Post-Genomics Analysis (Final Report)

Quaye E, George , Statistics(MSc)

Nkweteyim N, Raisa, Bioinfomatics (MSc)

Due: 12/10/2020

Contents

1 Introduction

1.1 About The Authors

George Ekow Quaye is a graduate in Actuarial Science of the University of Cape Coast, Ghana. He completed his bachelor's education in May 2016 and received his credential on June 2016 and became a teaching assistant in the University of Cape Coast. Currently he is pursuing graduate studies in statistics (MSc) at the University of Texas at El Paso due to his keen interest of been a statistician by profession. Academically, he is interested in mathematical computations, specifically statistical analysis and inferences. George desires to have his career to revolve around statistics and data, which he fondly calls, "the art of numbers", for he believes in the power of numbers to induce change personally and globally. The capacity of numbers to make someone happy or sad has impressed him so much. To him, statistical analyses are like math wherein one must master and know a workable formula to deliver the right inference or conclusion. Given his background in statistics and mathematics, His major contribution to this project is the statistical computations, suggestions and analysis to derive a meaningful conclusions and explanations with regard to the aims of this project.

Raisa Ntemenyi Nkweteyim is a bioinformatics master's student studying at The University of Texas at El Paso. She was born in Buea, Cameroon and speaks both English and French. Raisa achieved a bachelor's degree in biochemistry from the University of Buea with a minor in medical laboratory technology. Her interests in the computational sciences grew from her experiences working in a hospital laboratory and as a treasurer for the Women Techmakers, Buea organisation. She is a woman in STEM advocate, loves learning about people and different cultures, reading, and writing fictional stories and baking. Raisa plans on pursuing a PhD in either computational biology or bioinformatics after her master's degree. She hopes to one day work in laboratories involved in the drug discovery process.Her contributions in this project include explaining and providing information on the biological aspects of prostate cancer and next generation sequencing and using her expertise in bioinformatics in answering some of the questions posed.

Each other provided a valuable contribution to project by creating codes in R and Python to process the data, writing up the literature review, performing statistical analysis, interpreting the results and proof-reading the entire work.

2 Background Information

As reviewed by Wang et al. 2018 in a paper by Torre et al. 2015, “Prostate cancer is the most common non-cutaneous cancer in men worldwide, with an estimated $1,600,000$ cases and $366,000$ deaths annually.” Prostate cancer is a disease of old age as seen in data produced in the UK in 2015-2017 where on average each year 35% of new cases were in males aged 75 and over. Although not well understood, the occurrence of prostate cancer can be related to family history, increased BMI, smoking history, and ionizing/UV radiation (Cuzick et al. 2014). As a man ages, the prostate tends to increase in size. This can cause the urethra to narrow and decrease urine flow. This is called benign prostatic hyperplasia, and is different from prostate cancer (What Is Prostate Cancer?, CDC 2020).

2.1 Prostate Gland

The prostate is a dense fibromuscular gland that lies directly below the bladder and plays a supportive role in the male reproductive system. Its principal purpose is to secrete alkaline solution protective for sperm in the acidic environment of the vagina. The solution also contains supportive proteins and enzymes that provide nourishment to sperm (Singh and Bolla 2020).

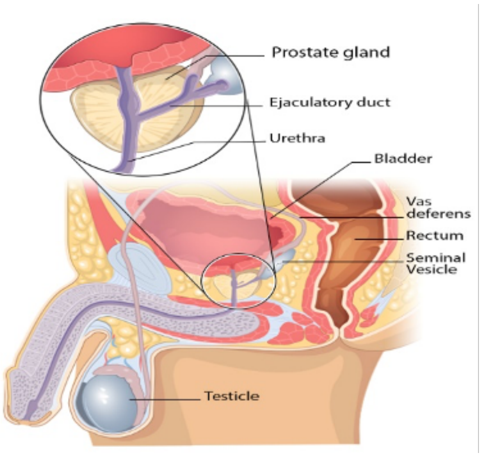


Figure 1: Location of the prostate.

Diagram above shows the location of the prostate, in front of the rectum and just below the bladder

(What Is Prostate Cancer? CDC, 2020).

On the cellular level, the human prostate contains a pseudostratified epithelium with three types of terminally differentiated epithelial cells: luminal, basal, and neuroendocrine (van Leenders and Schalken 2003; Shen and Abate- Shen 2010).

The luminal cells produce secretory proteins and are defined by expression of cytokeratin 8 (CK8), CK18 and androgen receptor (AR). The basal cells are nestled between the basal lamina and luminal cells and express high levels of CK5 and p63 and very low levels of AR. Neuroendocrine cells, a small population of endocrine–paracrine cells located on the basal cell layer, express neuroendocrine markers such as synaptophysin and chromogranin A and do not express AR (Wang et al. 2018).

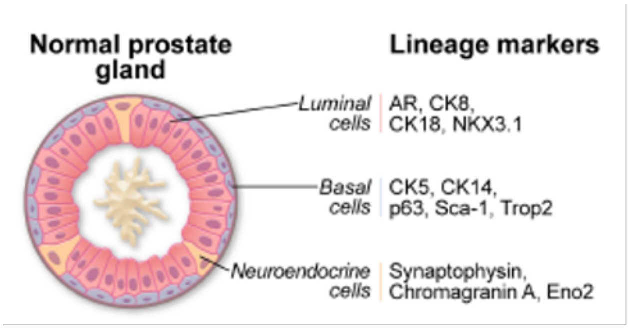


Figure 2: The normal prostate gland with lineage markers (Wang et al. 2018).

2.2 Molecular Biology of Prostate Cancer

Metastatic disease is the leading cause of prostate cancer associated deaths. The first site of metastases are usually the lymph nodes adjacent to the primary tumors (Datta et al. 2010), followed by metastases to the liver, lungs, and bones (Fig. 3) with bone as the most common site of metastasis. Human prostate cancer bone metastases often cause severe pain, hypercalcemia, and frequent fractures. In bone metastasis, there is a dynamic interaction between the cancer cells, osteoblasts, and osteoclasts, which results in a “vicious cycle” of bone formation and destruction—a process that supports cancer cell survival and tumor growth.

Huggins and Hodges (1941) first reported that castration led to tumor regression in prostate cancer patients. Since prostate cancer cells require androgen hormones such as testosterone to grow, androgen deprivation therapy (ADT) is now the standard of care for prostate cancer. Resistance to ADT can develop, resulting in primary castration-resistant prostate cancer (CRPC) or metastatic CRPC (mCRPC).

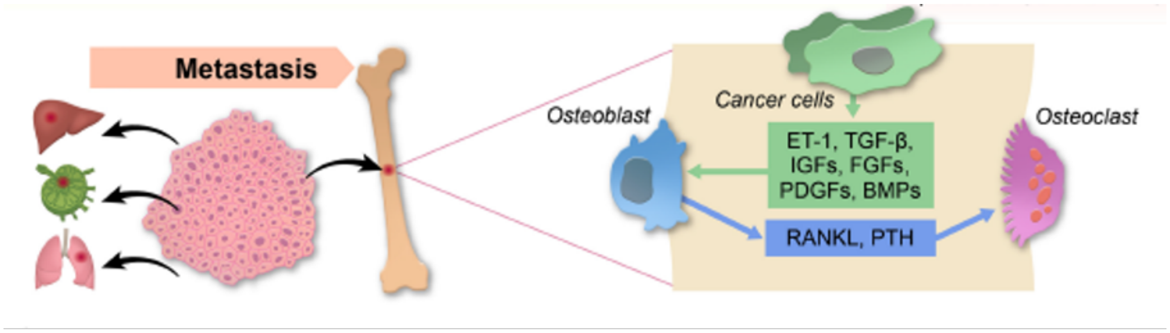


Figure 3: Progression of prostate cancer and the development of mCRPC (Wang et al. 2018).

2.3 Genetics of Prostate Cancer

Epidemiological studies have established that a family history of prostate cancer significantly increases risk (Goldgar et al. 1994; Lange 2010) with men of African descent having the highest rates of incidence and mortality (Shenoy et al. 2016), which may partially be attributed to genetic factors (Huang et al. 2017).

