	1.2%	0.0%	0.0%	0.6%	0.6%	1.2%	1.2%	13.2%
TCGA-CA-6719-01A-11D-1835-10	100.0%	0.0%	0.8%	3.3%	5.7%	34.4%	0.8%	0.0%	4.9%	0.8%	35.2%	0.0%	4.9%	0.0%	0.0%	0.0%	0.8%	2.5%	0.0%	0.0%	0.0%	5.7%

sample_ID	ALL	A_C	A_T	A_G	C_A	C_T	C_G	T_A	T_C	T_G	G_A	G_C	G_T	-_T	T_-	-_C	C_-	-_A	A_-	-_G	G_-	MNC
TCGA-DM-A282-01A-12D-A16V-10	100.0%	1.1%	1.1%	2.2%	8.9%	31.1%	1.1%	3.3%	2.2%	0.0%	32.2%	4.4%	5.6%	2.2%	1.1%	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	2.2%
TCGA-AD-6888-01A-11D-1924-10	100.0%	3.4%	1.7%	2.6%	5.1%	28.2%	2.6%	0.9%	1.7%	0.9%	39.3%	0.9%	5.1%	0.9%	0.9%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	5.1%
TCGA-CA-6716-01A-11D-1835-10	100.0%	0.5%	0.5%	4.8%	7.7%	35.6%	1.9%	0.5%	3.4%	2.4%	29.8%	1.9%	3.8%	1.0%	0.0%	1.0%	0.0%	1.9%	0.5%	0.5%	0.0%	2.4%
TCGA-CK-5912-01A-11D-1650-10	100.0%	0.0%	1.6%	1.6%	0.0%	46.0%	4.8%	1.6%	1.6%	1.6%	20.6%	3.2%	7.9%	1.6%	1.6%	0.0%	1.6%	0.0%	0.0%	0.0%	0.0%	4.8%
TCGA-CK-4950-01A-01D-1719-10	100.0%	2.5%	1.1%	5.5%	8.7%	22.5%	4.1%	1.4%	6.4%	3.2%	25.7%	4.6%	7.1%	0.0%	0.7%	0.0%	0.5%	0.2%	0.2%	0.0%	0.5%	5.0%
TCGA-A6-5660-01A-01D-1650-10	100.0%	2.6%	0.9%	5.2%	7.8%	29.3%	2.6%	0.9%	5.2%	0.9%	31.0%	3.4%	2.6%	0.0%	0.9%	0.9%	0.0%	0.9%	0.0%	0.9%	0.0%	4.3%
TCGA-AZ-6605-01A-11D-1835-10	100.0%	1.5%	0.7%	4.4%	5.9%	33.1%	2.9%	0.0%	7.4%	0.0%	29.4%	3.7%	5.1%	0.7%	0.0%	0.0%	0.7%	0.0%	0.0%	0.7%	0.0%	3.7%
TCGA-DM-A1D6-01A-21D-A152-10	100.0%	1.6%	1.6%	3.9%	2.3%	36.7%	3.1%	2.3%	3.1%	1.6%	29.7%	3.9%	4.7%	0.8%	0.8%	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%	3.1%
TCGA-CM-6165-01A-11D-1650-10	100.0%	0.0%	0.9%	8.0%	12.5%	23.2%	0.0%	1.8%	0.9%	0.0%	23.2%	4.5%	15.2%	2.7%	0.9%	0.0%	0.9%	0.0%	0.0%	0.0%	0.9%	4.5%
TCGA-D5-6541-01A-11D-1719-10	100.0%	1.1%	1.1%	3.2%	4.2%	32.6%	4.2%	1.1%	4.2%	1.1%	30.5%	0.0%	2.1%	1.1%	0.0%	0.0%	1.1%	1.1%	1.1%	1.1%	1.1%	8.4%
TCGA-CM-6677-01A-11D-1835-10	100.0%	0.8%	0.8%	2.3%	6.2%	34.6%	0.8%	1.5%	4.6%	1.5%	34.6%	0.8%	5.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	1.5%	0.0%	3.8%
TCGA-CM-4744-01A-01D-1408-10	100.0%	2.4%	0.6%	1.8%	8.4%	28.3%	0.6%	2.4%	2.4%	1.8%	28.9%	2.4%	7.2%	1.2%	0.0%	1.8%	0.6%	0.0%	0.0%	0.0%	0.6%	8.4%
TCGA-DM-A1D7-01A-11D-A152-10	100.0%	2.8%	5.0%	2.1%	14.2%	26.2%	0.7%	2.1%	2.8%	1.4%	29.1%	5.7%	6.4%	0.0%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.7%
TCGA-AA-3510-01A-01W-1461-10	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-A6-4105-01A-02D-1771-10	100.0%	0.0%	2.7%	4.7%	7.4%	32.2%	2.7%	2.7%	3.4%	2.0%	26.2%	4.7%	7.4%	0.7%	0.0%	0.7%	0.0%	0.0%	0.0%	0.0%	0.0%	2.7%
TCGA-G4-6306-01A-11D-1771-10	100.0%	0.0%	2.4%	5.6%	8.9%	37.1%	1.6%	0.8%	3.2%	0.8%	25.8%	0.0%	3.2%	0.0%	0.0%	0.0%	1.6%	0.8%	0.0%	0.8%	0.0%	7.3%
TCGA-D5-6535-01A-11D-1719-10	100.0%	3.1%	0.6%	4.4%	8.8%	26.9%	2.5%	2.5%	3.8%	3.1%	20.0%	6.9%	13.1%	0.0%	0.0%	0.0%	0.6%	0.6%	0.0%	0.0%	0.0%	3.1%
TCGA-CM-5349-01A-21D-1719-10	100.0%	2.9%	1.9%	6.7%	3.8%	38.5%	1.0%	3.8%	1.9%	1.9%	24.0%	1.0%	2.9%	1.9%	0.0%	0.0%	1.0%	0.0%	0.0%	0.0%	1.0%	5.8%
TCGA-A6-6138-01A-11D-1771-10	100.0%	0.9%	3.4%	3.4%	2.6%	35.0%	1.7%	0.0%	6.0%	2.6%	29.9%	1.7%	7.7%	0.9%	0.0%	0.0%	1.7%	0.9%	0.0%	0.0%	0.9%	0.9%
TCGA-A6-5656-01A-21D-1835-10	100.0%	0.9%	0.0%	0.9%	5.7%	41.5%	1.9%	1.9%	4.7%	0.0%	29.2%	1.9%	3.8%	0.9%	0.9%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	4.7%
TCGA-CM-4747-01A-01D-1408-10	100.0%	2.1%	0.0%	2.1%	6.3%	35.4%	1.0%	1.0%	6.3%	1.0%	36.5%	1.0%	2.1%	2.1%	1.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.1%
TCGA-D5-5540-01A-01D-1650-10	100.0%	3.0%	3.0%	0.8%	6.1%	34.1%	0.8%	4.5%	4.5%	1.5%	26.5%	2.3%	8.3%	0.8%	0.0%	1.5%	0.0%	0.0%	0.0%	0.0%	0.0%	2.3%
TCGA-CA-5796-01A-01D-1650-10	100.0%	0.0%	3.0%	7.6%	9.1%	33.3%	0.0%	1.5%	1.5%	3.0%	25.8%	1.5%	7.6%	0.0%	0.0%	1.5%	1.5%	1.5%	0.0%	0.0%	0.0%	1.5%
TCGA-D5-6539-01A-11D-1719-10	100.0%	3.0%	0.0%	9.1%	9.1%	30.3%	1.5%	3.0%	0.0%	4.5%	24.2%	4.5%	4.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.1%
TCGA-G4-6294-01A-11D-1806-10	100.0%	1.6%	5.4%	2.7%	6.0%	26.1%	7.6%	4.3%	4.3%	2.2%	15.2%	5.4%	10.3%	0.0%	0.5%	0.0%	0.0%	0.5%	0.5%	0.0%	0.5%	6.5%
TCGA-F4-6808-01A-11D-1835-10	100.0%	0.8%	2.4%	2.4%	8.1%	29.3%	3.3%	0.8%	2.4%	0.8%	31.7%	1.6%	8.9%	0.0%	2.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.9%
TCGA-F4-6704-01A-11D-1835-10	100.0%	0.0%	5.7%	0.0%	2.9%	20.0%	2.9%	5.7%	0.0%	0.0%	40.0%	2.9%	5.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	14.3%
TCGA-G4-6310-01A-11D-1719-10	100.0%	0.9%	1.7%	0.9%	4.3%	35.9%	2.6%	1.7%	2.6%	0.9%	32.5%	3.4%	8.5%	0.9%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.9%	2.6%
TCGA-DM-A28M-01A-12D-A16V-10	100.0%	0.9%	1.9%	3.7%	6.5%	29.6%	0.0%	1.9%	3.7%	0.9%	38.9%	0.9%	4.6%	1.9%	0.9%	0.0%	0.0%	0.9%	0.0%	0.0%	0.9%	1.9%
TCGA-A6-6654-01A-21D-1835-10	100.0%	0.0%	0.6%	4.2%	3.6%	42.3%	1.8%	2.4%	0.0%	0.6%	37.5%	0.0%	1.8%	0.6%	0.0%	0.0%	1.2%	0.6%	0.6%	0.0%	0.6%	1.8%
TCGA-D5-6927-01A-21D-1924-10	100.0%	1.0%	1.2%	6.3%	4.8%	24.1%	0.6%	0.4%	7.5%	0.7%	25.4%	0.6%	5.8%	1.4%	2.3%	0.6%	4.3%	1.2%	3.7%	1.9%	2.8%	3.2%
TCGA-A6-2671-01A-01D-1408-10	100.0%	3.4%	1.1%	2.3%	8.0%	22.7%	1.1%	1.1%	2.3%	1.1%	29.5%	2.3%	10.2%	1.1%	2.3%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	10.2%
TCGA-CM-4748-01A-01D-1408-10	100.0%	0.0%	0.0%	4.4%	10.3%	27.9%	2.9%	1.5%	2.9%	4.4%	23.5%	2.9%	8.8%	1.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	7.4%
TCGA-AA-3660-01A-01D-1719-10	100.0%	0.8%	0.0%	3.4%	6.8%	32.2%	0.8%	1.7%	4.2%	4.2%	32.2%	1.7%	2.5%	0.0%	0.0%	0.8%	0.8%	0.0%	0.0%	0.0%	0.0%	7.6%
TCGA-D5-6534-01A-21D-1924-10	100.0%	3.2%	3.2%	1.3%	11.0%	28.4%	1.9%	3.2%	1.3%	5.2%	24.5%	3.2%	7.7%	0.0%	0.0%	0.0%	0.0%	0.6%	0.6%	0.0%	0.0%	4.5%
TCGA-F4-6854-01A-11D-1924-10	100.0%	4.1%	2.4%	4.1%	8.9%	28.5%	1.6%	4.9%	4.1%	0.8%	26.0%	0.8%	5.7%	1.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	5.7%
TCGA-CK-4952-01A-01D-1719-10	100.0%	2.2%	1.5%	6.8%	4.3%	26.2%	2.5%	0.9%	8.3%	2.5%	29.9%	1.9%	4.6%	0.0%	1.2%	0.3%	0.3%	0.3%	0.3%	0.6%	0.0%	5.2%
TCGA-G4-6317-01A-11D-1719-10	100.0%	1.8%	0.0%	0.9%	7.0%	36.8%	1.8%	4.4%	2.6%	0.0%	32.5%	1.8%	3.5%	0.0%	0.0%	0.0%	0.9%	0.9%	0.0%	0.0%	0.0%	5.3%
TCGA-CK-4947-01B-11D-1650-10	100.0%	2.2%	3.2%	5.4%	6.5%	31.2%	5.4%	3.2%	5.4%	2.2%	26.9%	0.0%	4.3%	0.0%	2.2%	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	1.1%
TCGA-CM-6166-01A-11D-1650-10	100.0%	2.1%	0.0%	0.0%	13.5%	29.2%	5.2%	2.1%	2.1%	0.0%	29.2%	0.0%	5.2%	1.0%	0.0%	2.1%	0.0%	2.1%	0.0%	0.0%	0.0%	6.3%
TCGA-DM-A1D4-01A-21D-A152-10	100.0%	1.1%	2.2%	3.9%	11.0%	32.6%	2.2%	1.7%	2.2%	1.7%	29.3%	1.7%	6.6%	0.6%	0.0%	0.6%	0.0%	0.6%	0.6%	0.6%	0.0%	1.1%
TCGA-CM-5341-01A-01D-1408-10	100.0%	1.2%	2.4%	2.4%	3.6%	31.7%	1.8%	3.6%	3.6%	1.8%	28.1%	6.0%	6.6%	0.0%	0.0%	0.0%	1.8%	0.0%	0.0%	0.6%	0.0%	4.8%
TCGA-CM-6170-01A-11D-1650-10	100.0%	2.0%	1.0%	2.0%	2.0%	39.4%	4.0%	2.0%	3.0%	1.0%	32.3%	1.0%	3.0%	1.0%	1.0%	0.0%	0.0%	0.0%	0.0%	1.0%	0.0%	4.0%
TCGA-AZ-6603-01A-11D-1835-10	100.0%	0.8%	0.8%	7.6%	4.2%	22.7%	2.5%	0.8%	3.4%	3.4%	33.6%	1.7%	3.4%	0.8%	0.0%	0.0%	0.0%	0.0%	2.5%	0.0%	0.0%	11.8%
TCGA-A6-5662-01A-01D-1650-10	100.0%	1.5%	1.5%	1.5%	5.9%	29.4%	1.5%	4.4%	2.9%	0.0%	32.4%	1.5%	10.3%	1.5%	1.5%	1.5%	0.0%	0.0%	1.5%	0.0%	0.0%	1.5%
TCGA-AA-3655-01A-02D-1719-10	100.0%	0.0%	0.0%	5.7%	6.9%	31.0%	1.1%	1.1%	9.2%	0.0%	31.0%	4.6%	3.4%	0.0%	0.0%	0.0%	0.0%	1.1%	0.0%	1.1%	0.0%	3.4%
TCGA-CM-4746-01A-01D-1408-10	100.0%	0.3%	0.5%	1.4%	3.7%	37.5%	0.5%	0.2%	3.1%	0.3%	30.2%	0.7%	3.4%	1.3%	2.0%	1.2%	2.7%	1.5%	3.8%	0.9%	2.0%	2.9%
TCGA-A6-5666-01A-01D-1650-10	100.0%	2.9%	1.0%	4.8%	7.6%	29.5%	2.9%	1.0%	3.8%	2.9%	24.8%	1.0%	10.5%	1.9%	0.0%	1.0%	0.0%	1.9%	0.0%	0.0%	1.0%	1.9%
TCGA-CM-6161-01A-11D-1650-10	100.0%	0.0%	4.7%	6.5%	4.7%	33.6%	1.9%	1.9%	3.7%	0.9%	27.1%	2.8%	6.5%	0.9%	0.0%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	3.7%
TCGA-AD-6890-01A-11D-1924-10	100.0%	0.8%	0.0%	1.7%	8.5%	32.2%	0.8%	1.7%	0.0%	0.8%	32.2%	0.8%	7.6%	0.8%	1.7%	0.8%	0.0%	0.8%	0.8%	0.0%	0.8%	6.8%
TCGA-D5-6920-01A-11D-1924-10	100.0%	0.8%	1.7%	4.2%	5.9%	31.9%	4.2%	2.5%	2.5%	0.0%	31.1%	0.8%	3.4%	0.0%	0.8%	0.0%	0.0%	0.8%	0.8%	0.0%	1.7%	6.7%
TCGA-AZ-6607-01A-11D-1835-10	100.0%	5.8%	4.3%	2.9%	7.2%	29.0%	1.4%	2.9%	1.4%	0.0%	36.2%	0.0%	2.9%	0.0%	0.0%	0.0%	0.0%	1.4%	0.0%	0.0%	0.0%	4.3%
TCGA-CM-6676-01A-11D-1835-10	100.0%	1.1%	3.3%	2.2%	4.3%	46.7%	1.1%	2.2%	0.0%	1.1%	31.5%	0.0%	4.3%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%
TCGA-DM-A1D0-01A-11D-A152-10	100.0%	2.6%	5.1%	3.4%	3.4%	30.8%	6.0%	6.0%	6.0%	0.9%	16.2%	3.4%	8.5%	0.0%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.9%	6.0%
TCGA-AD-6965-01A-11D-1924-10	100.0%	0.7%	2.0%	2.7%	6.1%	25.9%	2.7%	2.0%	4.8%	0.7%	34.7%	2.7%	6.8%	1.4%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.7%	5.4%
TCGA-A6-6142-01A-11D-1771-10	100.0%	4.2%	2.8%	9.9%	5.6%	32.4%	0.0%	4.2%	4.2%	1.4%	21.1%	0.0%	2.8%	0.0%	1.4%	0.0%	0.0%	0.0%	4.2%	0.0%	0.0%	5.6%
TCGA-D5-6932-01A-11D-1924-10	100.0%	1.7%	4.1%	4.1%	11.6%	23.1%	5.0%	5.8%	2.5%	4.1%	22.3%	3.3%	7.4%	0.0%	0.0%	0.0%	1.7%	0.0%	0.8%	0.0%	0.0%	2.5%
TCGA-F4-6461-01A-11D-1771-10	100.0%	0.8%	2.4%	5.5%	4.7%	28.3%	1.6%	0.8%	5.5%	1.6%	31.5%	3.9%	7.1%	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%	0.8%	0.0%	4.7%

sample_ID	ALL	A_C	A_T	A_G	C_A	C_T	C_G	T_A	T_C	T_G	G_A	G_C	G_T	-_T	T_-	-_C	C_-	-_A	A_-	-_G	G_-	MNC
TCGA-A6-5657-01A-01D-1650-10	100.0%	2.3%	1.1%	0.0%	13.6%	39.8%	3.4%	0.0%	1.1%	0.0%	26.1%	1.1%	8.0%	0.0%	0.0%	1.1%	1.1%	1.1%	0.0%	0.0%	0.0%	0.0%
TCGA-DM-A1HA-01A-11D-A152-10	100.0%	0.0%	3.5%	1.2%	6.5%	26.5%	1.8%	2.9%	4.7%	0.6%	29.4%	7.1%	10.0%	0.6%	0.6%	0.0%	0.0%	1.2%	0.0%	0.0%	1.2%	2.4%
TCGA-CM-6680-01A-11D-1835-10	100.0%	1.7%	0.0%	6.2%	3.9%	27.0%	2.2%	2.8%	4.5%	2.2%	25.8%	2.2%	5.6%	0.0%	0.6%	0.0%	0.6%	1.1%	0.6%	0.0%	0.0%	12.9%
TCGA-A6-6651-01A-21D-1835-10	100.0%	0.9%	1.8%	5.4%	4.5%	19.8%	5.4%	2.7%	5.4%	0.9%	22.5%	2.7%	5.4%	0.0%	1.8%	0.9%	0.0%	0.0%	1.8%	0.0%	0.9%	17.1%
TCGA-D5-6922-01A-11D-1924-10	100.0%	2.6%	0.9%	2.6%	1.7%	33.6%	6.0%	0.9%	1.7%	0.9%	34.5%	4.3%	6.0%	0.9%	0.0%	0.0%	0.0%	0.9%	0.0%	0.0%	0.0%	2.6%
TCGA-F4-6459-01A-11D-1771-10	100.0%	3.6%	4.5%	2.7%	8.9%	25.9%	6.3%	2.7%	3.6%	0.0%	23.2%	6.3%	2.7%	0.0%	0.9%	1.8%	0.9%	0.0%	1.8%	0.0%	1.8%	2.7%
TCGA-AA-3697-01A-01D-1719-10	100.0%	2.7%	0.6%	10.3%	6.0%	19.0%	6.2%	1.5%	10.9%	2.8%	21.4%	5.1%	7.5%	0.1%	0.3%	0.3%	0.1%	0.1%	0.1%	0.3%	0.6%	3.9%
TCGA-AD-6963-01A-11D-1924-10	100.0%	1.1%	3.4%	2.2%	2.2%	34.8%	6.7%	5.6%	3.4%	1.1%	21.3%	6.7%	4.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.7%
TCGA-CM-5348-01A-21D-1719-10	100.0%	1.9%	2.5%	2.5%	3.1%	35.8%	1.9%	1.9%	5.0%	1.3%	30.2%	5.0%	4.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.6%	0.6%	3.1%
TCGA-DM-A0XD-01A-12D-A152-10	100.0%	1.1%	0.6%	1.1%	6.1%	34.3%	3.3%	1.7%	4.4%	1.1%	33.1%	3.3%	4.4%	0.0%	0.6%	0.6%	0.6%	0.6%	0.0%	0.0%	0.0%	3.3%
TCGA-AA-3511-01A-21D-1835-10	100.0%	2.5%	5.1%	0.8%	5.1%	29.7%	4.2%	5.1%	5.9%	5.9%	18.6%	4.2%	4.2%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%	6.8%
TCGA-AZ-6599-01A-11D-1771-10	100.0%	0.8%	0.4%	0.8%	1.2%	46.9%	1.2%	0.8%	1.2%	0.0%	39.1%	0.8%	3.1%	0.0%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	0.8%	2.7%
TCGA-CK-5914-01A-11D-1650-10	100.0%	1.4%	2.8%	5.6%	5.6%	29.6%	2.1%	2.8%	4.2%	1.4%	26.8%	4.2%	5.6%	0.0%	0.7%	0.0%	0.7%	0.7%	0.0%	0.0%	0.0%	5.6%
TCGA-CM-6675-01A-11D-1835-10	100.0%	1.1%	3.4%	5.7%	4.6%	26.4%	2.3%	1.1%	4.6%	3.4%	24.1%	1.1%	11.5%	3.4%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	5.7%
TCGA-AZ-6600-01A-11D-1771-10	100.0%	1.3%	2.7%	2.7%	6.7%	28.9%	3.4%	4.0%	8.7%	2.0%	22.1%	4.0%	6.7%	1.3%	0.0%	0.0%	0.0%	1.3%	0.0%	0.0%	0.0%	4.0%
TCGA-A6-6652-01A-11D-1771-10	100.0%	0.0%	1.1%	5.6%	10.0%	26.7%	3.3%	3.3%	2.2%	1.1%	28.9%	2.2%	5.6%	0.0%	1.1%	0.0%	3.3%	0.0%	0.0%	0.0%	1.1%	4.4%
TCGA-G4-6293-01A-11D-1719-10	100.0%	3.4%	1.7%	5.6%	6.4%	24.0%	2.6%	2.1%	6.4%	0.9%	27.0%	3.9%	3.9%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	0.4%	8.6%
TCGA-F4-6569-01A-11D-1771-10	100.0%	5.5%	0.0%	4.6%	3.7%	28.4%	3.7%	2.8%	2.8%	3.7%	23.9%	6.4%	8.3%	0.0%	0.9%	0.0%	0.9%	0.0%	0.9%	0.0%	0.9%	2.8%
TCGA-A6-6649-01A-11D-1771-10	100.0%	0.8%	1.6%	0.0%	3.1%	36.4%	3.9%	3.1%	5.4%	0.8%	34.9%	0.8%	4.7%	0.0%	0.0%	0.0%	0.0%	0.8%	0.0%	0.0%	0.0%	3.9%
TCGA-G4-6295-01A-11D-1719-10	100.0%	1.3%	4.7%	4.7%	6.0%	14.0%	10.0%	5.3%	7.3%	0.7%	25.3%	5.3%	10.7%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	0.0%	0.7%	3.3%
TCGA-CK-5913-01A-11D-1650-10	100.0%	0.4%	1.3%	6.3%	6.3%	24.6%	0.2%	0.8%	4.8%	0.8%	27.0%	0.6%	5.6%	1.4%	2.7%	0.9%	3.8%	1.8%	2.9%	2.1%	3.5%	2.0%
TCGA-CM-4752-01A-01D-1408-10	100.0%	2.0%	2.0%	5.3%	3.3%	29.6%	1.3%	0.0%	5.3%	0.0%	36.8%	2.0%	6.6%	0.0%	0.0%	0.0%	0.0%	0.0%	2.0%	0.0%	0.0%	3.9%
TCGA-AA-3502-01A-01D-1408-10	100.0%	1.8%	2.4%	1.2%	4.1%	32.9%	2.9%	1.2%	1.8%	1.2%	35.3%	4.1%	8.2%	0.0%	0.0%	0.0%	0.0%	1.2%	0.6%	0.0%	0.0%	1.2%
TCGA-D5-5538-01A-01D-1650-10	100.0%	0.9%	4.5%	6.3%	7.2%	27.0%	1.8%	2.7%	5.4%	0.9%	25.2%	3.6%	11.7%	0.0%	0.0%	0.9%	0.9%	0.0%	0.0%	0.0%	0.9%	0.0%
TCGA-CM-6678-01A-11D-1835-10	100.0%	3.0%	1.5%	1.5%	4.5%	44.7%	0.0%	0.8%	1.5%	0.0%	32.6%	0.8%	3.0%	1.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	3.8%
TCGA-AZ-4681-01A-01D-1408-10	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-CM-6674-01A-11D-1835-10	100.0%	0.3%	0.8%	3.0%	3.2%	32.3%	0.2%	0.9%	4.4%	0.7%	31.4%	0.3%	3.1%	1.7%	2.5%	1.3%	2.7%	0.9%	2.8%	1.0%	2.9%	3.8%
TCGA-AZ-5403-01A-01D-1650-10	100.0%	1.1%	4.6%	4.6%	5.7%	29.9%	3.4%	3.4%	1.1%	2.3%	29.9%	3.4%	4.6%	0.0%	1.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.6%
TCGA-AZ-4315-01A-01W-1461-10	100.0%	0.0%	0.0%	0.0%	33.3%	0.0%	0.0%	0.0%	0.0%	0.0%	66.7%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-A6-5667-01A-21D-1719-10	100.0%	1.4%	1.4%	8.8%	7.4%	21.2%	3.2%	2.8%	8.3%	2.3%	24.9%	4.6%	4.1%	0.0%	0.0%	0.5%	0.0%	0.9%	0.9%	0.0%	0.9%	6.5%
TCGA-DM-A28E-01A-11D-A16V-10	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	8.3%	0.0%	33.3%	8.3%	0.0%	50.0%
TCGA-G4-6625-01A-21D-1771-10	100.0%	1.1%	2.3%	1.1%	8.0%	20.5%	4.5%	3.4%	5.7%	0.0%	33.0%	5.7%	8.0%	1.1%	0.0%	1.1%	0.0%	2.3%	0.0%	0.0%	0.0%	2.3%
TCGA-G4-6299-01A-11D-1771-10	100.0%	3.4%	4.8%	2.7%	7.5%	26.0%	4.1%	2.1%	2.1%	1.4%	24.7%	5.5%	5.5%	0.0%	0.0%	0.0%	0.7%	2.1%	0.0%	0.0%	2.1%	5.5%
TCGA-DM-A0X9-01A-11D-A152-10	100.0%	2.2%	3.2%	2.7%	4.3%	34.9%	2.7%	3.8%	3.8%	1.6%	22.6%	1.6%	10.8%	1.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.5%	3.8%
TCGA-A6-5659-01A-01D-1650-10	100.0%	1.6%	1.6%	2.7%	12.8%	41.5%	1.1%	1.6%	2.7%	1.1%	22.3%	1.1%	6.4%	0.0%	0.0%	0.0%	0.0%	0.5%	0.5%	0.0%	0.5%	2.1%
TCGA-DM-A28C-01A-11D-A16V-10	100.0%	1.1%	1.1%	3.2%	9.5%	33.7%	4.2%	0.0%	3.2%	0.0%	22.1%	6.3%	3.2%	1.1%	0.0%	0.0%	1.1%	0.0%	1.1%	1.1%	2.1%	6.3%
TCGA-AA-3712-01A-21D-1719-10	100.0%	2.7%	0.7%	10.4%	5.5%	19.1%	9.2%	1.5%	9.4%	2.0%	18.1%	7.8%	8.9%	0.2%	0.0%	0.0%	0.2%	0.2%	0.2%	0.2%	0.2%	3.7%
TCGA-AZ-4616-01A-21D-1835-10	100.0%	3.2%	0.0%	3.2%	4.8%	32.3%	3.2%	3.2%	3.2%	6.5%	29.0%	4.8%	3.2%	1.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%
TCGA-D5-6931-01A-11D-1924-10	100.0%	7.8%	2.5%	6.3%	5.0%	25.6%	0.9%	2.5%	5.6%	9.4%	24.4%	0.6%	5.0%	0.6%	0.0%	0.3%	0.0%	0.0%	0.3%	0.3%	0.0%	2.8%
TCGA-G4-6309-01A-21D-1835-10	100.0%	0.7%	0.6%	3.0%	2.3%	31.0%	0.3%	0.6%	3.0%	0.6%	31.2%	0.3%	3.3%	2.0%	2.6%	1.3%	2.9%	1.6%	3.4%	0.8%	3.3%	5.2%
TCGA-CM-4746-01A-01W-1461-10	100.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-A6-6648-01A-11D-1771-10	100.0%	1.1%	1.1%	4.3%	4.3%	34.4%	3.2%	5.4%	3.2%	1.1%	31.2%	1.1%	3.2%	0.0%	0.0%	0.0%	0.0%	0.0%	1.1%	0.0%	1.1%	4.3%
TCGA-D5-6929-01A-31D-1924-10	100.0%	2.2%	1.5%	5.2%	11.9%	23.9%	3.7%	2.2%	4.5%	1.5%	21.6%	2.2%	9.7%	0.7%	1.5%	0.0%	0.0%	0.0%	0.7%	0.0%	0.0%	6.7%
TCGA-G4-6322-01A-11D-1719-10	100.0%	1.1%	0.0%	3.2%	7.4%	27.4%	2.1%	2.1%	3.2%	0.0%	33.7%	2.1%	8.4%	0.0%	0.0%	0.0%	1.1%	2.1%	0.0%	0.0%	0.0%	6.3%
TCGA-D5-6538-01A-11D-1719-10	100.0%	3.0%	2.2%	3.7%	3.0%	34.8%	0.0%	3.0%	3.7%	1.5%	34.1%	2.2%	3.7%	0.7%	0.7%	0.0%	0.0%	0.7%	0.0%	0.7%	0.0%	2.2%
TCGA-AZ-5407-01A-01D-1719-10	100.0%	3.8%	2.8%	9.8%	4.7%	24.1%	4.7%	1.3%	8.5%	1.9%	19.6%	5.7%	3.8%	0.0%	0.0%	0.0%	0.3%	0.3%	1.3%	0.0%	0.3%	7.0%
TCGA-G4-6323-01A-11D-1719-10	100.0%	1.2%	0.0%	4.9%	6.1%	30.5%	3.7%	2.4%	3.7%	0.0%	31.7%	3.7%	8.5%	0.0%	1.2%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.4%
TCGA-DM-A0XF-01A-11D-A152-10	100.0%	0.7%	3.7%	0.7%	7.4%	28.1%	3.7%	4.4%	5.2%	1.5%	20.7%	3.7%	11.1%	0.0%	0.0%	0.0%	0.0%	0.0%	1.5%	0.0%	1.5%	5.9%
TCGA-G4-6307-01A-11D-1719-10	100.0%	1.4%	0.0%	1.4%	16.2%	27.0%	1.4%	4.1%	4.1%	1.4%	20.3%	6.8%	8.1%	1.4%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	2.7%	4.1%
TCGA-G4-6298-01A-11D-1719-10	100.0%	1.6%	1.4%	9.5%	5.4%	23.1%	3.0%	1.4%	9.5%	2.2%	23.1%	4.3%	7.6%	0.0%	0.5%	0.0%	0.0%	0.5%	0.0%	0.3%	0.3%	6.3%
TCGA-D5-6924-01A-11D-1924-10	100.0%	1.1%	2.9%	0.6%	6.3%	27.0%	5.2%	5.7%	4.0%	4.0%	24.7%	3.4%	7.5%	0.0%	0.0%	0.0%	0.0%	1.1%	0.0%	0.0%	0.0%	6.3%
TCGA-F4-6463-01A-11D-1719-10	100.0%	0.8%	1.6%	10.2%	4.5%	22.0%	5.3%	2.0%	8.9%	2.4%	23.2%	4.9%	3.7%	0.4%	0.0%	0.4%	0.4%	0.0%	0.8%	0.4%	0.4%	7.7%
TCGA-DM-A285-01A-11D-A16V-10	100.0%	3.3%	2.5%	5.8%	11.6%	17.4%	2.5%	6.6%	1.7%	0.0%	28.9%	5.8%	6.6%	0.0%	1.7%	0.0%	0.8%	0.0%	0.0%	0.8%	0.0%	4.1%
TCGA-D5-5537-01A-21D-1924-10	100.0%	0.8%	1.6%	2.3%	6.2%	33.3%	1.6%	0.0%	3.1%	1.6%	30.2%	1.6%	9.3%	0.8%	0.0%	0.0%	0.0%	1.6%	0.0%	0.8%	0.8%	4.7%
TCGA-AY-6386-01A-21D-1719-10	100.0%	2.9%	1.7%	9.9%	4.8%	22.8%	4.1%	1.5%	13.3%	2.4%	20.3%	5.3%	4.6%	0.4%	0.4%	0.0%	0.0%	0.4%	0.0%	0.0%	0.0%	5.3%
TCGA-CM-6168-01A-11D-1650-10	100.0%	0.0%	0.0%	4.4%	7.3%	35.4%	4.4%	1.0%	6.3%	0.5%	23.8%	2.9%	10.2%	0.0%	0.5%	0.5%	0.0%	0.0%	1.0%	0.0%	0.5%	1.5%
TCGA-DM-A1D9-01A-11D-A152-10	100.0%	0.7%	6.3%	2.8%	6.3%	26.4%	4.2%	5.6%	1.4%	2.1%	20.8%	6.3%	6.9%	0.7%	1.4%	0.0%	0.0%	1.4%	0.7%	0.0%	0.0%	6.3%
TCGA-CM-5344-01A-21D-1719-10	100.0%	0.0%	3.1%	3.1%	6.3%	26.6%	4.7%	0.0%	0.0%	1.6%	34.4%	1.6%	10.9%	1.6%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	1.6%	4.7%

