Utilization of Next Generation Sequencing to Determine the Importance of Specific Non-Coding DNA Segments in Breast Cancer Development

Hrithik Jha

November 14, 2018

Contents

1 Abstract

An area of research that is garnering attention is the field of alternative splicing, the variance of splicing patterns and the inclusion of non-coding DNA segments into the transcription phase of the central dogma. These variances in splicing patterns have been theorized to be associated with tumor suppression or oncogenesis. This research aims to create a preemptive method of discovering breast cancer through the use of alternative splicing events.DNA was extracted from 1,126 patients. Out of these patients, 1,038 were tumor patients with breast cancer and 88 were normal patients. After the DNA of each patient was extracted, RNA-seq, a form of next-generation sequencing, was used to sequence a patient's DNA. The data collected from RNA-seq was stored in a binary alignment mapping (BAM) file, which was then converted into a sequence alignment mapping (SAM) file. Each DNA strand examined and profiled on the SAM file contained a read-ID, a Flag value, (which determined which strand - 5” - 3” OR 3” - 5" was read during sequencing), a CIGAR value (which was used to determine the start and end position of the introns), and start and ending positions of the nucleotide sequenced. Using python programming analysis, DNA fragments were analyzed for alternative splicing events and the intron associated with the splicing event was sequenced. These splicing events were compared with a refgene sequence to quantify the expression of introns within normal and tumor patient groups. Using a negative binomial distribution approach, genes were analyzed and differentially expressed introns, introns that are statistically inclined to promote tumorigenesis or tumor suppression, were determined after normalization of the data set. Analysis showed that introns located in loci of genes associated with phosphoprotein development and the cytoplasm were the most differentially expressed between tumor and normal patients. The results proved to be significant, as the p values calculated were much less than the 0.01 benchmark.

2 Introduction

Breast cancer is at the moment, the most commonly diagnosed cancer and the leading global cause of death in women, averaging over 1.38 million diagnoses and over 458,000 casualties each year[1]. Scientists have theorized that women with breast cancer contain certain risk factors that increase the chance of contraction. While the most common risk factors include usage of drugs, alcohol, and the hereditary history associated with a patient, there may be more at play than lifestyle, environmental, and genetic factors in contracting cancer.

The intron was discovered in 1977 by Richard Roberts and Phil Sharp. Considered an interruption located in the eukaryotic gene, introns seemed to have no purpose within the central dogma. These loci are removed via the use of spliceosomes when pre-mRNA (produced after transcription) is converted to mRNA prior to protein synthesis/translation. However, this method of removing introns, known as splicing, does not always remove the entire intron. Rather than exclusively coding exons for protein synthesis, there are instances in the human body when introns code for proteins. The inclusion of introns into mRNA is known as Alternative Splicing.

Alternative Splicing is a regulated process that increases the number of protein isoforms produced within one set of DNA. It is also known to play an important role in cellular differentiation. However, scientists have recently observed that alternative splicing activity may be responsible for tumorogenesis, or tumor development.

By determining specific introns associated with tumorogenesis and tumor suppression (due to their inclusion in mRNA through alternative splicing, the experimenter can be able to preemptively prevent breast cancer. The connection between intron retention throughout transcription and translation and increased oncogenesis, or tumor suppression, could suggest that differentially expressed introns (DEIs) could play a crucial part in cancer development.