As reviewed by Wang et al. 2018, genome-wide association studies (GWASs) have identified many prostate cancer susceptibility loci (Eeles et al. 2013; Takata et al. 2010; Schumacher et al. 2018), such as the risk-associated single-nucleotide polymorphism (SNP) rs339331 that increases expression of the cancer promoting RFX6 gene through a functional interaction with the prostate cancer susceptibility gene HOXB13 (Huang et al. 2014).

Research by The Cancer Genome Atlas Research Network in 2015 revealed that 15%–35% of mCRPC contain DNA repair defects, including in BRCA1/2, ATM, ATR, and RAD51. Mutations in BRCA1 and BRCA2 predispose individuals to breast, ovarian, and prostate cancers (Farmer et al. 2005).

Deregulation of genes controlling epigenetic processes involved in DNA modification such as methylation and histone modification can drive tumorigenesis in prostate cancer (Albany et al. 2011; Yegnasubramanian 2016).

2.4 Next Generation Sequencing

In the 1970s, two-time Nobel Laureate, Frederick Sanger and his colleagues developed a method for determining the nucleotide sequence of DNA called Sanger sequencing or the chain termination method. It became the prevailing DNA sequencing method for the next 30 years. Thanks to it, it ultimately enabled the completion of the first human genome sequence (human genome project) in 2004 (International Human Genome Sequencing Consortium, 2004).

The human genome project was an international scientific research project to determine the human genome sequence and map all its genes. After completing this project, it became evident that Sanger sequencing required vast amounts of time and resources, thus, faster, higher throughput, and cheaper technologies were required. This stimulated the development and commercialization of next-generation sequencing (NGS) technologies. NGS had the advantage of the preparation of NGS libraries in a cell free system rather than requiring bacterial cloning of DNA fragments, ability to run thousands-to-many-millions of sequencing reactions are in parallel and bypassing electrophoresis in sequencing output detection.

The first NGS technology to be released in 2005 was the pyrosequencing method by 454 Life Sciences (now Roche) (Margulies et al, 2005). The 454 Genome Sequencer generated about 200 000 reads (□20 millions of base pairs). A year later, the Solexa/Illumina sequencing platform was commercialized. The third technology to be released was Sequencing by Oligo Ligation Detection (SOLiD) by Applied Biosystems (now Life Technologies) in 2007 (Metzker, 2010). The Illumina and SOLiD sequencers generated much larger numbers of reads than 454 but the reads produced were only 35 bp long. In 2010, Ion Torrent (now Life Technologies) released the Personal Genome Machine (PGM). This system was developed by Jonathan Rothberg, the founder of 454, and resembles the 454 system.

Five platforms that have dominated the NGS market over the past decade included: 454, Illumina, SOLiD, Ion Torrent, and PacBio. Gene expression studies frequently changed from using microarrays to NGS-based methods, which enabled the identification and quantification of transcripts without prior knowledge of a particular gene and provided information regarding alternative splicing and sequence variation (Wang et al, 2009).

Illumina and SOLiD sequencing were more suitable than 454 for ChIP-seq (Park, 2009) (used to identify in vivo protein–DNA interactions), owing to their higher throughput. However, the reads were initially too short. Meanwhile, 454 was the preferred technology for de novo genome assemblies and had an important application in metagenomics. Now, Illumina technology can generate reads several hundred base pairs long, do de novo assembly and metagenomics as well. Illumina also offers the highest throughput per run and the lowest per-base cost (Liu, 2012) using their HiSeq X Ten, a set of ten HiSeq X sequencing machines. However, for their machines to be really cost effective, all HiSeq X machines run at full capacity.

There was a growing interest from the clinics to use NGS as a diagnostic tool. To meet this demand, Roche and Illumina launched compact bench-top sequencers. The fastest bench-top NGS platform remains Ion Torrent's PGM. The PacBio system is equally fast but is much more expensive and thus, less suitable for small laboratories and the clinical setting.

NGS has made whole genome sequencing (WGS) feasible and cost effective. WGS accelerates molecular diagnosis of monogenic illnesses and minimizes the duration of empirical treatment (Saunders et al., 2012). NGS has also facilitated transcriptome analysis through its use in RNA sequencing. It is useful in single cell genomics to reconstruct cell lineage trees using somatic mutations that arise due to DNA replication errors and is also used in whole-exome sequencing (WES), in which only the coding regions of the genome are sequenced (Choi et al., 2009).

With so many people's genomes sequenced, confidentiality becomes an important factor and ethical issues will probably emerge with a possibility of “genetic discrimination”. Still, there is no doubt of the advantages NGS has brought with its extensive data production and applications.

3 Data Description

The data was gotten from The Cancer Genome Atlas and it consists of 50 Variant Calling Format (VCF) files for prostate cancer (PrCa). Each of these VCF files contains a paired sample from the same patient with PrCa: one extracted from the primary tumor tissue and the other from the blood-derived normal sample.

The VCF files were converted to OnchoMiner Input format (OMI) files. OncoMiner is a bioinformatics pipeline developed at UTEP initially for mining local sets of WES data. It was implemented in 2016 as a web server (OncoMiner.utep.edu).

The columns of the OMI file include the following:

1. “chrom”: chromosome number

2. “left”: starting locus of Genetic Sequence Variants (GSVs)

3. “right”: ending locus of GSVs

4. “ref_seq”: reference base; base from human reference genome (GRCh38)

5. “alt”: mutated base, base from patient

6. “SubT1N1”: set of subjects who have specific variant in both tumor and normal tissue.

7. “SubT1N0”: set of subjects who have specific variant in the tumor but not normal tissue.

8. “SubT0N1”: set of subjects who have specific variant in the normal but not tumor tissue.

9. “SubT0N0”: set of subjects who do not have specific variant in the tumor and normal tissues.

10. “dbSNP”: link to dbSNP, if known

11. “Overlapped Gene: name of the gene (HGNC system) to which the variant is overlapped

12. “Annotation”: region GSV occurs

13. “genename”: the gene name

14. “where”: genomic region where GSV is found

15. “change_type”: change type of amino acid; that is, either synonymous or nonsynonymous

16. “Coding Score”: score computed by FATHMM-XF if variant occurs on CDS; ranges from zero to one.

17. “Non-Coding score”: score computed by FATHMM-XF if variant does not occur on CDS; ranges from 0-1.

18. “Further Information”: FATHMM prediction.

19. “AA var”: non-synonymous amino acid variation.

20. “provean score”: score computed by PROVEAN algorithm.

21. “prediction”: PROVEAN prediction.

4 Aims of Study

4.1 Major Aims

I. To identify possible genetic sequence variants (GSVs) that could be associated with the prostate cancer disease.

II. To identify genomic regions with high concentrations of GSVs.

4.2 Specific Goals

1. To obtain summary statistics of the data and test for significant differences in GSV counts between the tumor and normal tissues.

   Since Major aim II requires the determination of genomic regions with the highest concentration of GSVs, we believed comparing the percentages of GSVs per genomic region (which relies on GSV counts) with their concentration (relies on GSV counts and genomic region length) will give us an idea of the spread and variability of GSVs in the different genomic regions. For this reason, we selected the question below:

2. To determine what percentage of GSVs falls on genes, Exons, Introns, Protein coding regions,Intergenic regions and number of novel GSVs.

3. To determine the consistency of results of FATHMM and Provean predictions on non-synonymous novel GSVs.

   Both FATHMM and PROVEAN make predictions for non-synonymous variants based on different algorithms, scales and threshold values. It is however expected that their predictions are consistent.

5 Methods

5.1 Exploratory Data Analysis - Specific Aim 1

The data file is a Microsoft Excel file with three sheets labeled, “Tumor”, “Normal” and “Common”. Each of the 50 patient’s GSV count was computed based on the instructions of section 2.7 of the Data Processing file provided by the course instructor. Required counts for the normal and tumor samples that exist in the common sheet were extracted and added to each normal and tumor sheets respectively before the analysis was carried. The python and R codes used in doing this computation has been provided in the appendix folder.

The mean, standard deviation, 5-point summary were computed for the normal and tumor samples of all 50 patients using R (see appendix folder for code).

