Analysis of Carcinogenic Mutations on Chromosome 1 Using Probability Functions

Timothy Hedspeth

May 6^{th} 2021

Emmanuel College
Department of Mathematics

Abstract. Cancer is an evolutionary process, one that relies upon mutations at the cellular level. The utilization of single cell sequencing provides researchers the opportunity to examine the mutational landscape of populations of cancer cells. This project utilizes single cell sequencing data from cancer patients to explore the existence of mutations on chromosome 1 for individuals with two types of cancer. In addition to the exploration of the data, this project explores the application of probability functions presented by Petti et al. (2019) and by Anandakrishnan et al. (2019). In both instances the application of these functions to our chromosome 1 data yields results that are unexpected.

1 Introduction

Cancer is a horrible disease, one that touches all of us in one way or another through the course of our lives. Cancer is a complex condition, one that arises through a multitude of biological events. Cancer starts at the cellaur level, where mutations occur causing cells to lose their function resulting in unchecked cell proliferation. In the study of genetics there are many forms of mutations, for the purpose of our project we will look at changes to the known and expected alleles, which are refereed to as wild type. When these wild type alleles are replaced with a different allele, this is referred to as a variant. The scope of this research is to examine data complied in regard to point mutations, which are mutations that occur at a single base pair in the chromosome. In order to detect these mutations biologists sequenced the RNA in cells, and utilized unique molecular identifiers (UMIs) to quantify wild type and variant reads in a large sample [1].

In this paper we analyze read data from two types of cancer. The more abundant cancer type in our sample is Chronic Lymphotic leukemia, which is abbreviated as CLL. Leukemia is a class of cancer that originates in an individual's bone marrow [2]. CLL is the most common leukemia in adults, and grows slowly [2]. An individual can live with CLL for many years without being diagnosed and the cancer tends to be detected only after it metastasizes [2]. The other cancer type we have in our data is Diffuse Large B-cell Lymphoma, abbreviated as DLBL in this paper. This is an aggressive form of non-Hodgkin lymphoma and usually manifests as swelling of the lymph nodes [3]. Due to DLBL's aggressive nature people with this cancer usually require treatment quickly [3].

Single cell sequencing is a new field dating back approximately 10 years [4]. There are three methods of obtaining single cells, which are nanowells, droplets, and valves [5]. Once cells are captured, biologists are able to sequence multiple strands of RNA from the cells allowing them to see the distribution of mutations across the exon regions of the genome [5]. This sequencing allows for the derivation of patterns that could be helpful in understanding how genotypes of cells inform cancerous phenotypes[5]. The first goal of this project is to explore patterns of read coverage across chromosome 1 and use statistical analysis to verify the patterns that we observed.

Amplification bias, a phenomenon in sequencing of DNA or RNA that causes certain regions to have more read coverage than others, is often observed [5]. In particular the regions closer to the ends of the sequences, in our case the ends of chromosome 1, are expected to have more reads [5]. The sequencing that was conducted for our study utilized Unique Molecular Identifier, UMI for short [1]. UMIs act as identifier for short spans of nucleotide and are a tool that biologists use to attempt to reduce amplification bias [1]. A hypothesis test was run to see if UMIs helped reduce the amplification bias for our data.

The data we used was derived from strands of RNA in cells, which should only contain data that belongs to exon regions of DNA. This is due to the fact that intron regions are spliced out when DNA is transcribed into RNA. However,

the data had a few intron regions present, which according to Oikkenko and Lise arises in RNA data due to the fact that many biology tools are designed for DNA and not RNA [6]. Thus there was some level of error to be expected in our analysis as we should expect some data to map incorrectly [6]. To account for this potential error an R script that checks our basepair positions against the known the human genome browser is created to see how many basepairs were incorrectly mapped [7] [8].

We base our primary investigations on two recent publication by Petti et al. (2019) and by Anandakrishnan et al. (2019). The authors of the first paper sought to compute the probability of detection for a given heterozygous mutation from single cell sequencing data [9]. In order to compute the probability the authors created a function with a variety of inputs [9]. Through past analysis done by the researchers, it is known that the function presented by Petti does not always yield values that are consistent with a valid probability function, and thus we explored what happens with our chromosome 1 data.

The authors of the second paper seek to examine how many mutations are needed to cause carcinogenesis, which is the development of cancer [10]. Their approach was based on utilizing the number of accumulated somatic mutations and examining how this impacts incidences of cancer [10]. The model they developed is groundbreaking as they were some of the first to use mutations as the primary predictor of cancer, as most other models utilized an individual's age as the primary indication factor in cancer development [10]. We utilized this model with our data to see if we can accurately predict the distribution of mutations for our data.

2 Methods

2.1 Data Collection and Cleaning

All our data came from the 2020 DIMACS REU program. Data sets were obtained from single cell sequencing experiments with individuals that had one of two types of cancer. The first type is Chronic Lymphocytic Leukemia, and the second type is Diffuse large B-cell lymphoma. The data consists of 7637 cells sequenced for 39515 basepair positions. Eighty percent of the cells came from patients with Chronic Lymphocytic Leukemia, and twenty percent came from the patients with Diffuse large B-cell lymphoma. The data was in the form (Number of total UMI reads; Percent of reads that are UMI). We derived two data sets, one that contained the total amount of reads per cell and base pair, and one that contained the mutant reads. Once these data values were loaded into R we removed all rows that have no read coverage for our total reads [7].

2.2 Analysis

2.2.1 Descriptive Statistics

All data descriptions were done in R language [7]. We computed the total number of reads per base pair, by summing all of the rows for both the mutant and total data sets. The number of reads in each individual cell was computed by summing the columns of both the mutant and total data sets. Histograms and boxplots were created to give a visual representation of the distribution of the number of reads per base pair and reads per cell data. Since our reads per base pair were highly skewed to the right, data was split for the number of reads less than 100 reads, between 100 and 1000 reads, and greater than 1000. These categories were determined by looking at the distribution of data for each class, and the boundaries that gave the best visualization were selected. We then created histograms and boxplots for reads per sample. Along with these histograms and boxplots the mean, standard deviation, and five number summary were calculated.