sample_ID	ALL	A_C	A_T	A_G	C_A	C_T	C_G	T_A	T_C	T_G	G_A	G_C	G_T	-_T	T_-	-_C	C_-	-_A	A_-	-_G	G_-	MNC
TCGA-CM-6172-01A-11D-1650-10	100.0%	3.2%	1.1%	1.1%	3.2%	39.4%	3.2%	1.1%	1.1%	2.1%	39.4%	0.0%	5.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-A6-6782-01A-11D-1835-10	100.0%	3.5%	4.2%	2.8%	11.1%	27.1%	2.1%	3.5%	3.5%	0.7%	26.4%	4.9%	4.9%	0.0%	0.0%	0.0%	0.7%	0.7%	0.0%	0.0%	0.0%	4.2%
TCGA-G4-6303-01A-11D-1771-10	100.0%	2.9%	1.0%	1.0%	5.7%	29.5%	5.7%	1.0%	1.9%	1.9%	36.2%	0.0%	5.7%	1.0%	0.0%	0.0%	0.0%	1.9%	1.0%	0.0%	1.0%	2.9%
TCGA-A6-6137-01A-11D-1771-10	100.0%	1.7%	0.8%	0.8%	4.2%	33.1%	3.4%	2.5%	1.7%	2.5%	36.4%	1.7%	4.2%	2.5%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.8%	3.4%
TCGA-F4-6855-01A-11D-1924-10	100.0%	1.3%	1.3%	2.6%	11.7%	29.2%	2.6%	1.3%	5.2%	0.6%	27.9%	1.9%	6.5%	1.3%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	6.5%
TCGA-G4-6315-01A-11D-1719-10	100.0%	1.8%	1.2%	2.4%	6.6%	31.3%	1.2%	3.6%	3.0%	1.2%	27.1%	4.8%	9.0%	0.0%	0.0%	0.0%	0.0%	0.6%	1.2%	0.6%	0.6%	3.6%

Figure 16. Low Mutation Group Mutation Type Proportions

Columns are as in Figures 12 and 13, and the data are now shown as percentages instead of counts.

Cross comparing the indel percentages from Figure 17 and the MNC percentages from Figure 16, it seems they had very similar values, although not always matched up in magnitude within any one tumor. C_T and G_A mutations accounted for between 40%-60% of the mutations in most of the samples. C_A and G_T mutations were somewhat common single base changes, and accounted for about 10%-20% of the mutations in most of the samples. The A_G and T_C mutations were slightly less common than that with most ranging from 6%-20%, but tended to be low more often. The other three mutation type categories mostly ranged between 0% and 5%, and were closer to values between 3-5%.

sample_ID	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-CM-6169-01A-11D-1650-10	0.8%	2.3%	3.1%	62.8%	27.1%	3.9%	3.1%	58.1%	3.1%	4.7%	17.1%
TCGA-D5-5541-01A-01D-1650-10	5.0%	2.0%	6.9%	70.3%	17.8%	4.0%	3.0%	60.4%	3.0%	9.9%	7.9%
TCGA-G4-6314-01A-11D-1719-10	0.7%	0.0%	0.7%	63.6%	35.0%	4.9%	11.9%	56.6%	3.5%	7.0%	14.7%
TCGA-G4-6297-01A-11D-1719-10	1.0%	2.6%	3.6%	65.2%	25.2%	2.6%	8.5%	49.2%	2.6%	16.1%	11.5%
TCGA-F4-6703-01A-11D-1835-10	1.3%	7.6%	8.9%	67.3%	18.2%	2.7%	1.3%	55.8%	2.7%	11.6%	11.6%
TCGA-DM-A1DA-01A-11D-A152-10	0.0%	1.2%	1.2%	56.7%	39.0%	4.7%	8.4%	45.9%	4.9%	10.8%	20.9%
TCGA-F4-6806-01A-11D-1835-10	4.4%	0.0%	4.4%	70.0%	17.8%	2.2%	4.4%	66.7%	1.1%	3.3%	10.0%
TCGA-A6-6650-01A-11D-1771-10	0.7%	1.3%	2.0%	59.7%	29.5%	3.4%	4.7%	53.0%	4.7%	6.7%	16.8%
TCGA-F4-6805-01A-11D-1835-10	0.0%	3.8%	3.8%	60.4%	28.3%	5.7%	7.5%	47.2%	3.8%	13.2%	11.3%
TCGA-D5-6898-01A-11D-1924-10	3.9%	0.0%	3.9%	63.2%	31.6%	6.6%	6.6%	57.9%	1.3%	5.3%	17.1%
TCGA-CA-5797-01A-01D-1650-10	3.6%	3.6%	7.1%	60.7%	30.4%	7.1%	8.9%	57.1%	1.8%	3.6%	12.5%
TCGA-G4-6321-01A-11D-1719-10	0.0%	0.5%	0.5%	67.4%	26.0%	2.7%	8.2%	46.4%	4.2%	21.0%	10.9%
TCGA-CM-6679-01A-11D-1835-10	4.2%	0.0%	4.2%	79.2%	14.6%	8.3%	2.1%	77.1%	0.0%	2.1%	4.2%
TCGA-F4-6809-01A-11D-1835-10	2.4%	2.4%	4.8%	48.4%	37.1%	8.9%	8.1%	38.7%	1.6%	9.7%	18.5%
TCGA-CK-5915-01A-11D-1650-10	2.8%	0.7%	3.5%	76.2%	18.2%	2.8%	6.3%	66.4%	2.1%	9.8%	7.0%
TCGA-D5-6533-01A-11D-1719-10	0.6%	1.2%	1.8%	57.9%	35.7%	10.5%	7.0%	45.6%	4.1%	12.3%	14.0%
TCGA-AU-3779-01A-01D-1719-10	0.0%	0.0%	0.0%	68.3%	28.8%	6.7%	2.9%	59.6%	4.8%	8.7%	14.4%
TCGA-AD-6901-01A-11D-1924-10	1.2%	1.2%	2.4%	62.0%	32.8%	4.4%	8.8%	55.6%	3.2%	6.4%	16.4%
TCGA-D5-6529-01A-11D-1771-10	0.0%	2.6%	2.6%	76.1%	17.9%	2.6%	6.0%	69.2%	0.9%	6.8%	8.5%
TCGA-CM-6164-01A-11D-1650-10	1.7%	1.7%	3.3%	63.3%	30.0%	8.3%	3.3%	46.7%	1.7%	16.7%	16.7%
TCGA-DM-A28F-01A-11D-A16V-10	3.3%	1.7%	5.0%	54.8%	37.2%	6.3%	8.4%	47.3%	9.2%	7.5%	13.4%
TCGA-G4-6311-01A-11D-1719-10	0.3%	1.5%	1.8%	60.8%	30.9%	3.9%	7.4%	45.1%	5.3%	15.7%	14.2%
TCGA-D5-6532-01A-11D-1719-10	1.8%	2.7%	4.5%	60.4%	32.4%	4.5%	10.8%	55.0%	7.2%	5.4%	9.9%
TCGA-AZ-4682-01B-01D-1408-10	0.0%	1.0%	1.0%	78.1%	18.8%	4.2%	2.1%	75.0%	4.2%	3.1%	8.3%
TCGA-CM-5863-01A-21D-1835-10	0.0%	4.8%	4.8%	61.9%	22.6%	1.2%	6.0%	52.4%	2.4%	9.5%	13.1%
TCGA-D5-6536-01A-11D-1719-10	1.9%	1.0%	2.9%	63.1%	30.1%	5.8%	7.8%	56.3%	1.9%	6.8%	14.6%
TCGA-DM-A28H-01A-11D-A16V-10	3.3%	3.3%	6.7%	58.3%	27.5%	5.0%	10.0%	52.5%	5.0%	5.8%	7.5%
TCGA-F4-6460-01A-11D-1771-10	0.8%	0.8%	1.6%	59.8%	37.8%	7.9%	7.9%	52.8%	3.1%	7.1%	18.9%
TCGA-D5-6531-01A-11D-1719-10	0.5%	1.0%	1.5%	62.9%	31.0%	8.1%	6.1%	55.8%	4.1%	7.1%	12.7%
TCGA-G4-6304-01A-11D-1924-10	0.7%	4.4%	5.1%	75.5%	19.4%	4.4%	2.0%	66.0%	2.4%	9.5%	10.5%
TCGA-D5-6926-01A-11D-1924-10	0.8%	0.0%	0.8%	69.5%	23.7%	3.1%	6.9%	60.3%	2.3%	9.2%	11.5%
TCGA-CM-5868-01A-01D-1650-10	1.6%	2.1%	3.7%	60.3%	28.6%	6.3%	3.2%	44.4%	4.8%	15.9%	14.3%
TCGA-A6-2675-01A-02D-1719-10	1.2%	0.0%	1.2%	78.3%	16.9%	4.8%	4.8%	68.7%	2.4%	9.6%	4.8%
TCGA-DM-A1DB-01A-11D-A152-10	0.0%	1.3%	1.3%	67.5%	26.3%	1.3%	11.3%	57.5%	1.3%	10.0%	12.5%
TCGA-CM-4750-01A-01D-1408-10	1.0%	1.0%	2.0%	65.7%	25.3%	4.0%	0.0%	59.6%	4.0%	6.1%	17.2%
TCGA-AY-5543-01A-01D-1650-10	0.8%	0.8%	1.6%	70.6%	27.8%	2.4%	7.9%	61.1%	4.0%	9.5%	13.5%
TCGA-AD-6548-01A-11D-1835-10	1.0%	1.9%	2.9%	74.3%	19.0%	2.9%	2.9%	66.7%	2.9%	7.6%	10.5%
TCGA-AZ-4323-01A-21D-1835-10	0.0%	0.0%	0.0%	72.7%	24.2%	6.1%	6.1%	66.7%	3.0%	6.1%	9.1%
TCGA-CM-5864-01A-01D-1650-10	0.5%	1.8%	2.3%	58.4%	37.6%	5.0%	8.6%	49.3%	5.9%	9.0%	18.1%
TCGA-AA-3489-01A-21D-1835-10	2.4%	0.0%	2.4%	63.4%	28.0%	8.5%	2.4%	58.5%	2.4%	4.9%	14.6%
TCGA-A6-6140-01A-11D-1771-10	0.4%	0.0%	0.4%	74.7%	24.4%	3.6%	7.1%	64.6%	3.6%	10.1%	10.1%
TCGA-A6-5664-01A-21D-1835-10	0.0%	0.0%	0.0%	75.0%	15.4%	0.0%	0.0%	65.4%	3.8%	9.6%	11.5%
TCGA-CM-5860-01A-01D-1650-10	0.8%	0.0%	0.8%	74.6%	21.3%	0.8%	4.1%	67.2%	0.8%	7.4%	15.6%
TCGA-DM-A28K-01A-21D-A16V-10	4.3%	0.9%	5.2%	71.3%	16.5%	2.6%	4.3%	67.0%	1.7%	4.3%	7.8%
TCGA-F4-6807-01A-11D-1835-10	0.6%	0.6%	1.3%	48.8%	42.5%	5.6%	4.4%	43.1%	1.3%	5.6%	31.3%
TCGA-AD-6899-01A-11D-1924-10	0.6%	0.6%	1.3%	77.1%	19.7%	3.2%	6.4%	72.6%	0.6%	4.5%	9.6%
TCGA-CM-5862-01A-01D-1650-10	2.1%	1.1%	3.2%	70.2%	20.2%	2.1%	3.2%	58.5%	3.2%	11.7%	11.7%
TCGA-AY-6196-01A-11D-1719-10	0.5%	0.8%	1.3%	56.0%	39.1%	5.7%	13.5%	39.1%	6.2%	16.8%	13.7%
TCGA-CK-4948-01B-11D-1650-10	0.9%	0.0%	0.9%	67.6%	26.9%	4.6%	7.4%	60.2%	4.6%	7.4%	10.2%
TCGA-D5-5539-01A-01D-1650-10	2.8%	0.0%	2.8%	75.5%	20.8%	3.8%	7.5%	66.0%	1.9%	9.4%	7.5%
TCGA-G4-6626-01A-11D-1771-10	0.0%	0.0%	0.0%	72.6%	24.2%	5.1%	4.5%	65.6%	0.0%	7.0%	14.6%
TCGA-D5-6537-01A-11D-1719-10	0.3%	0.5%	0.8%	64.8%	26.9%	2.9%	5.9%	47.5%	4.8%	17.3%	13.3%
TCGA-DM-A28G-01A-11D-A16V-10	1.8%	0.0%	1.8%	67.5%	26.3%	4.4%	2.6%	62.3%	3.5%	5.3%	15.8%
TCGA-AA-3662-01A-01D-1719-10	0.6%	1.4%	1.9%	65.5%	28.5%	2.3%	9.9%	41.1%	5.4%	24.4%	10.9%
TCGA-DM-A1D8-01A-11D-A152-10	2.6%	0.0%	2.6%	53.2%	42.2%	7.8%	6.5%	46.8%	7.1%	6.5%	20.8%
TCGA-CM-6163-01A-11D-1650-10	0.0%	0.0%	0.0%	74.7%	23.2%	1.0%	5.1%	65.7%	1.0%	9.1%	16.2%
TCGA-D5-7000-01A-11D-1924-10	1.8%	3.0%	4.8%	65.3%	16.8%	2.4%	4.8%	57.5%	0.0%	7.8%	9.6%
TCGA-CA-6719-01A-11D-1835-10	2.5%	0.8%	3.3%	77.9%	13.1%	0.8%	0.8%	69.7%	0.8%	8.2%	10.7%

sample_ID	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-DM-A282-01A-12D-A16V-10	2.2%	2.2%	4.4%	67.8%	25.6%	4.4%	5.6%	63.3%	1.1%	4.4%	14.4%
TCGA-AD-6888-01A-11D-1924-10	1.7%	0.9%	2.6%	71.8%	20.5%	2.6%	3.4%	67.5%	4.3%	4.3%	10.3%
TCGA-CA-6716-01A-11D-1835-10	4.3%	0.5%	4.8%	73.6%	19.2%	1.0%	3.8%	65.4%	2.9%	8.2%	11.5%
TCGA-CK-5912-01A-11D-1650-10	1.6%	3.2%	4.8%	69.8%	20.6%	3.2%	7.9%	66.7%	1.6%	3.2%	7.9%
TCGA-CK-4950-01A-01D-1719-10	0.2%	1.8%	2.1%	60.1%	32.8%	2.5%	8.7%	48.2%	5.7%	11.9%	15.8%
TCGA-A6-5660-01A-01D-1650-10	2.6%	0.9%	3.4%	70.7%	21.6%	1.7%	6.0%	60.3%	3.4%	10.3%	10.3%
TCGA-AZ-6605-01A-11D-1835-10	1.5%	0.7%	2.2%	74.3%	19.9%	0.7%	6.6%	62.5%	1.5%	11.8%	11.0%
TCGA-DM-A1D6-01A-21D-A152-10	0.8%	1.6%	2.3%	73.4%	21.1%	3.9%	7.0%	66.4%	3.1%	7.0%	7.0%
TCGA-CM-6165-01A-11D-1650-10	2.7%	2.7%	5.4%	55.4%	34.8%	2.7%	4.5%	46.4%	0.0%	8.9%	27.7%
TCGA-D5-6541-01A-11D-1719-10	3.2%	3.2%	6.3%	70.5%	14.7%	2.1%	4.2%	63.2%	2.1%	7.4%	6.3%
TCGA-CM-6677-01A-11D-1835-10	1.5%	0.8%	2.3%	76.2%	17.7%	2.3%	1.5%	69.2%	2.3%	6.9%	11.5%
TCGA-CM-4744-01A-01D-1408-10	3.0%	1.2%	4.2%	61.4%	25.9%	3.0%	3.0%	57.2%	4.2%	4.2%	15.7%
TCGA-DM-A1D7-01A-11D-A152-10	0.7%	0.0%	0.7%	60.3%	38.3%	7.1%	6.4%	55.3%	4.3%	5.0%	20.6%
TCGA-AA-3510-01A-01W-1461-10	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
TCGA-A6-4105-01A-02D-1771-10	1.3%	0.0%	1.3%	66.4%	29.5%	5.4%	7.4%	58.4%	2.0%	8.1%	14.8%
TCGA-G4-6306-01A-11D-1771-10	1.6%	1.6%	3.2%	71.8%	17.7%	3.2%	1.6%	62.9%	0.8%	8.9%	12.1%
TCGA-D5-6535-01A-11D-1719-10	0.6%	0.6%	1.3%	55.0%	40.6%	3.1%	9.4%	46.9%	6.3%	8.1%	21.9%
TCGA-CM-5349-01A-21D-1719-10	1.9%	1.9%	3.8%	71.2%	19.2%	5.8%	1.9%	62.5%	4.8%	8.7%	6.7%
TCGA-A6-6138-01A-11D-1771-10	1.7%	2.6%	4.3%	74.4%	20.5%	3.4%	3.4%	65.0%	3.4%	9.4%	10.3%
TCGA-A6-5656-01A-21D-1835-10	0.9%	1.9%	2.8%	76.4%	16.0%	1.9%	3.8%	70.8%	0.9%	5.7%	9.4%
TCGA-CM-4747-01A-01D-1408-10	2.1%	1.0%	3.1%	80.2%	14.6%	1.0%	2.1%	71.9%	3.1%	8.3%	8.3%
TCGA-D5-5540-01A-01D-1650-10	2.3%	0.0%	2.3%	65.9%	29.5%	7.6%	3.0%	60.6%	4.5%	5.3%	14.4%
TCGA-CA-5796-01A-01D-1650-10	3.0%	1.5%	4.5%	68.2%	25.8%	4.5%	1.5%	59.1%	3.0%	9.1%	16.7%
TCGA-D5-6539-01A-11D-1719-10	0.0%	0.0%	0.0%	63.6%	30.3%	3.0%	6.1%	54.5%	7.6%	9.1%	13.6%
TCGA-G4-6294-01A-11D-1806-10	0.5%	1.6%	2.2%	48.4%	42.9%	9.8%	13.0%	41.3%	3.8%	7.1%	16.3%
TCGA-F4-6808-01A-11D-1835-10	0.0%	2.4%	2.4%	65.9%	26.8%	3.3%	4.9%	61.0%	1.6%	4.9%	17.1%
TCGA-F4-6704-01A-11D-1835-10	0.0%	0.0%	0.0%	60.0%	25.7%	11.4%	5.7%	60.0%	0.0%	0.0%	8.6%
TCGA-G4-6310-01A-11D-1719-10	0.9%	0.9%	1.7%	71.8%	23.9%	3.4%	6.0%	68.4%	1.7%	3.4%	12.8%
TCGA-DM-A28M-01A-12D-A16V-10	2.8%	1.9%	4.6%	75.9%	17.6%	3.7%	0.9%	68.5%	1.9%	7.4%	11.1%
TCGA-A6-6654-01A-21D-1835-10	1.2%	2.4%	3.6%	83.9%	10.7%	3.0%	1.8%	79.8%	0.6%	4.2%	5.4%
TCGA-D5-6927-01A-21D-1924-10	5.1%	13.2%	18.3%	63.3%	15.2%	1.6%	1.2%	49.5%	1.7%	13.8%	10.6%
TCGA-A6-2671-01A-01D-1408-10	1.1%	3.4%	4.5%	56.8%	28.4%	2.3%	3.4%	52.3%	4.5%	4.5%	18.2%
TCGA-CM-4748-01A-01D-1408-10	1.5%	1.5%	2.9%	58.8%	30.9%	1.5%	5.9%	51.5%	4.4%	7.4%	19.1%
TCGA-AA-3660-01A-01D-1719-10	0.8%	0.8%	1.7%	72.0%	18.6%	1.7%	2.5%	64.4%	5.1%	7.6%	9.3%
TCGA-D5-6534-01A-21D-1924-10	0.6%	0.6%	1.3%	55.5%	38.7%	6.5%	5.2%	52.9%	8.4%	2.6%	18.7%
TCGA-F4-6854-01A-11D-1924-10	1.6%	0.8%	2.4%	62.6%	29.3%	7.3%	2.4%	54.5%	4.9%	8.1%	14.6%
TCGA-CK-4952-01A-01D-1719-10	1.2%	1.9%	3.1%	71.3%	20.4%	2.5%	4.3%	56.2%	4.6%	15.1%	9.0%
TCGA-G4-6317-01A-11D-1719-10	0.9%	0.9%	1.8%	72.8%	20.2%	4.4%	3.5%	69.3%	1.8%	3.5%	10.5%
TCGA-CK-4947-01B-11D-1650-10	0.0%	3.2%	3.2%	68.8%	26.9%	6.5%	5.4%	58.1%	4.3%	10.8%	10.8%
TCGA-CM-6166-01A-11D-1650-10	5.2%	0.0%	5.2%	60.4%	28.1%	2.1%	5.2%	58.3%	2.1%	2.1%	18.8%
TCGA-DM-A1D4-01A-21D-A152-10	2.2%	0.6%	2.8%	68.0%	28.2%	3.9%	3.9%	61.9%	2.8%	6.1%	17.7%
TCGA-CM-5341-01A-01D-1408-10	0.6%	1.8%	2.4%	65.9%	26.9%	6.0%	7.8%	59.9%	3.0%	6.0%	10.2%
TCGA-CM-6170-01A-11D-1650-10	2.0%	1.0%	3.0%	76.8%	16.2%	3.0%	5.1%	71.7%	3.0%	5.1%	5.1%
TCGA-AZ-6603-01A-11D-1835-10	0.8%	2.5%	3.4%	67.2%	17.6%	1.7%	4.2%	56.3%	4.2%	10.9%	7.6%
TCGA-A6-5662-01A-01D-1650-10	2.9%	2.9%	5.9%	66.2%	26.5%	5.9%	2.9%	61.8%	1.5%	4.4%	16.2%
TCGA-AA-3655-01A-02D-1719-10	2.3%	0.0%	2.3%	77.0%	17.2%	1.1%	5.7%	62.1%	0.0%	14.9%	10.3%
TCGA-CM-4746-01A-01D-1408-10	4.9%	10.4%	15.3%	72.2%	9.6%	0.7%	1.2%	67.6%	0.7%	4.5%	7.1%
TCGA-A6-5666-01A-01D-1650-10	4.8%	1.0%	5.7%	62.9%	29.5%	1.9%	3.8%	54.3%	5.7%	8.6%	18.1%
TCGA-CM-6161-01A-11D-1650-10	0.9%	0.9%	1.9%	71.0%	23.4%	6.5%	4.7%	60.7%	0.9%	10.3%	11.2%
TCGA-AD-6890-01A-11D-1924-10	2.5%	3.4%	5.9%	66.1%	21.2%	1.7%	1.7%	64.4%	1.7%	1.7%	16.1%
TCGA-D5-6920-01A-11D-1924-10	0.8%	3.4%	4.2%	69.7%	19.3%	4.2%	5.0%	63.0%	0.8%	6.7%	9.2%
TCGA-AZ-6607-01A-11D-1835-10	1.4%	0.0%	1.4%	69.6%	24.6%	7.2%	1.4%	65.2%	5.8%	4.3%	10.1%
TCGA-CM-6676-01A-11D-1835-10	1.1%	0.0%	1.1%	80.4%	17.4%	5.4%	1.1%	78.3%	2.2%	2.2%	8.7%
TCGA-DM-A1D0-01A-11D-A152-10	0.9%	0.9%	1.7%	56.4%	35.9%	11.1%	9.4%	47.0%	3.4%	9.4%	12.0%
TCGA-AD-6965-01A-11D-1924-10	2.0%	0.7%	2.7%	68.0%	23.8%	4.1%	5.4%	60.5%	1.4%	7.5%	12.9%
TCGA-A6-6142-01A-11D-1771-10	0.0%	5.6%	5.6%	67.6%	21.1%	7.0%	0.0%	53.5%	5.6%	14.1%	8.5%
TCGA-D5-6932-01A-11D-1924-10	0.0%	2.5%	2.5%	52.1%	43.0%	9.9%	8.3%	45.5%	5.8%	6.6%	19.0%
TCGA-F4-6461-01A-11D-1771-10	1.6%	0.0%	1.6%	70.9%	22.8%	3.1%	5.5%	59.8%	2.4%	11.0%	11.8%

