Mutational Hotspot Detection in LGL Leukemia

Nikki Aaron
School of Data Science
University of Virginia
Charlottesville, Virginia
Email: na5zn@virginia.edu

Prabhjot Singh
School of Data Science
University of Virginia
Charlottesville, Virginia
Email: ps4fk@virginia.edu

Siddharth Surapaneni
School of Data Science
University of Virginia
Charlottesville, Virginia
Email: sss2ea@virginia.edu

Joseph Wysocki
School of Data Science
University of Virginia
Charlottesville, Virginia
Email: jw6mw@virginia.edu

***Abstract*—Cancer genomics has been focused primarily on identifying and studying mutations that are over-represented in known genes. This project applied methods to scan through entire chromosomes and label these loci as "genomic probabilistic hotspots" (GPHs). A GPH is defined as any area on a patient's chromosome where the observed rate of mutations over positions of a given chromosome window far exceeds what would be expected from random variation. The approach is then applied to 39 patients diagnosed with large granular lymphocyte (LGL) leukemia - a rare form of blood cancer. In order to calculate expected mutation rates in non-LGL patients, data were obtained from the 1000 Genome Project. A negative binomial test was employed to isolate specific GPHs where the distribution of mutations within the LGL patient sample was significantly high. The Negative Binomial approach identified a median of 1 to 2 patient hotspots per chromosome with a mean Jaccard's distance between patients being 0.90. The KDE method found a median of 40 hotspots with wider span resulting in a mean Jaccard's distance of 0.43. The results from the Negative Binomial approach indicated heterogeneity between hotspot locations, whereas KDE results were more homogeneous. Negative binomial is best for pinpointing the most significantly dense regions, whereas KDE is best for identifying all broad regions that are more mutated than a reference. These new, gene-agnostic approaches provide novel methods to search chromosomes for mutational abnormalities and can be generalized and scaled to any clinical syndrome. Future directions include extension of the GPH method across genomes, developing a robust library of disease- and/or model species-specific hotspot profiles. These may serve as reference guides in studies seeking to understand the exact biochemical processes driving the onset and progression of rare cancers.**

I. Introduction

Large granular lymphocyte (LGLs), a type of white blood cell, normally constitute 10-15% of a healthy person's white blood cells. Large Granular Lymphocyte Leukemia is a dysregulation in natural cell death resulting in an accumulation of LGLs and a decline in other types of blood cells necessary for normal bodily function. [1] The exact cause of this cancer is unknown, but mutations in several genes have been identified as likely to be playing a role. [2] The STAT3 mutation has been the most prominent of the genes observed and evaluated across the study of LGL Leukemia.

However, these mutations in STAT3 or other genes are not found in all patients or cells with LGL leukemia or even in all patients with the same subtype of LGL leukemia. [3]. There has been significant work demonstrated by other research projects on finding cancer-driving genes based on frequency of mutations seen near or in those genes. [2] Given how genes alone may not be contributing to the onset of disease, there is interest in evaluating the whole genome, including regions between genes known as intergenic regions. Intergenic regions, although they do not house known genes and their functions are largely unknown, have shown promise in playing important regulatory roles. In addition, increasing research, such as the ENCODE project has gone into understanding the implications of the whole genome, which includes these intergenic regions. [4]

Methodologies that have aimed at understanding the whole genome utilize some form of a sliding window approach or scan statistics. By sliding or scanning across the whole genome and locating areas that have a large amount of single nucleotide polymorphisms or SNPs, the goal is to search for regional variations in somatic mutation frequency. [5], [6], [7], [8] Building on these concepts, it is of significant interest to locate all region(s) across the whole genome that may be playing a role in the development of LGL Leukemia. Locating these regions may lead to the diagnostic and therapeutic innovation that could provide a breakthrough to the treatment of patients that have LGL Leukemia.

Specifically, this project aims to investigate the density of mutations across the whole genome and identify regions of high mutational density (aka "hotspots") if the rate of mutations exceeds what would be expected in a same sized region with a random or uniform distribution of mutations. This research will focus on defining an appropriate measurement of density across the genome, locating regions of high mutational density, and comparing these regions across the patient cohort. This project applied methods to scan through entire chromosomes to localize areas where the distribution of mutations within a single genome are high and label them as "genomic probabilistic hotspots" (GPH). One difficulty arising in this research was the inability to obtain non-LGL patient somatic SNP data to use as a control. Another issue was the computational challenges associated with handling whole genome sequence data.