5.2 Inferential Statistics- Specific aim 1

Wilcoxon Rank Sum Test was carried out to determine if there is any significant difference in GSV counts between the tumor and normal tissues.

1. Data type: The data is a paired sample data from given patient’s tumor and normal tissue. It is a discrete data with variables of two levels (i.e 0 or 1), hence it follows that a non-parametric test is appropriate for this analysis.

2. Test Statistics: Given that the data is a paired data and non-parametric, Wilcoxon Rank Sum Test was chosen for this analysis.

3. Hypothesis; where M is the median counts in the dataset.

$$H_0 : M_t = M_n$$

$$H_1 : M_t > M_n$$

4. Significant Level; The test is carried out with an α= 0.05 and n=50. We seek to accept or reject the H_0 at this significant level and make an honest conclusion.

5.3 Major aim I :Identifying possible genetic sequence variants (GSVs) that could be associated with the prostate cancer disease

All non-synonymous GSVs (nsGSV) for the 50 patients was compiled from the tumor sample data. McNemar test is used for the analysis if n is greater than 30 else Binomial test. Hence a binomial (McNemar) test was used since the sum of the discordance (b + c) = n < 30 for the paired data. First we identified those possible GSV's associated with the cancer by discarding those that less than 4% counts. A Bonferroni adjusted p-value was used to adjust any bias of the p-values and results from this analysis compared with FATHMM prediction and provean predictions to see if they augment one another.

Let, n=b+c, where b is the number of counts in T1N0 and c is the number of counts in T0N1 from the nsGSV's and with $\alpha = 0.05$

$$H_0 : \pi_1 = \pi_2$$

$$H_1 : \pi_1 > \pi_2$$

.

5.4 Major aim II : Identifying genomic regions with high concentrations of GSVS

The "where" column of the dataset includes information about the genomic region where each variant is found. The genomic regions of interest are "gene", "intron", "exon" and "intergenic region". To find the concentration of GSVs per genomic region, a python script was created to parse and count the number of occurrences of each genomic region in the "where" column and count the length of the genomic region where each GSV is situated. These computations were performed only on the first isoform if the variant had many isoforms.

The decision tree below was used to determine which genomic region a variant is found in.

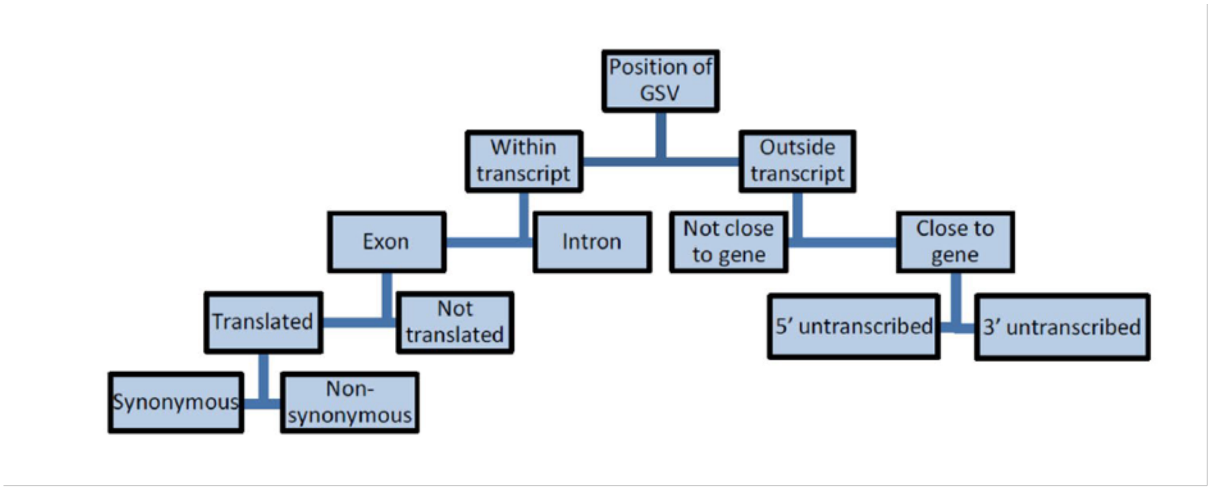


Figure 4: Decision tree (Leung et al 2016, Vasquez et al 2018).

According to the decision tree,

- gene (transcript) = intron + exon counts where exon = translated (CDS) + not translated (5' and 3' UTRs) counts
- intron = intron counts
- intergenic region = outside transcript (gene) where transcript = not close to gene + close to gene (5' untranscribed' + 3' untranscribed')

Out of these four genomic regions, the region with the highest concentration of GSVs is selected and explored in literature reviews.

5.5 Specific aim II

To calculate the percentage of GSVs, they were grouped according to their genomic region from the "where" column of the dataset, the number of GSVs in each group was counted and multiplied by 100 using R (Appendix folder). Genomic regions were determined according to the decision tree.

To compute the number of novel GSVs, the dbSNP column in the dataset was grouped into 2, based on if the dbSNP gene ID is available in the column or not. The number of GSVs in these two groups were counted and their percentage also computed.

5.6 Specific aim III (self formulated)

The non-synonymous tumor GSV data was filtered keeping the FATHMM and PROVEAN predictions and dbSNP columns. We decided to only look at unknown GSVs which did not have an ID in the dbSNP database. The predictions from FATHMM and provean was categorical where both tools classified the variants as pathogenic or benign. Frequencies were used to obtained counts in each category and a test was used to test the hypothesis:

H_0: Prediction of a GSV by FATHMM and Provean are independent.
H_a: Prediction of a GSV by FATHMM and Provean are related.

with an $\alpha = 0.05$. The resulting frequency table data follows the following assumptions;
1. The levels (or categories) of the variables are mutually exclusive
2. Each subject contribute data to one and only one cell in the χ_2.
3. The variables (FATHMM and Provean) both are measured as categories, usually at the nominal level.
4. The study groups (FATHMM and Provean) are independent.
As such, a Chi-Square test of dependency was appropriate.

6 Results and Discussion:

6.1 Exploratory Data Analysis

.

Table 1: Table of mean and SD for tumor and normal sample

	Normal	Tumor
Mean	6.920000	216.66000
Standard Deviation	1.575838	96.22987

Given the output above, On average, a patient is expected to have 6 GVS counts in the normal tissue with a standard deviation of 1.5758 and 216 in the tumor tissue also with 96.2299 in variation from the mean.

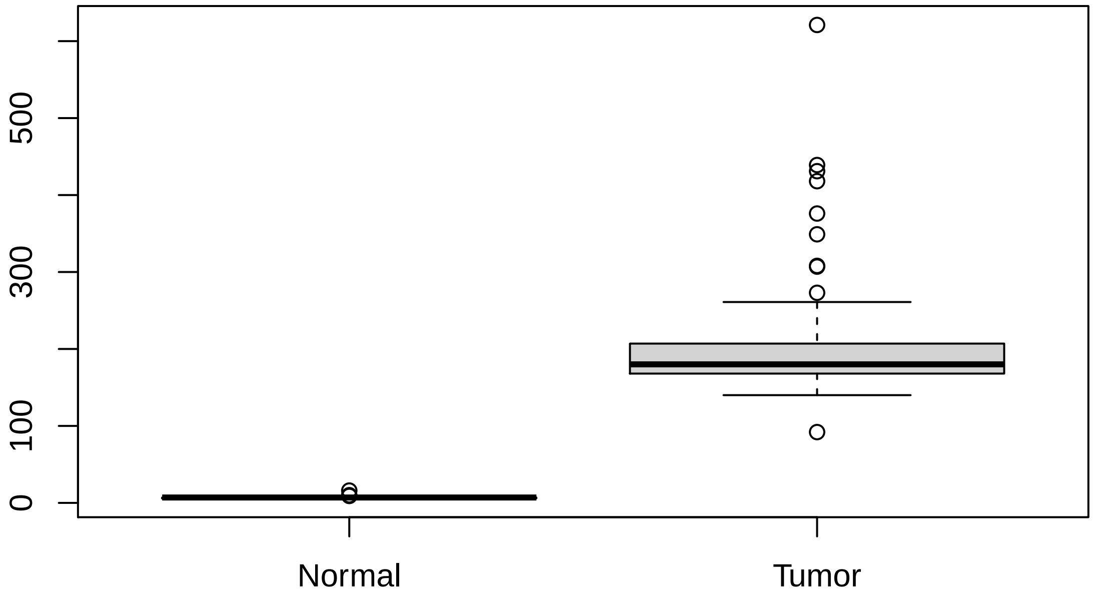
Table 2: The table of the 5-point summary statistics

	Normal	Tumor
Min.	6.00	92.00
1st Qu.	6.00	168.00
Median	7.00	180.00
Mean	6.92	216.66
3rd Qu.	7.00	205.25
Max.	16.00	621.00

The tables above displays the 5-point summary statistics for both the tumor and the normal samples .The lowest and highest GSV counts between the two samples are displayed in the table. It is revealed that the GSV counts of patients in the tumor sample is much greater than the GSV counts of patients in the normal sample.