sample_ID	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-A6-5657-01A-01D-1650-10	2.3%	1.1%	3.4%	67.0%	29.5%	1.1%	4.5%	65.9%	2.3%	1.1%	21.6%
TCGA-DM-A1HA-01A-11D-A152-10	1.8%	1.8%	3.5%	61.8%	32.4%	6.5%	8.8%	55.9%	0.6%	5.9%	16.5%
TCGA-CM-6680-01A-11D-1835-10	1.1%	1.7%	2.8%	63.5%	20.8%	2.8%	4.5%	52.8%	3.9%	10.7%	9.6%
TCGA-A6-6651-01A-21D-1835-10	0.9%	4.5%	5.4%	53.2%	24.3%	4.5%	8.1%	42.3%	1.8%	10.8%	9.9%
TCGA-D5-6922-01A-11D-1924-10	1.7%	0.0%	1.7%	72.4%	23.3%	1.7%	10.3%	68.1%	3.4%	4.3%	7.8%
TCGA-F4-6459-01A-11D-1771-10	1.8%	5.4%	7.1%	55.4%	34.8%	7.1%	12.5%	49.1%	3.6%	6.3%	11.6%
TCGA-AA-3697-01A-01D-1719-10	0.8%	1.2%	1.9%	61.6%	32.6%	2.2%	11.3%	40.4%	5.5%	21.2%	13.5%
TCGA-AD-6963-01A-11D-1924-10	0.0%	0.0%	0.0%	61.8%	31.5%	9.0%	13.5%	56.2%	2.2%	5.6%	6.7%
TCGA-CM-5348-01A-21D-1719-10	0.6%	0.6%	1.3%	73.6%	22.0%	4.4%	6.9%	66.0%	3.1%	7.5%	7.5%
TCGA-DM-A0XD-01A-12D-A152-10	1.1%	1.1%	2.2%	72.9%	21.5%	2.2%	6.6%	67.4%	2.2%	5.5%	10.5%
TCGA-AA-3511-01A-21D-1835-10	1.7%	0.0%	1.7%	55.1%	36.4%	10.2%	8.5%	48.3%	8.5%	6.8%	9.3%
TCGA-AZ-6599-01A-11D-1771-10	0.0%	1.2%	1.2%	88.0%	8.1%	1.2%	1.9%	86.0%	0.8%	1.9%	4.3%
TCGA-CK-5914-01A-11D-1650-10	0.7%	1.4%	2.1%	66.2%	26.1%	5.6%	6.3%	56.3%	2.8%	9.9%	11.3%
TCGA-CM-6675-01A-11D-1835-10	3.4%	1.1%	4.6%	60.9%	28.7%	4.6%	3.4%	50.6%	4.6%	10.3%	16.1%
TCGA-AZ-6600-01A-11D-1771-10	2.7%	0.0%	2.7%	62.4%	30.9%	6.7%	7.4%	51.0%	3.4%	11.4%	13.4%
TCGA-A6-6652-01A-11D-1771-10	0.0%	5.6%	5.6%	63.3%	26.7%	4.4%	5.6%	55.6%	1.1%	7.8%	15.6%
TCGA-G4-6293-01A-11D-1719-10	1.7%	1.7%	3.4%	63.1%	24.9%	3.9%	6.4%	51.1%	4.3%	12.0%	10.3%
TCGA-F4-6569-01A-11D-1771-10	0.0%	3.7%	3.7%	59.6%	33.9%	2.8%	10.1%	52.3%	9.2%	7.3%	11.9%
TCGA-A6-6649-01A-11D-1771-10	0.8%	0.0%	0.8%	76.7%	18.6%	4.7%	4.7%	71.3%	1.6%	5.4%	7.8%
TCGA-G4-6295-01A-11D-1719-10	0.7%	0.7%	1.3%	51.3%	44.0%	10.0%	15.3%	39.3%	2.0%	12.0%	16.7%
TCGA-CK-5913-01A-11D-1650-10	6.2%	12.9%	19.1%	62.8%	16.1%	2.1%	0.8%	51.6%	1.2%	11.2%	12.0%
TCGA-CM-4752-01A-01D-1408-10	0.0%	2.0%	2.0%	77.0%	17.1%	2.0%	3.3%	66.4%	2.0%	10.5%	9.9%
TCGA-AA-3502-01A-01D-1408-10	1.2%	0.6%	1.8%	71.2%	25.9%	3.5%	7.1%	68.2%	2.9%	2.9%	12.4%
TCGA-D5-5538-01A-01D-1650-10	0.9%	1.8%	2.7%	64.0%	33.3%	7.2%	5.4%	52.3%	1.8%	11.7%	18.9%
TCGA-CM-6678-01A-11D-1835-10	1.5%	0.8%	2.3%	80.3%	13.6%	2.3%	0.8%	77.3%	3.0%	3.0%	7.6%
TCGA-AZ-4681-01A-01D-1408-10	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
TCGA-CM-6674-01A-11D-1835-10	4.9%	10.8%	15.7%	71.0%	9.5%	1.7%	0.6%	63.6%	1.0%	7.4%	6.3%
TCGA-AZ-5403-01A-01D-1650-10	0.0%	1.1%	1.1%	65.5%	28.7%	8.0%	6.9%	59.8%	3.4%	5.7%	10.3%
TCGA-AZ-4315-01A-01W-1461-10	0.0%	0.0%	0.0%	66.7%	33.3%	0.0%	0.0%	66.7%	0.0%	0.0%	33.3%
TCGA-A6-5667-01A-21D-1719-10	1.4%	1.8%	3.2%	63.1%	27.2%	4.1%	7.8%	46.1%	3.7%	17.1%	11.5%
TCGA-DM-A28E-01A-11D-A16V-10	8.3%	41.7%	50.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
TCGA-G4-6625-01A-21D-1771-10	4.5%	0.0%	4.5%	60.2%	33.0%	5.7%	10.2%	53.4%	1.1%	6.8%	15.9%
TCGA-G4-6299-01A-11D-1771-10	2.1%	2.7%	4.8%	55.5%	34.2%	6.8%	9.6%	50.7%	4.8%	4.8%	13.0%
TCGA-DM-A0X9-01A-11D-A152-10	1.6%	0.5%	2.2%	64.0%	30.1%	7.0%	4.3%	57.5%	3.8%	6.5%	15.1%
TCGA-A6-5659-01A-01D-1650-10	0.5%	1.1%	1.6%	69.1%	27.1%	3.2%	2.1%	63.8%	2.7%	5.3%	19.1%
TCGA-DM-A28C-01A-11D-A16V-10	2.1%	4.2%	6.3%	62.1%	25.3%	1.1%	10.5%	55.8%	1.1%	6.3%	12.6%
TCGA-AA-3712-01A-21D-1719-10	0.5%	0.5%	1.0%	56.9%	38.3%	2.2%	17.0%	37.1%	4.8%	19.8%	14.3%
TCGA-AZ-4616-01A-21D-1835-10	1.6%	0.0%	1.6%	67.7%	29.0%	3.2%	8.1%	61.3%	9.7%	6.5%	8.1%
TCGA-D5-6931-01A-11D-1924-10	1.3%	0.3%	1.6%	61.9%	33.8%	5.0%	1.6%	50.0%	17.2%	11.9%	10.0%
TCGA-G4-6309-01A-21D-1835-10	5.7%	12.1%	17.8%	68.3%	8.7%	1.2%	0.7%	62.3%	1.3%	6.0%	5.6%
TCGA-CM-4746-01A-01W-1461-10	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
TCGA-A6-6648-01A-11D-1771-10	0.0%	2.2%	2.2%	73.1%	20.4%	6.5%	4.3%	65.6%	2.2%	7.5%	7.5%
TCGA-D5-6929-01A-31D-1924-10	0.7%	2.2%	3.0%	55.2%	35.1%	3.7%	6.0%	45.5%	3.7%	9.7%	21.6%
TCGA-G4-6322-01A-11D-1719-10	2.1%	1.1%	3.2%	67.4%	23.2%	2.1%	4.2%	61.1%	1.1%	6.3%	15.8%
TCGA-D5-6538-01A-11D-1719-10	2.2%	0.7%	3.0%	76.3%	18.5%	5.2%	2.2%	68.9%	4.4%	7.4%	6.7%
TCGA-AZ-5407-01A-01D-1719-10	0.3%	1.9%	2.2%	62.0%	28.8%	4.1%	10.4%	43.7%	5.7%	18.4%	8.5%
TCGA-G4-6323-01A-11D-1719-10	0.0%	1.2%	1.2%	70.7%	25.6%	2.4%	7.3%	62.2%	1.2%	8.5%	14.6%
TCGA-DM-A0XF-01A-11D-A152-10	0.0%	3.0%	3.0%	54.8%	36.3%	8.1%	7.4%	48.9%	2.2%	5.9%	18.5%
TCGA-G4-6307-01A-11D-1719-10	1.4%	2.7%	4.1%	52.7%	39.2%	4.1%	8.1%	47.3%	2.7%	5.4%	24.3%
TCGA-G4-6298-01A-11D-1719-10	0.8%	0.8%	1.6%	65.2%	26.9%	2.7%	7.3%	46.2%	3.8%	19.0%	13.0%
TCGA-D5-6924-01A-11D-1924-10	1.1%	0.0%	1.1%	56.3%	36.2%	8.6%	8.6%	51.7%	5.2%	4.6%	13.8%
TCGA-F4-6463-01A-11D-1719-10	1.2%	1.6%	2.8%	64.2%	25.2%	3.7%	10.2%	45.1%	3.3%	19.1%	8.1%
TCGA-DM-A285-01A-11D-A16V-10	0.8%	2.5%	3.3%	53.7%	38.8%	9.1%	8.3%	46.3%	3.3%	7.4%	18.2%
TCGA-D5-5537-01A-21D-1924-10	3.1%	0.8%	3.9%	69.0%	22.5%	1.6%	3.1%	63.6%	2.3%	5.4%	15.5%
TCGA-AY-6386-01A-21D-1719-10	0.7%	0.4%	1.1%	66.3%	27.3%	3.1%	9.4%	43.1%	5.3%	23.2%	9.4%
TCGA-CM-6168-01A-11D-1650-10	0.5%	1.9%	2.4%	69.9%	26.2%	1.0%	7.3%	59.2%	0.5%	10.7%	17.5%
TCGA-DM-A1D9-01A-11D-A152-10	2.1%	2.1%	4.2%	51.4%	38.2%	11.8%	10.4%	47.2%	2.8%	4.2%	13.2%
TCGA-CM-5344-01A-21D-1719-10	1.6%	1.6%	3.1%	64.1%	28.1%	3.1%	6.3%	60.9%	1.6%	3.1%	17.2%

sample_ID	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-CM-6172-01A-11D-1650-10	0.0%	0.0%	0.0%	80.9%	19.1%	2.1%	3.2%	78.7%	5.3%	2.1%	8.5%
TCGA-A6-6782-01A-11D-1835-10	0.7%	0.7%	1.4%	59.7%	34.7%	7.6%	6.9%	53.5%	4.2%	6.3%	16.0%
TCGA-G4-6303-01A-11D-1771-10	2.9%	1.9%	4.8%	68.6%	23.8%	1.9%	5.7%	65.7%	4.8%	2.9%	11.4%
TCGA-A6-6137-01A-11D-1771-10	2.5%	0.8%	3.4%	72.0%	21.2%	3.4%	5.1%	69.5%	4.2%	2.5%	8.5%
TCGA-F4-6855-01A-11D-1924-10	1.3%	0.0%	1.3%	64.9%	27.3%	2.6%	4.5%	57.1%	1.9%	7.8%	18.2%
TCGA-G4-6315-01A-11D-1719-10	1.2%	1.8%	3.0%	63.9%	29.5%	4.8%	6.0%	58.4%	3.0%	5.4%	15.7%

Figure 17. Low Mutation Group Mutation Type Category Proportions.

Columns are as in figures 14 and 15, but depict the data as percentages.

Figure 18 contains both the mutation type proportion table and the mutation category proportion table for the high mutation group. The C-T and G_A mutations seemed similar in the high group to what was found in the low group. The indels seemed to be somewhat raised in general (although some of these tumors did not experience this). The A_G and T_C mutations seemed to be a bit more common as well in this group. The other three types were a bit less common than in the low mutation group. This was relative to the total number of mutations, so the actual values were higher on average by quite a lot than in the low mutation group, but the ratios did seem to be a bit different.

sample_ID	ALL	A_C	A_T	A_G	C_A	C_T	C_G	T_A	T_C	T_G	G_A	G_C	G_T	-_T	T_-	-_C	C_-	-_A	A_-	-_G	G_-	MNC
TCGA-AA-3713-01A-21D-1719-10	100.0%	1.0%	1.3%	5.2%	3.7%	25.3%	1.6%	0.9%	4.7%	1.5%	24.6%	0.9%	5.5%	1.4%	2.4%	1.1%	5.2%	1.6%	3.1%	1.4%	4.3%	3.5%
TCGA-AD-6889-01A-11D-1924-10	100.0%	0.8%	1.2%	5.6%	6.0%	23.4%	0.4%	0.5%	5.4%	0.6%	26.1%	0.5%	5.2%	1.5%	3.0%	1.2%	3.9%	1.4%	4.1%	1.4%	4.6%	3.2%
TCGA-CM-6162-01A-11D-1650-10	100.0%	0.3%	0.3%	2.1%	2.5%	36.5%	0.0%	0.3%	2.7%	0.3%	40.8%	0.5%	2.3%	0.8%	1.4%	0.6%	1.7%	1.2%	1.5%	0.8%	2.0%	1.5%
TCGA-A6-6141-01A-11D-1771-10	100.0%	4.9%	0.5%	2.6%	18.7%	24.5%	0.2%	0.1%	2.5%	4.9%	25.0%	0.3%	15.4%	0.1%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.3%
TCGA-AZ-4615-01A-01D-1408-10	100.0%	1.4%	0.9%	8.5%	4.6%	23.5%	1.5%	0.8%	7.0%	1.1%	24.0%	1.9%	5.6%	0.6%	2.7%	0.9%	2.2%	1.0%	5.1%	0.4%	2.5%	3.7%
TCGA-A6-6781-01A-22D-1924-10	100.0%	0.2%	0.5%	2.6%	2.9%	35.3%	0.2%	0.2%	1.7%	0.3%	33.6%	0.5%	3.0%	1.5%	2.5%	0.7%	3.7%	1.1%	3.3%	0.5%	3.5%	2.4%
TCGA-G4-6302-01A-11D-1719-10	100.0%	1.5%	0.7%	7.4%	7.7%	28.3%	1.7%	1.1%	7.9%	1.4%	30.6%	2.0%	6.2%	0.0%	0.2%	0.0%	0.3%	0.0%	1.2%	0.1%	0.2%	1.6%
TCGA-AD-6964-01A-11D-1924-10	100.0%	1.0%	2.6%	7.9%	4.4%	28.0%	0.0%	2.5%	7.4%	0.2%	27.7%	0.3%	4.5%	1.2%	1.8%	0.9%	1.9%	1.2%	2.2%	0.2%	2.7%	1.5%
TCGA-F4-6570-01A-11D-1771-10	100.0%	1.2%	1.5%	4.6%	5.9%	27.3%	0.6%	1.2%	3.2%	0.7%	28.0%	0.5%	5.4%	1.2%	2.5%	1.0%	3.7%	1.1%	3.5%	0.4%	3.5%	2.8%
TCGA-G4-6586-01A-11D-1771-10	100.0%	1.3%	0.8%	5.9%	4.1%	27.8%	0.5%	0.5%	5.9%	1.2%	24.9%	0.7%	4.6%	0.7%	3.8%	0.8%	3.5%	0.9%	3.9%	1.2%	4.9%	2.4%
TCGA-CA-6717-01A-11D-1835-10	100.0%	8.5%	0.7%	5.8%	15.6%	20.0%	0.1%	0.9%	5.6%	8.3%	19.3%	0.2%	14.7%	0.1%	0.1%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%
TCGA-D5-6928-01A-11D-1924-10	100.0%	0.8%	0.8%	3.2%	6.0%	31.2%	0.2%	1.3%	3.8%	0.8%	33.7%	0.4%	7.6%	0.5%	1.3%	0.6%	2.0%	0.7%	1.4%	0.3%	1.5%	1.8%
TCGA-AD-5900-01A-11D-1650-10	100.0%	0.8%	0.8%	4.5%	5.9%	27.2%	0.1%	1.0%	4.7%	0.8%	28.4%	0.5%	4.6%	1.5%	2.1%	0.9%	3.7%	1.1%	2.8%	1.1%	4.3%	3.4%
TCGA-AA-3663-01A-01D-1719-10	100.0%	2.0%	0.7%	9.7%	5.8%	20.0%	1.2%	1.1%	10.7%	2.1%	19.9%	1.4%	6.6%	0.6%	3.1%	0.7%	3.8%	1.0%	2.6%	0.8%	3.3%	3.0%
TCGA-AZ-4315-01A-01D-1408-10	100.0%	5.4%	0.3%	5.3%	11.1%	27.8%	0.2%	0.4%	5.8%	5.3%	27.0%	0.2%	10.7%	0.1%	0.1%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.0%
TCGA-A6-6653-01A-11D-1771-10	100.0%	0.6%	0.3%	4.8%	4.2%	28.7%	0.5%	0.9%	4.2%	0.9%	30.1%	0.4%	3.4%	1.0%	3.3%	1.3%	4.0%	1.3%	2.4%	0.8%	3.5%	3.3%
TCGA-AZ-6598-01A-11D-1771-10	100.0%	0.6%	0.9%	4.6%	4.5%	26.0%	0.3%	0.8%	4.1%	0.8%	26.2%	0.5%	4.4%	1.1%	3.6%	1.1%	4.7%	1.3%	3.9%	1.4%	5.5%	3.9%
TCGA-A6-5665-01A-01D-1650-10	100.0%	1.4%	1.0%	8.1%	4.4%	22.3%	0.5%	1.1%	7.6%	0.9%	22.8%	0.4%	4.8%	1.2%	2.3%	2.6%	3.5%	2.1%	3.4%	2.1%	3.6%	3.9%
TCGA-CM-6171-01A-11D-1650-10	100.0%	0.9%	0.5%	4.9%	4.7%	28.3%	0.6%	0.7%	4.2%	0.6%	27.2%	0.7%	4.3%	1.2%	2.9%	1.6%	4.0%	0.7%	3.2%	1.1%	3.7%	3.9%
TCGA-AU-6004-01A-11D-1719-10	100.0%	1.0%	0.5%	6.5%	3.6%	24.8%	1.5%	0.5%	6.7%	1.2%	24.5%	1.3%	4.2%	0.9%	2.6%	1.3%	4.0%	1.3%	4.3%	0.8%	4.2%	4.0%
TCGA-CK-5916-01A-11D-1650-10	100.0%	0.7%	0.7%	5.4%	4.2%	26.1%	0.4%	0.7%	5.9%	0.5%	28.1%	0.7%	3.9%	1.4%	2.7%	0.9%	4.2%	1.0%	3.3%	1.1%	4.7%	3.4%
TCGA-AM-5820-01A-01D-1650-10	100.0%	2.0%	1.8%	11.8%	3.9%	24.1%	3.9%	1.4%	11.1%	2.6%	25.6%	3.9%	3.7%	0.2%	0.4%	0.2%	0.4%	0.2%	0.5%	0.1%	0.5%	1.4%
TCGA-D5-6540-01A-11D-1719-10	100.0%	1.5%	1.2%	2.5%	6.5%	26.9%	0.8%	0.7%	2.5%	1.1%	27.4%	0.2%	6.2%	1.2%	2.4%	1.1%	5.3%	1.2%	3.7%	1.0%	4.3%	2.2%
TCGA-G4-6320-01A-11D-1719-10	100.0%	0.5%	0.7%	2.3%	6.0%	34.6%	0.3%	0.8%	2.6%	0.3%	32.3%	0.7%	6.0%	0.5%	1.3%	0.8%	3.5%	0.6%	1.9%	0.5%	2.5%	1.6%
TCGA-CM-4743-01A-01D-1719-10	100.0%	1.2%	0.8%	6.9%	5.3%	22.1%	0.4%	0.7%	4.8%	0.8%	26.5%	0.9%	6.0%	1.7%	2.1%	1.0%	4.6%	0.8%	3.8%	1.2%	5.2%	3.1%
TCGA-D5-6930-01A-11D-1924-10	100.0%	0.5%	0.6%	1.8%	4.6%	32.5%	0.5%	1.1%	2.8%	0.6%	33.3%	0.2%	4.3%	1.1%	2.6%	0.6%	3.8%	0.8%	2.8%	0.5%	2.9%	2.1%
TCGA-CM-5861-01A-01D-1650-10	100.0%	1.4%	0.9%	11.0%	6.3%	22.6%	0.8%	1.1%	9.2%	1.8%	22.0%	0.2%	5.0%	1.1%	2.1%	1.2%	3.3%	1.6%	2.1%	1.0%	3.7%	1.9%
TCGA-AD-6895-01A-11D-1924-10	100.0%	0.6%	4.3%	3.8%	3.0%	32.0%	1.0%	5.0%	4.1%	0.6%	29.1%	0.8%	3.3%	0.8%	1.7%	0.8%	1.9%	1.1%	2.0%	0.3%	1.4%	2.3%
TCGA-F4-6856-01A-11D-1924-10	100.0%	0.5%	1.4%	4.3%	3.2%	28.8%	0.3%	0.9%	3.7%	1.1%	30.9%	0.5%	2.9%	1.4%	2.6%	1.1%	3.5%	1.4%	2.6%	0.9%	3.8%	3.9%
TCGA-AZ-6601-01A-11D-1771-10	100.0%	0.4%	0.5%	2.6%	6.1%	38.8%	0.3%	0.6%	3.0%	0.7%	39.3%	0.2%	4.3%	0.3%	0.4%	0.4%	0.4%	0.1%	0.4%	0.2%	0.7%	0.4%
TCGA-G4-6588-01A-11D-1771-10	100.0%	1.0%	0.9%	7.0%	5.1%	24.4%	0.6%	1.1%	7.4%	0.7%	25.6%	0.4%	5.2%	0.9%	2.9%	1.6%	4.6%	1.1%	2.4%	0.6%	4.1%	2.3%
TCGA-A6-6780-01A-11D-1835-10	100.0%	1.5%	1.4%	6.8%	5.6%	20.4%	0.9%	0.3%	5.7%	1.7%	23.5%	0.6%	4.4%	1.6%	3.7%	1.4%	4.9%	1.6%	5.6%	1.1%	3.3%	3.9%
TCGA-CA-6718-01A-11D-1835-10	100.0%	8.1%	0.4%	4.0%	18.6%	19.2%	0.1%	0.6%	3.5%	8.5%	19.2%	0.2%	17.0%	0.1%	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%	0.0%	0.1%
TCGA-AA-3492-01A-01D-1408-10	100.0%	0.6%	1.5%	5.3%	4.4%	27.8%	0.6%	1.3%	5.0%	1.1%	27.9%	0.6%	4.5%	0.9%	2.2%	0.8%	2.7%	1.0%	4.4%	0.8%	3.6%	3.0%
TCGA-A6-5661-01A-01D-1650-10	100.0%	0.5%	0.9%	3.4%	6.7%	28.8%	0.4%	1.0%	3.9%	0.4%	28.9%	0.6%	4.7%	1.8%	2.0%	1.1%	2.7%	2.7%	3.1%	1.0%	2.5%	2.8%
TCGA-AM-5821-01A-01D-1650-10	100.0%	2.0%	1.9%	12.3%	3.4%	26.4%	3.5%	1.5%	12.1%	2.0%	25.6%	4.1%	3.4%	0.0%	0.0%	0.0%	0.1%	0.1%	0.1%	0.0%	0.1%	1.5%
TCGA-AY-6197-01A-11D-1719-10	100.0%	0.9%	1.0%	10.0%	4.9%	22.3%	1.3%	1.3%	9.9%	1.5%	20.9%	1.4%	5.3%	1.1%	2.1%	0.4%	3.4%	1.0%	4.0%	1.1%	2.7%	3.2%
TCGA-AA-3510-01A-01D-1408-10	100.0%	4.6%	0.2%	2.5%	15.5%	27.0%	0.0%	0.6%	2.4%	4.3%	27.1%	0.1%	15.4%	0.0%	0.0%	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%	0.1%
TCGA-G4-6628-01A-11D-1835-10	100.0%	0.4%	0.8%	3.1%	5.8%	27.6%	0.8%	0.8%	3.5%	0.5%	30.8%	0.6%	5.1%	2.7%	3.1%	0.7%	3.2%	2.3%	3.3%	0.7%	2.5%	2.1%

sample_ID	INS	DEL	INDEL	TRANSITION	TRANSVERSION	A-T|T-A	C-G|G-C	C-T|G-A	A-C|T-G	A-G|T-C	C-A|G-T
TCGA-AA-3713-01A-21D-1719-10	5.5%	15.0%	20.5%	59.7%	16.3%	2.1%	2.5%	49.8%	2.5%	9.9%	9.1%
TCGA-AD-6889-01A-11D-1924-10	5.5%	15.6%	21.1%	60.6%	15.1%	1.6%	0.9%	49.5%	1.4%	11.1%	11.3%
TCGA-CM-6162-01A-11D-1650-10	3.3%	6.6%	9.9%	82.1%	6.5%	0.5%	0.5%	77.3%	0.7%	4.8%	4.8%
TCGA-A6-6141-01A-11D-1771-10	0.1%	0.1%	0.2%	54.6%	45.0%	0.6%	0.5%	49.5%	9.7%	5.0%	34.2%
TCGA-AZ-4615-01A-01D-1408-10	2.9%	12.6%	15.5%	63.0%	17.8%	1.7%	3.4%	47.5%	2.5%	15.5%	10.2%
TCGA-A6-6781-01A-22D-1924-10	3.8%	13.0%	16.8%	73.2%	7.6%	0.7%	0.7%	68.9%	0.4%	4.3%	5.9%
TCGA-G4-6302-01A-11D-1719-10	0.1%	1.9%	2.0%	74.0%	22.3%	1.7%	3.7%	58.8%	3.0%	15.2%	13.9%
TCGA-AD-6964-01A-11D-1924-10	3.6%	8.5%	12.1%	70.9%	15.5%	5.1%	0.3%	55.6%	1.2%	15.3%	8.9%
TCGA-F4-6570-01A-11D-1771-10	3.7%	13.2%	17.0%	63.1%	17.2%	2.8%	1.1%	55.3%	1.9%	7.8%	11.4%
TCGA-G4-6586-01A-11D-1771-10	3.5%	16.1%	19.6%	64.4%	13.5%	1.3%	1.2%	52.7%	2.4%	11.7%	8.7%
TCGA-CA-6717-01A-11D-1835-10	0.2%	0.1%	0.3%	50.6%	49.0%	1.5%	0.4%	39.2%	16.8%	11.4%	30.3%
TCGA-D5-6928-01A-11D-1924-10	2.1%	6.3%	8.3%	71.8%	18.0%	2.1%	0.6%	64.9%	1.6%	7.0%	13.6%
TCGA-AD-5900-01A-11D-1650-10	4.5%	13.0%	17.5%	64.7%	14.5%	1.8%	0.7%	55.6%	1.6%	9.1%	10.5%
TCGA-AA-3663-01A-01D-1719-10	3.1%	12.8%	15.9%	60.2%	20.9%	1.8%	2.6%	39.8%	4.0%	20.4%	12.4%
TCGA-AZ-4315-01A-01D-1408-10	0.4%	0.1%	0.5%	65.8%	33.6%	0.7%	0.4%	54.7%	10.7%	11.1%	21.9%
TCGA-A6-6653-01A-11D-1771-10	4.4%	13.2%	17.7%	67.9%	11.1%	1.1%	1.0%	58.9%	1.5%	9.0%	7.6%
TCGA-AZ-6598-01A-11D-1771-10	4.8%	17.7%	22.5%	60.9%	12.6%	1.7%	0.7%	52.2%	1.3%	8.7%	8.8%
TCGA-A6-5665-01A-01D-1650-10	8.1%	12.8%	20.8%	60.7%	14.6%	2.2%	0.9%	45.1%	2.3%	15.6%	9.2%
TCGA-CM-6171-01A-11D-1650-10	4.6%	13.8%	18.4%	64.6%	13.1%	1.2%	1.3%	55.5%	1.5%	9.1%	9.0%
TCGA-AU-6004-01A-11D-1719-10	4.4%	15.0%	19.4%	62.6%	14.0%	1.1%	2.8%	49.4%	2.2%	13.2%	7.9%
TCGA-CK-5916-01A-11D-1650-10	4.3%	14.9%	19.3%	65.5%	11.9%	1.4%	1.1%	54.2%	1.2%	11.3%	8.1%
TCGA-AM-5820-01A-01D-1650-10	0.7%	1.8%	2.5%	72.7%	23.3%	3.3%	7.8%	49.7%	4.6%	23.0%	7.6%
TCGA-D5-6540-01A-11D-1719-10	4.6%	15.7%	20.2%	59.3%	18.2%	2.0%	1.0%	54.2%	2.5%	5.1%	12.7%
TCGA-G4-6320-01A-11D-1719-10	2.4%	9.1%	11.5%	71.7%	15.3%	1.5%	1.0%	66.8%	0.8%	4.8%	12.0%
TCGA-CM-4743-01A-01D-1719-10	4.7%	15.8%	20.5%	60.2%	16.1%	1.5%	1.3%	48.6%	2.0%	11.7%	11.3%
TCGA-D5-6930-01A-11D-1924-10	2.9%	12.1%	15.0%	70.4%	12.5%	1.7%	0.8%	65.8%	1.1%	4.6%	8.9%
TCGA-CM-5861-01A-01D-1650-10	4.9%	11.1%	16.0%	64.7%	17.4%	2.0%	1.0%	44.6%	3.1%	20.2%	11.3%
TCGA-AD-6895-01A-11D-1924-10	3.0%	7.1%	10.0%	69.0%	18.6%	9.3%	1.8%	61.1%	1.2%	7.9%	6.3%
TCGA-F4-6856-01A-11D-1924-10	4.9%	12.5%	17.5%	67.8%	10.8%	2.3%	0.8%	59.8%	1.6%	8.0%	6.1%
TCGA-AZ-6601-01A-11D-1771-10	0.9%	1.9%	2.8%	83.6%	13.1%	1.1%	0.4%	78.0%	1.2%	5.6%	10.5%
TCGA-G4-6588-01A-11D-1771-10	4.3%	14.0%	18.3%	64.4%	14.9%	2.0%	1.0%	50.0%	1.7%	14.4%	10.3%
TCGA-A6-6780-01A-11D-1835-10	5.7%	17.6%	23.3%	56.5%	16.3%	1.7%	1.4%	44.0%	3.2%	12.5%	10.0%
TCGA-CA-6718-01A-11D-1835-10	0.3%	0.0%	0.4%	45.9%	53.5%	1.0%	0.4%	38.4%	16.5%	7.5%	35.6%
TCGA-AA-3492-01A-01D-1408-10	3.5%	12.9%	16.4%	66.0%	14.6%	2.8%	1.2%	55.7%	1.7%	10.3%	8.9%
TCGA-A6-5661-01A-01D-1650-10	6.6%	10.4%	17.0%	64.9%	15.3%	2.0%	1.0%	57.6%	0.9%	7.3%	11.4%
TCGA-AM-5821-01A-01D-1650-10	0.1%	0.3%	0.4%	76.3%	21.8%	3.4%	7.6%	52.0%	4.0%	24.4%	6.8%
TCGA-AY-6197-01A-11D-1719-10	3.7%	12.2%	15.9%	63.2%	17.7%	2.4%	2.8%	43.2%	2.4%	20.0%	10.2%
TCGA-AA-3510-01A-01D-1408-10	0.2%	0.0%	0.2%	59.0%	40.6%	0.8%	0.1%	54.1%	8.9%	4.9%	30.8%
TCGA-G4-6628-01A-11D-1835-10	6.4%	12.0%	18.5%	64.9%	14.5%	1.6%	1.3%	58.3%	0.8%	6.5%	10.8%

Figure 18. High Mutation Group Mutation Type (and Category) Proportions

Columns here are as shown in tables 9-14. These values are represented as proportions of total mutation counts per sample.