We divided the number of mutant reads by the number of total reads to find percent of variants reads (VAF) for each base pair, and for each sample. For the reads per base pair it was observed that there were many VAF that had a value of 1, which caused an extreme skew of the distribution. With this in mind the data set was reduced to not include samples with a VAF of 1. The percent of reads per mutated cell was shown to have minimal variability and a histogram and boxplot were constructed. We noted that since the sequencing conducted was not looking for mutations in the Diffuse large B-cell lymphoma group, we expected these cells to have lower percentages of mutant reads per cell than those that are Chronic Lymphocytic Leukemia. With this in mind the samples were split with the lowest twenty percent of cells in terms of percent mutant reads were placed into the DLBL group and the rest were placed into a CLL group. From these groups we created histograms and boxplots to see the distribution of data for assumed cancer types.

2.2.2 Hypothesis Testing

All statistical tests were conducted with an $\alpha = .05$ level of significance. We used a 1-sample wilcoxon rank sum test to compare our average VAF that we calculated against a benchmark value. We also tested data regarding amplification bias for basepairs: we split our data set into two equal groups, the start and end basepairs for the chromosome and the middle. Following this we did a wilcoxon rank sum test and an F test for the comparison of median and variance of these groups.

We were interested to see if there is any difference between our assumed cell types, for mutant read coverage. In order to test this we ran a wilcoxon rank sum test for comparison of the medians and an F test for the equality of variance for these groups.

2.2.3 Probability function, Petti et al.

The Petti et al. paper presented an equation that the authors utilized to calculate the probability a specific heterozygous mutation occurs in a sample [9]. The equation they used is defined as [9]:

$$P(m) = na(f[1 - (1 - ct)^r] + (1 - f)e) \tag{1}$$

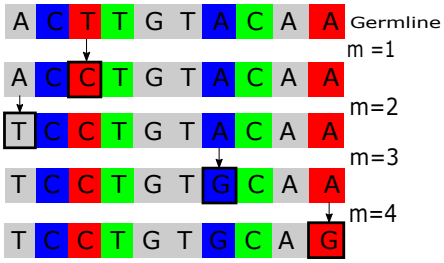
Where n is the number of cells in the sample,
a the percent of cells that are mutant,
f is twice the variant allele frequency due to the fact that RNA is single stranded and the mutation could have occurred and was not transcribed into the RNA,
c is the percent of cells that have coverage at the given basepair,
t is the expression level of the gene,
r is the average number of mutant UMIs per cell,
and e is the site specific error rate which is the frequency that a wild type read is called a mutant [9].

The terms of equation (1) can be better understood when broken down. The authors first calculate the probability that the mutant is expressed identified by using the complement rule they found: $P(\text{at least one of } r \text{ mutants is expressed}) = 1 - P(\text{none of } r \text{ mutants are expressed}) = [1 - (1 - ct)^r]$ [9]. The probability that the mutant was expressed is then multiplied by f, as this is the frequency at which the mutation occurs in the cell [9]. This term computes the probability of correctly identifying a mutant for a cell in the sample [9].

They also took into account potential error in which a wild type allele is incorrectly a mutant allele [9]. They found the probability of this error using the term $(1 - f)e$, where $(1 - f)$ is the wild type allele frequency [9]. Adding these terms yields the probability that the mutant is detected [9]. Given the computed probability applies to a single cell, the authors multiplied by na to derive the probability that this heterozygous mutation is detected in the sample [9].

2.2.4 Probability function, Anandakrishnan et al.

The second function has the parameters G which is all the known somatic mutations that can become fixed in a population, h which is the number of required mutations for carcinogenesis, k is the number of possible combinations of carcinogenic genes, and m is all of the somatic mutations [10]. The authors also assume that mutations are equally likely [10].



: 1

Figure 1: A visual representation of the Anandakrishnan et al. model, adapted from figure 1 in their paper [10]

The above figure (1) gives a visual representation for their model with $k = 3$ and $h = 2$ hit model [10]. The base pairs shaded blue, red, and green are basepairs of interest, the assumption being that when one of these pairs becomes occupied by mutants cancer starts to develop [10]. As generations pass the germ line sequence accumulates one somatic mutations per generation until one of the pairs of mutations for carciogenesis occurs [10]. Figure 1 shows a case in which only 4 mutations occurred before carciongensis, as the red base pairs accumulated mutations which resulted in carciogenesis[10] .

Through this paper the number of permutations of y items from a list of x-distinct items with repetitions allowed is denoted as $\text{Perm}(x, y)$

The authors begin by taking the permutation to find all ways to arrange m somatic mutations across G possible sites as

$$\text{Perm}(G, m) = G^m \tag{2}$$

The permutations that exclude particular i of h carcinogenic mutations, was given by:

$$\text{Perm}(G - i, m) = (G - i)^m \tag{3}$$

The authors used the inclusion-exclusion principle to find the number of permutations with exactly h hits as:

$$\text{Perm}_h(G, m) = \text{Perm}(G, m) - \sum_{i=1}^{h} (-1)^{i+1} \binom{h}{i} \text{Perm}(G - i, m) \tag{4}$$

To find the probability that there were exactly h hits, the ratio of permutations with exactly h hits to the total number of permutations was used, which gives the equation

$$P(h, m) = \frac{\text{Perm}(G, m) - \sum_{i=1}^{h} (-1)^{i+1} \binom{h}{i} \text{Perm}(G - i, m)}{\text{Perm}(G, m)} \tag{5}$$

this is simplified to

$$P(h,m) = 1 - \sum_{i=1}^{h} (-1)^{i+1} \binom{h}{i} \left(\frac{G-i}{G}\right)^m \ (h \leq m) \tag{6}$$

The authors let $P_k(m)$ be the probability that at least one of the combinations of h hit

$$P_k(m) \approx 1 - [1 - P(h,m)]^k \tag{7}$$

The above probability is an approximation that is used to find the probability that the h^{th} hit occurs on the m^{th} somatic mutation which is defined as

$$P(m) = P_k(m) - P_k(m-1) \tag{8}$$

2.2.5 Modification of Anandakrishnan et al. function

Some values of m made the equation (8), negative for our data. We believed that the value G^m over counted the spots available for mutations as m increases. This was due to the fact that if a mutation occurs at one spot it is not likely that another carcinogenic mutation will occur and replace this mutation. We changed the Anandakrishnan et al. function to utilize permutations without repetition. Since there are G spots with m mutations we us Perm $= \frac{G!}{(G-m)!}$ instead of G^m.