II. Data

University of Virginia's LGL leukemia registry is currently the largest registry of LGL leukemia specimens with clinical data [9]. From these specimens, somatic genome sequencing of the cancer cells was obtained for 39 patients, primarily of European descent and a subset of this data was compiled

consisting only of chromosome positions where mutations may be present. For each patient, this initial data set provided a binary indicator of mutated / not mutated for 395,531 nucleotide positions across 23 chromosomes. Of these positions, each patient has between 2,292 and 106,984 mutations, with 10,291 mutations on average.

For the majority of chromosome positions, only a single patient possesses a mutation. Occasionally, a mutation is shared between two or more patients, with a maximum of eight patients sharing a mutated position. Since shared mutations are so rare in the patient cohort, it can be inferred that finding a specific mutation that causes LGL is unlikely.

An additional dataset was compiled from the 1000 Genomes project repository as a reference set. This set contained germline genome sequencing for over 400 patients, because a somatic dataset was not available. However, this research was undertaken with the assumption that while germline data has a much higher number of mutations across the chromosomes than somatic, the rate and distribution of non-LGL data can show where there is non-random variation in the genome of the general population. The 1000 Genomes data set was limited to patients of European descent to more closely match the LGL patient cohort. A similar strategy as used by Przytycki & Singh in their somatic mutation research [10]. After this filtering, 1000 Genomes data contained between 1.05 million (chromosome 22) and 6.78 million (chromosome 2) positions for each chromosome, and covered 404 patients.

To ensure that the reads for each mutation were correctly placed to the proper chromosome position, mappability scores for sequences in the genome from the UCSC genome browser were integrated into the patient data. The source for mappability was the Duke Uniqueness 35bp track which contained mappability scores that were assigned a value between 0 and 1 [11]. The mappability score for a read sequence is calculated by dividing 1 with the number of possible placements in the genome. A mappability score of 1 indicated that the position of that particular 35bp sequence is known, and could not possibly be located in other areas of the genome. For each mutation in the patient data, binary mappability value of 1 was assigned when the mappability score was 1, and 0 when it was less than 1. In order to make sure that the location of where the mutation happened was correctly identified, only rows of patient data that were assigned a mappability score of 1 were used, all others were dropped. This removed 15% of rows in the data set.

This transformation of the patient data now gives the dataset a binary indicator of mutated/not mutated for 334,226 nucleotide positions. On average patients had 6,440 mutations, excluding one outlier patient G03 who had 93,101 mutations. However, methods were made to obtain independent patient genomic probabilistic hotspots, so patient G03 was included in analysis.

A copy of this data set was also saved with each patient's mutation distribution column standardized to sum to one. By standardizing this way, the data was changed to represent the percentage of mutations rather than the count. When looking at a window of positions, all rows in the window would be summed and divided by the total over the entire chromosome to get the percentage of mutations within that window. This allows direct comparison between rates of mutation within a window amongst patients who have vastly different overall mutation rates across the chromosomes.

For use as a control for comparison, the 1000 Genomes data set was also annotated with a column for mappability and unmappable rows were dropped. Then all 400+ patients' binary mutation columns were averaged across patients. Next, the resulting single mutation distribution column was standardized to sum to 1.

III. Methodologies

A. Discrete Window Analysis

A discrete window approach was used to gain a general understanding of the distribution of mutations across patients. This process segmented each chromosome into 1 MB windows and then counted the number of mutations present in each window. Because this was a preliminary approach towards understanding the whole genome, mappability criteria was not integrated in analysis.

Three metrics were used to summarize mutational features across the whole genome. Mutational occurrence represents the total number of mutations observed within each window across all LGL patients. Mutational density measures the distance covered by the center 50% of mutations. Lastly, mutation coverage highlights the proportion of patients who have at least one mutation in the window.

B. Negative Binomial Distribution

One method that was used to detect GPHs was to model the mutational processes of the patients using a Negative Binomial Distribution. The negative binomial distribution has been used previously [12], [13] Like Weinhold et al., this project used a negative binomial test to detect clusters. However it diverges from their approach in the use of independent patient mutation profiles as opposed to aggregated patient data [14].