This assertion seems to follow since the lowest count of 92 in the tumor sample is greater in value than the lowest counts of 6 in the normal sample. Also, the highest counts of 621 in the tumor sample is greater in value than the highest counts of 16 in the normal sample. Again by considering the upper one-fourth, upper half, and upper three-fourths instead of just the lowest and highest counts, we still concluded that the GSV counts of patients in the tumor sample is more than that of the normal sample.

fig a : Boxplot of GSV counts for the tumor and normal sample



The box-and-whisker plots above shows that the lowest count, highest count, median, Q1,and Q3 of the tumor sample is greater than that of the normal sample, so GVS counts in the tumor sample is quite larger than the normal sample as indicated earlier in the table.

6.2 Inferential Statistics

Given that n=50, α=0.05 and the two samples tumor and normal, we performed a Wilcoxon Rank Sum Test on the following;

$$H_0 : M_t = M_n$$

$$H_1 : M_t > M_n$$

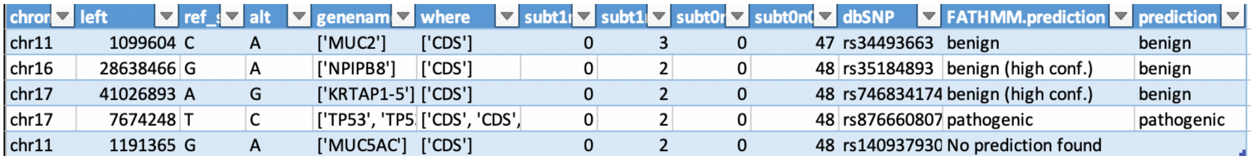
```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  final$Tumor and final$Normal
## V = 1275, p-value = 3.884e-10
## alternative hypothesis: true location shift is greater than 0
```

Since the $P-Value = 7.768e-10$ ($< \alpha-value = 0.05$), we reject H_0 and conclude that median counts of GSV in tumor sample $>$ median counts of GSV in normal sample. Implying that for the given dataset the median GSV counts in tumor tissue is statistically significantly greater than the median GSV counts in the normal tissue.

6.3 Main objective I: Identifying possible genetic sequence variants (GSVs) that could be associated with the prostate cancer disease

After the compilation of the nonsynonymous variants (unGSV's) in the 50 patients data set we found about 10000 unGSV's from 1665 genes. To filter the data, GSVs occurring in less than 4% of patients were discarded.

The table above is the resulting table after applying the threshold of 4% to obtain possible ns-GSV's associated with the cancer disease. An analysis table was created based on these for further investigation.

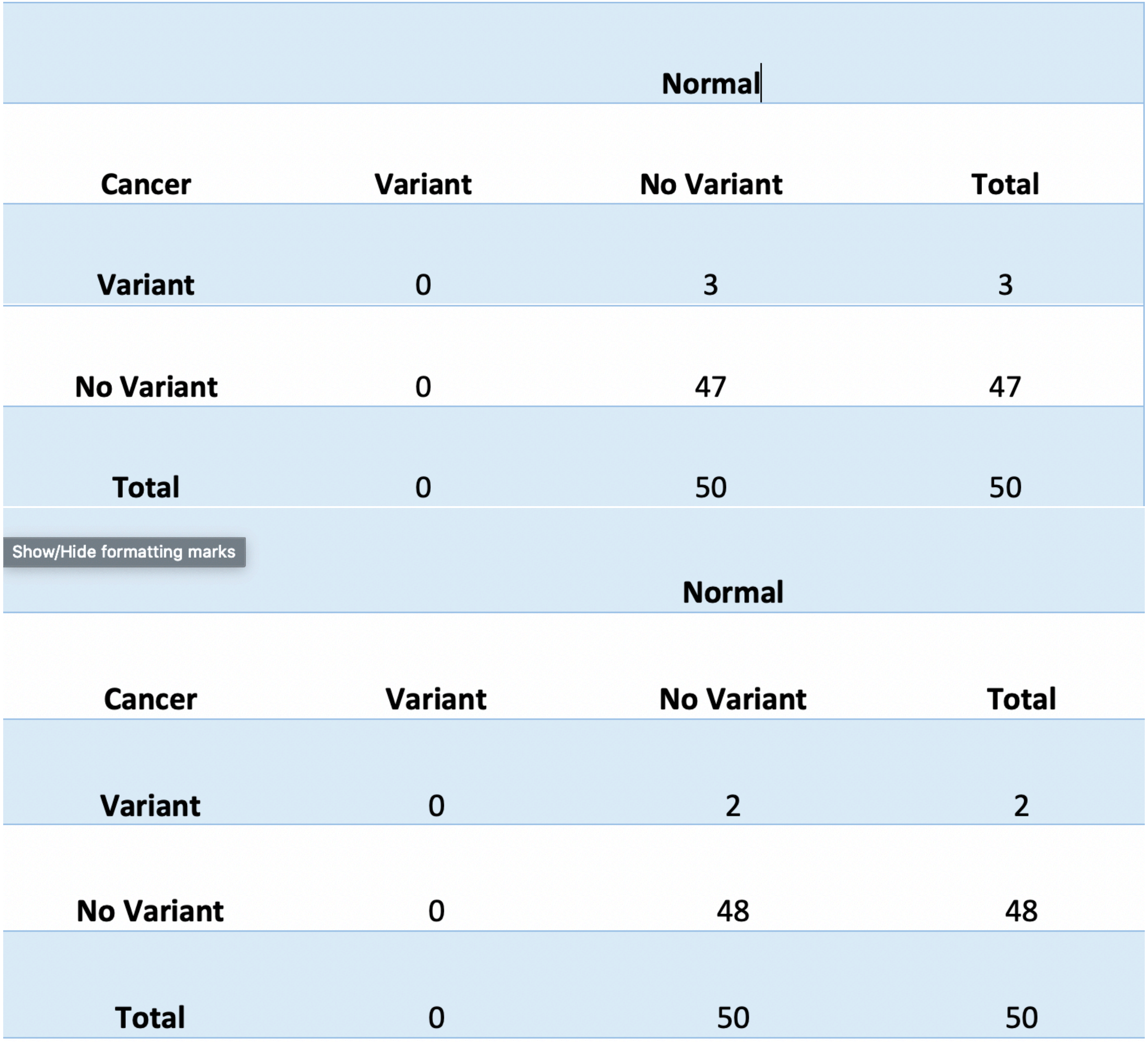


chror	left	ref_	alt	genenam	where	subt1	subt1	subt0r	subt0n(	dbSNP	FATHMM.prediction	prediction
chr11	1099604	C	A	['MUC2']	['CDS']	0	3	0	47	rs34493663	benign	benign
chr16	28638466	G	A	['NPIPB8']	['CDS']	0	2	0	48	rs35184893	benign (high conf.)	benign
chr17	41026893	A	G	['KRTAP1-5']	['CDS']	0	2	0	48	rs746834174	benign (high conf.)	benign
chr17	7674248	T	C	['TP53', 'TP5	['CDS', 'CDS',	0	2	0	48	rs876660807	pathogenic	pathogenic
chr11	1191365	G	A	['MUC5AC']	['CDS']	0	2	0	48	rs140937930	No prediction found	

Figure 5: 5 nsGSV's with > 4% relative frequency.