This changes our equations (1-6) as we now have

$$\text{Perm}(G,m) = \frac{G!}{(G-m)!} \tag{9}$$

Our equation 2 becomes

$$\text{Perm}(G-i,m) = \frac{(G-i)!}{(G-i-m)!} \tag{10}$$

Our equation 3 becomes

$$\text{Perm}_h(G,m) = \text{Perm}(G,m) - \sum_{i=1}^{h} (-1)^{i+1} \binom{h}{i} \text{Perm}(G-i,m) \tag{11}$$

Our equation 5 becomes

$$P(h,m) = \frac{\text{Perm}(G,m)}{\text{Perm}(G,m)} - \sum_{i=1}^{h} (-1)^{i+1} \frac{\text{Perm}(i,m)}{\text{Perm}(G,m)} \tag{12}$$

This results in

$$P(h,m) = 1 - \sum_{i=1}^{h} (-1)^{i+1} \binom{h}{i} \left(\frac{\frac{(G-i)!}{(G-i-m)!}}{\frac{G!}{(G-m)!}}\right) (h \leq m) \tag{13}$$

We note that $\frac{\frac{(G-i)!}{(G-i-m)!}}{\frac{G!}{(G-m)!}}$ is equivalent to

$$\frac{(G-i)!}{(G-i-m)!} * \frac{(G-m)!}{G!} \tag{14}$$

This can be broken down and expressed as

$$\frac{(G-m)(G-m-1)(G-m-2)...(G-m-i+1)}{G(G-1)(G-2)...(G-i+1)} \tag{15}$$

After this correction we still utilized equations (7-8) [10]. We took the resulting probabilities from each $P(m)$ and multiplied by the number of samples to get the expected distribution of mutations for each m in our somatic mutation list. In our data set we had a maximum number of 600 mutant reads per samples. The percent of cells that fall into each 50 read segment up to 600 were calculated and plotted as a histogram. The curves of the expected mutations were calculated for $h = 1 - 5$ and superimposed onto the histograms for the percent of samples. We used the midpoint of each class and their associated $P(m)$ from equation (8) to calculate the RMSD.

3 Results

3.1 Chromosome wide Analysis

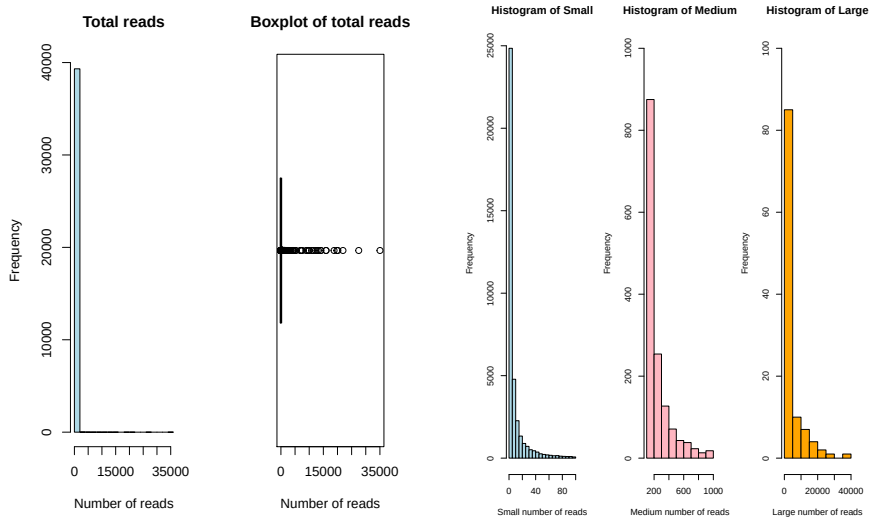
3.1.1 Descriptive Statistics

Type	Median	Mean	SD	Min	Q_1	Q_3	Max
Mutant	2	24.05	292.858	0	1	8	23869
Total	3	30.7	400.27	1	1	10	35022

Table 1 contains descriptive statistics for the number of mutant and total reads for our basepair data.
We can see that even when we remove all of the 0 rows from the total UMI read data set some mutant rows still have no read coverage. These rows with no mutant coverage were not interesting as mutations allow for inferences to be made. We notice there is large variability in regard to the total and mutant reads, which was exemplified by the large standard deviation. We divided the mean number of mutant reads by the mean number of total reads to find an expected average VAF of approximately $\frac{24}{31} = .77$. This VAF is close to the average VAF from of all tabulated VAFs, which was .84. Thus showing that our data set had a lot of detected mutations, an expected results given that most of our cells come from Chronic Lymphocytic Leukemia, the cancer type for which mutations were known.

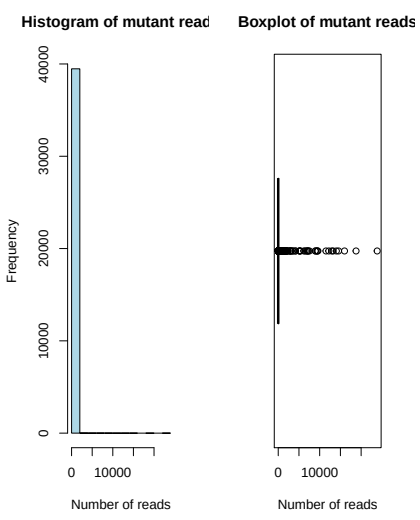
Type	Median	Mean	SD	Min	Q_1	Q_3	Max
Mutant	123	124.6	110.1	1	9	208	566
Total	149	155.2	141	1	10	259	825

Table 2 contains descriptive statistics for the number of mutant and total reads per cell in our data.
There is less variability in the reads per cell compared to the base pair analysis. This result makes sense as these cells were all derived from a small sample of individuals with cancer. Thus we did expect to see a similar distribution of UMI coverage for cells in our samples. Due to the fact that the sequencing of the cells was not done in regard to DLBL mutations we expected the cells of this type to primarily have normal reads. We believed this allowed for us to see more coverage for total reads compared to mutant.