In this approach, each position on a chromosome is treated as an independent Bernoulli trial where the position being a SNP is considered is modeled as a success. The total number of mappable mutations in the genome for a patient is divided by the total number of mappable positions to represent the patient's mutation rate for a Bernoulli trial. This is to model the assumption that most of the genome does not consist of SNPs. To identify the GPHs for each patient, each SNP for each chromosome was scanned to see if at that particular position if there are 2 or more SNPs within a specified number of base pairs (the window size parameter), for that SNP. This is considered a cluster of mutations.Any overlapping clusters get merged , and this merged cluster is considered a potential GPH if it has more than 3 SNPs. Each potential GPH is evaluated using a negative binomial test where the probability of observing the number of the non-mutated positions for the given number of SNPs in the GPH was assessed. The number of non-mutated positions is calculated by counting the number

of non-SNPs between the first and last SNP in the potential GPH. The mutation rate used for each test was the patient specific mutation rate determined above.

The cumulative distribution function of the negative binomial distribution is used to calculate the p-value for a given cluster. The scipy.stats package nbinom function in Python was used for these tests [15]. A Benjamin Hochberg Correction was then applied to the p-values of each GPH to adjust for false positives using a family-wise error rate of 0.05. The statsmodel package in Python was used to apply this multiple test correction [16]. This correction yielded the specific GPHs for each patient. This method was a probabilistic approach to detect local dense GPHs for a given patient. Window sizes of 10k, 50k, 100k, and 500k base pairs were investigated.

C. Kernel Density Estimation

Another approach this research investigated was Kernel Density Estimation (KDE). This method produces a smoothed estimate for mutation density at each point along the chromosome. Unlike the negative binomial method, KDE has the benefit that control data can be used to identify areas of the chromosome that are more densely mutated in LGL patients than non-LGL germline average. However, KDE alone cannot indicate whether differences in density are statistically significant.

Density plots were created for each patient from their binary mutation data at mappable positions. To find out which areas are more densely mutated than the healthy average, a KDE was created from the standardized, mappable 1000 Genomes data. After subtracting this KDE from the KDE for each LGL patient, areas of interest are areas where the result is greater than 0. Window sizes between 10k and 10million base pairs wide were tried on the KDE difference results to investigate the best window size to use. The boundaries of each hotspot were defined by changing any negative values of the KDE difference to 0, then finding bands that start at either 0 or a local minimum, increase to a local maximum, then decrease to a local minimum of 0. These are the visible "peaks" in the distribution. To determine if a hotspot was statistically significant, the KDE of that hotspot boundary was compared between the healthy and LGL data sets using the Kolmogorov-Smirnov test.

IV. Results

A. Discrete Window Analysis

Mutational density, occurrence, and coverage were visualized across the whole genome in Figure 1. The top two images show that mutations are frequent along the chromosomes, but that areas of concentration can be identified. The lower image shows that there are several regions of high density shared across the patient cohort. In this aggregated approach, patients that simply had a tendency for more SNPs tend to skew the data. This bias was extremely evident with patient G03. Several regions that were initially thought to have been of interest were dense solely due to SNPs belonging to patient G03.

Beyond this, the discrete and sliding window also utilized an arbitrary size of windows, which led to rather arbitrary cut-offs of regions that were being categorized as belonging to the same group. However, the evidence from this exploration aided the development of more statistically robust and patient specific methods of locating areas densely packed with SNPs, methods able to detect regional variations across the whole genome for each patient independently.

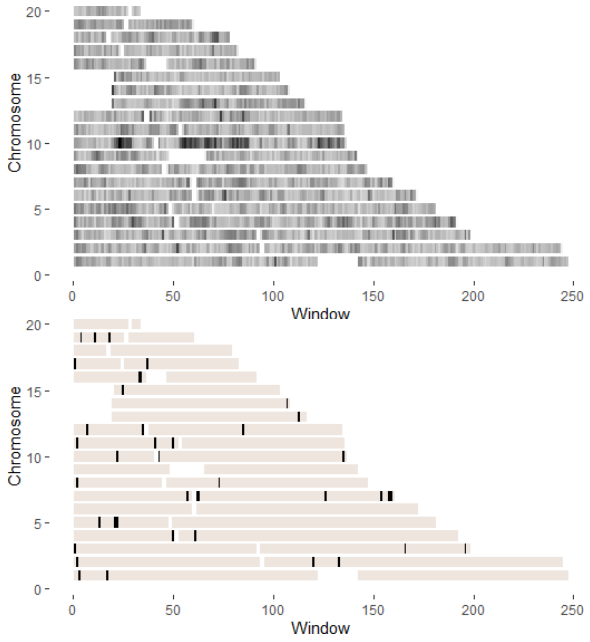

Fig. 1. Mutation Mapping: Occurrence (top) shows regions where the center 50% of mutations within each window occur. Coverage (bottom) shows the proportion of patients with at least one mutation in each window