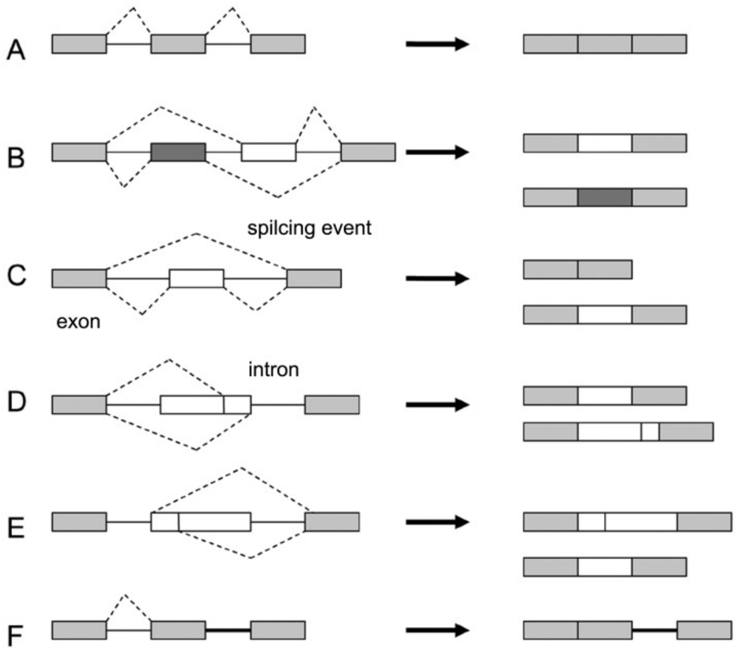


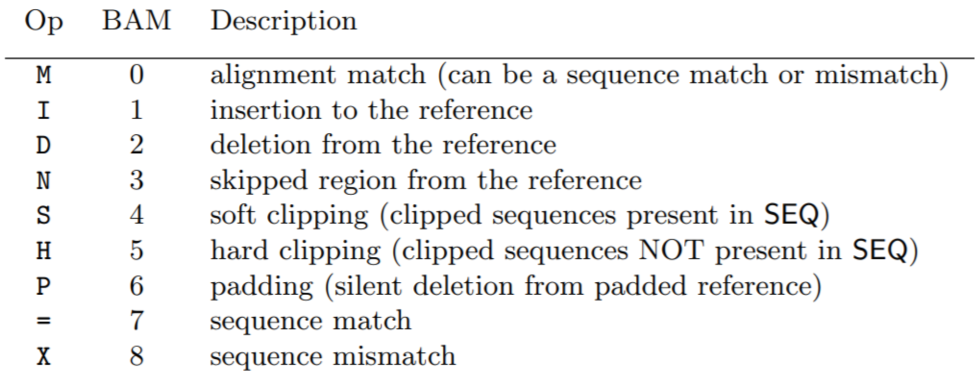
Figure 1: Alternative Splicing events results in multiple proteins being produced from a single gene. [1]

3 Materials, Methods, and Procedures

In order to test the hypothesis that differentially expressed introns could be useful in determining cancer development, a four step procedure was created. First and foremost, data was collected from 1,126 patients, 1038 of whom were tumor patients and 88 of whom were normal patients void of breast cancer. The patients then provided their DNA for sequencing through RNA-seq.

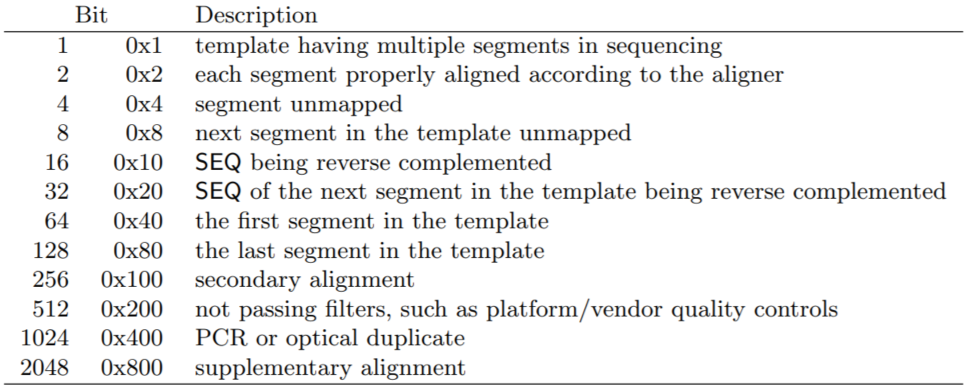
RNA-seq utilizes deep sequencing technologies to convert a population of RNA to a library of cDNA fragments. The nucleotides are sequenced and then aligned with a reference genome or transcriptome, and are classified as junction reads, exonic reads, or poly(A) end reads. RNA-seq offers advantages relative to DNA microarrays and can provide more accurate sequencing. The sequencing tool also detects variances that are otherwise rendered obsolete.