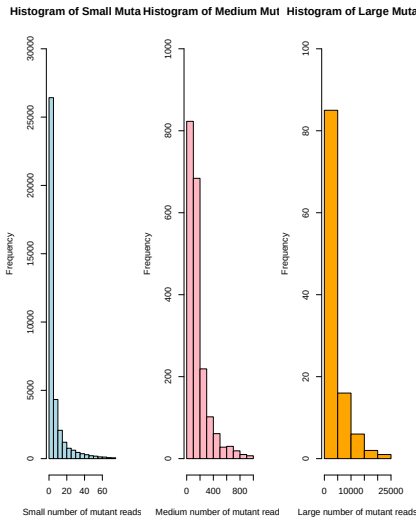


: 1 a : 1 b

Figure 2: (a) Contains the histogram and boxplot for the distribution of total reads per base pair. (b) This figure shows the number of total reads per base pair broken into 3 classes of read coverage.
The figures (2a-2b) above show the distribution of total reads for basepairs. Figure 1 shows the distribution of reads, in which we observed that most of the data fell into the 0 to 500 read range, with some extreme outliers that were barley apparent. To better see the distribution of reads per base pair we broke our data into 3 classes 0-100, 100-1000, and >1000, shown in figure (2b). We can see from the histograms of these different brackets all showed similar right skewed distributions.



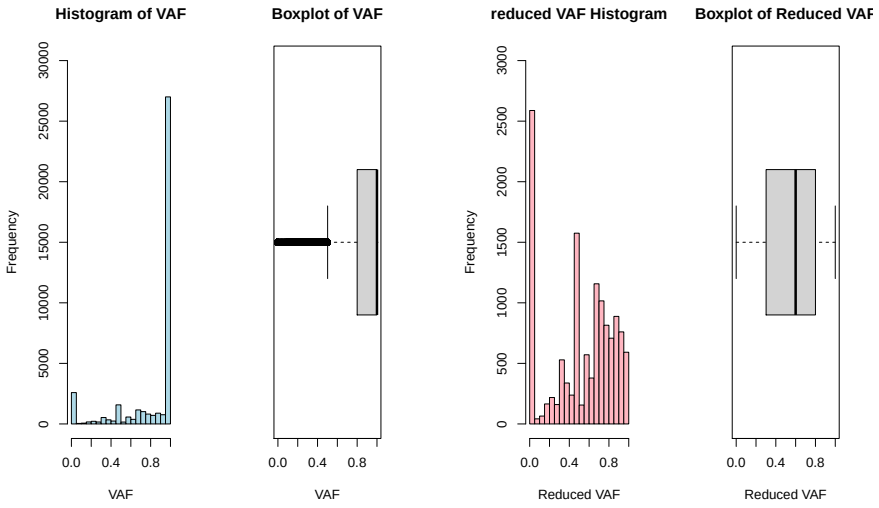
: 2a



: 2b

Figure 3: (a) Contains the histogram and boxplot for the distribution of mutant reads per base pair. (b) This figure shows the number of mutant reads per base pair broken into 3 classes of read coverage.

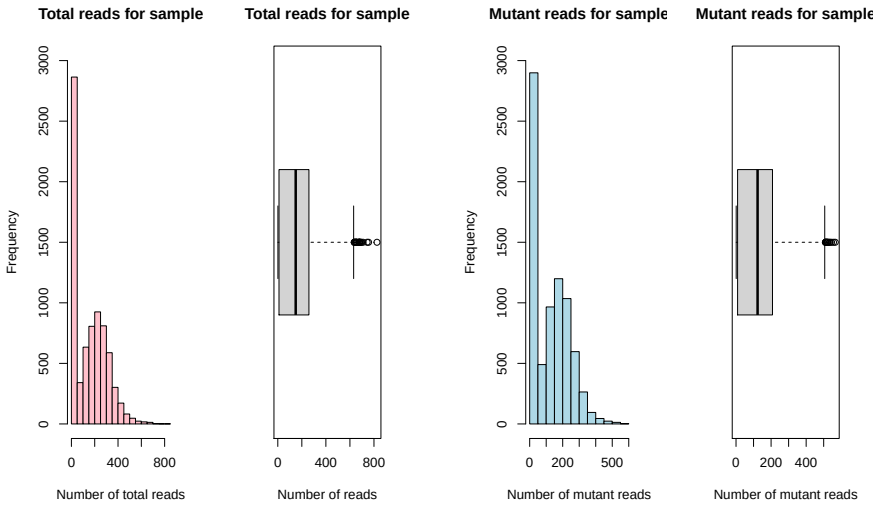
We made histograms and boxplots for the mutant reads per base pair (figure 3a). We saw that the distribution of mutant reads was very similar to that of the total reads. Due to these extreme skews we broke the number of mutant reads into 3 groups, 0-75, 75-1000, >1000, shown in figure (3b). These groups were similar to the ones for total reads, but due to a higher concentration in the first class the lower bound was switched to 75 to give a better visual representation of the data. We observed that there is a higher concentration of data in the first two histograms in figure (3b), indicating that we likely had less extreme outliers for our mutants than total.



: 3a : 3b

Figure 4: (a) Contains the histogram and boxplot of the distribution of VAF for our base pairs. (b) Shows the histogram and boxplot for the distribution of VAFs when VAF values of 1 were exlcuded.

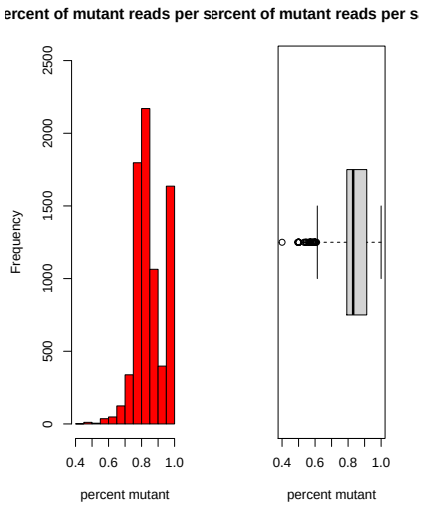
We looked at the VAF that was calculated for each basepair, shown in figure 4. We can see that there are many reads for the value of 1 indicating that we have many base pairs that only have mutant reads (figure 4a). When these basepairs were removed, we observed that the variability is reduced, and the boxplot shows a median closer to .7 (figure 4b).



: 4a : 4b

Figure 5: (a) Shows the histogram and boxplot of the distribution of total reads per cell data. (b) Shows the histogram and boxplot of the distribution of mutant reads per cell.