Chromosome (chr) 11 had 3 variant counts in the tumor sheet while the other chromosomes had 2 each. Two variants occurred on chromosomes 11 and 17 each and one of the variants on chr 17 had a pathogenic prediction from both FATHMM and PROVEAN.

Tables for Binomial (McNemar's) test for for chr11 (table a - MUC2 gene) and other remaining chromosomes (table b).



		Normal	
Cancer	Variant	No Variant	Total
Variant	0	3	3
No Variant	0	47	47
Total	0	50	50

		Normal	
Cancer	Variant	No Variant	Total
Variant	0	2	2
No Variant	0	48	48
Total	0	50	50

The binomial McNemar test for paired data was carried out on the two contingent tables above as follows;

Test;

Given that n=b+c, where b is the number of counts in T1N0 and c is the number of counts in T0N1. Hence from table 1 , c= 0 and b= 3 for chr 11 resulting to n= 0+3 = 3. Since n<30, a binomial McNemar test is used.

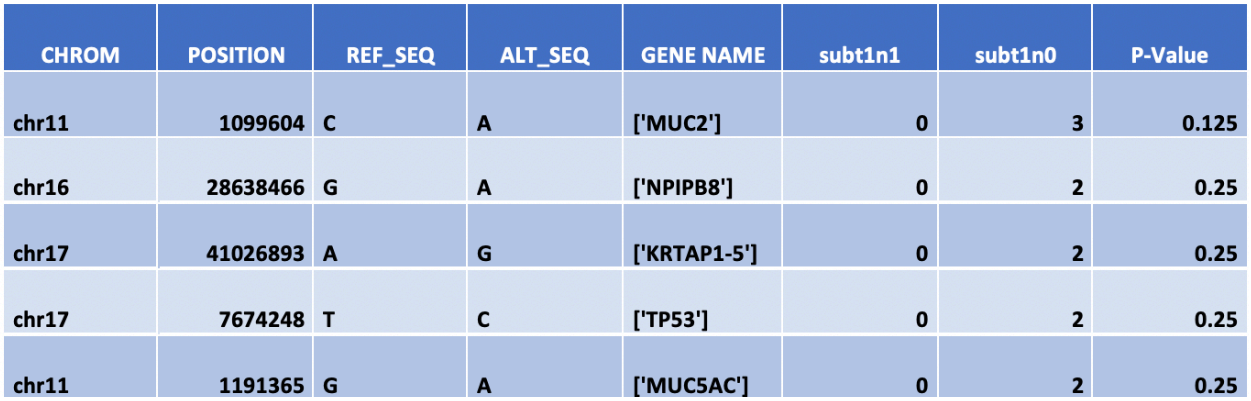
$$H_0 : \pi_t = \pi_n$$

.

$$H_1 : \pi_t > \pi_n$$

. P-value = $\sum_{i=b}^{n} \binom{n}{i} 0.5^i (1-0.5)^{n-i}$=$0.5^n \sum_{i=b}^{n} \binom{n}{i}$

Table c : Results from the Binomial (McNemar) Test.

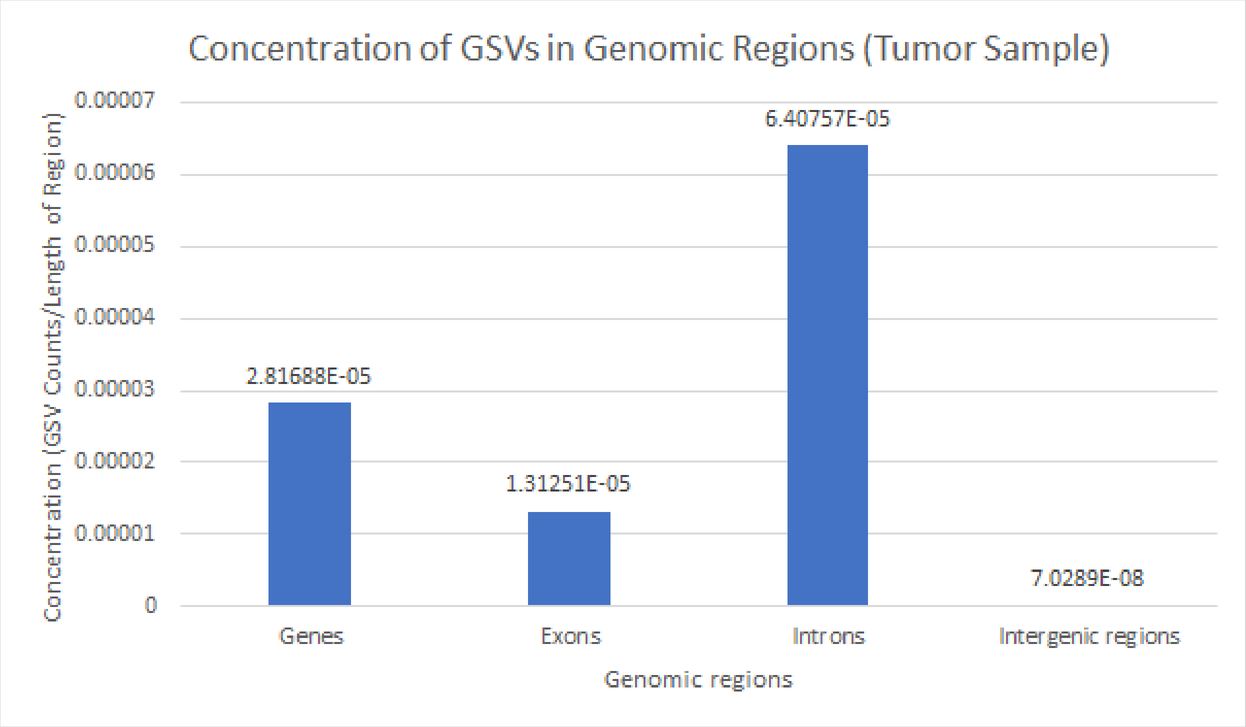


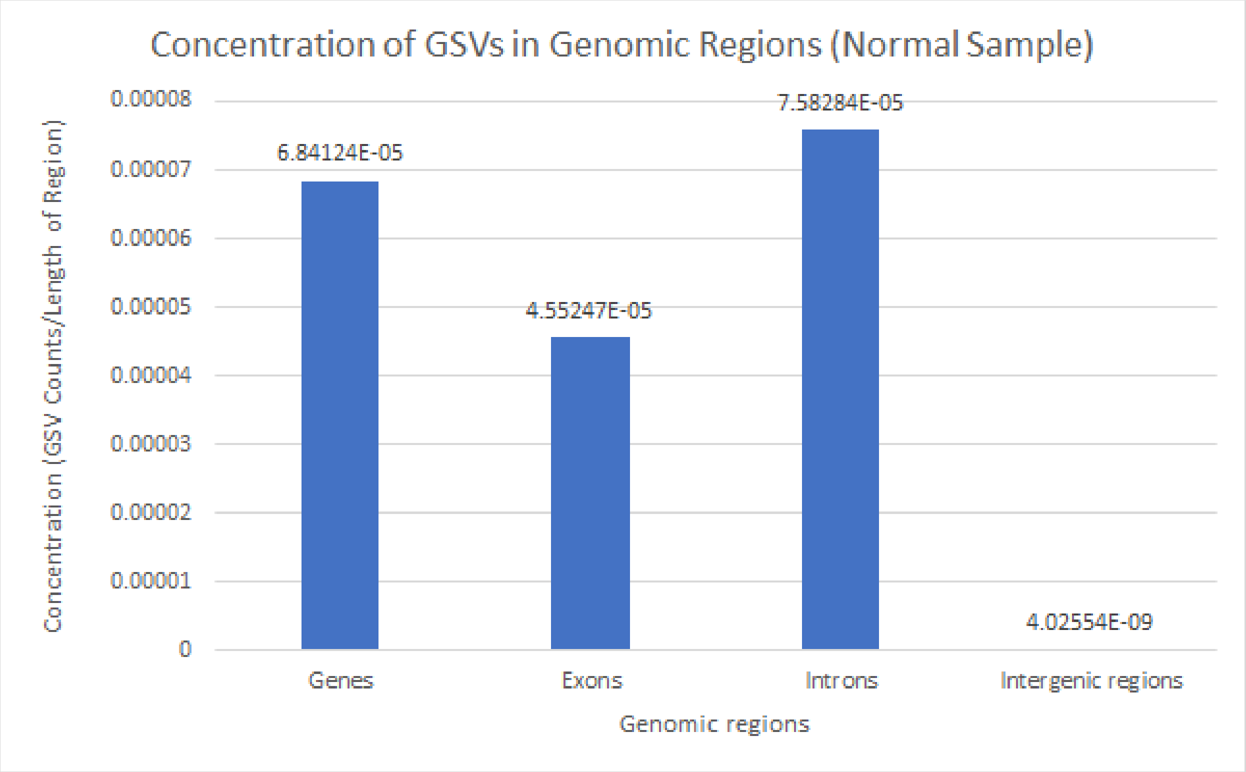
CHROM	POSITION	REF_SEQ	ALT_SEQ	GENE NAME	subt1n1	subt1n0	P-Value
chr11	1099604	C	A	['MUC2']	0	3	0.125
chr16	28638466	G	A	['NPIPB8']	0	2	0.25
chr17	41026893	A	G	['KRTAP1-5']	0	2	0.25
chr17	7674248	T	C	['TP53']	0	2	0.25
chr11	1191365	G	A	['MUC5AC']	0	2	0.25