I then performed a t test on both the raw data and the proportions. C_G mutations did not pass requirements for significance of difference base on counts, while its chemical equivalent G_C did, but only barely. The combined category for these, also did not pass, having a p value just slightly higher than 5%. All the rest of the count values were significantly different, as expected.

In terms of proportions, several mutation types counted separately did not pass 5% requirement for significance, including A_C, C_A, G_A, G_T, T_G. In the categories, the A_C and T_G, the C_A and G_T, and the transitions categories did not pass requirements for significance

Raw Counts							
mut_type	statistic	parameter	p.value	conf.int_low	conf.int_high	estimate_low	estimate_high
A_C	-2.800	38.019	0.0080	-85.108	-13.685	2.911	52.308
A_G	-4.734	38.093	0.0000	-181.938	-72.947	8.250	135.692
A_T	-4.515	38.180	0.0001	-27.377	-10.429	2.994	21.897
C_A	-4.320	38.030	0.0001	-207.774	-75.172	10.322	151.795
C_G	-2.000	38.260	0.0526	-33.536	0.197	4.767	21.436
C_T	-8.589	38.236	0.0000	-629.368	-389.313	48.839	558.179
G_A	-8.800	38.227	0.0000	-638.331	-399.614	46.489	565.462
G_C	-2.036	38.227	0.0487	-35.740	-0.105	5.078	23.000
G_T	-4.365	38.040	0.0001	-198.238	-72.615	10.522	145.949
T_A	-4.961	38.212	0.0000	-25.523	-10.732	3.411	21.538
T_C	-4.716	38.101	0.0000	-176.219	-70.382	8.622	131.923
T_G	-2.861	38.019	0.0068	-86.086	-14.742	2.817	53.231
-_A	-9.490	38.661	0.0000	-19.384	-12.571	0.894	16.872
-_C	-8.033	38.458	0.0000	-16.126	-9.637	0.606	13.487
-_G	-7.060	38.744	0.0000	-14.136	-7.839	0.628	11.615
-_T	-9.006	38.833	0.0000	-17.812	-11.278	0.917	15.462
A_-	-8.868	38.457	0.0000	-49.836	-31.317	1.372	41.949
C_-	-8.901	38.363	0.0000	-54.869	-34.541	1.167	45.872
G_-	-8.549	38.278	0.0000	-54.830	-33.838	1.256	45.590
T_-	-8.694	38.394	0.0000	-39.445	-24.549	1.106	33.103
MNC	-7.988	39.250	0.0000	-43.437	-25.886	7.467	42.128

mut_type	statistic	parameter	p.value	conf.int_low	conf.int_high	estimate_low	estimate_high
INS	-9.450	38.716	0.0000	-66.036	-42.746	3.044	57.436
DEL	-9.114	38.379	0.0000	-197.499	-125.727	4.900	166.513
INDEL	-9.460	38.472	0.0000	-262.209	-169.799	7.944	223.949
TRANSITIONS	-7.732	38.173	0.0000	-1613.908	-944.205	112.200	1391.256
TRANSVERSIONS	-4.237	38.052	0.0001	-662.522	-234.141	42.822	491.154
A_C|T_G	-2.833	38.018	0.0073	-171.133	-28.489	5.728	105.538
A_G|T_C	-4.730	38.094	0.0000	-358.052	-143.434	16.872	267.615
A_T|T_A	-4.789	38.156	0.0000	-52.681	-21.380	6.406	43.436
C_A|G_T	-4.345	38.032	0.0001	-405.920	-147.879	20.844	297.744
C_G|G_C	-2.022	38.231	0.0502	-69.222	0.039	9.844	44.436
C_T|G_A	-8.703	38.229	0.0000	-1267.448	-789.178	95.328	1123.641

Proportions							
mut_type	statistic	parameter	p.value	conf.int_low	conf.int_high	estimate_low	estimate_high
A_C	-0.140	46.058	0.8893	-0.007	0.006	0.016	0.016
A_G	-3.327	53.199	0.0016	-0.026	-0.006	0.039	0.055
A_T	5.445	121.327	0.0000	0.006	0.012	0.019	0.010
C_A	0.153	51.781	0.8792	-0.013	0.015	0.064	0.063
C_G	10.126	133.347	0.0000	0.016	0.024	0.027	0.007
C_T	3.158	133.665	0.0020	0.013	0.054	0.302	0.269
G_A	0.858	105.044	0.3927	-0.011	0.029	0.283	0.274
G_C	10.136	133.169	0.0000	0.017	0.025	0.029	0.008
G_T	0.286	49.530	0.7760	-0.011	0.014	0.062	0.060
T_A	7.480	122.404	0.0000	0.010	0.017	0.023	0.010
T_C	-2.883	54.825	0.0056	-0.022	-0.004	0.040	0.054
T_G	-0.345	46.249	0.7318	-0.008	0.006	0.015	0.017
A_-	-4.801	56.413	0.0000	-0.008	-0.003	0.004	0.010
C_-	-5.194	57.249	0.0000	-0.007	-0.003	0.003	0.008
G_-	-4.107	79.849	0.0001	-0.006	-0.002	0.003	0.007
T_-	-4.036	62.829	0.0001	-0.007	-0.002	0.005	0.009
A_-	-6.008	90.533	0.0000	-0.025	-0.013	0.006	0.025
C_-	-8.467	43.149	0.0000	-0.029	-0.018	0.004	0.028
G_-	-8.372	40.787	0.0000	-0.028	-0.017	0.004	0.027
T_-	-7.998	43.690	0.0000	-0.019	-0.012	0.004	0.020
MNC	6.069	205.785	0.0000	0.016	0.031	0.047	0.024

mut_type	statistic	parameter	p.value	conf.int_low	conf.int_high	estimate_low	estimate_high
INS	-5.515	46.332	0.0000	-0.026	-0.012	0.015	0.034
DEL	-8.376	44.936	0.0000	-0.100	-0.061	0.019	0.100
INDEL	-7.917	44.137	0.0000	-0.125	-0.075	0.034	0.134
TRANSITIONS	0.908	70.879	0.3670	-0.015	0.041	0.665	0.652
TRANSVERSIONS	3.502	48.849	0.0010	0.027	0.101	0.255	0.191
A_C|T_G	-0.248	43.500	0.8057	-0.015	0.011	0.031	0.033
A_G|T_C	-3.193	50.506	0.0024	-0.048	-0.011	0.080	0.109
A_T|T_A	7.268	96.036	0.0000	0.016	0.028	0.042	0.020
C_A|G_T	0.223	45.537	0.8247	-0.023	0.028	0.126	0.123
C_G|G_C	11.066	109.381	0.0000	0.033	0.048	0.056	0.015
C_T|G_A	2.480	67.067	0.0156	0.008	0.076	0.585	0.543

Figure 19. Results of Two Sided Welch Two Sample T Test on Mutation Type Data

The first pair of tables is the result of the t test on the raw count values. The second pair of tables is the result of the t test using proportions as input instead of the raw counts. Within each pair the first table contains the results for the individual mutation types and the second table for the categories.

Differences in Mutated Genes

Plotting the mutation counts binned according to gene for the entire population of tumors resulted in a scatterplot that was quite messy. It was apparent however that there was a collection of a small number of genes that mutated significantly more than their similarly sized counterparts. In the first plot, showing all of the points, one can see several genes that mutated to the most extreme levels, and with the slightly more zoomed in scatterplot, there were quite a number of highly mutated genes that deviated from the general trend to a lesser, but still quite obvious, extent. I became interested in the appearance of this kind of plot when produced using the split populations.

Figure 20. Scatterplot of Gene based counts from non-split data

These two plots both depict the total sum of all mutations per gene for the entire set of samples. The first plot has a larger maximum on the x and y axes in order to see the values for the most mutated genes and the largest genes. The second plot is more zoomed in to give a better view that is not possible in the first image.

The one difference that became obvious right away was the number of mutations. The population size of the high mutation group was 39, while the low mutation group was 180. These counts had not been normalized by population size, and yet the plots were covering approximately the same region in terms of count values. Additionally, there were not quite as many obvious outliers on the low-length side of this scatterplot in the higher mutation count group. The dot near the top right corner, which happened to be TTN, seemed to have approximately twice the number of mutations, indicating that it garnered several mutations in several of the tumors.

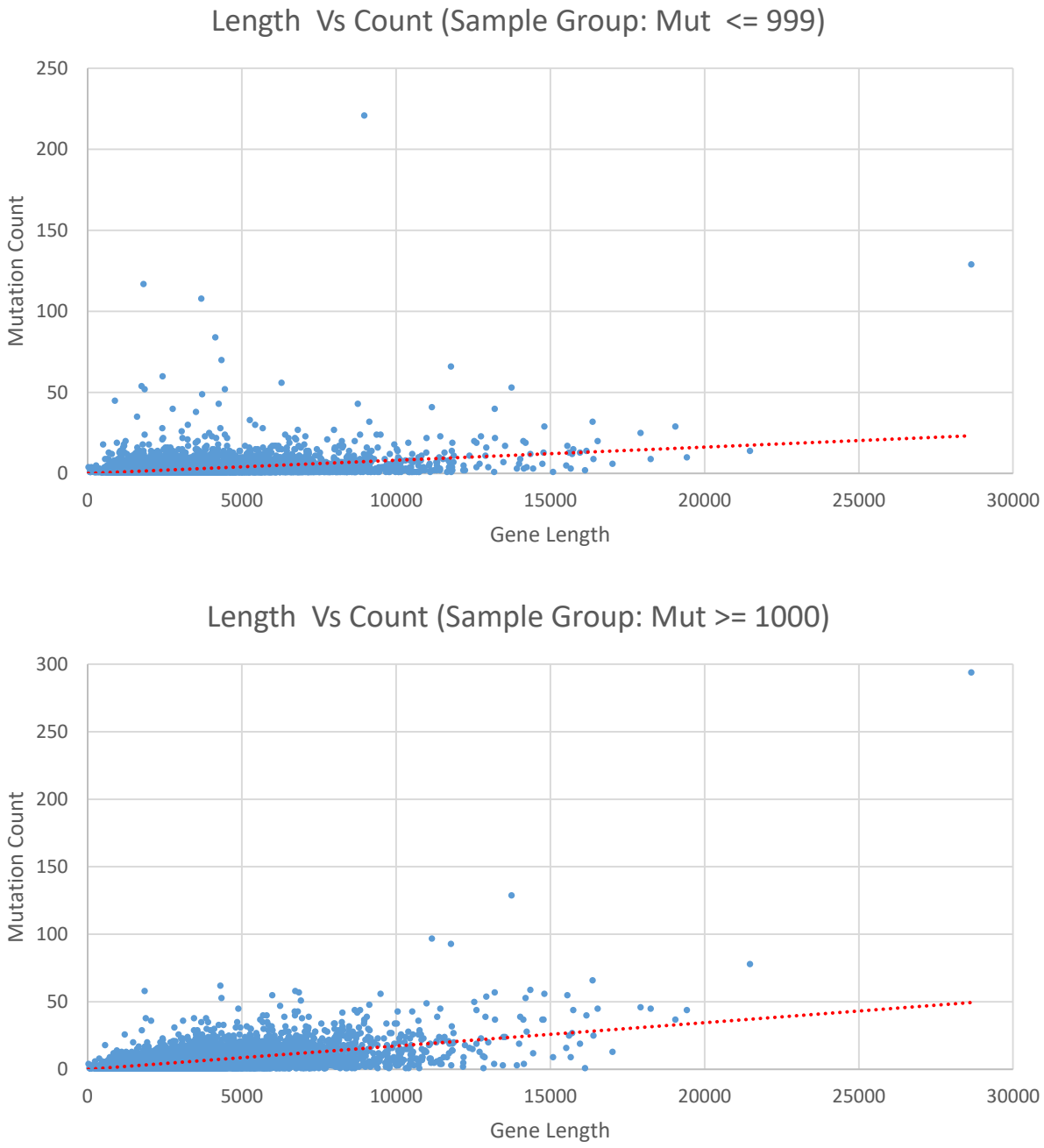

Figure 21. Gene Mutation Counts Derived from split data

These two plots are produced similarly to the first plot in Figure 20. Each is produced from one of the mutation count groups. The first plot is the low mutation group and the second plot is the high mutation group.

Zooming in more closely around the bulk of the points, I observed that the high mutation group again had more high counts trending toward the right side of the plot, and also that the main blob of plotted points was more closely packed in addition to being “taller”. The bulk of the population of genes within the low mutation count group were below a value of 10, and the densely packed region was fairly flat from about length 8000 and below, with a region of slightly less density between 8000 and 10000. The high mutation group, in addition to having a “taller” dense collection of points, also had its dense region peaking at genes approximately of length 6500, with its less dense region spanning 6500 to 8500. This less dense area trailed off in both the positive Y direction and positive X direction. The overall impression was that the mutations in the high population were more associated with gene length than those in the low population, for at least a visibly noticeable sub-population of genes.

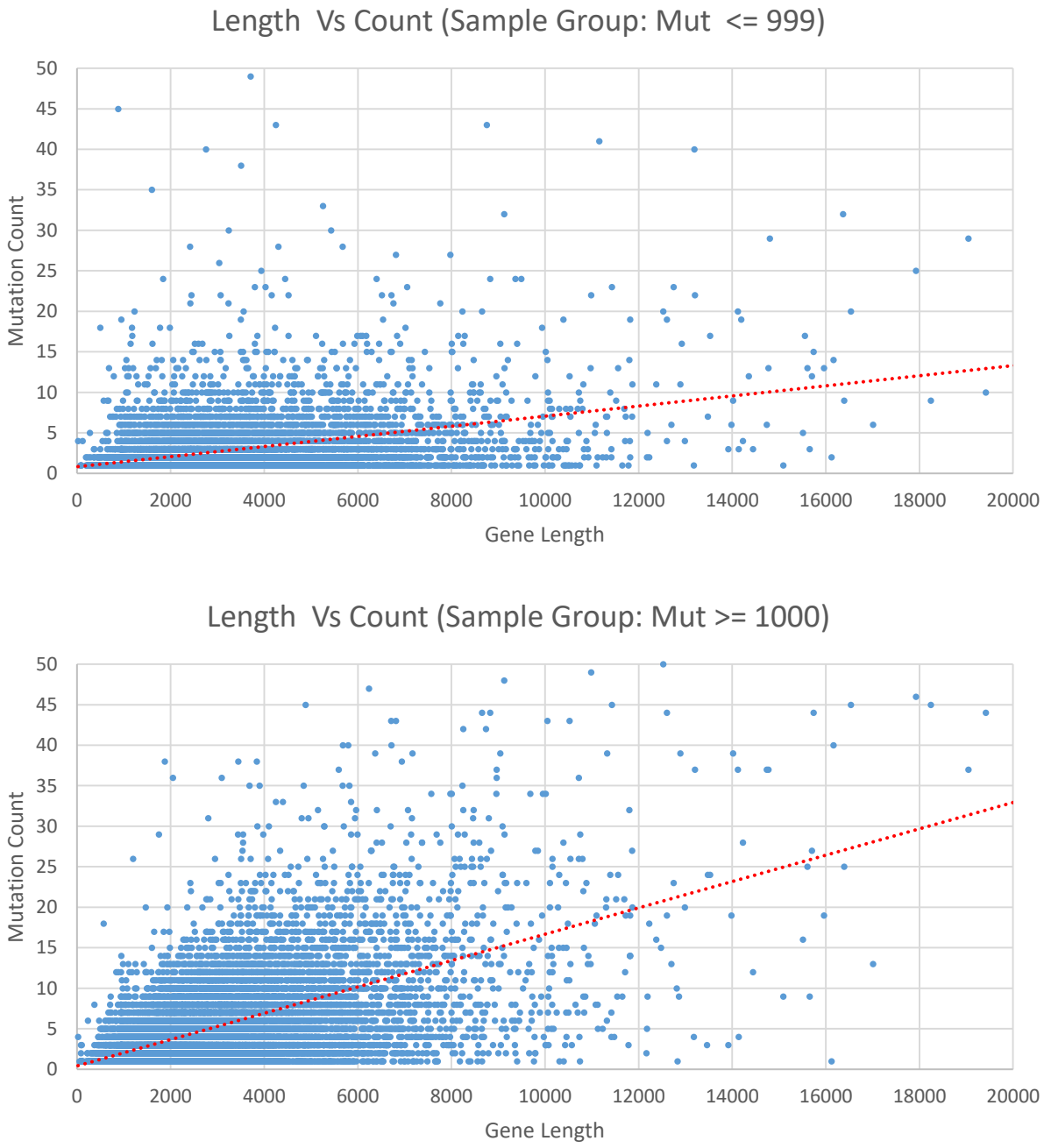

Figure 22. Gene Mutation Counts Derived from split data Zoomed in

These two plots are produced similarly to the second, more zoomed in plot in figure 20. Each is produced from one of the mutation count groups. The first plot is the low mutation group and the second plot is the high mutation group.

The next question was which genes were most significantly mutated. After I returned to the list previously examined, and added in gene lengths, it became apparent that there was a difference between the two lists in terms of gene lengths. The high mutation population count list, when sorted by count value from high to low, showed a stronger trend of the length values following a high to low order than the low mutation count list. There were several genes that bucked this trend, which might be due to mutational hot spots or the effects of clonal selection.

Mutations <= 999 #Samples 180				Mutations >= 1000 #Samples 39		
Gene Symbol	Mutation Count	Gene Length		Gene Symbol	Mutation Count	Gene Length
APC	221	8962		TTN	294	28638
TTN	129	28638		MUC16	129	13734
TP53	117	1809		SYNE1	97	11156
NEFH	108	3681		OBSCN	93	11772
KRAS	84	4132		CCDC168	78	21470
DSPP	70	4331		RYR2	66	16365
OBSCN	66	11772		SSPO	62	4300
ERICH6B	60	2421		MUC17	59	14350
ZNF814	56	6283		DNAH17	58	1846
IRF5	54	1746		NEB	58	6722
MUC16	53	13734		XIRP2	57	6847
TBP	52	1844		FAT4	57	13189
ATXN1	52	4441		CSMD3	56	9489
PIK3CA	49	3709		PLEC	56	14799
KRTAP4-5	45	879		DNAH8	55	5984
MUC4	43	4245		RYR3	55	15551
PRIC285	43	8754		GPR98	54	12918
SYNE1	41	11156		DSPP	53	4331
TMPRSS13	40	2755		DNAH11	53	14188
FAT4	40	13189		DCHS2	51	6903
PHF2	38	3505		RYR1	50	12524
GPRIN2	35	1600		PCLO	49	10983
CACNA1B	33	5251		ZFHX4	48	9124
ZFHX4	32	9124		FAT1	47	6234
RYR2	32	16365		MUC5B	46	17924
ZFPM1	30	3241		LAMA1	45	4883
MAML2	30	5428		LAMA5	45	11426
PLEC	29	14799		LRP1B	45	16531
FAT3	29	19047		AHNAK2	45	18243
KCNN3	28	2414		CSMD1	44	8655
SSPO	28	4300		DNAH5	44	8827
CACNA1H	28	5675		USH2A	44	12599
ABCA13	27	6810		LRP2	44	15735
RP1L1	27	7977		MLL2	44	19419
AR	26	3039		DNAH3	43	6709
SOX9	25	3934		ABCA13	43	6810
MUC5B	25	17924		HMCN1	43	10046
TPRX1	24	1839		PCNT	43	10522
CRIPAK	24	4443		DNAH2	42	8249
CROCC	24	6396		SYNE2	42	8735

Mutations <= 999 #Samples 180			Mutations >= 1000 #Samples 39		
Gene Symbol	Mutation Count	Gene Length	Gene Symbol	Mutation Count	Gene Length
DNAH5	24	8827	TNXB	40	5682
SMAD4	24	9365	ANK3	40	5794
CSMD3	24	9489	NCOR2	40	6715
PCDHA7	23	3796	FREM2	40	16162
COL18A1	23	4024	FRAS1	39	6371
TCHH	23	7051	DST	39	7165
LAMA5	23	11426	WDR87	39	9043
FLG	23	12747	MACF1	39	11323
KRT1	22	2440	CMYA5	39	12892
ANKLE1	22	3069	PKHD1L1	39	14017
IRF2BPL	22	4157	DNAH14	38	1876
KNDC1	22	4515	DNAH10	38	3446
SDK1	22	6510	EYS	38	3837
NCOR2	22	6715	ASPM	38	6940
PCLO	22	10983	ALPK2	37	5589
ZNF469	22	13203	APC	37	8962
OGFR	21	2417	ZNF469	37	13203
FBXW7	21	3235	APOB	37	14121
KIF26A	21	6757	MYCBP2	37	14735
COL6A3	21	7761	ZFHX3	37	14770
TCF15	20	1227	FAT3	37	19047
NUMBL	20	3557	DNAH12	36	2048
EP400	20	8232	PDE4DIP	36	3089
CSMD1	20	8655	FBN3	36	8963
RYR1	20	12524	PKD1	36	10717
APOB	20	14121	NEFH	35	3681
LRP1B	20	16531	LRP1	35	3898
KRTAP4-3	19	942	UBR4	35	4842
MEFV	19	3499	CACNA1H	35	5675
PPM1E	19	6537	TEX15	35	5816
CSMD2	19	10391	EP400	35	8232
PTPRT	19	11817	DNAH9	34	7569
USH2A	19	12599	RP1L1	34	7977
DNAH11	19	14188	MEGF8	34	8004
CRCT1	18	497	SACS	34	8957
PTPLA	18	1173	LAMA2	34	9684
OPRD1	18	1774	AHNAK	34	9952
ZNF837	18	1982	WDFY4	34	10013
ESPNL	18	4226	MUC4	33	4245
DSCAM	18	7013	RNF43	33	4395
TRPS1	18	9932	KIAA1109	33	5852
KRTAP4-1	17	1179	PKD1L1	32	5148

Mutations <= 999 #Samples 180				Mutations >= 1000 #Samples 39		
Gene Symbol	Mutation Count	Gene Length		Gene Symbol	Mutation Count	Gene Length
GRIN3B	17	3254		HERC2	32	5961
ABCA7	17	3852		LRRK2	32	6406
ADAMTS7	17	4518		NBEA	32	7068
WNK2	17	5101		HSPG2	32	8253
DNAH8	17	5984		DNAH1	32	8473
ANK2	17	6054		MYO15A	32	11800
LAMC3	17	6098		COL6A5	31	2802
SCN5A	17	6177		ATM	31	4799
HECW1	17	6839		PCDH15	31	4940
UNC13C	17	8140		POLE	31	5934
GLI3	17	8280		CACNA1A	31	7156
UNC80	17	13526		MLL4	31	8460
RYR3	17	15551		RNF213	31	8794
KRTAP4-8	16	1142		ABCA7	30	3852
NTSR2	16	1608		NLRC5	30	4096
MAP1S	16	2516		CUBN	30	5280
PCDHB8	16	2592		CEP192	30	5283
ADAMTS2	16	2677		FMN2	30	5696
POLRMT	16	3800		VPS13B	30	6697
ATP10A	16	5234		MUC6	30	8003
TNXB	16	5682		FLNC	30	9088
FAM123B	16	5925				
NOTCH3	16	6197				
ODZ4	16	6455				
DCHS2	16	6903				
PCDH17	16	8007				
HSPG2	16	8253				
WDR87	16	9043				
ZNF831	16	9403				
GPR98	16	12918				

Figure 23. Gene Counts with Lengths from Split Populations.

This is the same list as from Figure 5, but with the calculated gene sizes added.

I decided to try some methods to reduce the effect of gene length, and after speaking to Yu-Bo Wang and Dr. James Grady, both experts in statistics, I went with using a linear regression to obtain studentized residuals. The values of these residuals should reflect the degree to which the count was an outlier from the trend line for that population, with negative values indicating a negative deviation from the trend and positive values indicating a positive deviation from the trend. I compiled a list of these computed residual values and sorted the list from high to low. The 100 highest values from each population are listed in Figure 24.

A more advanced statistical model than a linear regression might be desirable for future analyses, as there were multiple overlapping effects other than just gene length. Some genes would exhibit negative selection for damaging mutations, due to cell viability needs and either show low mutation counts or zero counts. Other genes would have positive selection in tumors from any deactivating mutations, due to not requiring subtle mutations in order to contribute to the tumor phenotype.

Another set of genes would exhibit positive selection on very specific subtle mutations, but negative selection or no selection at all (depending on the tumor's existing genetic background) for any deactivating mutations, making some mutations within this class of genes passenger mutations despite the gene being a cancer-related gene.

If a tumor was already carrying a subtle mutation in one of these genes, then a destructive change to the function of that gene would be unlikely to persist within the tumor unless it occurred within the second copy of the gene on the other chromosome for a gene where losing the remaining normal copy would not be detrimental to cell viability. The chances of such a double mutation within the same gene on the same chromosome

were also remote for most genes, however this effect might also impact other required genes downstream in the functional pathway from the subtly affected gene.