We examined the variability of the reads per cell, for total and mutant reads. Overall we saw that both distributions are both bimodal with prominent clusters of reads at the start (figure 5a-b). Though the distributions are very similar, the total reads histogram approaches approximately 800 reads while the mutant reads has a maximum of approximately 600 (figure 5a-b). The boxplots also show these patterns with longer right tails and a significant amount of outliers (figure 5a-b). The data from these samples did show less variability compared to the read data for basepairs (figures 2-3). However, some basepairs had more mutations than others given that only a few mutations are needed to be driver genes for cancer development.

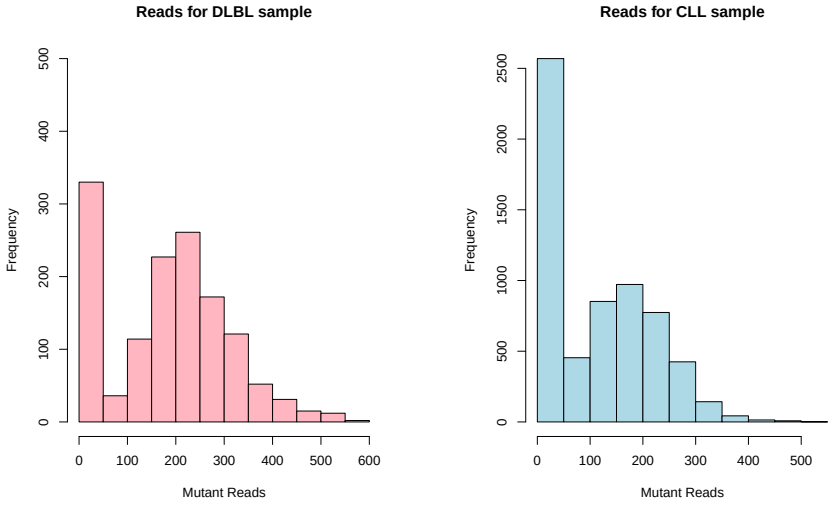


: 5

Figure 6: Shows the histogram and boxplot for the percent of mutant reads per cell.

We looked at the percent of reads per each sample that were designated as a mutant, we can see that most most of the data falls into the range of .6 to 1 (Figure 6). This indicates that most of the samples have a lot of mutant reads. We assumed that samples that had low expression of mutant reads were likely not from our Chronic Lymphocytic Leukemia patients, and were rather likely from the Diffuse large B-cell lymphoma patients.

We examined the read coverage on the ends of the chromosome and compared it to the middle and the two categories were similar with mean total reads from the middle being 29.5 and the mean reads from the start/end group being 30.5. We do see very large in each group variability, with standard deviations of 408 reads from the start and end, and the standard deviation of 392 reads from the middle. These results gave us a strong indication that there may not be any difference between the number of total reads between the two groups.



: 7a : 7b

Figure 7: (a) Shows the histogram for mutant reads of suspected DLBL cells. (b) Shows the histogram for mutant reads suspected to be CLL.

The last result we had was in regard to the read distribution from cells assumed to be Chronic Lymphocytic Leukemia, versus cells assumed to be Diffuse large B-cell lymphoma (Figure 7a-b). We observed that these distributions are fairly similar in overall appearance, with Chronic Lymphocytic Leukemia having a more reads as it contains approximately eighty percent of the sample. We observe that there is a noticeable difference between the mean number of mutant reads for Diffuse large B-cell lymphoma, (184), compared to the mean number of mutant reads for Chronic Lymphocytic Leukemia (112) . This is an unexpected result due to the fact that we expect Chronic Lymphocytic Leukemia to have more mutant reads. But we suspected that this disparity in reads was due to the fact that cells that had less reads overall, had a higher percent of mutant reads.

3.1.2 Hypothesis Testing for Basepairs

We first want to test a hypothesis that the variant allele frequencies from our data set are greater than the expected VAF of .5 [11]. We hypothesized:

$$\begin{aligned} H_0 &: p = .5 \\ H_a &: p > .5 \end{aligned} \tag{H1}$$

In order to test this we do a test of equal proportion. We find a $\chi^2(1) = 8305.7$ with a p-value $< .0001$ which indicates that our mean VAF is greater than

.5. This result is rather intuitive, as we were looking at data that comes from cancer patients, and should have expected their cells to have more mutations.

It is commonly known, when sequencing cellular data, there is a possibility for amplification bias: closer to the end of the chromosomes we will have more read coverage [5]. Biologists try to control for this through the use of UMIs [5]. Since our data is using UMI coverage, we wanted to see if there was still a significant difference in the reads from the ends of the chromosome compared to the middle. For this analysis we used the total reads, as this gives us a more accurate picture of potential read bias. We hypothesized:

$$\begin{aligned} H_0 &: \text{Med}_{end} = \text{Med}_{mid} \\ H_a &: \text{Med}_{end} > \text{Med}_{mid} \end{aligned} \tag{H2}$$

A wilcoxon rank test was run to compare population medians and we found W=196,163,283 with $p - value$=.01. We concluded that there was a difference in the distribution of reads throughout the chromosome, with more reads at the ends. We next tested to see if there is an equal variability of total reads in these regions. We hypothesized:

$$\begin{aligned} H_0 &: \sigma^2_{end} = \sigma^2_{mid} \\ H_a &: \sigma^2_{end} > \sigma^2_{mid} \end{aligned} \tag{H3}$$

We find an $F(19684, 19683)$=1.0834 with $p - value < .001$, which indicates that the variance for these samples is not the same. But we do not have normality or independence, so the result of this $F-$test, was not necessarily reliable. This result showed us that even though we were using UMIs end bias was still present in our data.

3.1.3 Hypothesis testing for Sample

We examined the samples by splitting the samples into two different groups, cells suspected to be Chronic Lymphocytic Leukemia, and cells suspected to be Diffuse large B-cell lymphoma. We hypothesised that approximately 1500 cells were Diffuse large B-cell lymphoma, and that DLBL cells had lower VAF than CLL cells. But since we have different cells we do assume independence between them. We hypothesized:

$$\begin{aligned} H_0 &: \sigma^2_{CLL} = \sigma^2_{DLBL} \\ H_a &: \sigma^2_{CLL} < \sigma^2_{DLBL} \end{aligned} \tag{H4}$$

With this in mind we ran a test of equality of variance for these two samples, which produces an $F(6254, 1372)$=.64361, and has an associated $p-value<$.0001. We next ran a wilcoxin test to see if the two samples have similar mutant reads per cell. We hypothesized:

$$H_0 : \text{Med}_{CLL} = \text{Med}_{DLBL}$$
$$H_0 : \text{Med}_{CLL} < \text{Med}_{DLBL} \quad \text{(H5)}$$

This test yielded a W=2,917,196 and a $p - value$<.0001. From these results we concluded that there are significant differences in the number of reads and their variability for the two assumed types of cancer. Though we did not have normality, so our F statistic may not have been reliable.