Given the p-values from the table c, we reject the H_0 and conclude that they are statistically insignificant since the p-values are all less than the $\alpha - value$ of 0.05. Hence the identified genes possibly associated with the disease (cancer) are all benign based on the 50 patients sampled and the thresholds used.

6.4 Main objective II: Identifying genomic regions with high concentrations of GSVs

The counts of GSVS in each genomic region and the length of the genomic region per variant were computed with a python script and bar charts plotted using Microsoft Excel.





It was observed that for both the normal and tunor samples, the concentration of GSVs were highest in the introns second by the genes. The high concentration of GSVs in the genes might just be a reflection of the high concentration of GSVs in the introns since genes are a combination of introns and exons.

The presence of introns in a genome is believed to impose substantial burden on the host. First, unlike self-splicing introns, the excision of spliceosomal introns requires a spliceosome, which is among the largest molecular complexes in the cell, comprising 5 snRNAs and more than 150 proteins (Wahl et al., 2009). Intron-bearing genomes must code for all these proteins and snRNAs. It is estimated that more than 50% of human genetic disorders are caused by disruption of the normal splicing pattern (Lopez-Bigas et al., 2005; Wang and Cooper, 2007). In addition, malfunction of any of the snRNAs and proteins that are necessary for proper splicing will have a general detrimental effect on the cell.

In observing evolution, it appeared that introns had no function, however, the mere existence of transcribed gene parts, that are free from selective constraints triggered an increase in genetic diversity that eventually led to the gain of many intron-related functions. An example of intronic function in contemporary eukaryotes is the increase in protein abundance of intron-bearing genes.

Some introns are so efficient in boosting expression levels, that they are regularly included in constructs in order to guarantee high expression (Clark et al., 1993). Intron-bearing genes in mammals were shown to have higher and broader expression than intronless genes (Shabalina et al., 2010). Reconstruction of the intron–exon evolutionary history in 19 eukaryotes revealed that highly expressed genes tend to have higher intron gain rates (Carmel et al., 2007a).

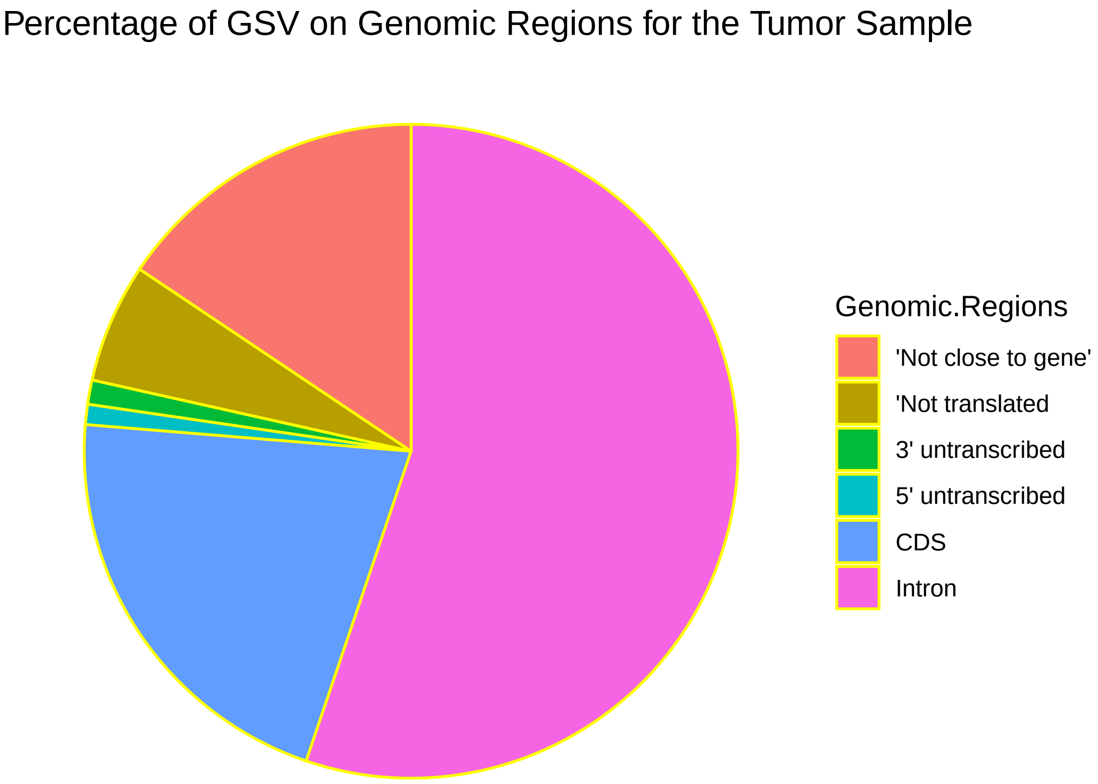
Dysregulation in pre-mRNA alternative splicing is emerging as a hallmark of cancer (Sveen, 2016). Since the initial discovery of frequent point mutations in the core spliceosome subunits in myelodysplastic syndromes and, later, in hematological malignancies (Sveen, 2016), splicing dysregulation has been appreciated as a major contributor to cancer phenotypes.

Zhang et al, (2020) reported that intron retention represented the most salient and consistent feature across the spectrum of prostate cancer entities and positively correlates with prostate cancer stemness and aggressiveness. They also reported that splicing abnormalities impact prostate cancer biology, partially, via switching the isoform expression of key cancer-related genes. Despite this, splicing misregulation can also be exploited therapeutically for treating castration-resistant prostate cancer.