B. Negative Binomial

For each window size (10k, 50k, 100k, 500k base pairs) tested, we obtained varying number and length of hotspots from both the Negative Binomial and Kernel Density Estimation methods. The total number of hotspots summed over all patients and chromosomes is shown in Table I.

TABLE I

Window Size	Hotspots Detected Negative Binomial	Hotspots Detected KDE
10k bp	4,373	83,168
50k bp	7,344	81,250
100k bp	8,339	76,456
500k bp	17,864	27,234

Table 1. Hotspot Counts: Aggregated count of hotspots for all patients and chromosomes at varying Window Sizes.

Table II compares the quantity of total and shared hotspots across the patient cohort using window size of 50k bp. In order to obtain total hotspots, intervals of overlap between patients were merged. Shared hotspots were then calculated by filtering for hotspots with 2 or more patients. For example,

Chromosome 1 had 485 hotspots after merging, with 28 of these hotspots being shared. The procedure was done for Chromosomes 1 to 22. The difference between total hotspots and shared hotspots for each chromosome showed that there were quite a number of hotspots unique to an individual patient.

TABLE II

Chromosome	Total Hotspots After Merging	Total Hotspots Shared
1	485	28
2	556	30
3	469	28
4	521	52
5	447	30
6	405	19
7	414	24
8	393	26
9	264	16
10	469	52
11	320	20
12	322	14
13	279	19
14	199	12
15	187	5
16	219	8
17	176	4
18	240	16
19	149	13
20	128	9
21	98	8
22	86	7

Table 2. After merging intervals of hotspots that overlapped between patients, total number of merged hotspots and shared hotspots were listed. The table indicates that many of the hotspots are not shared across the patient group.

C. Kernel Density Estimation

The first resulting plots indicated that there was consistently low mutational density near the centromere, start, and endpoints of the chromosome. See Figure 2 for results from chromosome 1. These plots were created using the tidyverse package in R with the default bin width (window size) selection method used, nrd0 [17]. The bin width estimates for each patient were between 11,713,118 base pairs to 27,279,638 base pairs. The average bin width for the default selection was 20,679,009 base pairs. The highest variability amongst patients in chromosome 1 is seen between positions 50 million and 100 million. But, this method needed refinement as the high window size chosen by default smoothed the data too much to be useful. Furthermore, it is undesirable to use different window sizes for different patients as results between patients should be compared with the same parameters.

From here a kernel density estimation procedure using standardized germline mutations from the human genome project was executed. The resulting plots showed that bin widths in the millions smoothed data too much, and bin widths less than 100k were identical as most mutations are spaced apart by approximately this width. Small window size will identify a large number of very short hotspots. Large window size will identify fewer hotspots, but each with more length. See Figure 3 for KDE differences between averaged LGL and non-LGL datasets at two different window size. For the remainder of the KDE research, a window size of 500k base pairs was used as this appeared in plots to be an acceptable choice for representing dense clusters in the LGL and control data.

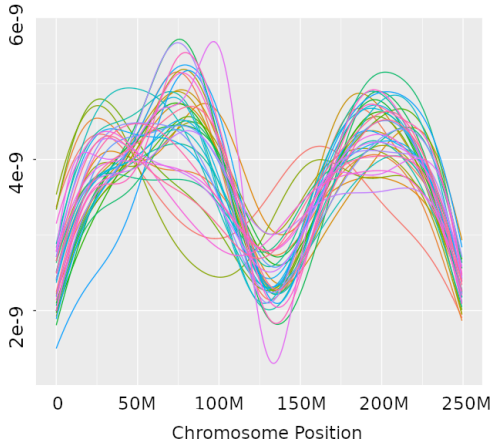

Fig. 2. Mutation Densities - Chromosome 1: Plotting mutational density for each patient on a separate line shows areas of shared low density at the edges and center. Between these regions, there is higher and more varied density.