3.2 Probability functions

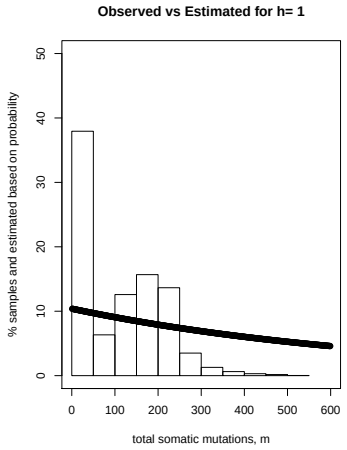
3.2.1 Petti et al. Function

In this section we examined the results of the equation (1) presented by Petti et al. for our data. Our parameters were
$n = 7637$
$a = 0.8$
$f = 2(.75)$
$c = 0.0003932879$
$t = 13.78$
$r = 124.6$
$e = .00039$
Due to the fact that we had no available data in regard to site specific mutation rates, we utilized the maximum rate of error of .0039 found by Petti [9]. In our data set we only had three basepairs that were belonged to our known mutation data set. These basepairs belonged to the genes SRGAP2C, NOTCH2NL, NBPF15. Though we had three base pairs the only one with all the parameters known was NBPF15. This function did not produce results that were consistent with a valid probability function, as the resulting value was 2936.21. Even when the value of e was set to 0, the function still output a value of 2930. Thus this function did not work with our data. It seems that this function was designed for smaller sample sizes, and thus did not work well with our large data set.

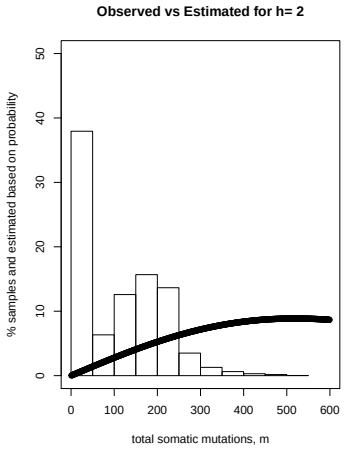
3.2.2 Anandakrishnan et al. function

In this section we discuss the probability function created by Anandakrishnan et al., equation (8). Given that G is the number possible somatic mutations, we noted that in chromosome 1, we had 39515 basepairs and a mutation could have arisen at any. In each location a mutation could have occurred and been passed on from the DNA, so the total is multiplied by a value of two, and also by 1.1 as this is the average number of mutations that can fix at a base pair [10]. This lead to our parameter being $G = 39515(2)(1.1)$. The next parameter we had to find was our k. The probability that a mutation will occur in a cancerous genome was computed to be $1 * 10^{-4}$ [12]. With our expected rate of mutation we multiplied this by the number of possible mutation sites for our chromosome which is 39515*2, which yielded the expected number of mutations

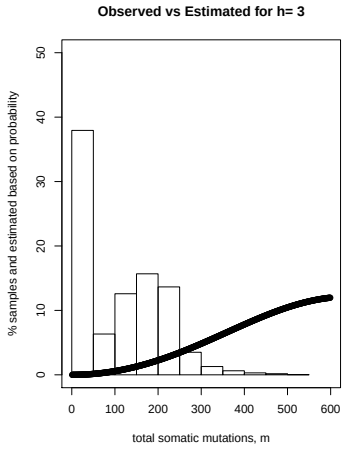
in the chromosome. We also approximated that there were 15 genes that could be mutated as it is expected to find between 5-20 mutated genes in exon regions of CLL patients [13]. These values were multiplied together and raised to the power of h. Which gave us a parameter $k = (1 * 10^{-4} * 39515 * 2 * 15)^h$. [10]. Our data showed that for chromosome 1 there were 1790 known somatic mutations. The total amount of mutant reads per cell in our data set ranged from 0-600, so our m, the total number of somatic mutations was set to 600. We used the row sums and computed the percent of samples that fall in each 50 mutant read interval.



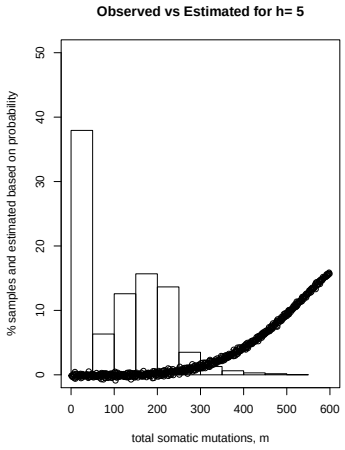
: 8a



: 8b



: 8c



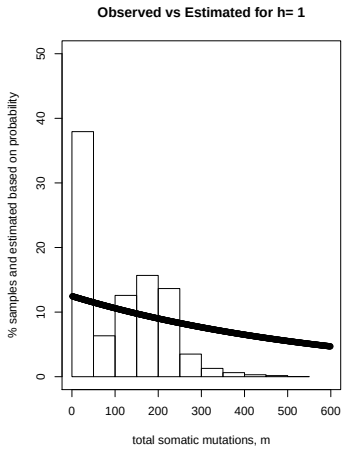
: 8d

Figure 9: (a) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of equation (8) for $h = 1$ superimposed upon the histogram. (b) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of equation (8) for $h = 2$ superimposed upon the histogram. (c) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of equation (8) for $h = 3$ superimposed upon the histogram. (d) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of equation (8) for $h = 5$ superimposed upon the histogram.