In addition to these effects from subtle mutations and disabling damaging mutations in cancer related genes, there were also a very large number of mutations that had little to no effect on the genes they occur in, due to being a redundant codon swap, or causing a functionally synonymous amino acid change. These kinds of mutations would inherently be passenger mutations, and it is important to note, they could also occur within cancer related genes. There would also be a collection of mutations that were destructive in terms of gene function, but which occurred in genes that were not particularly important for tumor cell survival or competition. These mutations, too, would be passenger mutations from a cancer genomics perspective. Realizing the possibility of cancer genes experiencing mutations that ought to be classified as passenger mutations rather than driver mutations, I became interested in looking at distribution of mutations within single-genes.

These competing selective effects, in addition to other potential confounding influences, made the data fairly noisy as a result. There was clearly a positive correlation of gene length with mutation counts, but the scatterplot did not pack very tightly around the trend line in either population.

Mutations <= 999 #Samples 180			
Gene Symbol	Length	Mutation Count	Studentized Residual
APC	8962	221	57.453
TP53	1809	117	27.963
TTN	28638	129	26.967
NEFH	3681	108	25.342
KRAS	4132	84	19.248
DSPP	4331	70	15.792
OBSCN	11772	66	13.718
ERICH6B	2421	60	13.660
IRF5	1746	54	12.319
ZNF814	6283	56	12.124
TBP	1844	52	11.825
ATXN1	4441	52	11.439
PIK3CA	3709	49	10.831
MUC16	13734	53	10.314
KRTAP4-5	879	45	10.297
MUC4	4245	43	9.323
TMPRSS13	2755	40	8.830
PRIC285	8754	43	8.661
PHF2	3505	38	8.245
SYNE1	11156	41	7.836
GPRIN2	1600	35	7.815
FAT4	13189	40	7.303
CACNA1B	5251	33	6.805
ZFPM1	3241	30	6.391
MAML2	5428	30	6.071
KCNN3	2414	28	6.040
ZFHX4	9124	32	6.003
SSPO	4300	28	5.764
CACNA1H	5675	28	5.563
AR	3039	26	5.477
FAT3	8758	29	5.348
TPRX1	1839	24	5.181
ABCA13	6810	27	5.161
SOX9	3934	25	5.110
RP1L1	7977	27	4.991
RYR2	16365	32	4.950
CRIPAK	4443	24	4.800
PCDHA7	3796	23	4.659
COL18A1	4024	23	4.625
KRT1	2440	22	4.622
ANKLE1	3069	22	4.530
CROCC	6396	24	4.514
PLEC	14799	29	4.469
OGFR	2417	21	4.389

Mutations >= 1000 #Samples 39			
Gene Symbol	Length	Mutation Count	Studentized Residual
TTN	28638	294	56.067
MUC16	13734	129	22.318
SYNE1	11156	97	16.358
OBSCN	11772	93	15.300
SSPO	4300	62	11.332
DNAH17	1846	58	11.331
NEB	6722	58	9.674
DSPP	4331	53	9.442
XIRP2	6847	57	9.423
DNAH8	5984	55	9.299
CCDC168	21470	78	8.864
CSMD3	9489	56	8.321
DCHS2	6903	51	8.155
RYR2	16365	66	8.082
LAMA1	4883	45	7.590
FAT1	6234	47	7.549
MUC17	14350	59	7.305
FAT4	13189	57	7.280
DNAH14	1876	38	7.153
ZFHX4	9124	48	6.781
GPR98	12918	54	6.748
DNAH12	2048	36	6.679
DNAH10	3446	38	6.622
DNAH3	6709	43	6.558
PLEC	14799	56	6.530
ABCA13	6810	43	6.524
EYS	3837	38	6.490
PCLO	10983	49	6.362
PDE4DIP	3089	36	6.327
TNXB	5682	40	6.282
ANK3	5794	40	6.244
DNAH11	14188	53	6.112
CSMD1	8655	44	6.109
RYR3	15551	55	6.068
DNAH5	8827	44	6.051
RYR1	12524	50	6.050
NCOR2	6715	40	5.933
NEFH	3681	35	5.920
LRP1	3898	35	5.847
FRAS1	6371	39	5.842
DNAH2	8249	42	5.831
ALPK2	5589	37	5.691
SYNE2	8735	42	5.667
DST	7165	39	5.574

Mutations <= 999 #Samples 180				Mutations >= 1000 #Samples 39			
Gene Symbol	Length	Mutation Count	Studentized Residual	Gene Symbol	Length	Mutation Count	Studentized Residual
IRF2BPL	4157	22	4.370	UBR4	4842	35	5.528
TCF15	1227	20	4.328	ASPM	6940	38	5.443
KNDC1	4515	22	4.318	HMCN1	10046	43	5.432
FBXW7	3235	21	4.270	COL6A5	2802	31	5.387
TCHH	7051	23	4.183	LAMA5	11426	45	5.382
DNAH5	8827	24	4.160	DNAH6	1747	29	5.329
KRTAP4-3	942	19	4.134	MUC4	4245	33	5.315
SMAD4	9365	24	4.081	PCNT	10522	43	5.272
CSMD3	9489	24	4.063	RNF43	4395	33	5.264
SDK1	6510	22	4.027	CACNA1H	5675	35	5.247
NCOR2	6715	22	3.997	TEX15	5816	35	5.200
NUMBL	3557	20	3.987	WDR87	9043	39	4.941
CRCT1	497	18	3.964	HLA-A	1198	26	4.892
PTPLA	1173	18	3.865	ABCA7	3852	30	4.826
OPRD1	1774	18	3.777	PKD1L1	5148	32	4.803
MEFV	3499	19	3.760	USH2A	12599	44	4.780
KIF26A	6757	21	3.755	KIAA1109	5852	33	4.773
ZNF837	1982	18	3.747	DNHD1	3440	29	4.758
KRTAP4-11	1179	17	3.629	NLRC5	4096	30	4.744
COL6A3	7761	21	3.609	LOXHD1	3548	29	4.721
LAMA5	11426	23	3.546	ATM	4799	31	4.714
ESPNL	4226	18	3.419	PCDH15	4940	31	4.666
KRTAP4-8	1142	16	3.399	FAT3	8758	37	4.623
PCLO	10983	22	3.375	IGSF10	3980	29	4.576
FLG	12747	23	3.353	APC	8962	37	4.554
NTSR2	1608	16	3.331	HERC2	5961	32	4.529
GRIN3B	3254	17	3.326	CDH23	3548	28	4.514
PPM1E	6537	19	3.317	DNAH9	7569	34	4.402
EP400	8232	20	3.305	EP400	8232	35	4.385
CSMD1	8655	20	3.243	LRRK2	6406	32	4.379
ABCA7	3852	17	3.238	FBN3	8963	36	4.346
MAP1S	2516	16	3.198	CUBN	5280	30	4.344
PCDHB8	2592	16	3.187	CEP192	5283	30	4.343
ADAMTS2	2677	16	3.175	POLE	5934	31	4.331
ADAMTS7	4518	17	3.141	DOCK2	3533	27	4.312
MUC5B	17924	25	3.072	BRAF	2946	26	4.303
WNK2	5101	17	3.056	RP1L1	7977	34	4.264
ZNF469	13203	22	3.052	MEGF8	8004	34	4.255
DSCAM	7013	18	3.012	FMN2	5696	30	4.204
POLRMT	3800	16	3.011	MACF1	11323	39	4.173
ZNF707	2458	15	2.971	NBEA	7068	32	4.156
B3GNT6	2492	15	2.966	PHF2	3505	26	4.114
KRTAP4-6	1055	14	2.941	MYH9	3709	26	4.046

Mutations <= 999 #Samples 180				Mutations >= 1000 #Samples 39			
Gene Symbol	Length	Mutation Count	Studentized Residual	Gene Symbol	Length	Mutation Count	Studentized Residual
DNAH8	5984	17	2.927	PLXNB2	4336	27	4.041
ANK2	6054	17	2.917	ITPR3	5853	29	3.944
LAMC3	6098	17	2.910	SACS	8957	34	3.934
SP8	2938	15	2.901	PREX2	4078	26	3.921
SCN5A	6177	17	2.899	CACNA1A	7156	31	3.919
SCN9A	3061	15	2.883	MYH13	5992	29	3.897
UBXN11	1700	14	2.847	VPS13B	6697	30	3.867
ZNF787	1807	14	2.831	ERICH6B	2421	23	3.859
SCARF2	3500	15	2.819	HSPG2	8253	32	3.757
LOXHD1	3548	15	2.812	PKD1	10717	36	3.756
HECW1	6839	17	2.802	LRP2	15735	44	3.724
ATP10A	5234	16	2.801	LAMA2	9684	34	3.689
SPHK1	2054	14	2.795	DNAH1	8473	32	3.683

Figure 24. 100 Genes with highest studentized residuals from each population

This list was produced by sorting the resulting table of genes with studentized residuals produced from a linear regression. The lengths were a calculated value as described in methods, and the counts were produced by the same counting program used previously.

Discussion

In terms of the mutation types, there did appear to be some significant differences. Some specific types of mutations did not pass significance tests such as the proportions of C_A|G_T and A_C|T_G, and A_C, C_A, G_A, G_T, T_G proportions individually, and the sum of all transitions as well as the non-proportional counts of the C_G|G_C mutations and the C_G mutations individually (while G_C with a p value of 0.0487 just barely passed), but most did, even when checking if the proportions were the same. Oddly, the p value for G_A proportions was 0.39 while the p value for C_T proportions was much lower at 0.002. This was probably the result of a relatively small number of samples within the low population that had very few reported mutations in the MAF file, resulting in 0% for several categories. Since these categories were fairly high in percentage (between 20-35%), these zero values may have lowered the mean values of the proportions enough to cause a statistical significance between the populations.

When looking at which genes were affected by mutation, here again there were differences. A primary driver of these differences appeared to be gene size, but when sorting the lists by mutation counts and looking manually at the gene sizes, and when using a linear regression and looking at the genes with the highest residuals there were some differences that were not driven purely by gene size.

Having shown that there were indeed differences an obvious question that follows is: “What is causing these differences”? I speculate that there are probably a combination of structural, biological, and biochemical mechanisms behind these differences (40,76–83). In addition to clonal selection, there may be mutational hot spots, such as microsatellites within some genes and not others. These genes would be prone to mutate

more frequently if a condition causing microsatellite instability were to affect the cell. Depending on the mutation or regulatory problem that led to the condition causing increased mutation retention, certain types of DNA damage may become harder for the cells to detect or to repair leading to a rise in mutations that result from that kind of damage (45). In addition to this, expression seems to negatively correlate with mutation rates across the genome (76). There is also the expected effect of larger genes being bigger targets.

Other causes might broadly increase all types of mutation, due to affecting detection of DNA damage or weakening the ability of the DNA mismatch repair mechanism to locate and repair multiple types of damage (67,84–86).

Chapter 4

Kurtosis of mutation locations as a Possible Mutation Survey Method and Detailed Analysis of Potentially Interesting Genes

Introduction

I next wanted to determine if the distinction between the two classes of tumors based on the number of mutated genes carried within the tumors had an effect on the mutations that were within the genes themselves. Were certain mutations within a given gene more frequently found in the highly mutated group or the less mutated group? Were the mutations more random in the high frequency group while those in the low frequency group were more specific? The thinking behind my hypothesis is that mutations that act as drivers in driver genes should be enriched in the low mutation group than the high, while passenger mutations in driver genes should be relatively enriched in the high mutation group.

An extension of this thinking is that within a specific cancer-associated gene, specific mutations that enhance an activity that assists in producing the cancer phenotype should exhibit an enrichment in the low mutation group and a diluted frequency in the high mutation group.

I expected to see a decrease in specificity of mutations in the high mutation group, such that even when oncogenes are mutated in the high mutation group, the correct gain-of-function mutation would not be hit frequently, and that existing tumors or the pre-tumor somatic lineage would instead acquire mutations in oncogenes at non-tumorigenic locations. In contrast, inactivating mutations in tumor suppressor genes were far less precise than the gain-of-function mutations in oncogenes. Thus a random mutation event was far more likely to produce a tumor assisting (inactivating) mutation within a suppressor than in an oncogene (assuming the given gene will be hit somewhere along

its length), and I expected that this effect would also be borne out if the locations of mutations within genes were examined between the two groups. In simple terms, I expected no dramatic difference in targeting of mutations to specific sites in tumor suppressors between the low and high mutation groups (except that the low mutation group may be enriched in nonsense or frameshift mutations relative to the high mutation group), but that there would be a significant difference in distribution of mutations in oncogenes between the low and high mutation groups with low mutation group showing targeting of specific activating mutations and a higher degree of randomness in the high mutation group.

Types of Cancer Driver Genes

Some cancer driver mutations achieve their effects by shutting down a gene which when functionally normal acts against pathways that favor tumor formation or survival. These genes with anti-tumor effects are classified as tumor suppressors (31,87). In genes with this kind of function one will expect higher mutation rates to result in a broad targeting of the gene sequence across all exons in any location where a disabling change can occur. There may be peaks in codons that only require a single base change to become STOP codons, especially if those require mutations are achievable with the most common variety of mutations. Due to the less targeted nature of the cancer contributing changes in tumor suppressors as compared to oncogenes, higher mutation rates would favor more of these mutations having a possible effect just by chance.

Other cancer driver mutations result in a functional change in the function/activity of a gene. These effects require that the gene still be expressed, but the mutation changes the function of the gene in some way that supports tumor survival or helps create the

cancer phenotype. These genes which can acquire pro-tumor functional changes are classified as oncogenes (31,32,87). Due to the requirement of conversion of proto-oncogenes to oncogenes being a gain of function mutation, these genes will often exhibit a mutational profile with specific locations where point mutations can have an activating effect (87,88). Due to the highly specific nature of cancer driving mutations in these genes, these genes would not be expected to be particularly enriched in tumors with a mutator phenotype, as the chance that a mutation would land outside of the specific activating targets that are pro-cancer is considerably larger than that of hitting the right spot with the right mutation.

Assessing the Nature of Mutational Specificity: Kurtosis

Kurtosis (a word derived from Greek, meaning “curved, arching”) is a statistical computation that is usually used to get a numeric value that indicates something about the overall shape of a graph (89,90). It and skewness are often used when determining whether a collection of data seem to fit a normal distribution or not (91). It is a measure of “tailedness” of the probability distribution of a real-valued random variable. A peak at a point would not be sufficient to raise kurtosis if a normal distribution surrounded that point. A normal distribution with very small standard deviation would still have an excess kurtosis of 0 (excess kurtosis is a calculation where the raw kurtosis value for the normal distribution, usually 3, is subtracted from the raw kurtosis value). A value is leptokurtotic (high kurtosis value) when there is a very strong peak that looks like a pointy spike, and/or heavy tails. The opposing situation, platykurtosis, would exist in a distribution where there was no pointy peak, and the peak was surrounded by a wide and rounded distribution, or in more extreme cases, a very wide and very flat distribution. Mesokurtosis refers to

values that are less extreme. Normal distributions are mesokurtotic and have an excess kurtosis of 0. A Bernoulli distribution with p=1/2, (a coin flip) has a kurtosis of -2. A discrete uniform distribution (dice roll) has a kurtosis that varies according to the number of possible values (number of faces on the die). At N=3 the excess kurtosis is -1.5, and it approaches -1.2 asymptotically as N increases.

When dealing with genomic coordinates of mutations within a gene, I was not dealing with a direct measure of probability of a single random variable, but those genes in which there are specific hot spots could still be expected to produce a higher kurtosis value than genes where there are no hot spots and mutations are randomly distributed throughout the entire length of the gene. Dr. Yu-Bo Wang suggested trying to use kurtosis values due to lack of a more obvious tailor-made method for detecting such differences, so I decided to see what resulted from these calculations for the mutated genes in these populations of tumors.

Kurtosis is a measure of the shape of a distribution (89,90). Technically, it is a measurement of how weighty the tails of a distribution are, though in practical terms, for distributions that have a bulk of their probability at the center of the distribution and that are symmetrical, it can also be an indicator of sharpness, or pointy-ness.

There are three terms that generally describe graphs with different kinds of kurtosis. The normal distribution, shown in gray on Figure 25, actually has a true kurtosis value of 3, but is used to define the concept of excess kurtosis, which subtracts the value of the kurtosis of the normal distribution from the true kurtosis, which is what is usually reported as kurtosis by various types of software, and is what many kurtosis estimation formulas return, and as such the normal distribution is considered to have an excess kurtosis value

of 0. When I refer to kurtosis values in the latter parts of this document, I am actually referring to excess kurtosis returned by an algorithm.

Leptokurtic distributions generally have a sharper peak, and heavier tails, such as the laplace distribution shown in blue in Figure 25. Platykurtic distributions generally have flatter peaks and lighter tails. The raise cosine distribution shown in yellow and the uniform distribution shown in orange on Figure 25 are both platykurtic. The raised cosine distribution has the typical qualities mentioned, and the uniform distribution has no peak at all, and is all tail.

To explain this point, there is a probability distribution known as a U-quadratic distribution, that is a continuous probability distribution over a defined range from a to b. The quadratic that describes this distribution would be a U shaped parabola (with a focus pointing in the positive direction) that is symmetrical about the midpoint between a and b, and has total area under the curve of 1 between a and b. The excess kurtosis for such a function is $3/112*(b-a)^4$. The longer the distance from b to a, the more kurtosis this distribution would have, and yet it has no peak in the middle. Kurtosis values for this distribution can approach zero for very small a to b intervals, but it will never be negative.

This does suggest that strongly positive kurtosis values obtained from mutation locations across a gene would need to be examined for similar heavy tailed distributions without a centralized probability peak. I do not expect that mutation distributions fitting a very long and heavy tailed distribution with no peaks in the middle such as this would be particularly common.

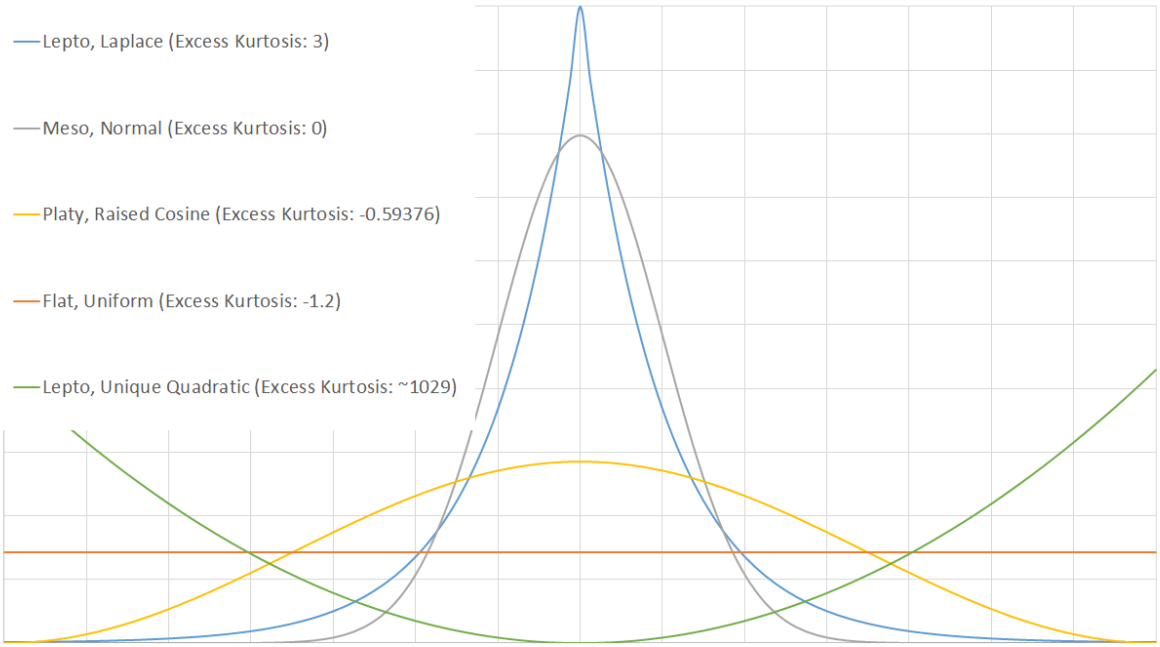

Figure 25. Illustrations of kurtosis shapes.

Several probability distributions were produced to fit within the same x value ranges (0-14) to illustrate how kurtosis relates to shape. The blue example is a Laplace distribution (Leptokurtic). The gray example is a Normal distribution (Mesokurtic). The yellow example is a Raised Cosine distribution (Platykurtic). The orange horizontal flat line is a continuous Uniform Distribution (Platykurtic). The green example is a unique quadratic distribution (Leptokurtic). Excess kurtosis values are shown in the legend

Hypothesis

Within any cancer driver oncogene, specific activating or gain-of-function mutations should be enriched in the low mutation group of tumors while non-specific or passenger mutations should be enriched in the high mutation group of tumors. To detect this difference, I examined the mutations within specific genes. I expected that within oncogenes there are a small number of locations where subtle mutations lead to constitutive activation of the resulting protein or other subtle alteration in its structure or behavior that leads to a tumorigenic effect. This means that there ought to be a very specific set of mutations in oncogenes that are capable of contributing to the cancer phenotype within the selected tumor population. I expected therefore to see a very small number of locations with high mutation counts in these genes when they present selective advantage.

In contrast, I expected that tumor suppressor genes would show a profile where mutations appear throughout the gene at any sites where they can deactivate the function of the gene, perhaps with some hot spots where it is easier to create damaging mutations, but with less precision than I would expect within oncogenes.

There is also the effect of the mutation rates to consider. My model is that the highly increased rate of mutation in the high mutation group would create possibilities for random mutations to occur outside the tumorigenesis process that wouldn't exist at the lower mutation rates in the low mutation group. As a result of this increased rate of mutations in the high mutation group, there is an increased possibility of redundantly affecting signaling pathways via multiple hits to more easily disabled gene targets in the high mutation group instead of requiring a precise hit on a canonically important cancer-

related gene target. Also, the effect of the enhanced mutation rate would generally cause a lot more passenger mutations. Therefore, in the high mutation group I would expect the mutation accumulation peaks to be less pronounced in oncogenes, and for the oncogenes that were mutated, there would be an increased number of non-tumorigenic mutations, which would effectively be passenger mutations, in a significantly increased number of samples. In contrast, I expect that there would not be very much difference to the spectrum of mutations found in most tumor suppressor genes between the two groups except possibly an enrichment of missense or silent mutations in the high mutation group.

In summary, the mutations found within many genes that require precise mutations will likely be found to have more accurate targeting in the low mutation group, and less accurate targeting in the high mutation group, while genes that simply require deactivation will not experience much difference in the distribution of their mutations between the two groups.

Given that mutations in hot-spots will create peaks at a specific point rather than at an “average location”, these would tend to produce a leptokurtic scenario. If there are multiple hot spots that are not localized to a particular section of the gene this would trend toward producing a mesokurtic result, especially if there are also off target mutations spread through the gene, or spanning multiple exons. If the gene has no hot spots and no selective locations producing oncogenic traits, and sports very long “tails” of off target mutations spread throughout the gene, this would be expected to produce platykurtic values.

Methods

For a selected set of oncogenes with known effective mutation sites and a set of suppressor genes, I examined the mutation locations that appeared within the two sample groups. Counts were tallied per genomic location, and these counts were used in conjunction with annotations of known effective sites to test the hypothesis.

I sought advice on how to discriminate between mutation profiles containing strong peaks and mutation profiles with wide dispersal of mutations from statistics experts Drs. James Grady and Yu-Bo Wang. Yu-Bo suggested that I try computing the kurtosis of the genomic positions, and see if the kurtosis values signaled the position of strong peaks. Based on my observation of some of the mutations lists and the resulting kurtosis values this method has resulted in at least a partial success, although there are some caveats to the method. Some distributions that have no peaks result in kurtosis values that suggest there should be strong peaks. It did produce some interesting and potentially useful results though.

I accomplished the kurtosis calculations using a python program to control function calls in R using a module called rpy2. The kurtosis function from the R package e1071 was used. A list of start positions for each mutation within each gene was compiled, and this list of numbers was fed to the kurtosis function. The resulting numbers were compiled into a tab separated table as output.

For a selected set of genes, I pulled a list of the mutations out of the MAF file using command line tools. I queried for the lines containing the requested gene symbol and pulled out the relevant fields for the start and end locations. This was sorted and duplicates were counted, producing a set of tables showing the number of mutations at

each position. For better comparison between the high and low mutation groups, I added columns that divided the count values by the tumor population size.

Results

As expected, mutations in oncogenes showed a strong preference for specific sites to a much higher degree than the mutations in tumor suppressors within the low mutation population. The high mutation population showed a higher degree of randomness and significantly less specificity. However, the results were not completely clean. In both the low and high mutation groups, tumor suppressor genes still showed statistical preferences for some sites. It is possible that this preference might have been due to ease of mutational deactivation at these sites by creating premature stop codons or significantly disrupting the action of the gene by disrupting normal folding or structural integrity of the peptide at these enriched sites.

An exception to the higher degree of randomness in the mutations in the high mutation group was found for the gene BRAF. This decreased randomness was consistent with the previous observation that specific mutations in BRAF were associated with and advantageous to the mutator phenotype (92–94). Thus I confirmed the possibility of there being some exceptions to my hypothesis on a genome-wide basis if certain mutations within specific oncogenes are strongly selective within the mutator phenotype.

Vogelstein et al. composed a list of cancer related genes classified as tumor suppressor genes and oncogenes according to a scoring algorithm (31). I wanted to see the mutation profiles of these genes and whether the kurtosis values for these genes obeyed my model based on their Vogelstein classification. I found that the kurtosis values

were consistent with my model for both tumor suppressors and oncogenes and between the low and high mutation frequency populations.

APC, short for Adenomatous polyposis coli, is a well-known tumor suppressor gene that has been found to have very prominent involvement in colon cancer (95), and was given a name based on that relationship. APC had high kurtosis values in both populations, with a very strong value for the high population. This was not unexpected, given that destructive mutations often cluster in important functional motifs, and due to the very strong association of this gene with colon cancer. The increase in kurtosis in the high mutation group could be due merely to the very high mutation rate, or might be the result of random fluctuation in the locations where the mutations hit. Perhaps the gene has hot spots for mutation that are more likely to mutate under the conditions present in a cell with a mutator phenotype.

BRAF an oncogene with known association to tumors with DNA repair disorders, (92,94) had very high kurtosis in the high mutation group, an expected result in this population.

KRAS does not share an association with mutator phenotypes or DNA repair disorders (96–98). It had high kurtosis in the low mutation group, showing very strong selection for the oncogenic mutation locations. In contrast, KRAS was rarely mutated at the oncogenic sites in the high mutation group.

Compared purely in terms of accumulated mutation count, APC had 84 mutations in a population of 180 tumors in the low group (0.46 mutations/sample), and 15 mutations out of a population of 39 (0.38 mutations/sample), but to focus on this alone would have missed a very important aspect of how oncogenes like KRAS mutate. KRAS only

experiences oncogenic mutations as a result of specific changes at specific locations in its sequence.

Position chr12:25398284 (in GRCh37/hg19), corresponding to the 2nd base in the 12th codon of KRAS, was the most mutated site, with 39 mutations. The codon at this site is usually GGU (Glycine). There were 3 mutation variants found at this site. C to T (19 counts), changes the codon to GAU (Aspartic acid). C to A (15 counts), changes the mRNA codon to GUU (Valine). C to G (5 counts), changes this codon to GCU (Alanine). A site immediately adjacent at position chr12:25398285 which is the 1st base within the same codon, had 11 mutations. C to T (5 counts) results in codon AGU (Arginine). C to A (4 counts) results in codon UGU (Cysteine). C to G (2 counts) results in codon CGU (Serine).

Another site, at position chr12:25398281 corresponds to the 2nd base in the 13th codon which is usually GGC. This C to T mutation (16 counts) results in GAC (Aspartic acid). The first base of this codon also experienced a C to A mutation to produce a UGC codon (cysteine) in one sample.

It is unclear whether all of these amino acid changes result in an oncogenic protein. Changes to codon 12 and 13 tend to be oncogenic (98,99), but mutations other known locations were not particularly common in this type of tumor.

Limitations of the analysis

Kurtosis appeared to function well as a positive identifier, but its power in terms of negative results was unreliable due to the relatively low mutation counts some genes experience and the fact that some genes' involvement in cancer is related to the tissue of origin, and that some genes might exhibit a very platykurtic distribution if they are tumor

suppressors lacking hot spots. Thus its utility in identifying interesting genes lies solely in comparative results (which group shows a higher value) and even so requires more detailed follow-up analysis to interpret any positive results. It therefore can serve as a useful flag for taking a further look at genes that show high kurtosis values.