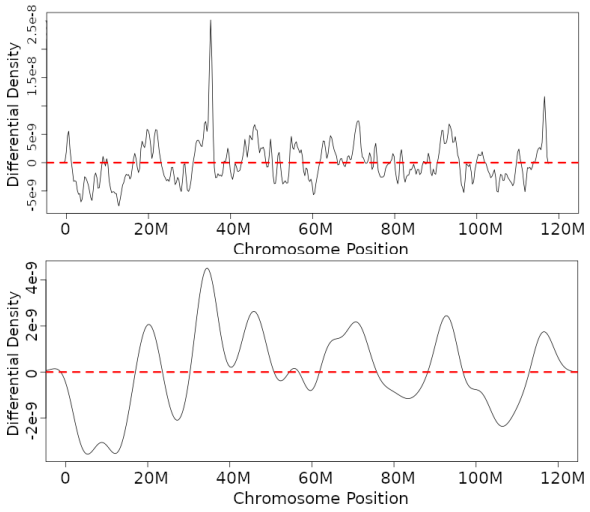

Fig. 3. Differential KDEs - Chromosome 1: Window sizes smaller than 250k become more erratic and computationally intensive (top), but a window size of 500k or higher is overly smoothed (bottom).

A sample of the hotspot results from the KDE method from patient A03 on chromosome 1 is shown in Figure 4 with hotspot centers shown in red. The figure indicates that non-zero regions cover a large portion of the chromosome length, and many of them occur consecutively with no gap in between. This may be due to the sparsity of the LGL somatic data in comparison to the healthy 1000 Genomes germline data. A single mutation of sparse data represents a much larger ratio of total mutations, so mutation rates tend to be higher when

compared to dense germline data. This method achieves its goal of identifying where mutation rate is higher, but would be better suited for comparing data sets of the same nature.

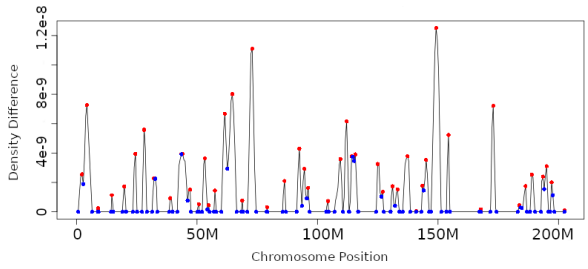

Fig. 4. Zero-floored differential density - Chromosome 1, Patient A03: Red dots indicate the peak center of hotspots. A large proportion of chromosome positions have densities higher than the reference genome.

D. Comparison of Results

The number of hotspots found in each chromosome varied by patient. The median and standard deviation of this variation is shown in Table III for Negative Binomial method using a 50k bp window size between patients and Table IV for KDE method using a 500k bp window size. Using these hotspot locations, Jaccard's distance was used to calculate the dissimilarity between each possible pair of patients. This number indicated the proportion of hotspots that are not shared between patients. The distribution of these results are also shown in the tables.

TABLE III
HOTSPOT RESULTS - NEGATIVE BINOMIAL

Chromosome	Median ± MAD Hotspot Counts Negative Binomial	Mean ± StdDev Jaccard's Distance Negative Binomial
1	2±3.0	0.95±0.22
2	2±3.0	0.98±0.14
3	2±3.0	0.98±0.14
4	2±3.0	0.95±0.22
5	2±3.0	0.95±0.22
6	2±1.5	0.96±0.19
7	3±3.0	0.97±0.15
8	2±3.0	0.92±0.26
9	2±3.0	0.89±0.31
10	2±3.0	0.94±0.22
11	1±1.5	0.96±0.19
12	1±1.5	0.96±0.17
13	2±1.5	0.96±0.19
14	1±1.5	0.89±0.31
15	0±0.0	0.69±0.43
16	1±1.5	0.88±0.33
17	1±1.5	0.89±0.31
18	1±1.5	0.88±0.33
19	1±1.5	0.84±0.37
20	0±0.0	0.74±0.43
21	1±1.5	0.82±0.37
22	0±0.0	0.74±0.44

Table 3. For each chromosome, the median number of hotspots detected using the Negative Binomial approach across patients along with the median absolute deviation was calculated. In addition the mean Jaccard's distance between patients was also calculated.