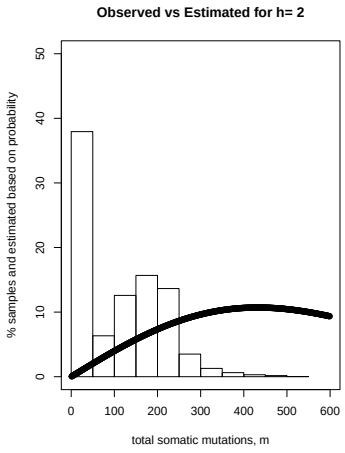
h	root mean square difference
1	9.77
2	13.5
3	14.3
4	14.5
5	14.4

Table 4 contains the root mean square difference for the output of the original Anandakrishnan et al. function with $h : 1 - 5$.
Above are the resulting plots (figures 9a-d) for the original function presented in the hits paper using our data superimposed on the histogram of percent of samples in each 50 read interval [10]. We noted that this function output plausible values for $h = 1 - 4$. However, we noted in figure 8d, that when $h = 5$ this function starts outputting negative values for smaller values of m. Since the curves imposed on the graphs are based on probability a negative output should not ensue. Table 4 above indicates that $h = 1$ had the lowest RMSD, so we concluded that $h = 1$ is the best fit for our data.

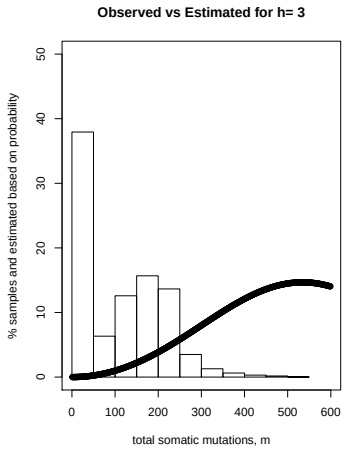
This function was then modified as an issue from the biological standpoint was noted. The original function allowed for the repetition of positions where mutations can occur. This does not make sense from a biological standpoint as we do not expect these mutations to be overwritten thus instead of using permutations with repetitions we utilized permutations without repetitions. The changed model is shown in the methods section above (equations 9-13).



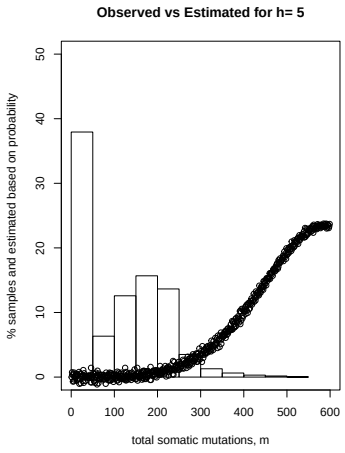
: 9a



: 9b



: 9c



: 9d

Figure 10: (a) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of the modified function for $h = 1$ superimposed upon the histogram. (b) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of the modified function for $h = 2$ superimposed upon the histogram. (c) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of the modified function for $h = 3$ superimposed upon the histogram. (d) Shows the histogram with the percent mutant reads in each 50 basepair interval with the values of the modified function for $h = 5$ superimposed upon the histogram.

h	root mean square difference
1	9.31
2	13.75
3	15.32
4	16.1
5	16.28

Table 5 contains the root mean square difference for the modified model values of h ranging from 1 to 5.
When we looked at our new plots (figure 10a-d) we noted that modifying the function to exclude repetitions yielded very similar results to the original function's output (figure 9a-d). There are subtle differences, as figures 10a-d showed more pronounced curvature than figures 9a-d, and figure 10a had a higher y intercept (≈ 14) when compared to figure 9a which had a y-intercept of (≈ 10). The modified model did allow for the lowest RMSD value to be obtained, a value of 9.31 for $h = 1$ (Table 5). Since $h = 1$ for the modified function yielded the lowest out of all trials, shown in table 4 and table 5, we concluded that this would be the best model for the data [10].

Though the modified function does minimize the error for $h = 1$, it increases the error for the values of h from 2 to 5 (table 4 and table 5). The modification to the model also did not resolve the issue of outputting negative values for the case $h = 5$ (figure 9d). Thus modifying this function to exclude repetitions was not the issue with the function.

3.2.3 Example of function

In order to understand the behavior of the Anandakrishnan et al. function we utilized small values for the parameters. In this our parameters were
$G = 10$
$m = 3$
$h = 1, 2, 3$
$k = 1, 2, 3$
In our analysis we began by utilizing equation (6) from methods to find the probability of exactly h hits,
$P(1,3) = 1 - (-1)^{1+1}\binom{1}{1}\left[\frac{10-1}{10}\right]^3 = 1 - (.9)^3 = .271$
$P(2,3) = 1 - (-1)^{1+1}\binom{2}{1}\left[\frac{10-1}{10}\right]^3 + (-1)^{2+1}\binom{2}{2}\left[\frac{10-2}{10}\right]^3 = 1 - 2 * (.9)^3 + (.8)^3 = .054$
$P(3,3) = 1 - (-1)^{1+1}\binom{3}{1}\left[\frac{10-1}{10}\right]^3 + (-1)^{2+1}\binom{3}{2}\left[\frac{10-2}{10}\right]^3 + (-1)^{3+1}\binom{3}{3}\left[\frac{10-3}{10}\right]^3 = 1 - 3 * (.9)^3 + 3 * (.8)^3 - .7^3 = .006$
We noted that as we increase h that the probability decreases, an intuitive result as the more places that have to be hit, the more uncommon hitting all of them is. We plugged the values we derived into equation (7) from methods.
For $h = 1$
$P_1(3) = 1 - [1 - .271]^1 = .271$
$P_2(3) = 1 - [1 - .271]^2 = .469$
$P_3(3) = 1 - [1 - .271]^3 = .613$

$P_4(3) = 1 - [1 - .271]^4 = .7176$
$P_5(3) = 1 - [1 - .271]^5 = .794$
For $h = 2$
$P_1(3) = 1 - [1 - .054]^1 = .054$
$P_2(3) = 1 - [1 - .054]^2 = .105$
$P_3(3) = 1 - [1 - .054]^3 = .153$
$P_4(3) = 1 - [1 - .054]^4 = .199$
$P_5(3) = 1 - [1 - .054]^5 = .242$
For $h = 3$
$P_1(3) = 1 - [1 - .006]^1 = .006$
$P_2(3) = 1 - [1 - .006]^2 = .012$
$P_3(3) = 1 - [1 - .006]^3 = .018$
$P_4(3) = 1 - [1 - .006]^4 = .024$
$P_5(3) = 1 - [1 - .006]^5 = .030$
This allowed for the conclusion that as the value of k increases that the probability of of hitting one of the possible combinations increases.