#Gene Symbol	Classification*	Kurtosis LOW	Kurtosis HIGH
AR	Oncogene	6.641	-1.849
KRAS	Oncogene	1.515	-1.519
PIK3CA	Oncogene	-0.249	-1.917
GATA2	Oncogene	-0.766	-2.333
GNAS	Oncogene	-0.869	-1.989
DNMT1	Oncogene	-1.698	-2.072
SMO	Oncogene	-2.395	-2.750
IDH1	Oncogene	-2.431	-2.750
SETBP1	Oncogene	-2.333	3.960
BRAF	Oncogene	-0.134	5.112
FGFR3	Oncogene	-2.333	1.079
MED12	Oncogene	-2.333	0.064
DNMT3A	Oncogene	-2.750	-0.523
ERBB2	Oncogene	-2.363	-0.447
ABL1	Oncogene	-2.750	-1.045
JAK2	Oncogene	-2.750	-1.116
CBL	Oncogene	-2.750	-1.319
SF3B1	Oncogene	-2.051	-0.921
RET	Oncogene	-1.950	-0.894
EGFR	Oncogene	-2.750	-1.750
PPP2R1A	Oncogene	-2.750	-1.786
PTPN11	Oncogene	-2.750	-1.893
KLF4	Oncogene	-2.750	-1.918
AKT1	Oncogene	-2.750	-2.003
ALK	Oncogene	-1.959	-1.234
GNA11	Oncogene	-2.750	-2.047
KIT	Oncogene	-2.245	-1.802
HRAS	Oncogene	-2.750	-2.333
FGFR2	Oncogene	-2.333	-1.969
PDGFRA	Oncogene	-2.131	-1.811
JAK3	Oncogene	-1.969	-1.818
CARD11	Oncogene	-1.602	-1.469
TSHR	Oncogene	-2.333	-2.226
CTNNB1	Oncogene	-1.848	-1.795
FBXW7	TSG	7.245	-1.260
TP53	TSG	5.697	-1.394
MLL3	TSG	2.850	-1.027
FAM123B	TSG	0.422	-1.297
SMAD4	TSG	0.052	-1.533
NOTCH1	TSG	-0.233	-1.459
CREBBP	TSG	-0.637	-1.427
SOX9	TSG	-1.629	-2.180
BAP1	TSG	-2.333	-2.750
TRAF7	TSG	-2.333	-2.750
MLH1	TSG	-1.774	-2.083

#Gene Symbol	Classification*	Kurtosis LOW	Kurtosis HIGH
WT1	TSG	-2.405	-2.427
BCOR	TSG	-2.090	-2.097
RNF43	TSG	0.457	13.021
ARID1A	TSG	-2.022	10.235
BRCA1	TSG	-2.750	2.085
PTCH1	TSG	-2.750	0.452
APC	TSG	3.411	6.380
ARID1B	TSG	-2.333	0.278
RB1	TSG	-2.750	-0.482
EP300	TSG	-2.045	0.219
BRCA2	TSG	-1.580	0.568
PAX5	TSG	-2.750	-0.807
ATRX	TSG	-2.015	-0.084
NCOR1	TSG	-1.803	-0.094
SMARCB1	TSG	-2.750	-1.142
AXIN1	TSG	-2.750	-1.329
CIC	TSG	-1.956	-0.581
TSC1	TSG	-2.750	-1.578
CDH1	TSG	-2.333	-1.202
PTEN	TSG	0.362	1.354
NF1	TSG	-2.750	-1.762
MSH6	TSG	-0.709	0.241
PIK3R1	TSG	-1.052	-0.108
SETD2	TSG	-2.263	-1.323
ACVR1B	TSG	-2.333	-1.502
MLL2	TSG	-1.570	-0.833
CASP8	TSG	-2.333	-1.607
PRDM1	TSG	-2.750	-2.121
SMAD2	TSG	-2.333	-1.754
GATA3	TSG	-2.750	-2.214
B2M	TSG	-1.688	-1.160
SMARCA4	TSG	-1.746	-1.243
ASXL1	TSG	-2.113	-1.626
KDM5C	TSG	-2.750	-2.305
KDM6A	TSG	-2.750	-2.333
STK11	TSG	-2.750	-2.333
ATM	TSG	-1.784	-1.532
FUBP1	TSG	-2.333	-2.239
PBRM1	TSG	-1.726	-1.713

Figure 26. Kurtosis values for Vogelstein Subtly Mutated Gene List for TCGA data. Genes that were found in the supplemental data of the 2013 Vogelstein paper on Genomic Landscapes in Cancer (31) but which had no counts or not enough counts to be able to compute kurtosis for either population were omitted from this table. The table was sorted by the Vogelstein classification first, and as a secondary sorting rule the genes with higher values in the LOW population than the HIGH population were sorted toward the top.

Examining Outliers and Potentially Interesting Genes in More Detail

I examined a subset of genes that seemed particularly out of place in regard to their size vs their mutation counts. I included a calculation for the absolute value of the difference for the kurtosis value between the groups and sorted according to that difference. My hope was the genes with the most different values would have significant shifts in their mutation distribution, and perhaps would have a functional association with cancer phenotypes. They might also have targeted locations that mutate and become strongly positively selected. I included TTN on this list because despite its massive size, it had many more mutations than its size ought to dictate based on the general trend, and it is also a very good example of a gene that was computed to have low kurtosis on its mutation locations, and actually deserves the negative result. The mutations in TTN do not pile up at any particular spots. The mutations are spread out quite a bit, with all of the locations only having 1 or 2 mutations, with counts of 2 being infrequent. Due to this very spread out and sparse mutation pattern, I opted not to show TTN's mutation locations in detail.

In Figure 27, I highlighted in green those genes that were also contained within the Vogelstein list of oncogenes and tumor suppressor genes known to be associated with cancer phenotypes (31). Their results may also help in understanding these mutation profiles and how the mutation rates affect them. There are clearly some genes with very strong kurtosis values, and also some genes that only pile up enough to get positive kurtosis in one group or the other. Interestingly, a good number of the genes showed a higher degree of "pointiness", as scored by kurtosis, in the high mutation group, where

one would usually expect an increase in randomness. These genes may provide a benefit of some kind to mutator phenotypes when mutated.

The first table contains those genes which had a positive kurtosis in one group and negative kurtosis in the other. The second table was produced from the remaining genes. Kurtosis could not be computed for TPRX1 in the high group, so I left it in the second list.

SOX9 is a good example of a gene where the kurtosis is higher on one side, but for which additional information is required to determine if this corresponds to a meaningful difference in mutation location or is a result of random differences. There are more mutations in the low population (25) than in the high population (7), which could lead to there being more mutations in a similar region merely due to the difference in size between the populations. The mutations don't pile up at a single location, but since SOX9 acts a tumor suppressor gene (100,101), this is not surprising. They may still be piling up within protein domains where a disruption is more likely to destroy the function of the gene, and this might lead to a relatively higher kurtosis when there are more mutations to use in computation.

Old_Name	GENE_SYMBOL	LOW_GROUP	HIGH_GROUP	ABS_VAL_DIFFERENCE
ZNF814	ZNF814	27.702	-0.970	28.672
FBXW7	FBXW7	7.245	-1.260	8.504
AR	AR	6.641	-1.849	8.491
KRTAP4-3	KRTAP4-3	7.580	-0.727	8.307
MUC17	MUC17	-0.380	6.781	7.161
KCNN3	KCNN3	5.479	-1.632	7.111
TCF15	TCF15	4.310	-2.306	6.615
ZFPM1	ZFPM1	4.355	-1.897	6.251
GPR98	GPR98	-1.629	4.336	5.964
HMCN1	HMCN1	-1.362	2.992	4.354
KRTAP4-5	KRTAP4-5	3.224	-1.049	4.273
RP1L1	RP1L1	-1.696	2.450	4.146
RYR2	RYR2	-0.887	2.855	3.742
FRAS1	FRAS1	-1.975	1.570	3.545
TCHH	TCHH	2.619	-0.776	3.396
KRAS	KRAS	1.515	-1.519	3.035
ANKLE1	ANKLE1	2.280	-0.421	2.701
PCLO	PCLO	-0.390	1.874	2.264
SDK1	SDK1	-0.251	1.897	2.148
PCDHA7	PCDHA7	0.300	-1.763	2.063
FAT1	FAT1	-1.688	0.365	2.054
CCDC168	CCDC168	-1.105	0.756	1.861
CACNA1B	CACNA1B	0.465	-1.317	1.782
SMAD4	SMAD4	0.052	-1.533	1.585
CSMD3	CSMD3	-0.793	0.738	1.532
MUC16	MUC16	-0.178	0.462	0.640
CRIPAK	CRIPAK	-0.582	0.051	0.633

Old_Name	GENE_SYMBOL	LOW_GROUP	HIGH_GROUP	ABS_VAL_DIFFERENCE
TPRX1	TPRX1	3.007	--	#VALUE!
XIRP2	XIRP2	2.129	37.066	34.937
NEFH	NEFH	39.123	18.489	20.634
ATXN1	ATXN1	21.173	1.151	20.022
FAM194B	ERICH6B	24.785	15.264	9.521
MAML2	MAML2	9.797	2.079	7.718
PDE4DIP	PDE4DIP	0.485	8.169	7.684
PLEC	PLEC	13.160	7.204	5.956
TBP	TBP	8.381	3.570	4.811
OGFR	OGFR	-0.063	-2.076	2.013
PIK3CA	PIK3CA	-0.249	-1.917	1.667
CRCT1	CRCT1	-0.701	-2.253	1.552
IRF2BPL	IRF2BPL	-0.146	-1.624	1.478
LAMA1	LAMA1	-1.738	-0.417	1.321
KNDC1	KNDC1	-0.056	-1.193	1.137
CROCC	CROCC	-0.227	-1.347	1.120
KRT1	KRT1	-1.037	-2.054	1.018
SYNE2	SYNE2	-2.195	-1.218	0.977
NEB	NEB	-1.717	-0.744	0.974
LAMA5	LAMA5	0.955	0.034	0.921
DNAH12	DNAH12	-2.103	-1.204	0.899
FAT3	FAT3	-0.355	-1.243	0.888
DST	DST	-1.823	-0.966	0.857
DNAH17	DNAH17	-1.914	-1.070	0.844
COL6A5	COL6A5	-1.527	-0.696	0.831
DNAH6	DNAH6	-1.745	-0.967	0.777
ALPK2	ALPK2	-1.656	-0.889	0.767
DSPP	DSPP	0.162	0.910	0.748
COL18A1	COL18A1	-1.580	-0.879	0.701
DNAH2	DNAH2	-1.870	-1.218	0.652
MUC4	MUC4	1.233	0.626	0.607
ABCA13	ABCA13	0.618	0.038	0.580
SOX9	SOX9	-1.629	-2.180	0.551
IRF5	IRF5	-1.140	-1.688	0.548
DNAH10	DNAH10	-1.743	-1.268	0.476
ZFHX4	ZFHX4	-1.134	-1.609	0.476
DNAH3	DNAH3	-1.490	-1.108	0.382
UBR4	UBR4	-1.692	-1.327	0.365
NCOR2	NCOR2	-0.155	-0.515	0.360

DNAH5	DNAH5	-0.975	-1.298	0.322
NUMBL	NUMBL	-1.246	-0.931	0.315
ASPM	ASPM	-1.710	-1.402	0.308
DCHS2	DCHS2	-1.707	-1.428	0.279
DNAH14	DNAH14	-1.011	-1.258	0.246
EYS	EYS	-1.062	-1.256	0.194
DNAH8	DNAH8	-0.524	-0.382	0.142
FAT4	FAT4	-1.642	-1.726	0.084
TTN	TTN	-1.509	-1.584	0.075
GPRIN2	GPRIN2	-1.589	-1.649	0.060
TNXB	TNXB	-1.046	-1.096	0.051
CACNA1H	CACNA1H	-1.632	-1.595	0.037
SSPO	SSPO	-1.376	-1.395	0.019
LRP1	LRP1	-0.546	-0.565	0.018

Figure 27. Kurtosis Values for selected genes from linear regression table.

I picked a number of genes with relatively smaller sizes from the residual-sorted tables This table was sorted to place the values which were negative on one side and positive on the other first. These values were placed into the first table. The rest were placed into the second table. The genes were then sorted by the absolute value of the difference in kurtosis between the high and low groups. Two dashes signify that kurtosis could not be calculated for that gene in that population. Genes from the Vogelstein list (31) are highlighted in green.

The cancer gene list (31), as a whole, contained genes that were not necessarily expected to be associated with colon cancer, and thus resulted in many genes having low kurtosis in both groups. It was interesting that so many genes had higher kurtosis for the HIGH mutation group, where I generally expected more randomness. I then decided to take a look at the genes canonically associated with colon cancer.

I compiled a table (shown in Figure 28) based on the list of mutations and genes on mycancergenome.org (102). While it is a complex table, I structured it such that only known mutations would be counted. It would show a dash for any value that was not defined in the known mutation part of the table.

As mentioned previously, KRAS showed a strong preference for mutations in codon 12 and 13, and to a lesser extent, codons 61 and 146, while other locations were rarely mutated.

In BRAF there was strong tendency for changing codon 600 to code for Valine. This BRAF mutation is highly associated with defects in DNA repair (92,93,96). This targeting of the mutation to a specific location (V600E) was captured quite readily by the difference in kurtosis.

PIK3CA is strongly associated with colon cancer (103,104) and showed a strong tendency for mutation at codon 545, changing from Glycine to Lysine. There was a concordant rise in kurtosis, but the number of off target mutations (and perhaps their distance from this mutation peak), was sufficient to cause the resulting kurtosis to be somewhat low. This is why the change in value between the same gene in both populations is more important than the actual value of the kurtosis.

PTEN is another gene associated with microsatellite-stable colon cancer (104,105). Its mutation spectrum was particularly interesting. It showed NONE of the three known cancer-associated mutations (102), though one was very similar (there was a deletion mutation at a position very close to the listed one found in both populations). There were also a few other frame shift mutations at different locations and some nonsense mutations as well. These could possibly all be oncogenic due to destroying the function of the resulting PTEN protein or initiating nonsense-mediated decay (106) and resulting in no protein being made at all.

							L_Counts					H_Counts						AA Subs				
Gene	Chr	CDS(c.#)	codon	Pos_GRCH37	Pos_GRCH38	Ref base	A	C	G	T	-	A	C	G	T	-	Ref_AA	A	C	G	T	-
NRAS	1	34	12	115258748	114716127	C	0	-	-	0	-	0	-	-	0	-	G	C	-	-	S	-
NRAS	1	35	12	115258747	114716126	C	1	-	0	2	-	0	-	0	0	-	G	V	-	A	D	-
NRAS	1	181	61	115256530	114713909	G	-	-	-	3	-	-	-	-	0	-	Q	-	-	-	K	-
NRAS	1	182	61	115256529	114713908	T	-	1	-	-	-	-	0	-	-	-	Q	-	R	-	-	-
AKT1	14	49	17	105246551	104780214	C	0	0	0	1	0	0	0	0	2	0	E	-	-	-	K	-
BRAF	7	1397	466	140481411	140781611	C	1	-	-	-	-	0	-	-	-	-	G	V	-	-	-	-
BRAF	7	1406	469	140481402	140781602	C	0	-	1	0	-	0	-	0	0	-	G	V	-	A	E	-
BRAF	7	1781	594	140453154	140753354	T	0	0	-	-	-	0	0	-	-	-	D	V	G	-	-	-
BRAF	7	1786	596	140453149	140753349	C	-	-	0	-	-	-	-	0	-	-	G	-	-	R	-	-
BRAF	7	1799	600	140453136	140753336	A	-	-	-	5	-	-	-	-	20	-	V	-	-	-	E	-
PIK3CA	3	1624	542	178936082	179218294	G	7	-	-	-	-	1	-	-	-	-	E	K	-	-	-	-
PIK3CA	3	1633	545	178936091	179218303	G	19	0	-	-	-	0	0	-	-	-	E	K	Q	-	-	-
PIK3CA	3	1634	545	178936092	179218304	A	-	-	0	0	-	-	-	0	0	-	E	-	-	G	V	-
PIK3CA	3	1636	546	178936094	179218306	C	0	-	0	-	-	0	-	0	-	-	Q	K	-	E	-	-
PIK3CA	3	1637	546	178936095	179218307	A	-	0	1	0	-	-	0	0	0	-	Q	-	P	R	L	-
PIK3CA	3	1645	549	178936103	179218315	G	0	-	-	-	-	0	-	-	-	-	D	N	-	-	-	-
PIK3CA	3	3140	1047	178952085	179234297	A	-	-	2	0	-	-	-	4	0	-	H	-	-	R	L	-
PTEN	10	477	159	89692993	87933236	G	-	-	-	0	-	-	-	-	0	-	R	-	-	-	S	-
PTEN	10	697	233	89717672	87957915	C	-	-	-	0	-	-	-	-	0	-	R	-	-	-	*	-
PTEN	10	800	267	89717775	87958018	A	-	-	-	-	0	-	-	-	-	0	K	-	-	-	-	A
SMAD4	18	989	330	48591826	51065456	A	-	0	-	-	-	-	0	-	-	-	E	-	A	-	-	-
SMAD4	18	1051	351	48591888	51065518	G	0	0	-	-	-	0	0	-	-	-	D	N	H	-	-	-
SMAD4	18	1065	355	48591902	51065532	C	0	-	-	-	-	0	-	-	-	-	D	E	-	-	-	-
SMAD4	18	1081	361	48591918	51065548	C	0	-	-	2	-	0	-	-	1	-	R	S	-	-	C	-
SMAD4	18	1082	361	48591919	51065549	G	3	-	-	-	-	1	-	-	-	-	R	H	-	-	-	-
SMAD4	18	1609	537	48604787	51078417	G	-	-	-	0	-	-	-	-	0	-	D	-	-	-	Y	-
KRAS	12	34	12	25398285	25245351	C	4	-	2	5	-	0	-	0	0	-	G	C	-	R	S	-
KRAS	12	35	12	25398284	25245350	C	15	-	5	19	-	1	-	0	2	-	G	V	-	A	D	-
KRAS	12	37	13	25398282	25245348	C	1	-	0	0	-	0	-	0	0	-	G	C	-	R	S	-
KRAS	12	38	13	25398281	25245347	C	0	-	0	16	-	0	-	0	2	-	G	V	-	A	D	-
KRAS	12	64	22	25398255	25245321	G	-	-	-	0	-	-	-	-	1	-	Q	-	-	-	K	-
KRAS	12	181	61	25380277	25227343	G	0	-	-	-	-	0	-	-	-	-	Q	K	-	-	-	-
KRAS	12	182	61	25380276	25227342	T	0	1	1	-	-	0	0	0	-	-	Q	L	R	P	-	-
KRAS	12	183	61	25380275	25227341	T	0	0	2	-	-	0	0	0	-	-	Q	H	R	H	-	-
KRAS	12	351	117	25378647	25225713	T	-	-	1	0	-	-	-	0	0	-	K	-	-	N	N	-
KRAS	12	436	146	25378562	25225628	C	-	-	0	4	-	-	-	0	4	-	A	-	-	P	T	-
KRAS	12	437	146	25378561	25225627	G	1	-	-	-	-	0	-	-	-	-	A	V	-	-	-	-

	LOW			HIGH		
	Known	Total	Ratio	Known	Total	Ratio
NRAS	7	7	1.000	0	2	0.000
AKT1	1	2	0.500	2	7	0.286
BRAF	7	10	0.700	20	26	0.769
PIK3CA	29	49	0.592	5	22	0.227
PTEN	0	13	0.000	0	12	0.000
SMAD4	5	24	0.208	2	8	0.250
KRAS	77	84	0.917	10	15	0.667

	LOW_GROUP	HIGH_GROUP	ABS_VAL_DIF
NRAS	-2.204	--	--
AKT1	-2.750	-2.003	0.747
BRAF	-0.134	5.112	5.246
PIK3CA	-0.249	-1.917	1.667
PTEN	0.362	1.354	0.992
SMAD4	0.052	-1.533	1.585
KRAS	1.515	-1.519	3.035

Figure 28. Genes known to be associated with colon adenocarcinoma (102), their known mutations, and their kurtosis values in the high and low mutation populations.

The first table lists genomic position, codon number, CDS position, reference base, original amino acid, the resulting amino acid for known mutations, and counts of specified mutations within the high and low mutation groups.

Of the smaller tables, the one on the left shows counts of known mutations the total number of mutations for that gene and a ratio of known mutations to total mutations.

The smaller table on the right shows the kurtosis values for these genes. (-- is a stand in for a value that was unable to be computed due to there not being enough mutations)

The Table in Figure 28 demonstrates a key point about the differences between tumor suppressor genes and oncogenes. For the tumor suppressor genes, as long as their function is diminished or destroyed, the mutation can be assumed to contribute to oncogenesis. For oncogenes, a gain of function is required. Thus, in most cases, for oncogenes, nonsense mutations and frameshifts, while certainly destructive to gene function, will not likely contribute to oncogenesis. Mutations involving the splice site and an in-frame insertion or deletion are also more likely to be non-oncogenic in these genes, depending on their effects. Silent mutations would be expected to be ineffective in both types of genes. Thus, using algorithms that determine disruptiveness of the mutation and/or the ontology of the gene where a mutation occurs as the determinant for whether the mutation is expected to be pathogenic in terms of cancer, would be a mistake if one does not take into account what sort of mutation would actually produce an oncogenic outcome.

			Missense_Mutation	Nonsense_Mutation	Silent	Frame_Shift_Ins	Frame_Shift_Del	In_Frame_Del	In_Frame_Ins	Splice_Site
NRAS	Oncogene	L	7	0	0	0	0	0	0	0
		H	0	0	0	0	0	0	0	2
AKT1	Oncogene	L	2	0	0	0	0	0	0	0
		H	5	0	2	0	0	0	0	0
BRAF	Oncogene	L	9	0	0	0	0	0	0	1
		H	21	2	1	1	1	0	0	0
PIK3CA	Oncogene	L	47	1	0	0	0	1	0	0
		H	21	1	0	0	0	0	0	0
PTEN	TSG	L	3	1	1	0	6	0	0	2
		H	5	3	0	1	3	0	0	0
SMAD4	TSG	L	18	3	0	0	2	0	0	1
		H	5	2	0	0	1	0	0	0
KRAS	Oncogene	L	81	0	2	0	0	1	0	0
		H	13	0	2	0	0	0	0	0
Sub_Total	Oncogene	L	146	1	2	0	0	2	0	1
		H	60	3	5	1	1	0	0	2
Sub_Total	TSG	L	21	4	1	0	8	0	0	3
		H	10	5	0	1	4	0	0	0
ALL	ALL	L	167	5	3	0	8	2	0	4
		H	70	8	5	2	5	0	0	2
Total	Total	-	237	13	8	2	13	2	0	6

Figure 29. COAD Mutations Classified by types

Missense mutations change the amino acid. silent mutations result in the same amino acid. Nonsense mutations result in a stop codon. Frame shifts resulting from insertion or deletion are shown, as well as in-frame deletions and insertions. Mutations involving the splice site are counted as well. Subtotals for the low and high groups as well as an overall total are at the bottom.

The genes in Figure 30 were of particular interest. BRAF, EGFR, HRAS, KRAS, RB1, and TP53 are on this list due to being known for bearing canonical cancer driver mutations. TTN was an example of a gene that was not expected to have any specific mutations related to cancer. KRTAP4-3 (107,108), KRTAP4-5 (107,109) FAM194B (110,111) (which has had its official symbol changed to ERICH6B), and DSPP (112–114) were mutation count trend outliers that looked to have potentially interesting mutation distributions.

HUGO_Symbol	Count_LOW	Count_HIGH	Kurtosis_LOW	Kurtosis_HIGH
BRAF	10	26	-0.134	5.112
EGFR	2	10	-2.750	-1.750
HRAS	2	3	-2.750	-2.333
KRAS	84	15	1.515	-1.519
RB1	2	6	-2.750	-0.482
TP53	117	13	5.697	-1.394
TTN	129	294	-1.509	-1.584
KRTAP4-3	19	14	7.580	-0.727
KRTAP4-5	45	12	3.224	-1.049
FAM194B	60	23	24.785	15.264
DSPP	70	53	0.162	0.910

Figure 30. Selected Genes of Interest

Counts were obtained from the same program as used previously. Kurtosis values were obtained using R as described in methods.

KRAS had three heavily targeted mutation locations within the TCGA dataset which are known sites (98,102), one of which is the most frequent (Figures 29 and 33). BRAF has one specific oncogenic mutation site (92,93,96) (Figure 31). In both of these genes the oncogenic sites were found to be hit at these sites much more frequently than any other locations. KRAS had a strong preference for its oncogenic site in the low mutation population, but this preference significantly diminishes in the high mutation population. BRAF shows the opposite. It barely showed any mutations within the low mutation group and had very high rate of mutation at its oncogenic target in the high mutation population. BRAF has a known association with tumors that exhibit DNA mismatch repair defects (92–94,96), so these results are expected. It is interesting that oncogenic KRAS had such a strongly diminished representation within the high mutation population of tumors. KRAS followed my expectation for oncogenes, but BRAF, due to its association and apparent selective effects within that population, actually followed the opposite pattern from my expectation. It did however still have a very strong kurtosis in its associated population, due to the strongly targeted mutation.

TP53 is also associated with colon cancer (53,115–117) and had several peaks, but it also had a great many mutations throughout its sequence (Figure 31). In that respect, it follows my expectation for tumor suppressors. It does not follow my expectation in respect to mutation frequency between the two populations. In the high mutation population, the representation of TP53 mutations drops by about 50%. In the low mutation population containing 180 tumors, there were 117 mutations in TP53 (0.65 mutations/sample), where in the high mutation population of 39 tumors, there were only 13 mutations in TP53 (0.33 mutations/sample). This was consistent with previous reports

that colon tumors with microsatellite instability due to mutations in MMR genes are less likely to have mutations in either KRAS or TP53 (72).

KRAS					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
25362777	25362777	0	1	0.000	0.026
25362805	25362805	0	1	0.000	0.026
25362833	25362835	1	0	0.006	0.000
25368481	25368481	0	1	0.000	0.026
25378561	25378561	1	0	0.006	0.000
25378562	25378562	4	4	0.022	0.103
25378647	25378647	1	0	0.006	0.000
25380246	25380246	1	0	0.006	0.000
25380275	25380275	2	0	0.011	0.000
25380276	25380276	2	0	0.011	0.000
25380277	25380277	1	1	0.006	0.026
25380278	25380278	1	1	0.006	0.026
25398214	25398214	1	0	0.006	0.000
25398218	25398218	1	0	0.006	0.000
25398255	25398255	0	1	0.000	0.026
25398262	25398262	1	0	0.006	0.000
25398281	25398281	16	2	0.089	0.051
25398282	25398282	1	0	0.006	0.000
25398284	25398284	39	3	0.217	0.077
25398285	25398285	11	0	0.061	0.000

BRAF					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
140449170	140449170	0	1	0.000	0.026
140453136	140453136	5	20	0.028	0.513
140453155	140453155	1	0	0.006	0.000
140453193	140453193	1	0	0.006	0.000
140476881	140476881	0	1	0.000	0.026
140477847	140477847	0	1	0.000	0.026
140481402	140481402	1	0	0.006	0.000
140481411	140481411	1	0	0.006	0.000
140482926	140482927	0	1	0.000	0.026
140482927	140482927	0	1	0.000	0.026
140508768	140508768	0	1	0.000	0.026
140508796	140508796	1	0	0.006	0.000

TP53					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
7574003	7574003	2	0	0.011	0.000
7574017	7574017	1	0	0.006	0.000
7574018	7574018	1	0	0.006	0.000
7576852	7576852	3	0	0.017	0.000
7577017	7577017	1	0	0.006	0.000
7577022	7577022	1	2	0.006	0.051
7577036	7577036	0	1	0.000	0.026
7577085	7577085	1	0	0.006	0.000
7577094	7577094	8	0	0.044	0.000
7577106	7577106	2	0	0.011	0.000
7577114	7577114	1	0	0.006	0.000
7577117	7577117	1	0	0.006	0.000
7577120	7577120	7	0	0.039	0.000
7577121	7577121	1	1	0.006	0.026
7577124	7577124	2	0	0.011	0.000
7577138	7577138	1	1	0.006	0.026
7577141	7577141	2	0	0.011	0.000
7577505	7577505	1	0	0.006	0.000
7577506	7577506	1	0	0.006	0.000
7577538	7577538	12	0	0.067	0.000
7577539	7577539	5	0	0.028	0.000
7577548	7577548	2	1	0.011	0.026
7577561	7577561	1	0	0.006	0.000
7577565	7577565	1	0	0.006	0.000
7577574	7577574	1	0	0.006	0.000
7577586	7577586	1	0	0.006	0.000
7577594	7577595	0	1	0.000	0.026
7578177	7578177	1	0	0.006	0.000
7578190	7578190	2	0	0.011	0.000
7578208	7578208	1	0	0.006	0.000
7578210	7578210	0	1	0.000	0.026
7578211	7578211	1	0	0.006	0.000
7578212	7578212	4	1	0.022	0.026
7578217	7578217	2	0	0.011	0.000
7578235	7578235	1	0	0.006	0.000
7578240	7578241	1	0	0.006	0.000
7578253	7578253	1	0	0.006	0.000
7578256	7578257	1	0	0.006	0.000
7578257	7578257	1	0	0.006	0.000
7578263	7578263	1	0	0.006	0.000
7578280	7578280	1	0	0.006	0.000
7578369	7578369	1	0	0.006	0.000
7578370	7578370	1	0	0.006	0.000
7578384	7578401	1	0	0.006	0.000
7578388	7578388	1	0	0.006	0.000

TP53					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
7578394	7578394	1	0	0.006	0.000
7578403	7578403	2	0	0.011	0.000
7578406	7578406	13	2	0.072	0.051
7578407	7578407	1	0	0.006	0.000
7578411	7578412	1	0	0.006	0.000
7578442	7578442	1	0	0.006	0.000
7578445	7578445	1	0	0.006	0.000
7578449	7578449	1	0	0.006	0.000
7578454	7578454	1	0	0.006	0.000
7578455	7578455	2	0	0.011	0.000
7578457	7578457	1	0	0.006	0.000
7578461	7578461	1	0	0.006	0.000
7578471	7578478	1	0	0.006	0.000
7578503	7578503	1	0	0.006	0.000
7578526	7578526	2	0	0.011	0.000
7578550	7578550	1	0	0.006	0.000
7579312	7579312	0	1	0.000	0.026
7579368	7579368	1	0	0.006	0.000
7579389	7579389	1	0	0.006	0.000
7579406	7579406	0	1	0.000	0.026
7579415	7579415	1	0	0.006	0.000
7579451	7579451	1	0	0.006	0.000

Figure 31. Known Cancer Related Genes

KRAS (Kirsten Rat Sarcoma Viral Oncogene Homolog) is on chromosome 12 (118). BRAF (B-Raf Proto-Oncogene, Serine/Threonine kinase) is on chromosome 7 (119). TP53 (Transformation-Related Protein 53) is on chromosome 17(120).