TABLE IV
HOTSPOT RESULTS - KDE

Chromosome	Median ± MAD Hotspot Counts	Mean ± StdDev Jaccard's Distance
1	55±4.4	0.5±0.05
2	58±4.4	0.6±0.06
3	47±4.4	0.6±0.06
4	45±4.4	0.6±0.07
5	44±3.0	0.5±0.07
6	46±4.4	0.4±0.07
7	40±4.4	0.5±0.07
8	38±4.4	0.6±0.08
9	29±3.0	0.5±0.08
10	34±4.4	0.7±0.06
11	36±3.0	0.5±0.07
12	35±3.0	0.4±0.08
13	26±3.0	0.5±0.10
14	23±3.0	0.5±0.10
15	23±1.5	0.4±0.11
16	21±1.5	0.5±0.10
17	23±3.0	0.3±0.09
18	20±3.0	0.5±0.10
19	17±3.0	0.4±0.12
20	17±3.0	0.0±0.00
21	11±1.5	0.0±0.00
22	11±1.5	0.0±0.00

Table 3. For each chromosome, the median number of hotspots detected using the KDE approach across patients along with the median absolute deviation was calculated. In addition the mean Jaccard's distance between patients was also calculated.

Comparing the results from both methods, the KDE method was able to identify hotspots at each chromosome for each of the patients unlike the Negative Binomial method which only found hotspots for only some of the patients at each chromosome. Further comparing these methods, analysis showed that KDE hotspots did not completely encompass the hotspots found by the Negative Binomial. This indicated that there were hotspots that were distinct to only one of the two methods.

V. DISCUSSION

The project explored various approaches in the attempt to locate genomic probabilistic hotspots. A discrete window analysis of all mutations was initially used to gain a global view of the whole genome and was useful in visualizing potentially mutationally dense regions across the whole patient cohort. In an attempt to locate hotspots more precisely across each patient individually, the negative binomial method utilized patient specific mutation rates and compared that to regions across the whole genome that consist of densely packed SNPs. On the other hand, the kernel density estimation method utilized additionally acquired data of germline mutations, which were standardized and then compared to the distribution of somatic SNPs to find peaks that were only seen in the LGL Leukemia patient group.

Window size has interesting implications on the degree of generalization imputed on the genome. A smaller window size, although leading to more precise intervals of genomic probabilistic hotspots, may lead to blind-spots on the genome, where crucial SNP(s) may be playing a role in the development

of a disease. On the other hand, a larger window size, although more tolerant in the allowance of a hotspot, will lead to an oversimplification of the genome that will not aid in the whole genome discovery process. Therefore, depending on the exact research goals, results from various window sizes could be of interest.

Furthermore, the results obtained from the negative binomial indicate potential heterogeneity between patient genomic probabilistic hotpots. As seen from the table III, most patient hotspots didn't overlap. Elaborating on this, the degree of dissimilarity between patients was analyzed using the Jaccard's distance, which indicated an average distance of greater than 0.80 for most chromosomes. Ultimately this furthermore supports the idea that most patients seem to be having their own unique hotspots.

The kernel density estimation method also provided interesting insights into the genome. The results are useful for pinpointing the center of regions across the genome that have higher mutation rates in LGL cells than healthy germline cells. It also has the advantage that it can compare standardized data from patients of varying overall mutation rates, and is therefore not as sensitive to outlier samples with many more mutations than others. This method's drawback is that it cannot filter out regions where density was insignificantly higher, so the identified regions are overly wide. The Kolmogorov-Smirnov methodology was not productive for this type of filtering, because the test calculated that every hotspot is significantly different between data sets. This could be a genuine result, or could mean that the method is not applicable in this particular situation. It could also be retried with more comparable data in future research.

The findings of this paper suggest that the heterogeneity of patient hotspots makes it difficult to appropriately find relevant signals related to LGL Leukemia. However, by reducing the high dimensional feature space of SNPs to genomic probabilistic hotspots, this project has enabled future research to understand the implications of the whole genome on specific attributes related to LGL Leukemia. Specifically, future research could investigate the differences between hotspots of patients with and without the STAT3 mutation. Any findings from such an endeavor could reveal differing biological pathways for patients that have LGL Leukemia. Furthermore, the project has enabled a whole-genomic approach towards understanding a complicated disease, which hopefully will allow for a more comprehensive understanding of LGL Leukemia within the context of the whole genome.