4 Discussion

This project allowed for some interesting discoveries to be made. Through our exploration of the data, we were able to observe that our cancer cells have a higher average VAF compared to normal cells, which was expected given that cancer is a mutation based disease. We also discovered that amplification bias was still present in our sample. The most unexpected result being the cells that were suspected to be DLBL had more mutant reads than CLL. These results were interesting, but due to the fact that this analysis was done on a single chromosome it limits our ability to generalize these results to the rest of the genome. A future direction of this project would be to expand these analyses to the rest of the genome.

In our analysis our main undertaking was investigating probability functions from Petti et al. (2019) and by Anandakrishnan et al. (2019). Though both of the functions utilized were in regard to probability they were used to explore different questions. The Petti et al. function computed given a particular heterozygous mutation what is the probability that at least one cell in the sample had that mutant detected in its sequencing [9]. While the Anandakrishnan et al. function computed the probability that the m^{th} somatic mutation results in one of the combination of mutations required for carcinogenesis occurring [10].

The Petti et al. function was a function that has presented issues in the past. The function, given by equation (1) required a multitude of inputs that are specific to the basepair that was being studied. This resulted in missing information in our study, as when these samples were collected and sequenced the site specific error rate was not computed, which left us with missing information. Due to this lack of information we were constrained to simulations. Also due to the fact that there were so many inputs into this function there was only one basepair that had all the other information necessary to run utilize

this function. Thus we were not able to do an in depth exploration into this function and its behavior.

Past analysis of the function found that this function had output negative results. In this study there were no negative results that ensued from equation (1), but the experiments did lead to values »1, which should not have occurred given that this function was computing probability. We suspect that the main issue with this function was multiplying the probability by the number of cells caused the issue, as we were multiplying by a number > 6000. If the sample of cells was not of such a large magnitude we would expect that this function could have better results. In order to better understand this function in the future conducting another biological experiment to derive all the appropriate information would be beneficial. It could also be favorable to use targeted sequencing for certain genes on small samples of the cells in the tumor population and then apply this function.

The Anandakrishnan et al. (2019) function did not initially present issues, as the parameters were able to be approximated for our data. We found that the function had constraints for our data as we increased the number of mutations required for carciongenesis (h), we observed that equation (8) began outputting negative values, which should not have occurred given that this was a probability distribution. Even when the function's inputs were modified equations (9-13) negative results were still output. We believe that this could be due to the fact that k is not dependent upon m. The number of possible combinations for the mutations is very large when h is greater than 4. Thus we would likely need more than 600 somatic mutations to be present for these larger h.

In the future it would be interesting to explore this Anandakrishnan et al. function on different chromosomes that are known to have a lot of mutations associated with cancer. On chromosomes with genes related to this specific cancer we would expect to observe maximum number of somatic mutations greater than 600. The authors also found great success in applying this function in applying this function to other cancer types, so in the future this function and its modified version could be applied to other cancer types.

5 Glossary

Somatic mutation- a mutation that occurs within an individual, that is not passed on to their offspring, but is passed on when cells go through divisions.
Variant Allele Frequency (VAF)- the percent of reads that are not mapped to the wild type allele
Germline mutation - a mutation that occurs in an individuals reproductive line and thus is a heritable mutation
UMI- unique molecular identifier, a technique in which strands of DNA are labeled before sequencing in order to get a more accurate read [1]

6 Acknowledgements

I would like to thank Dr. Dementieva and Dr. Sample for all of their help and guidance on this project
I would like to thank Dr. Khiabanian and the members of the Khiabanian lab

References

[1] Dominic Grün and Alexander vannbsp;Oudenaarden. "Design and Analysis of Single-Cell Sequencing Experiments". In: *Cell* 163.4 (2015), pp. 799–810. DOI: 10.1016/j.cell.2015.10.039.

[2] *What Is Chronic Lymphocytic Leukemia?* May 2018. URL: https://www.cancer.org/cancer/chronic-lymphocytic-leukemia/about/what-is-cll.html.

[3] *Diffuse large B-cell lymphoma*. Jan. 2021. URL: https://www.leukaemia.org.au/blood-cancer-information/types-of-blood-cancer/lymphoma/non-hodgkin-lymphoma/diffuse-large-b-cell-lymphoma/.

[4] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-cell RNA sequencing technologies and bioinformatics pipelines". In: *Experimental amp; Molecular Medicine* 50.8 (2018), pp. 1–14. DOI: 10.1038/s12276-018-0071-8.

[5] Sanjay M. Prakadan, Alex K. Shalek, and David A. Weitz. "Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices". In: *Nature Reviews Genetics* 18.6 (2017), pp. 345–361. DOI: 10.1038/nrg.2017.15.

[6] Laura Oikkonen and Stefano Lise. "Making the most of RNA-seq: Preprocessing sequencing data with Opossum for reliable SNP variant detection". In: *Wellcome Open Research* 2 (2017), p. 6. DOI: 10.12688/wellcomeopenres.10501.1.

[7] *The R Project for Statistical Computing*. URL: https://www.r-project.org/.

[8] URL: https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg19&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chrX%3A15578261-15621068&hgsid=1089726871_FS0nhEtrOZnVntmDY8v76MIo1DtU.

[9] Allegra A. Petti et al. "A general approach for detecting expressed mutations in Aml cells using single CELL RNA-SEQUENCING". In: *Nature Communications* 10.1 (2019). DOI: 10.1038/s41467-019-11591-1.

[10] Ramu Anandakrishnan et al. "Estimating the number of genetic mutations (hits) required for carcinogenesis based on the distribution of somatic mutations". In: *PLOS Computational Biology* 15.3 (2019). DOI: 10.1371/journal.pcbi.1006881.

[11] Samuel P. Strom and Samuel P. Strom. "Current practices and guidelines for clinical next-generation sequencing oncology testing". In: *Cancer Biology amp; Medicine* 13.1 (2016), pp. 3–11. DOI: 10.20892/j.issn.2095-3941.2016.0004.

[12] I. P. Tomlinson, M. R. Novelli, and W. F. Bodmer. "The mutation rate and cancer". In: *Proceedings of the National Academy of Sciences* 93.25 (1996), pp. 14800–14803. DOI: 10.1073/pnas.93.25.14800.

[13] Nisar A. Amin and Sami N. Malek. "Gene mutations in chronic lymphocytic leukemia". In: *Seminars in Oncology* 43.2 (2016), pp. 215–221. DOI: 10.1053/j.seminoncol.2016.02.002.