Start and End columns refer to the genomic locations of the mutations.

Count column refers to the number of mutations found at the coordinates.

The Normalized columns were produced by dividing the counts by the sample population size of each group (39 for the HIGH group, and 180 for the low group).

Both KRTAP4 genes (107–109) seemed to have a strong mutation peak in the low group at a single location (data shown in Figure 32). KRTAP4-5 also had some mutations at 3 other locations that might be selective, with 8, 4, and 6 mutations each. These mutations are relatively close to each other, and might affect the same structure or protein domain. At least one other keratin associated protein has been found to be involved in cancer (KRTAP5-5 (121)). KRTAP4-3 has 47 references listed on its COSMIC (Catalogue Of Somatic Mutations In Cancer) page, and KRTAP4-5 has 53. However, it is reasonable to be suspicious of these genes as possible drivers or contributing in some meaningful way to the cancer phenotype. Both KRTAP genes result in fairly small transcripts (879 for KRTAP4-5 and 942 for KRTAP4-3). Neither appeared to be mutated as broadly as TP53 or DSPP. Despite the presence of mutational peaks, these genes were not quite as preferentially mutated in either population as KRAS and BRAF. They had a prominent kurtosis in the low population, and not so much in the high population. The profile appeared to follow my expectations for an oncogene in respect to their mutation peaks, but they did show significant numbers of mutations at off target sites, which was more similar to my expectations of tumor suppressors. Thus, whether or not my predictions for the behavior of those tumor gene classifications turns out to be correct, they do appear to be worth treating as candidate driver genes.

KRTAP4-5					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
39305769	39305769	1	0	0.006	0.000
39305773	39305773	2	0	0.011	0.000
39305774	39305774	3	0	0.017	0.000
39305775	39305775	2	0	0.011	0.000
39305775	39305776	16	4	0.089	0.103
39305779	39305780	0	1	0.000	0.026
39305785	39305785	8	2	0.044	0.051
39305800	39305800	4	1	0.022	0.026
39305800	39305814	6	2	0.033	0.051
39305820	39305820	0	1	0.000	0.026
39305837	39305837	1	0	0.006	0.000
39305911	39305912	2	0	0.011	0.000
39305956	39305956	0	1	0.000	0.026
KRTAP4-3					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
39323916	39323916	0	1	0.000	0.026
39323998	39323998	0	1	0.000	0.026
39324041	39324041	0	1	0.000	0.026
39324139	39324139	1	0	0.006	0.000
39324229	39324230	3	1	0.017	0.026
39324333	39324333	9	3	0.050	0.077
39324346	39324347	1	1	0.006	0.026
39324347	39324347	0	2	0.000	0.051
39324348	39324348	2	1	0.011	0.026
39324349	39324349	2	3	0.011	0.077
39324367	39324367	1	0	0.006	0.000

Figure 32. Two keratin associated proteins found on chromosome 17 that were adjacent to each other (107–109), and similar in size. Columns are the same as in Figure 31.

ERICH6B, formerly called FAM194B, has been reported as mutated in cancers (110). Its COSMIC entry (which is still under its older name FAM194B), has 62 entries in its references list. It showed a very targeted set of mutation peaks within a small genomic region (data shown in Figure 33). These were about 11 bp in size, with 4 locations being targets. There were not very many mutations within this gene outside of the targeted location. It also seemed to be mutated at these targets to a proportionally larger degree in the high mutation group. Its locus spans 81218 bases, but its exonic size is fairly small containing ~2421 bases and resulting in a protein 696 amino acids in size. The profile does seem to be more similar to a tumor suppressor than to my expectations for an oncogene, but it is hard to use my model to suggest which classification this gene might fit. The relative lack of mutations throughout the gene could simply be a function of how small the spliced transcript is. The range of the mutation target suggests that the mutation may be disrupting a structure in that area or a function that relies on that 11 bp region. That domain might become inactivated due to these mutations, but I do not know how that impacts the function of the gene as a whole. It still presents itself as an interesting driver gene candidate.

ERICH6B (FAM194B)					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
46118944	46118944	0	1	0	0.025641026
46137868	46137868	1	0	0.005555556	0
46142524	46142524	1	0	0.005555556	0
46154078	46154079	0	1	0	0.025641026
46170599	46170599	0	1	0	0.025641026
46170656	46170656	0	1	0	0.025641026
46170719	46170719	3	0	0.016666667	0
46170726	46170726	15	5	0.083333333	0.128205128
46170728	46170728	11	4	0.061111111	0.102564103
46170735	46170735	15	6	0.083333333	0.153846154
46170737	46170737	10	3	0.055555556	0.076923077
46170742	46170742	2	0	0.011111111	0
46170744	46170744	2	0	0.011111111	0
46171109	46171109	0	1	0	0.025641026

Figure 33. ERICH6B / FAM194B. This gene, from chromosome 13, is named ERICH6B. It was named FAM194B within the MAF file obtained from TCGA due to the older annotation data the project used. The name has since been changed. The new name stands for Glutamate rich 6B, while the old name stood for family with sequence similarity 194, member B. Columns are the same as in Figure 31.

ZNF814 is a gene found on chromosome 19. It has a generic name based on its high number of zinc finger motifs, and not much appears to be known about it. A ten base pair region showed a high concentration of mutations in both groups, with 3 additional locations relatively close to that region also showing involvement. It is unknown what affect disrupting this region of the gene will do, but it does appear potentially interesting as a candidate driver gene deserving further research.

ZNF814					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
58384672	58384672	0	1	0	0.025641026
58384908	58384908	0	1	0	0.025641026
58384952	58384952	0	1	0	0.025641026
58384955	58384955	0	1	0	0.025641026
58384956	58384956	0	1	0	0.025641026
58384964	58384964	0	1	0	0.025641026
58385222	58385222	0	1	0	0.025641026
58385223	58385223	0	1	0	0.025641026
58385598	58385598	0	1	0	0.025641026
58385748	58385748	1	2	0.005555556	0.051282051
58385762	58385762	3	2	0.016666667	0.051282051
58385790	58385790	10	1	0.055555556	0.025641026
58385793	58385793	10	2	0.055555556	0.051282051
58385798	58385798	14	3	0.077777778	0.076923077
58385799	58385799	12	3	0.066666667	0.076923077
58385869	58385869	5	2	0.027777778	0.051282051
58385953	58385954	0	1	0	0.025641026
58386040	58386040	0	1	0	0.025641026
58386076	58386076	0	1	0	0.025641026
58386126	58386126	1	0	0.005555556	0

Figure 34. ZNF814. This gene is located on chromosome 19. “ZNF” is an abbreviation of “zinc finger”, a type of motif that this gene contains in several locations. Columns are the same as in Figure 31.

DSPP has also been associated with cancer in previous work (113,122). There are 93 entries in its COSMIC references page indicating it may have some involvement in cancer. DSPP had a higher kurtosis value in the high mutation group. It showed a peak of mutations in a 12 bp region targeting 4 locations (data shown in Figure 35). There was also another peak that showed up within the low population 621 bp away from the other cluster of mutations. This population lacked a peak at the second location, which might be partly responsible. It demonstrated preferred sites, but the peaks were pretty consistent within the high group and there were a lot of off target mutation locations all through the gene as well. Its mutation profile superficially resembled that of TP53.

There were some targeted locations, but these were mostly silent mutations, and there were also a very large number of mutation peak locations all through the gene. Thus the fairly wide spread of mutation locations indicated that I should expect the gene to act as a tumor suppressor if it has any involvement in cancer, but it turns out that a large number of these mutations were silent. Due to the mutations mostly being silent, the gene is less promising as a candidate driver gene, but it does deserve a deeper look given that it has some degree of mutation frequency other than the mutation count peaks.

DSPP					
START	END	COUNT_LOW	COUNT_HIGH	Norm_LOW	Norm_HIGH
88533665	88533665	0	1	0.000	0.026
88534082	88534082	0	1	0.000	0.026
88534230	88534230	0	1	0.000	0.026
88534235	88534235	0	1	0.000	0.026
88534399	88534399	0	2	0.000	0.051
88535233	88535233	1	0	0.006	0.000
88535346	88535346	0	1	0.000	0.026
88535442	88535442	0	1	0.000	0.026
88535460	88535460	0	1	0.000	0.026
88535868	88535868	0	1	0.000	0.026
88535987	88535987	0	1	0.000	0.026
88536291	88536291	0	1	0.000	0.026
88536317	88536319	0	1	0.000	0.026
88536361	88536361	0	1	0.000	0.026
88536448	88536448	8	5	0.044	0.128
88536451	88536451	9	9	0.050	0.231
88536457	88536457	7	5	0.039	0.128
88536460	88536460	8	5	0.044	0.128
88536471	88536471	2	0	0.011	0.000
88536472	88536472	5	0	0.028	0.000
88536475	88536475	1	0	0.006	0.000
88536553	88536553	1	0	0.006	0.000
88536681	88536681	0	1	0.000	0.026
88536929	88536937	0	1	0.000	0.026
88537073	88537081	1	0	0.006	0.000
88537078	88537078	1	1	0.006	0.026
88537081	88537081	9	1	0.050	0.026
88537087	88537087	1	0	0.006	0.000
88537107	88537107	0	1	0.000	0.026
88537205	88537213	1	0	0.006	0.000
88537240	88537240	1	0	0.006	0.000
88537288	88537288	0	1	0.000	0.026
88537303	88537303	0	2	0.000	0.051
88537306	88537306	3	3	0.017	0.077
88537315	88537315	1	0	0.006	0.000
88537389	88537389	0	1	0.000	0.026
88537398	88537398	0	1	0.000	0.026
88537420	88537420	0	1	0.000	0.026
88537435	88537435	2	0	0.011	0.000
88537441	88537441	2	0	0.011	0.000
88537444	88537444	1	0	0.006	0.000
88537456	88537456	3	0	0.017	0.000
88537476	88537476	0	1	0.000	0.026
88537513	88537513	2	0	0.011	0.000

Figure 35. DSPP (Dentin sialophosphoprotein) is found on chromosome 4 (112). Columns are the same as in Figure 31.

TP53, DSPP, ERICH6B, KRTAP4-3, and KRTAP4-5 were all found to be mutated more often than their size should dictate, and have mutation patterns suggesting there is a selective effect at some of these loci in tumors within the colon TCGA sample set. These genes have also been previously reported in lists of affected genes or in cancer studies previously. Many of the mentions involving these genes have been the result of genome wide association studies in large lists of results as opposed to specific research into these genes to characterize their function and possible contribution to cancer phenotypes.

Discussion

In summary, I found that some genes had differences in their mutation distribution pattern between these groups. The kurtosis values as a survey method do appear to be somewhat useful, but require careful analysis to interpret. My hypothesis regarding targeted mutations concentrating the in the low mutation group did not hold for all genes, but in the broad sense the expectation was still valid. Some of the genes known to be associated with colon adenocarcinoma possessed many mutations in known locations, however genes such as PTEN did not mutate in known locations at all and as predicted by my model, there were more passenger mutations in several of the genes that I examined closely in the high mutation group than there were in the low mutation group. At least one gene, BRAF, had its targeted mutations concentrated in the high mutation group, presumably because they offer these tumors some kind of selective benefit, and either do not affect other types of tumors or present a selective detriment.

Alternative approaches

A more ideal method would be to develop a sophisticated bioinformatics model to carry out this analysis. A statistical test for each site or for groupings of sites (if there are multiple effective spots, vs useless spots in the case of oncogenes) could be programmed, and each gene would require a designation as an oncogene (with an associated list of effective mutation sites, and types of mutations that are known achieve the tumor driving effects), or as a tumor suppressor, or a passenger (reasonable to conclude as uninvolved with tumor phenotype) or an unknown status (where not enough is known about a gene to make a determination). From there it could analyze the patterns of mutations found within the gene against expectation values to produce an ability to cluster samples according to similarity of mutation patterns, or even perhaps to set up a neural network to recognize these patterns in future data sets.

Once the profiles of these genes are better understood, a similar algorithm could be engineered to use these types of profiles to predict if a candidate gene is likely an oncogene or a suppressor, based on the distribution of mutations within population splits based on total mutation count in large datasets. A potential confounding factor is how to differentiate preferred sites based on deactivation from those that are activating mutations. The number of such sites within a gene tends to be much smaller for oncogenes, so the probability could be assumed to be inversely correlated with the number of codons affected, though genes with complicated folding pattern, or that participate in multi-protein complexes could have multiple locations throughout their sequences that participate in important enzymatic activities, ability of the protein to recognize and bind to substrates, influence folding and structure, or even influence the protein's ability to bind to a

multiprotein complex. My method of using kurtosis over the full set of mutation locations might be unable to properly signal peaks in genes where these preferred sites are more distant from one another. Perhaps a different statistic would be better, and perhaps computing kurtosis, or a more specialized statistic, using a sliding window of some number of nucleotides might prove to be a more powerful application of this idea to identify locations where there is a mutation peak. This is an important improvement to seek, because there may be cases where a range of nucleotides within a protein domain could be modified and achieve similar effects rather than just a single location of small handful of locations. For example, a gene that promotes cell survival or growth might be converted to being more active by blocking the binding of an inhibitor by changing its binding site via mutation. This could possibly be achieved by changing multiple nucleotides in the sequence of its binding domain, while retaining the overall structure of the protein and preserving its ability to carry out its biological activity. In such a case there wouldn't just be one peak, but rather a generally raised mutation rate over a wider genomic region. Whether kurtosis calculations would pick up on this pattern would depend partly on the size of the domain, how spread out the mutation sites are, and on the number and location of any passenger mutations exist in the population being studied, but perhaps a more sophisticated statistical model would be more sensitive and able to pick up on such a pattern more easily.

If protein structure data is available, perhaps the 3-dimensional distance between different amino acids and their associated codons could be used to determine if clusters of mutation peaks at different locations within the transcript are affecting the same relative physical location in a folded protein structure. Structure based algorithms would be

significantly more complex than simple sequence analysis, but could prove fruitful. Unfortunately, this would require data that is not always available for genes of interest, and would require an analytical method requiring a highly complex model and algorithm that is not currently available.

Chapter 5

Technical Difficulties Encountered During the Analysis

Technical Difficulties Encountered During the Analysis

Database Issues

There were some issues that I encountered during this project that were related to the way that the data was collected and formatted in the mutation database. When I was performing the length and mutation count analysis in chapter 3, I had to obtain gene lengths to use. I decided to use the transcript length since the MAF files did not include mutations in intronic regions. For genes with multiple transcript lengths I opted for a simple average. Since the MAF file was not making transcript distinctions, I would have had to obtain and parse through a list of alternatively spliced genes and their exon locations. This would have complicated the analysis quite a bit, and for a gain in accuracy that did not seem to be very significant.

I used the UCSC table browser to pull down a table containing the information I was looking for. However, when I tried to relate the information in the MAF files to the information in the gene tables from UCSC I quickly noticed that there was a problem. The gene symbols did not match for a number of the genes from the TCGA project. There had been changes to quite a few of them. I used a combination of methods including checking the current ID for the ENTREZ ID in the MAF file and checking databases of gene symbols for previous and new names. To parse all of this output, I tagged on an Updated Symbols column to the table. After doing this I was able to perform the length collection, and averaging where necessary.

This would have been made significantly easier if there were a TSV file available that I could have parsed through for old symbols and new symbols, but due to having to

cover for weaknesses and missed genes in each approach, it became a bit more difficult than it ought to have been.

Problems with the COSMIC database

I had looked at the COSMIC (Catalogue Of Somatic Mutations In Cancer) database, and noticed that there was a pair of columns in the table that included FATHMM score and FATHMM prediction. I looked up the FATHMM software, and its webpage states that it uses Hidden Markov models to score the probability that single nucleotide coding and noncoding variants have functional consequences. It has options for inherited diseases, cancer, and other disease specific options.

Intrigued by this, I set about writing a program to match up the entries from the MAF file to the COSMIC database. A few things became apparent fairly quickly. The first was that again, the gene symbols were not matching up, as had been the problem in pulling the transcript lengths out of the genome database table. I used a similar solution for the COSMIC file as I had used with the lengths. The next was ensuring that I was matching up the correct mutation. The COSMIC database encodes its mutations in CDS format. This format proved quite frustrating to read using programmed parsing, and eventually I gave up trying in favor of another choice. I noticed that the VCF format COSMIC files encoded the genomic location, in addition to the from and to nucleotide sequence changes.

Due to a few quirks of the VCF format, I had to work out some minor mismatches (under certain circumstances this format includes leading sequence homology of at least 1 base, and adjusts the sequences displayed and the indexes accordingly). It was still far simpler to do this text manipulation and simple math than to match up the CDS formatting.

To my disappointment, I found that the FATHMM results appeared to be purely structural. As an example, destructive mutations in known oncogenes were still being scored as pathogenic, rendering the results significantly less useful than they could have been.

There were also cases of multiple entries for the same genomic location with the same mutation. The COSMIC database is in need of some periodic manual or automated curation to identify entries that ought to be merged. Perhaps there is a reason these entries are being kept separated, such as being from tumors of different tissue origin, despite being the same mutation. I am not aware of what justifications there may be, only that it is not something that I had expected to find in such a large, ostensibly important database for genomic research. It would also be nice if they added additional columns that contained genomic coordinates and the “FROM” and “TO” DNA base characters rather than requiring that one decipher the CDS string or read the VCF files.

As a consequence of these technical problems, I chose not to pursue this line of analysis any further. Recall the classifications of tumorigenic genes that Vogelstein et. al produced. Some genes were classified as tumor suppressors, and would be expected to become tumorigenic via any disabling mutations, while others, called oncogenes, required specific activating mutations to become tumorigenic. One possible way to approach this problem with COSMIC classifications might be to try to use the Vogelstein classifications and the COSMIC predictions together to come up with a merged result. The problem in attempting to do this is that the one would need a list of gain-of-function mutations for each oncogene in order to know the difference between the destructive mutations and the gain-of-function mutations.

What might actually be better than trying to merge these results would be a more strictly curated sub-database specifically dedicated to tracking known functional mutations in cancer instead of all reported mutations. A database that could track the class of cancer gene would help in making predictions of mutation consequence more accurate and more comprehensive. This database could also usefully curate the classification of the mutation's effects, such as specifying gain-of-function, loss-of-function, loss of ability to respond to regulatory signals, loss of critical protein domains or other types of protein truncation, etc.

Chapter 6

Future Directions and Conclusion

Implications and Future Directions

In this last chapter, I would like to describe some future directions in which my thesis work could be extended. These include both novel analytical and functional avenues to examine the notion of differential targeting of genes based on mutation frequency in the individual tumors.

Obtain access to the rest of the Samples, and Germline data, and use SOLiD sequencing results

There were 54 samples in a separate MAF file that resulted from SOLiD sequencing. I did not think the benefit of the small increase in numbers would warrant the possible confounding influence by including these results so I did not utilize this file in this analysis. However, according to the GDC data portal there are now 463 colon tumor samples in the database. When I originally pulled the available data from the portal that the TCGA project had previously made available, data from some of these samples were not included. Perhaps analysis had not been completed on all of the samples, or perhaps not all of the samples were made publicly available without a security agreement at the time. Regardless of the reason, data from 193 samples was unavailable. In addition, none of the germline data was available to me, since neither I nor the UCHC had gone through the security agreement process in order to assure the TCGA that I will be able to adhere to minimum security requirements. The data for 458 of the samples, including the samples that were previously missing, has now become available from the GDC portal.

Before proceeding into other tissue types, the obvious follow-up to my work would be to go through the required steps to gain access to the germline mutations and to include these in the analysis. Additionally, the location of the mutation data for the

remaining 193 samples should be included. The germline mutations, when combined with expression data also available in the database, would give clues as to which samples were the result of heritable conditions like Lynch Syndrome or of sporadic mutation or repression of known repair genes. There are also 172 cases of rectal cancer with sample data in the database, of which 158 samples have nucleotide variation data available, that could reasonably be combined into an analysis of colorectal cancer as a whole.

Extend analysis into other tumor types typically associated with either MMR or other kinds of genomic instability.

A very common cause of tumors with very high mutation rates is a defect in one of the genes that contributes to the DNA mismatch repair (MMR) pathway. Heritable mutations in these genes lead to a condition known as hereditary nonpolyposis colorectal cancer, or Lynch Syndrome. Somatic defects in these genes, including mutations and aberrant epigenetic silencing, also occur.

Tumors of this type occur most commonly in colon and rectal tissue, but the heritable conditions also cause a significant rise in tumors of other tissue types, which would also have a chance of suffering sporadic mutations producing the same mutation rate increasing condition. Expanding this analysis to include other tissue types might be beneficial as it could help reveal the impact of tissue specificity of certain tumor genes, and would also increase the number of samples significantly. The obvious first choice would be to include the rectal tumor samples. With the larger dataset it would become more feasible to see if the raised mutation rate tumors all have the same mutation profile, or whether there are subtypes within the category. One possible distinction would be tumors with defects in MMR vs those that have a large number of mutations despite an

apparent lack of defect in the MMR system. Classifying tumors this way would require access to the germline mutation data to detect Lynch Syndrome, and would also require scanning the expression data for the genes known to play a part in the mismatch repair system. The colon tumors are already characterized in terms of microsatellite stability, which is another diagnostic clue, but it is possible that not all of the other tumors would have the results of such analysis included in their clinical data files.

Other tumors have been associated with Lynch Syndrome including gastric cancer, endometrial and ovarian cancer, prostate cancer, hepatocellular cancer, pancreatic cancer, urinary tract cancers, kidney and bile duct cancers and brain tumors. It would be interesting to compare the frequency and patterns of mutations in these other Lynch Syndrome-related tumors to the mutations that were found in the colon tumors. It is likely that the same pattern of high and low frequency mutations would be detected in these tumors and that like the colon tumors, there would be differences in the types of genes that are altered in the high and low frequency mutation tumors. It would be interesting to compare the mutations between these tumors and the colon tumors to see if MMR-related tumors show preferences for types of genes that are mutated. It would also be interesting to compare the types of genes mutated in the tumors with low numbers of mutations to see what common pathways occur in the non-MMR-related examples of these tumors. This would potentially be a way to discover common driver mutations that are associated with sporadic forms of these tumors.

Extend analysis into other tumor types not typically associated with MMR

Finally, it would also be interesting to look at cancers that are outside of the Lynch Syndrome cluster of tumors. Lung cancer, breast cancer, as well as the soft tissue and

bony sarcomas typically are not associated with Lynch Syndrome or mutations in the MMR pathway. An analysis of these tumors would be interesting both to see if there are differences in the frequency of mutations within the tumor samples for a particular cancer and whether those tumors would also show a difference in the types of genes that are mutated in the high and low frequency mutation groups.

Replace kurtosis with a more accurate scoring model

While kurtosis did turn out to be somewhat useful, it is also apparent that it has strong weaknesses. It is liable to being confounded by spread out mutations and multiple independent hot spots. While it could possibly detect clustering of mutations within a protein domain, this is both a good and a bad thing. It would be more ideal if I could differentiate what was contributing to the score. Peaks within specific codons and mutations clustering within a small genomic region are different phenomena that would be useful to identify.

As such, it would be ideal to construct a set of bioinformatics models to use for scoring these mutations that would identify these different conditions. Different models for detecting domain preference and specific codon preference would be needed. Use ontology data and known mutation information where appropriate and available. Ideally one would include gene ontology and classifications like whether a gene was an oncogene, tumor suppressor gene, or of unknown status, in the analysis of mutations. Then I could replace kurtosis with these scoring algorithms and gain a much more useful and specific result.

Try to develop self-clustering methods based on mutation patterns

Early on in this project I became interested in the potential application of self-directed clustering on these mutation patterns. One key problem is that most clustering algorithms depend on numerical distance scores, and much of the mutation data is Boolean in the sense of whether a mutation has occurred within a designated region, or with a matching base change if one got specific enough. Different models for scoring mutations between any two samples or any grouped set of samples and another sample could be developed that calculate distance values based on this information. One part of the algorithm could look at the mutation patterns in terms of the chemistry of which bases are changing to which other bases, and another module could examine for selective genes using information about the codons, known cancer mutations, and gene ontology and systems biology pathway information. This could aid in discovering patterns in the data that would be difficult to manually expose, and that might have biological or clinical significance.

Identify new driver genes based on outlier status in the analysis

One potential future extension of this analysis and expansions of it into other tissues, would be to use the results of this analysis in an effort to identify new tumor driver genes from the positively deviating outliers in the mutation versus length analysis. The kurtosis values or a replacement scoring algorithm might also provide additional clues in terms of identifying genes with mutational hot spots that may prove to be interesting candidates.

When these mutations at specific sites are identified as statistically common, one could then induce or engineer cell lines to have the same mutations and examine what

effects these mutations have on cell phenotype, regulatory pathways, and expression patterns.

Examine differences between selected genes and mutation incidence.

One of the potentially most interesting applications of this work might be to examine the differences between potential driver and passenger genes and the mutation incidence over time. Dr. Richard Lenski has been working on an ongoing long-term evolution experiment in E. coli (123). This project, that began in 1988, involves keeping 12 populations of E. coli bacteria in continuous culture, storing samples of the strains at regular intervals, and examining their genomes for mutations and changes in genotype frequency over time. These 12 populations were initially the same strain. They have since diverged and produced many new mutations and new traits. A similar experiment examining mutations in human cell lines rather than just the overall effects of selection under different culture conditions, might be useful to examine the effect of mutation rate on mutation incidence and patterns.

With the addition of more samples from other tumor types, it would become more feasible to try to differentiate between effects of selection on genes due to their benefit to cancer cell survival, and the raw patterns of mutation that are imposed by the regulatory or genetic conditions creating the higher mutation rates. Perhaps growing immortalized cells with these mutation rate increasing disorders in culture and examining the mutations that occur over time, similar to how the Lenski experiment in bacteria were performed, would enable researchers to observe mutations in the cell populations as they occur. The key difficulty in this is that if the mutations are not selective, they would either be neutral or deleterious, and in either case may not spread to many cells, so the ability to detect

these changes in potentially one or two of thousands of cells requires very low sequencing error, so that the detected base changes remain above minimal quality acceptance criteria. Some of these cell lines are already heavily mutated, so it might be better to start with the youngest stock cultures available, in terms of number of generations of cells since collection from the patient or derivation of the cell line.

To some extent this experiment has already been done, albeit in an much less controlled and methodical way than with the Lenksi E.coli project, with the long-term use of immortalized human cell cultures such as HeLa cells. Samples from various time points in the culturing of HeLa cells exist in labs all over the world, as does sequencing data for many of these samples. This existing data could be mined to reconstruct a picture of what has been happening, genomically, within these cells as they have been used in experiments over the decades they have been in use. This data could also be used in conjunction with a follow-up project keeping these cell lines in a more methodical experimental setup. Some of the cultures could be given a starting DNA repair deficit by deleting or mutating one or more of several known repair-related genes in such a way as to significantly raise unrepaired mutation rates. Then the results between the different cultures could be compared over time.