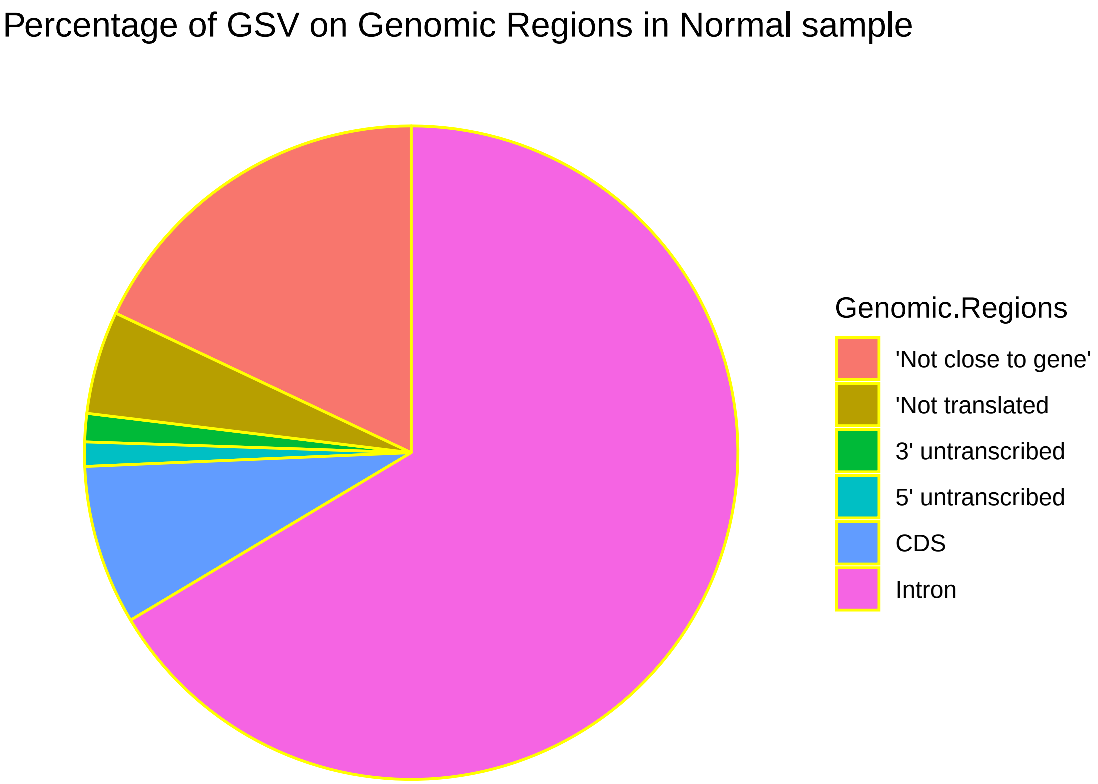
The concentration of GSVs in the genes and exons of the normal sample is a lot higher than that in the tumor sample. This is rather surprising since it could be assumed that the tumor sample would have a higher concentration of variants in the exons. These results imply that the lengths of the genomic regions where the variants in the normal sample are located is shorter than those in the tumor sample. It is not certain why the genomic region length is shorter in the normal sample than tumor sample.

6.5 Specific aim 2 - Percentage of GSV count on each genomic region

The pie chart below showcases the percentage of GSVs per genomic region for the tumor sample of all 50 patients.



Percentage of GSV on Genomic Regions for the Tumor Sample



The pie chart above displays the percentage of GSVs per genomic region for the normal sample of all 50 patients. From the plots it clearly seen that majority of the GSVs on falls in the intronic region, with the least GSV counts found on the 5’ untranscribed region.

Table 3: Percengtage GSV counts on Genomic Regions

Genomic_Regions	Tumor	Percentage_tumor	Normal	Percentage_normal
Genes	8195	82.20	393	79.50
Exons	2691	27.00	64	13.00
Introns	5504	55.20	329	66.50
Protein coding region	2101	21.10	39	7.90
Intergenic Region	1781	17.80	102	20.60
Novel	6720	67.36	176	35.56

The table above displays the percentage of of GSVs found in genomic regions of interest.

Considering the tumor sample from the table above, Genes made up of 8195.0(82.2%) of the GSVs with Intergenic Regions comprising 1781.0(17.8%). Of the total gene GSV counts, 5504(55.2%) were on the Introns and 2691(27%) on the Exons. Also indicated in the table is 2101(21.1%) of the GSVs lying within the Protein Coding Regions. Novel GSV counts from the tumor sample data is 6720(67.4%)

Now considering the normal sample from the table above, Genes made up of 393(79.5%) of the GSV's with Intergenic Regions comprising 102(20.6%). Of the total gene GSV counts, 329(66.5%) were on the Introns and 64(13%) on the Exons. Also indicated in the table is 39(7.9%) of the GSVs lying within the Protein Coding Regions. Novel GSV counts from the tumor sample data is 176(35.6%)

Looking across the table, it is observed that the percentages of GSVs in each genomic region of both the normal and tumor samples are similar in value except for the protein coding region and exons.

The higher percentage of GSVs in the tumor sample in the protein coding region (21.1%) compared to the normal (7.9%) might be a link to explain why the prostate cancer cells became cancerous. The exisitence of many variants in the protein coding region might possibly lead to a malfunction in the way a protein folds; especially if the variant led to a non-synonymous mutation.

Since the exons comprises a region made of the protein coding region and untranslated regions (3' and 5' UTRs), the large difference in the tumor and normal samples percentage for this region is still due to the large difference in protein coding region GSV percentage.

The highest percentage of variants in both tumor and normal samples existed in the genes. This is unsurprising as genes are made up of exons and introns, so will take a higher percentage. The second highest percentage came from the introns which seems to tie with the results gotten from main ojective II (genomic region with highest concentration of GSVs). It is still uncertain why introns are so ubiquitous in the human genome.

Relatively low percentages of GSVs occurred in intergenic regions. Intergenic regions close to a gene have been known to regulate gene expression by turning a gene on and off with the help of transcription factors binding to those regions.

6.6 Specific aim III (self formulated)

Considering the two tools used in the dataset for predicting the disease nature of a GSV, that is FATHMM and PROVEAN, the counts were obtained through cross-tabulation as shown below.

```
##
##              benign pathogenic
##   benign        321        197
##   pathogenic    112        368
```

From the table, there is different distribution of values for each prediction method. A Chi-square test for association was then performed with the results shown below.

Chi-square test for association in consistency in prediction by FATHMM and Provean.

H_0: Prediction of a GSV by FATHMM and Provean are independent.

H_a: Prediction of a GSV by FATHMM and Provean are related.

$\alpha = 0.05$

```
##
##  Pearson's Chi-squared test
##
## data:  data1$FATHMM.prediction and data1$prediction
## X-squared = 151.41, df = 1, p-value < 2.2e-16
```

Since the p-value=2.2e-16 < $\alpha = 0.05$, we reject the null hypothesis H_0 and conclude that the prediction from FATHMM and PROVEAN are related. Meaning that both FATHMM and PROVEAN used in making predictions for non-synonymous variants based on their different algorithms, scales and threshold values have a related outcome. However, using one prediction tool is risky and multiple tools should be used for greater confidence.

7 Conclusion and Suggestions

A significantly greater number of GSVs were seen in the tumor sample than normal sample with an average of 6 GSV counts in the normal sample and 216 in the tumor sample. In identifying possible genetic sequence variants (GSVs) that could be associated with the prostate cancer by performing the McNemar's test, our results did not return any significant variant. However, a variant on chromosome 17 on the TP53 gene had a pathogenic prediction from both FATHMM and PROVEAN. If this computation is repeated with a larger sample size, the results may be different and significant variants may be identified. The highest concentration of GSVs in both the normal and tumor samples were in the intronic region. Previous research has shown that intron retention represented the most salient and consistent feature across the spectrum of prostate cancer entities and positively correlates with prostate cancer stemness and aggressiveness (Zhang et al, 2020) and our results seem to be consistent with this finding. Zhang et al also reported that splicing abnormalities impact prostate cancer biology, partially, via switching the isoform expression of key cancer-related genes. We were surprised with the clear difference GSV concentration on genes in the tumor sample had with the normal sample. The results results implied that the lengths of the genomic regions where the variants in the normal sample are located in is shorter those in the tumor sample. It is not certain why the genomic region length is shorter in the normal sample than tumor sample. We observed also that the highest percentage of GSVs occurred in the intronic region (as seen in the pie chart). This result is similar to the results gotten when determining the genomic region with the highest concentration of GSVs in which the intronic region also had the highest concentration of GSVs. The intergenic regions had a considerably low percentage of GSVs. A possible explanation could be that intergenic regions close to a gene have been known to regulate gene expression by turning a gene on and off with the help of transcription factors binding to those regions, so would less likely have GSVs. A chi-square test for independence shown that there was an association between FATHMM and PROVEAN predictions, but we agreed that multiple tools should be used for greater confidence.

Further research can be done to determine why the concentration of GSVs in genes in the normal sample is much higher than that in tumor samples. It appears that there was a massive insertion of nucleotides in the tumor sample for the tumor sample to have a longer genomic length than the normal sample. More research also needs to be done to understand the mechanism of alternative splicing to better understand why there is intron retention in prostate cancer samples.