Data from the RNAseq process is stored in a Binary Alignment Mapping (BAM) file. The BAM file is converted to a Sequence Alignment Mapping File, which contains specific information about each strand sequenced through RNA-seq. Of the groups of information provided, the CIGAR value, FLAG Value, and Position of the strands is used to find the introns associated with a DNA fragment.



Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch

Figure 2: The CIGAR value of a DNA fragment consists of a -M-N-M format. The M's represent a match between the reference genome and the DNA fragment while the N's represent a mismatch. CIGAR is used to find alternative splicing and intron inclusion in sequenced fragments. [3]



Bit		Description
1	0x1	template having multiple segments in sequencing
2	0x2	each segment properly aligned according to the aligner
4	0x4	segment unmapped
8	0x8	next segment in the template unmapped
16	0x10	SEQ being reverse complemented
32	0x20	SEQ of the next segment in the template being reverse complemented
64	0x40	the first segment in the template
128	0x80	the last segment in the template
256	0x100	secondary alignment
512	0x200	not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	supplementary alignment

Figure 3: The FLAG value of a DNA Fragment consists of a base 10 number. Upon Conversion of the number to base 2, each digit shows properties of the strand. [3]

After the fragments are analyzed through python programming, they were compared to a refgene file. The refgene file contains a list of introns and the genes they are located in. Using python and UNIX, all introns found to be alternatively spliced were counted per file and a matrix between matched samples (patients that donated both tumor and normal tissue for sequencing) was created. 81 patients, in this experiment, donated both a tumor and normal tissue sample for sequencing.

Before showing results, the data provided was prepared for analysis on edger. Intron events that were found to have less than 2 counts per million (cpm) were removed from analysis. Normalization Factors were calculated to reduce bias caused by variations in total RNA ouput in a per-cell basis. The calcNormFactors function in edger normalizes for RNA composition by finding a set of scaling factors for the library sizes that minimize the log-fold changes between the samples for most genes. A design matrix was created from the matched matrix file and dispension factors, or the biological coefficient of variation for each intron was calculated and plotted. A negative binomial distribution was utilized to analyze the data as opposed to a typical Poisson distribution due to the variance found with groups (not all tumor samples are similar and not all normal samples are similar).

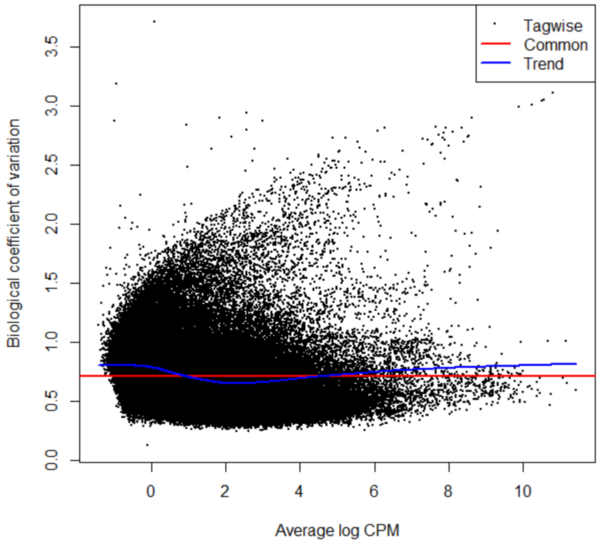


Figure 4: The BCV CPM relationship found in each of the samples follows a negative binomial distribution [4]

4 Results

In order to visualize the matched matrix data (which contains over 20,000 rows of introns), a plotMDS command was implemented. Also known as multidimensional scaling, the plotMDS algorithm aims to show N dimension data into 2 dimensions while maintaining distance as much as possible. Because plotMDS employs non-linear dimensionality reduction, axes cannot be labeled. However, contrary to its linear counterpart Principal Component Analysis (PCA), multidimensional scaling is much more versatile and is generally more accurate and precise.