VI. Acknowledgment

The authors would like to thank our sponsors Mary Poss, David Feith, and Aakrosh Ratan of the University of Virginia for lending their mentorship and expertise. LGL leukemia patient samples were obtained from the LGL Leukemia Registry at the University of Virginia, and non-LGL data was obtained from the 1000 Genomes Project with the assistance of Aakrosh Ratan. Thanks also to Professor Jack Van Horn for his support during this research, and guidance on the conference submission.

References

[1] K. Oshimi, "Clinical features, pathogenesis, and treatment of large granular lymphocyte leukemias," *Internal Medicine*, vol. 56, no. 14, p. 1759–1769, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5548667/

[2] T. Lamy, A. Moignet, and T. P. Loughran, "Lgl leukemia: from pathogenesis to treatment," *Blood*, vol. 129, no. 9, p. 1082–1094, Mar 2017. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/28115367/

[3] K. B. Moosic, et al., "Genomics of lgl leukemia and select other rare leukemia/lymphomas," *Best Practice Research Clinical Haematology*, vol. 32, no. 3, p. 196–206, Sep 2019. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S1521692619300428

[4] T. E. P. Consortium, "Identification and analysis of functional elements in 1% of the human genome by the encode pilot project," *Nature*, vol. 447, no. 7146, p. 799–816, Jun 2007. [Online]. Available: https://www.nature.com/articles/nature05874/

[5] Y. V. Sun, D. M. Jacobsen, and S. L. R. Kardia, "Chromoscan: a scan statistic application for identifying chromosomal regions in genomic studies," *Bioinformatics*, vol. 22, no. 23, p. 2945–2947, Oct 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/17032677/

[6] J. Hoh and J. Ott, "Scan statistics to scan markers for susceptibility genes," *Proceedings of the National Academy of Sciences*, vol. 97, no. 17, p. 9615–9617, Aug 2000. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/10931953/

[7] Y. V. Sun, A. M. Levin, E. Boerwinkle, H. Robertson, and S. L. Kardia, "A scan statistic for identifying chromosomal patterns of snp association," *Genetic Epidemiology*, vol. 30, no. 7, p. 627–635, Nov 2006. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16858698/

[8] Z. Li, X. Li, Y. Liu, J. Shen, H. Chen, H. Zhou, A. C. Morrison, E. Boerwinkle, and X. Lin, "Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies," *The American Journal of Human Genetics*, vol. 104, no. 5, p. 802–814, May 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6507043/

[9] 2015. [Online]. Available: https://giving.uvahealth.com/article/creating-hope-and-positive-outcomes

[10] P. F. Przytycki and M. Singh, "Differential analysis between somatic mutation and germline variation profiles reveals cancer-related genes," *Genome Medicine*, vol. 9, no. 1, Aug 2017. [Online]. Available: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0465-6

[11] UCSC, "Mappability or uniqueness of reference genome from encode," 2012. [Online]. Available: https://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19g=wgEncodeMapability

[12] I. A. Klein, W. Resch, M. Jankovic, T. Oliveira, A. Yamane, H. Nakahashi, M. Di Virgilio, A. Bothmer, A. Nussenzweig, D. F. Robbiani, and et al., "Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in b lymphocytes," *Cell*, vol. 147, no. 1, p. 95–106, Sep 2011. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3190307/

[13] I. T. Silva, R. A. Rosales, A. J. Holanda, M. C. Nussenzweig, and M. Jankovic, "Identification of chromosomal translocation hotspots via scan statistics," *Bioinformatics*, vol. 30, no. 18, p. 2551–2558, May 2014. [Online]. Available: https://academic.oup.com/bioinformatics/article/30/18/2551/2475616

[14] N. Weinhold, A. Jacobsen, N. Schultz, C. Sander, and W. Lee, "Genome-wide analysis of noncoding regulatory mutations in cancer," *Nature Genetics*, vol. 46, no. 11, p. 1160–1165, Sep 2014. [Online]. Available: https://www.nature.com/articles/ng.3101

[15] Scipy, 2021. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.nbinom.html

[16] Statsmodels, 2019. [Online]. Available: https://www.statsmodels.org/dev/generated/statsmodels.stats.multitest.multipletests.html

[17] R Kernel Density Estimation Bandwidth Selectors, 2021. [Online]. Available: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/bandwidth.html