A shortcoming of this research is that we cannot guarantee the accuracy of the code created in this project until a bug comes up to indicate a problem with our programs. However, we hope more students can partake in this project to obtain relevant findings.

8 References

Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A. Global cancer statistics, 2012. CA Cancer J Clin. 2015 Mar;65(2):87-108. doi: 10.3322/caac.21262. Epub 2015 Feb 4. PMID: 25651787.

Wang G, Zhao D, Spring DJ, DePinho RA. Genetics and biology of prostate cancer. Genes Dev. 2018;32(17-18):1105-1140. doi:10.1101/gad.315739.118

Cancer Research UK. Prostate Cancer Incidence Statistics [Internet]. 2014 [cited 2020 Nov 22]. Available from: http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/prostatecancer/incidence#heading-One

Singh O, Bolla SR. Anatomy, Abdomen and Pelvis, Prostate. [Updated 2020 Aug 23]. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020 Jan-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK540987/

Cuzick J, Thorat MA, Andriole G, Brawley OW, Brown PH, Culig Z, Eeles RA, Ford LG, Hamdy FC, Holmberg L, Ilic D, Key TJ, La Vecchia C, Lilja H, Marberger M, Meyskens FL, Minasian LM, Parker C, Parnes HL, Perner S, Rittenhouse H, Schalken J, Schmid HP, Schmitz-Dräger BJ, Schröder FH, Stenzl A, Tombal B, Wilt TJ, Wolk A. Prevention and early detection of prostate cancer. Lancet Oncol. 2014 Oct;15(11):e484-92.

van Leenders GJ, Schalken JA. 2003. Epithelial cell differentiation in the human prostate epithelium: implications for the pathogenesis and therapy of prostate cancer. Crit Rev Oncol Hematol 46 Suppl: S3–S10.

Centers for Disease Control and Prevention. 2020. What Is Prostate Cancer?. [online] Available at: https://www.cdc.gov/cancer/prostate/basic_info/what-is-prostate-cancer.htm [Accessed 22 November 2020].

Datta K, Muders M, Zhang H, Tindall DJ. 2010. Mechanism of lymph node metastasis in prostate cancer. Future Oncol 6: 823–836.

Huggins C, Hodges CV. 1941. Studies on prostatic cancer. I. The effect of castration, of estrogen and of androgen injection on serum phosphatases in metastatic carcinoma of the prostate. Cancer Res 1: 293–297.

Goldgar DE, Easton DF, Cannon-Albright LA, Skolnick MH. 1994. Systematic population-based assessment of cancer risk in first-degree relatives of cancer probands. J Natl Cancer Inst 86: 1600–1608.

Lange EM. 2010. Identification of genetic risk factors for prostate cancer: analytic approaches using hereditary prostate cancer families. In Male reproductive cancers: epidemiology, pathology and genetics (ed. Foulkes WD, Cooney KA), pp. 203–228. Springer, New York.

Shenoy D, Packianathan S, Chen AM, Vijayakumar S. 2016. Do African-American men need separate prostate cancer screening guidelines? BMC Urol 16: 19.

Huang FW, Mosquera JM, Garofalo A, Oh C, BacoM, Amin-Mansour A, Rabasha B, Bahl S, Mullane SA, Robinson BD, et al. 2017. Exome sequencing of African-American prostate cancer reveals loss-of-function ERF mutations. Cancer Discov 7: 973–983.

Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, Ghoussaini M, Luccarini C, Dennis J, Jugurnauth-Little S, et al. 2013. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 45: 385–391.

Takata R, Akamatsu S, Kubo M, Takahashi A, Hosono N, Kawaguchi T, Tsunoda T, Inazawa J, Kamatani N, Ogawa O, et al. 2010. Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population. Nat Genet 42: 751–754.

Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, Dadaev T, Leongamornlert D, Anokian E, Cieza- Borrella C, et al. 2018. Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet 50: 928–936.

Huang Q, Whitington T, Gao P, Lindberg JF, Yang Y, Sun J, Vaisanen MR, Szulkin R, Annala M, Yan J, et al. 2014. A prostate cancer susceptibility allele at 6q22 increases RFX6 expression by modulating HOXB13 chromatin binding. Nat Genet 46: 126–135.

Farmer H, McCabe N, Lord CJ, Tutt AN, Johnson DA, Richardson TB, Santarosa M, Dillon KJ, Hickson I, Knights C, et al. 2005. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. Nature 434: 917–921. The Cancer Genome Atlas Research Network. 2015. The molecular taxonomy of primary prostate cancer. Cell 163: 1011–1025.

Albany C, Alva AS, Aparicio AM, Singal R, Yellapragada S, Sonpavde G, Hahn NM. 2011. Epigenetics in prostate cancer. Prostate Cancer 2011: 580318.

Yegnasubramanian S. 2016. Prostate cancer epigenetics and its clinical implications. Asian J Androl 18: 549–558.

Margulies, M., Egholm, M., Altman, W. et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376–380 (2005). https://doi.org/10.1038/nature03959

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature 431, 931–945 (2004). https://doi.org/10.1038/nature03001

Metzker, M. Sequencing technologies — the next generation. Nat Rev Genet 11, 31–46 (2010). https://doi.org/10.1038/nrg2626

Wang, Z., Sun, B. Annular multiphase flow behavior during deep water drilling and the effect of hydrate phase transition. Pet. Sci. 6, 57–63 (2009). https://doi.org/10.1007/s12182-009-0010-3

Park, P. ChIP–seq: advantages and challenges of a maturing technology. Nat Rev Genet 10, 669–680 (2009). https://doi.org/10.1038/nrg2641

L. Liu, et al. Comparison of next-generation sequencing systems J. Biomed. Biotechnol., 2012 (2012), p. 251364

Saunders, C., Miller, N., Soden, S., Dinwiddie, D., Noll, A., Alnadi, N., Andraws, N., Patterson, M., Krivohlavek, L., Fellis, J., Humphray, S., Saffrey, P., Kingsbury, Z., Weir, J., Betley, J., Grocock, R., Margulies, E., Farrow, E., Artman, M., Safina, N., Petrikin, J., Hall, K. and Kingsmore, S., 2012. Rapid Whole-Genome Sequencing for Genetic Disease Diagnosis in Neonatal Intensive

Care Units. Science Translational Medicine, [online] 4(154), pp.154ra135-154ra135. Available at: https://stm.sciencemag.org/content/4/154/154ra135.

Choi, M., Scholl, U., Ji, W., Liu, T., Tikhonova, I., Zumbo, P., Nayir, A., Bakkaloğlu, A., Özen, S., Sanjad, S., Nelson-Williams, C., Farhi, A., Mane, S. and Lifton, R., 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proceedings of the National Academy of Sciences, [online] 106(45), pp.19096-19101. Available at: https://www.pnas.org/content/106/45/19096 [Accessed 8 December 2020].

van Dijk, E., Auger, H., Jaszczyszyn, Y. and Thermes, C., 2014. Ten years of next-generation sequencing technology. Trends in Genetics, [online] 30(9), pp.418-426. Available at: https://www.sciencedirect.com/science/article/pii/S0168952514001127#glo0005.

Chorev, M., & Carmel, L. (2012). The function of introns. Frontiers in genetics, 3, 55. https://doi.org/10.3389/fgene.2012.00055

Sveen A, Kilpinen S, Ruusulehto A, Lothe RA, Skotheim RI. Aberrant RNA splicing in cancer; expression changes and driver mutations of splicing factor genes. Oncogene. 2016;35:2413–2427. doi: 10.1038/onc.2015.318.

Zhang, D., Hu, Q., Liu, X., Ji, Y., Chao, H. P., Liu, Y., Tracz, A., Kirk, J., Buonamici, S., Zhu, P., Wang, J., Liu, S., & Tang, D. G. (2020). Intron retention is a hallmark and spliceosome represents a therapeutic vulnerability in aggressive prostate cancer. Nature communications, 11(1), 2089. https://doi.org/10.1038/s41467-020-15815-7.

9 Appendix

The McNemar test was done manually by the formula and excel used for filtering the relative percentage count.

Attached to the report is a separate zip file containing the R and Python codes used for this report.