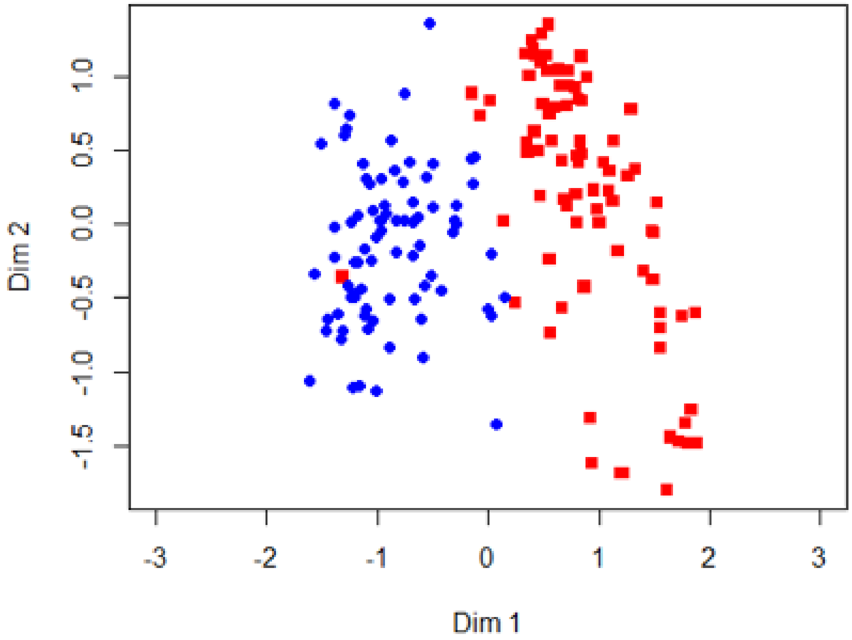
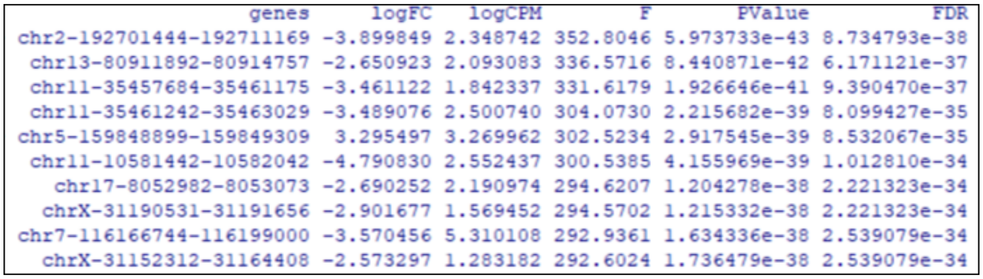


Figure 5: Multi-Dimensional Scaling plot on Differentially Expressed Introns [5]

Qualitatively, there already seems to be a distinct difference between intron splicing events in tumor (blue) and normal(red) samples.

A Quasi-Likelihood F Test was performed on all introns. Introns with a p value less than 0.01 were shown to be differentially expressed and FDR values, or False Discovery Rates of less than 0.05 were shown to be statistically significant.

The log fold change value, the second column from the results of the QLFTests results is a ratio between the expression of the intron in tumor trials over normal trials. If the intron



```
                    genes    logFC  logCPM        F       PValue          FDR
chr2-192701444-192711169 -3.899849 2.348742 352.8046 5.973733e-43 8.734793e-38
 chr13-80911892-80914757 -2.650923 2.093083 336.5716 8.440871e-42 6.171121e-37
 chr11-35457684-35461175 -3.461122 1.842337 331.6179 1.926646e-41 9.390470e-37
 chr11-35461242-35463029 -3.489076 2.500740 304.0730 2.215682e-39 8.099427e-35
chr5-159848899-159849309  3.295497 3.269962 302.5234 2.917545e-39 8.532067e-35
 chr11-10581442-10582042 -4.790830 2.552437 300.5385 4.155969e-39 1.012810e-34
   chr17-8052982-8053073 -2.690252 2.190974 294.6207 1.204278e-38 2.221323e-34
  chrX-31190531-31191656 -2.901677 1.569452 294.5702 1.215332e-38 2.221323e-34
chr7-116166744-116199000 -3.570456 5.310108 292.9361 1.634336e-38 2.539079e-34
  chrX-31152312-31164408 -2.573297 1.283182 292.6024 1.736479e-38 2.539079e-34
```

Figure 6: QLFTest results [6]

is expressed more in tumor samples then normal samples, the logFC value will be positive. The opposite is true when the log fold change value is negative.