Concluding Thoughts

I have shown that there were differences in mutation pattern between tumors with a lot of mutations and those will smaller numbers of mutations. I have examined some of these differences in greater detail, and identified the genes that were most affected by the conditions that led to this difference in amount of total mutation. Some of these genes were clearly more affected due to size, but there also appeared to be a subset of cancer

related genes that were affected by selection in one population or the other. This phenomenon needs further study to determine causative mechanisms of these differences. It is also necessary to attempt to explain what is driving the genomic instability of the microsatellite stable tumors with very high mutation counts.

I have also highlighted the importance of determining whether a mutation that is structurally deleterious actually contributes to cancer. It is not sufficient for a gene to be known to be cancer related, or part of a known tumorigenic pathway, or to have mutations causing structural alterations, even destructive ones to be classified as a driver. For example, in oncogenes, which require an activating mutation to cause cancer, a deactivating mutation would simply be a passenger mutation. Many existing predictive analysis programs lack the ability to make this important distinction and assign misleading predictions based on gene ontology and/or structural predictions, but without consideration for which specific mutations are actually capable of being tumor drivers and which are not.

To this end there must be a greater effort to determine which genes appearing to be frequently mutated in cancer are actually capable of being cancer drivers, whether they are oncogenes or tumor suppressors, and where the activating mutations of oncogenes are located.

Bibliography

1. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature [Internet]. Nature Publishing Group; 2001 Feb 15 [cited 2016 Aug 15];409(6822):860–921. Available from: http://www.nature.com/doifinder/10.1038/35057062
2. Human Genome Sequencing Consortium I. Finishing the euchromatic sequence of the human genome. Nature [Internet]. Nature Publishing Group; 2004 Oct 21 [cited 2016 Aug 15];431(7011):931–45. Available from: http://www.nature.com/doifinder/10.1038/nature03001
3. Kolata G. SCIENTIST AT WORK: Francis S. Collins; Unlocking the Secrets of the Genome. The New York Times [Internet]. 1993; Available from: http://www.nytimes.com/1993/11/30/science/scientist-at-work-francis-s-collins-unlocking-the-secrets-of-the-genome.html?pagewanted=all
4. Belkin L. Splice Einstein and Sammy Glick. Add a Little Magellan. The New York Times [Internet]. 1998; Available from: http://www.nytimes.com/1998/08/23/magazine/splice-einstein-and-sammy-glick-add-a-little-magellan.html?pagewanted=all
5. Williams H, Cohen W, Doyle J, Fetter D, Gentzkow M, Murray F, et al. Intellectual property rights and innovation: Evidence from the human genome * the WEAI meetings, and IFPRI. Several individuals from Celera, the Human Genome Project, and related institutions provided invaluable guidance, including. 2010;
6. Celera Genomics Company Profile, Information, Business Description, History, Background Information on Celera Genomics [Internet]. Available from:

http://www.referenceforbusiness.com/history/Ca-Ch/Celera-Genomics.html

7. The Human Genome Project Race | Genomics Institute [Internet]. Available from: https://genomics.soe.ucsc.edu/research/hgp_race
8. Joint Statement - President Clinton & Prime Minister Blair [Internet]. 2000. Available from: https://clinton4.nara.gov/WH/EOP/OSTP/html/00314.html
9. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science [Internet]. American Association for the Advancement of Science; 2001 Feb 16 [cited 2016 Aug 15];291(5507):1304–51. Available from: http://www.ncbi.nlm.nih.gov/pubmed/11181995
10. Release: International Consortium Completes HGP [Internet]. 2003. Available from: https://www.genome.gov/11006929/2003-release-international-consortium-completes-hgp/
11. Genome Reference Consortium [Internet]. Available from: http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/
12. França LTC, Carrilho E, Kist TBL. A review of DNA sequencing techniques. Q Rev Biophys. 2002;35(2):169–200.
13. Gužvić M. The History of DNA Sequencing / ISTORIJAT SEKVENCIRANJA DNK. J Med Biochem [Internet]. De Gruyter Open; 2013;32(4):301–12. Available from: http://10.2478/jomb-2014-0004\nhttp://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=91556396&site=ehost-live
14. Sequencing Platforms Archives - AllSeq [Internet]. Available from: http://allseq.com/knowledge-bank/kb-category/sequencing-platforms/

15. Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM, editors. Bioinformatics for High Throughput Sequencing [Internet]. New York, NY: Springer New York; 2012 [cited 2016 Aug 15]. Available from: http://link.springer.com/10.1007/978-1-4614-0782-9

16. França LTC, Carrilho E, Kist TBL. A review of DNA sequencing techniques. Q Rev Biophys. Cambridge University Press; 2002;35(2):169–200.

17. Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A [Internet]. 2009;106(45):19096–101. Available from: http://www.pnas.org/content/106/45/19096.full

18. Bashiardes S, Veile R, Helms C, Mardis ER, Bowcock AM, Lovett M. Direct genomic selection. Nat Methods. 2005;2(1).

19. Jiang Z, Zhou X, Li R, Michal JJ, Zhang S, Dodson M V., et al. Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. Cell Mol Life Sci [Internet]. Springer Basel; 2015 Sep 28 [cited 2016 Aug 15];72(18):3425–39. Available from: http://link.springer.com/10.1007/s00018-015-1934-y

20. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly. Annu Rev Genomics Hum Genet. 2015;16:153–72.

21. Staden R. A strategy of DNA sequencing employing computer programs. Nucleic Acids Res [Internet]. 1979 [cited 2016 Aug 15];6(7):2601–10. Available from: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/6.7.2601

22. Sequencing Systems | Sequencer Comparison Table [Internet]. Available from:

http://www.illumina.com/systems/sequencing.html

23. Cancers Selected for Study - TCGA [Internet]. Available from: http://cancergenome.nih.gov/cancersselected

24. Hanahan D, Weinberg RA, Adams JM, Cory S, Aguirre-Ghiso JA, Ahmed Z, et al. Hallmarks of Cancer: The Next Generation. Cell [Internet]. Elsevier; 2011 Mar [cited 2016 Aug 15];144(5):646–74. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0092867411001279

25. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, et al. Genetic Alterations during Colorectal-Tumor Development. N Engl J Med [Internet]. Massachusetts Medical Society ; 1988 Sep [cited 2016 Aug 15];319(9):525–32. Available from: http://www.nejm.org/doi/abs/10.1056/NEJM198809013190901

26. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell [Internet]. Cell Press; 1990 Jun [cited 2016 Aug 15];61(5):759–67. Available from: http://linkinghub.elsevier.com/retrieve/pii/009286749090186I

27. Vogelstein B, Kinzler KW. The multistep nature of cancer. Trends Genet [Internet]. Elsevier Current Trends; 1993 Apr [cited 2016 Aug 15];9(4):138–41. Available from: http://linkinghub.elsevier.com/retrieve/pii/016895259390209Z

28. Ragusa M, Barbagallo C, Statello L, Condorelli AG, Battaglia R, Tamburello L, et al. Non-coding landscapes of colorectal cancer. World J Gastroenterol [Internet]. Baishideng Publishing Group Inc; 2015 Nov 7 [cited 2016 Aug 15];21(41):11709–39. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26556998

29. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, et al. The

consensus coding sequences of human breast and colorectal cancers. Science [Internet]. 2006 Oct 13 [cited 2014 Jul 19];314(5797):268–74. Available from: http://www.sciencemag.org/content/314/5797/268.full

30. Wood LD, Parsons DW, Jones S, Lin J, Sjöblom T, Leary RJ, et al. The genomic landscapes of human breast and colorectal cancers. Science. 2007;318(5853):1108–13.

31. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Kinzler KW, et al. Cancer genome landscapes. Science [Internet]. American Association for the Advancement of Science; 2013 Mar 29 [cited 2016 Aug 15];339(6127):1546–58. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23539594

32. Vogelstein B, Kinzler KW. The Path to Cancer — Three Strikes and You're Out. N Engl J Med [Internet]. Massachusetts Medical Society; 2015 Nov 12 [cited 2016 Aug 15];373(20):1895–8. Available from: http://www.nejm.org/doi/10.1056/NEJMp1508811

33. Bauer B, Siebert R, Traulsen A. Cancer initiation with epistatic interactions between driver and passenger mutations. J Theor Biol. 2014;358:52–60.

34. Lynch HT, Lynch PM, Lanspa SJ, Snyder CL, Lynch JF, Boland CR. Review of the Lynch syndrome: history, molecular genetics, screening, differential diagnosis, and medicolegal ramifications. Clin Genet [Internet]. NIH Public Access; 2009 Jul [cited 2016 Aug 15];76(1):1–18. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19659756

35. Poulogiannis G, Frayling IM, Arends MJ. DNA mismatch repair deficiency in sporadic colorectal cancer and Lynch syndrome. Histopathology [Internet].

Blackwell Publishing Ltd; 2010 Jan [cited 2016 Aug 15];56(2):167–79. Available from: http://doi.wiley.com/10.1111/j.1365-2559.2009.03392.x

36. Horvat M, Stabuc B. Microsatellite instability in colorectal cancer. Radiol Oncol. 2011;45(2):75–81.

37. Timmermann B, Kerick M, Roehr C, Fischer A, Isau M, Boerno ST, et al. Somatic mutation profiles of MSI and MSS colorectal cancer identified by whole exome next generation sequencing and bioinformatics analysis. PLoS One. 2010;5(12):1–10.

38. What are the key statistics about colorectal cancer? [Internet]. [cited 2015 Apr 16]. Available from: http://www.cancer.org/cancer/colonandrectumcancer/detailedguide/colorectal-cancer-key-statistics

39. Zaanan A, Meunier K, Sangar F, Fléjou J-F, Praz F. Microsatellite instability in colorectal cancer: from molecular oncogenic mechanisms to clinical implications. Cell Oncol (Dordr). 2011;34(3):155–76.

40. Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. Nature [Internet]. Nature Research; 2015 Feb 23 [cited 2016 Aug 16];521(7550):81–4. Available from: http://www.nature.com/doifinder/10.1038/nature14173

41. de las Alas MM, Aebi S, Fink D, Howell SB, Los G. Loss of DNA mismatch repair: effects on the rate of mutation to drug resistance. J Natl Cancer Inst [Internet]. 1997 Oct 15 [cited 2016 Aug 16];89(20):1537–41. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9337351

42. Lengauer C, Kinzler KW, Vogelstein B. Genetic instability in colorectal cancers. Nature [Internet]. 1997 Apr 10 [cited 2015 Mar 14];386(6625):623–7. Available from: http://dx.doi.org/10.1038/386623a0

43. Li SKH, Martin A. Mismatch Repair and Colon Cancer: Mechanisms and Therapies Explored. Trends Mol Med. 2016;22(4):274–89.

44. Torgovnick A, Schumacher B. DNA repair mechanisms in cancer development and therapy. Front Genet [Internet]. Frontiers; 2015 Apr 23 [cited 2016 Aug 16];6:157. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2015.00157/abstract

45. Richman S. Deficient mismatch repair: Read all about it (Review). Int J Oncol. Spandidos Publications; 2015;47(4):1189–202.

46. Zhang T, Boswell EL, McCall SJ, Hsu DS. Mismatch repair gone awry: Management of Lynch syndrome. Crit Rev Oncol Hematol. 2015;93(3):170–9.

47. About TCGA - TCGA [Internet]. Available from: http://cancergenome.nih.gov/abouttcga

48. Peltomaki P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum Mol Genet [Internet]. 2001;10(7):735–40. Available from: http://hmg.oxfordjournals.org/content/10/7/735.abstract

49. Grady WM. Genomic instability and colon cancer. Cancer Metastasis Rev [Internet]. 2004;23(1-2):11–27. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15000146

50. Sieber OM, Heinimann K, Tomlinson IPM. Genomic instability — the engine of tumorigenesis? Nat Rev Cancer. 2003;3(9):701–8.

51. Ashktorab H, Schäffer A a., Daremipouran M, Smoot DT, Lee E, Brim H. Distinct genetic alterations in colorectal cancer. PLoS One. 2010;5(1).

52. Hampel H, Frankel WL, Martin E, Arnold M, Khanduja K, Kuebler P, et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). N Engl J Med [Internet]. 2005 May 5 [cited 2015 Mar 22];352(18):1851–60. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15872200

53. Lynch HT, de la Chapelle a. Genetic susceptibility to non-polyposis colorectal cancer. J Med Genet. 1999;36(11):801–18.

54. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med [Internet]. 2003 Mar 6 [cited 2015 Mar 16];348(10):919–32. Available from: http://www.ncbi.nlm.nih.gov/pubmed/12621137

55. Liu B, Nicolaides NC, Markowitz S, Willson JK, Parsons RE, Jen J, et al. Mismatch repair gene defects in sporadic colorectal cancers with microsatellite instability. Vol. 9, Nature genetics. 1995. p. 48–55.

56. Swede H, Bartos JD, Chen N, Shaukat A, Dutt SS, McQuaid D a., et al. Genomic profiles of colorectal cancers differ based on patient smoking status. Cancer Genet Cytogenet. 2006;168(2):98–104.

57. Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci U S A [Internet]. National Academy of Sciences; 1996 Dec 10 [cited 2016 Aug 16];93(25):14800–3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/8962135

58. Heng HH. Debating cancer : the paradox in cancer research. World Scientific Publishing Company; 2015. 464 p.

59. Pennisi E. Genomics. ENCODE project writes eulogy for junk DNA. Science [Internet]. American Association for the Advancement of Science; 2012 Sep 7 [cited 2016 Aug 17];337(6099):1159, 1161. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22955811

60. Fractals and Chaos Theory [Internet]. Available from: http://faculty.mc3.edu/cvaughen/chaos/index.html

61. Green E. Table of Contents [Internet]. 1998. Available from: http://pages.cs.wisc.edu/~ergreen/honors_thesis/contents.html

62. Lanius C. Cynthia Lanius' Lessons: A Fractals Lesson - Introduction [Internet]. 1996. Available from: http://math.rice.edu/~lanius/frac/

63. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res [Internet]. 2009 Jan [cited 2016 Aug 16];37(1):1–13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19033363

64. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc [Internet]. 2009 [cited 2016 Aug 16];4(1):44–57. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19131956

65. AJCC - Desk References [Internet]. Available from: https://cancerstaging.org/references-tools/deskreferences/Pages/default.aspx

66. Fedier A, Fink D. Mutations in DNA mismatch repair genes: Implications for DNA damage signaling and drug sensitivity (Review). Int J Oncol. Spandidos Publications; 2004;24(4):1039–47.

67. Silva FCC da, Valentin MD, Ferreira F de O, Carraro DM, Rossi BM. Mismatch repair genes in Lynch syndrome: a review. Sao Paulo Med J [Internet]. Associação Paulista de Medicina; 2009 Jan [cited 2016 Aug 16];127(1):46–51. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1516-31802009000100010&lng=en&nrm=iso&tlng=en

68. Peltomäki P. DNA mismatch repair gene mutations in human cancer. Environ Health Perspect [Internet]. National Institute of Environmental Health Science; 1997 Jun [cited 2016 Aug 16];(Suppl 4):775–80. Available from: http://www.ncbi.nlm.nih.gov/pubmed/9255561

69. Li G-M. Mechanisms and functions of DNA mismatch repair. Cell Res [Internet]. Nature Publishing Group; 2008 Jan [cited 2016 Aug 16];18(1):85–98. Available from: http://www.nature.com/doifinder/10.1038/cr.2007.115

70. Linkowska K, Jawień A, Marszałek A, Skonieczna K, Grzybowski T. Searching for association of the CAG repeat polymorphism in the mitochondrial DNA polymerase gamma gene (POLG) with colorectal cancer. Acta Biochim Pol. 2015;62(3):625–7.

71. Linkowska K, Jawień A, Marszałek A, Malyarchuk BA, Tońska K, Bartnik E, et al. Mitochondrial DNA Polymerase γ Mutations and Their Implications in mtDNA Alterations in Colorectal Cancer. Ann Hum Genet [Internet]. 2015 Sep [cited 2016 Aug 16];79(5):320–8. Available from: http://doi.wiley.com/10.1111/ahg.12111

72. Boland CR, Goel A. Microsatellite Instability in Colorectal Cancer. Gastroenterology [Internet]. Elsevier Inc.; 2010;138(6):2073–87.e3. Available from: http://dx.doi.org/10.1053/j.gastro.2009.12.064

73. Hatch SB, Lightfoot HM, Garwacki CP, Moore DT, Calvo BF, Woosley JT, et al. Microsatellite instability testing in colorectal carcinoma: Choice of markers affects sensitivity of detection of mismatch repair-deficient tumors. Clin Cancer Res. 2005;11(6):2180–7.

74. Scherer S. Gene Structure [Internet]. 2010. Available from: http://www.cshlp.org/ghg5_all/section/gene.shtml

75. Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature [Internet]. Nature Research; 2010 May 27 [cited 2016 Aug 16];465(7297):473–7. Available from: http://www.nature.com/doifinder/10.1038/nature09004

76. Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature [Internet]. Nature Research; 2013 Jun 16 [cited 2016 Aug 16];499(7457):214–8. Available from: http://www.nature.com/doifinder/10.1038/nature12213

77. Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature [Internet]. Nature Research; 2012 Jul 22 [cited 2016 Aug 16];488(7412):504–7. Available from: http://www.nature.com/doifinder/10.1038/nature11273

78. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, et al. Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res [Internet]. Cold Spring Harbor Laboratory Press; 2010

Apr 1 [cited 2016 Aug 16];20(4):447–57. Available from: http://genome.cshlp.org/cgi/doi/10.1101/gr.098947.109

79. Sima J, Gilbert DM. Complex correlations: replication timing and mutational landscapes during cancer and genome evolution. Curr Opin Genet Dev. 2014;25:93–100.

80. Stamatoyannopoulos JA, Adzhubei I, Thurman RE, Kryukov G V, Mirkin SM, Sunyaev SR. Human mutation rate associated with DNA replication timing. Nat Genet [Internet]. 2009 [cited 2016 Aug 16];41(4):393–5. Available from: http://www.nature.com/doifinder/10.1038/ng.363\nhttp://www.ncbi.nlm.nih.gov/pubmed/19287383\nhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC2914101

81. Waters LS, Walker GC. The critical mutagenic translesion DNA polymerase Rev1 is highly expressed during G2/M phase rather than S phase. Proc Natl Acad Sci [Internet]. National Academy of Sciences; 2006 Jun 13 [cited 2016 Aug 16];103(24):8971–6. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0510167103

82. Liu L, De S, Michor F. DNA replication timing and higher-order nuclear organization determine single-nucleotide substitution patterns in cancer genomes. Nat Commun [Internet]. Nature Research; 2013 Feb 19 [cited 2016 Aug 16];4:1502. Available from: http://www.nature.com/doifinder/10.1038/ncomms2502

83. Swami M. Mutation: It’s the CpG content that counts. Nat Rev Genet [Internet]. Nature Publishing Group; 2010 Jun 8 [cited 2016 Aug 16];11(7):456–7. Available

from: http://www.nature.com/doifinder/10.1038/nrg2820

84. Harris RS, Lawrence M, Stojanov P, Polak P, Kryukov G, Cibulskis K, et al. Cancer mutation signatures, DNA damage mechanisms, and potential clinical implications. Genome Med [Internet]. BioMed Central; 2013 [cited 2016 Aug 16];5(9):87. Available from: http://genomemedicine.biomedcentral.com/articles/10.1186/gm490

85. Clancy S (2008) DD& RM for MDI. DNA Damage & Repair: Mechanisms for Maintaining DNA Integrity. Nat Educ [Internet]. 2008;1(1):103. Available from: http://www.nature.com/scitable/topicpage/dna-damage-repair-mechanisms-for-maintaining-dna-344

86. Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. Nat Rev Cancer [Internet]. NIH Public Access; 2014 Dec [cited 2016 Aug 16];14(12):786–800. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25568919

87. Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. Proto-Oncogenes and Tumor-Suppressor Genes. In: Molecular Cell Biology 4th edition [Internet]. 4th ed. New York: W. H. Freeman; 2000 [cited 2016 Aug 16]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK21662/

88. McClean P. The Eukaryotic Cell Cycle and Cancer [Internet]. 1997. Available from: https://www.ndsu.edu/pubweb/~mcclean/plsc431/cellcycle/cellcycl5.htm

89. Runkel P. Why Kurtosis is Like Liposuction. And Why it Matters. [Internet]. 2014. Available from: http://blog.minitab.com/blog/statistics-and-quality-data-analysis/why-kurtosis-is-like-liposuction-and-why-it-matters

90. How skewness and kurtosis affect your distribution.

91. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. Int J Endocrinol Metab [Internet]. Kowsar Medical Institute; 2012 [cited 2016 Aug 16];10(2):486–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23843808

92. Kumar K, Brim H, Giardiello F, Smoot DT, Nouraie M, Lee EL, et al. Distinct BRAF (V600E) and KRAS mutations in high microsatellite instability sporadic colorectal cancer in African Americans. Clin Cancer Res [Internet]. 2009 Feb 15 [cited 2015 Apr 14];15(4):1155–61. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2713502&tool=pmcentrez&rendertype=abstract

93. Brim H, Mokarram P, Naghibalhossaini F, Saberi-Firoozi M, Al-Mandhari M, Al-Mawaly K, et al. Impact of BRAF, MLH1 on the incidence of microsatellite instability high colorectal cancer in populations based study. Mol Cancer. 2008;7:68.

94. Thiel A, Heinonen M, Kantonen J, Gylling A, Lahtinen L, Korhonen M, et al. BRAF mutation in sporadic colorectal cancer and Lynch syndrome. Virchows Arch [Internet]. Springer Berlin Heidelberg; 2013 Nov 21 [cited 2016 Aug 15];463(5):613–21. Available from: http://link.springer.com/10.1007/s00428-013-1470-9

95. Aoki K, Taketo MM. Adenomatous polyposis coli (APC): a multi-functional tumor suppressor gene. J Cell Sci [Internet]. The Company of Biologists Ltd; 2007 Oct 1 [cited 2016 Aug 17];120(Pt 19):3327–35. Available from:

http://www.ncbi.nlm.nih.gov/pubmed/17881494

96. Gonsalves WI, Mahoney MR, Sargent DJ, Nelson GD, Alberts SR, Sinicrope FA, et al. Patient and tumor characteristics and BRAF and KRAS mutations in colon cancer, NCCTG/Alliance N0147. J Natl Cancer Inst [Internet]. Oxford University Press; 2014 Jul [cited 2016 Aug 17];106(7). Available from: http://www.ncbi.nlm.nih.gov/pubmed/24925349

97. Hutchins G, Southward K, Handley K, Magill L, Beaumont C, Stahlschmidt J, et al. Value of Mismatch Repair, KRAS, and BRAF Mutations in Predicting Recurrence and Benefits From Chemotherapy in Colorectal Cancer. J Clin Oncol [Internet]. American Society of Clinical Oncology; 2011 Apr 1 [cited 2016 Aug 17];29(10):1261–70. Available from: http://jco.ascopubs.org/cgi/doi/10.1200/JCO.2010.30.1366

98. Yoon HH, Tougeron D, Shi Q, Alberts SR, Mahoney MR, Nelson GD, et al. KRAS codon 12 and 13 mutations in relation to disease-free survival in BRAF-wild-type stage III colon cancers from an adjuvant chemotherapy trial (N0147 alliance). Clin Cancer Res [Internet]. Clinical Cancer Research; 2014 Jun 1 [cited 2016 Aug 17];20(11):3033–43. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24687927

99. COSMIC: Gene analysis - KRAS [Internet]. Available from: http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=KRAS

100. Wang H, Lian P, Zheng P-S, Wang H, Lian P, Zheng P-S. SOX9, a potential tumor suppressor in cervical cancer, transactivates p21 WAF1/CIP1 and suppresses cervical tumor growth. Oncotarget. Impact Journals;

2015;6(24):20711–22.

101. Prévostel C, Rammah-Bouazza C, Trauchessec H, Canterel-Thouennon L, Busson M, Ychou M, et al. SOX9 is an atypical intestinal tumor suppressor controlling the oncogenic Wnt/ß-catenin signaling. Oncotarget. Impact Journals; 2016;5(0).
102. Molecular Profiling of Colorectal Cancer - My Cancer Genome [Internet]. Available from: https://mycancergenome.org/content/disease/colorectal-cancer/
103. Cathomas G. PIK3CA in Colorectal Cancer. Front Oncol [Internet]. Frontiers Media SA; 2014 [cited 2016 Aug 17];4:35. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24624362
104. Naguib A, Cooke JC, Happerfield L, Kerr L, Gay LJ, Luben RN, et al. Alterations in PTEN and PIK3CA in colorectal cancers in the EPIC Norfolk study: associations with clinicopathological and dietary factors. BMC Cancer [Internet]. BioMed Central; 2011 Dec 7 [cited 2016 Aug 17];11(1):123. Available from: http://bmccancer.biomedcentral.com/articles/10.1186/1471-2407-11-123
105. Nassif NT, Lobo GP, Wu X, Henderson CJ, Morrison CD, Eng C, et al. PTEN mutations are common in sporadic microsatellite stable colorectal cancer. Oncogene [Internet]. Nature Publishing Group; 2004 Jan 15 [cited 2016 Aug 17];23(2):617–28. Available from: http://www.nature.com/doifinder/10.1038/sj.onc.1207059
106. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nat Struct Mol Biol [Internet]. Nature Publishing Group; 2009 Feb [cited 2016 Aug 17];16(2):107–13. Available from:

http://www.nature.com/doifinder/10.1038/nsmb.1550

107. Rogers MA, Langbein L, Winter H, Ehmann C, Praetzel S, Korn B, et al. Characterization of a Cluster of Human High/Ultrahigh Sulfur Keratin-associated Protein Genes Embedded in the Type I Keratin Gene Domain on Chromosome 17q12-21. J Biol Chem [Internet]. American Society for Biochemistry and Molecular Biology; 2001 May 25 [cited 2016 Aug 17];276(22):19440–51. Available from: http://www.jbc.org/cgi/doi/10.1074/jbc.M100657200
108. KRTAP4-3 Gene - GeneCards [Internet]. [cited 2016 Jan 8]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=KRTAP4-3
109. KRTAP4-5 Gene - GeneCards [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=KRTAP4-5
110. COSMIC: Gene analysis - FAM194B [Internet]. Available from: http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=FAM194B
111. ERICH6B Gene - GeneCards [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=ERICH6B
112. DSPP Gene - GeneCards [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=DSPP
113. TAWFIK A, EDEH N, HSU S, DICKINSON D, OGBUREKE KU. DSPP upregulation in oral squamous cell carcinoma suppresses caspase-14 expression. 2010;
114. MacDougall M, Simmons D, Luan X, Nydegger J, Feng J, Gu TT. Dentin Phosphoprotein and Dentin Sialoprotein Are Cleavage Products Expressed from a Single Transcript Coded by a Gene on Human Chromosome 4: DENTIN

PHOSPHOPROTEIN DNA SEQUENCE DETERMINATION. J Biol Chem [Internet]. American Society for Biochemistry and Molecular Biology; 1997 Jan 10 [cited 2016 Aug 17];272(2):835–42. Available from: http://www.jbc.org/cgi/doi/10.1074/jbc.272.2.835

115. Cancer T, Atlas G. Integrated genomic analyses of ovarian carcinoma. Nature. 2011;474(7353):609–15.
116. Zilfou JT, Lowe SW. Tumor suppressive functions of p53. Cold Spring Harb Perspect Biol [Internet]. Cold Spring Harbor Laboratory Press; 2009 Nov [cited 2016 Aug 17];1(5):a001883. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20066118
117. Bieging KT, Mello SS, Attardi LD. Unravelling mechanisms of p53-mediated tumour suppression. Nat Rev Cancer [Internet]. Nature Research; 2014 Apr 17 [cited 2016 Aug 17];14(5):359–70. Available from: http://www.nature.com/doifinder/10.1038/nrc3711
118. KRAS Gene - GeneCard [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=KRAS
119. BRAF Gene - GeneCards [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=braf
120. TP53 Gene - GeneCards [Internet]. Available from: http://www.genecards.org/cgi-bin/carddisp.pl?gene=tp53
121. Berens EB, Sharif GM, Schmidt MO, Yan G, Shuptrine CW, Weiner LM, et al. Keratin-associated protein 5-5 controls cytoskeletal function and cancer cell vascular invasion. Oncogene [Internet]. Nature Publishing Group; 2016 Jul 4

[cited 2016 Aug 17]; Available from: http://www.nature.com/doifinder/10.1038/onc.2016.234

122. COSMIC: Gene analysis - DSPP [Internet]. Available from: http://cancer.sanger.ac.uk/cosmic/gene/analysis?ln=DSPP

123. Tenaillon O, Barrick JE, Ribeck N, Deatherage DE, Blanchard JL, Dasgupta A, et al. Tempo and mode of genome evolution in a 50,000-generation experiment. Nature [Internet]. Nature Research; 2016 Aug 1 [cited 2016 Aug 12];536(7615):165–70. Available from: http://www.nature.com/doifinder/10.1038/nature18959