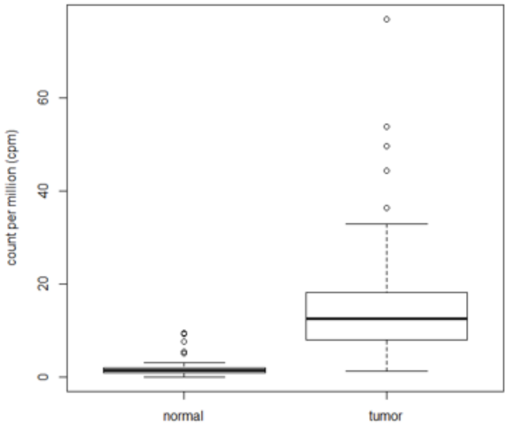
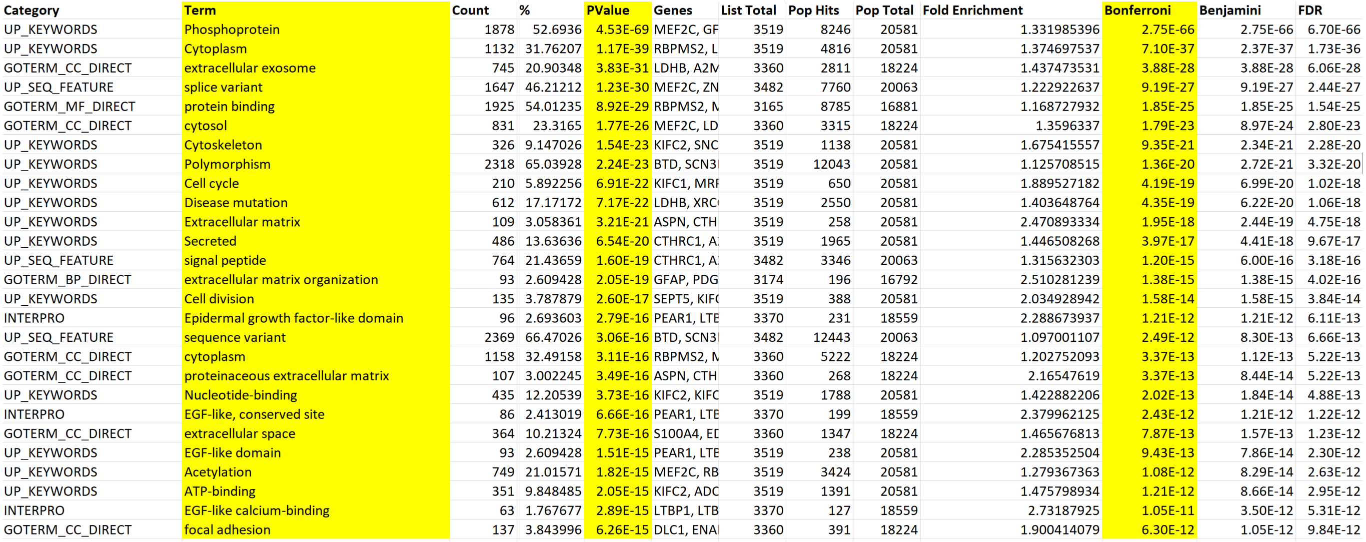


Figure 7: Log-fold box plot visualization [7]

The genes that contained these differentially expressed introns was found. These genes are then filtered out. Genes that are already differentially expressed are removed from the gene list; the experiment aims to find differentially expressed introns and which genes they are located in. If a gene is already differentially expressed in tumor cells than cancer cells or vice versa, then the inclusion of the gene would cause non-differentially expressed introns to seem differentially expressed.

A Gene Ontology is performed to show the functions of the genes which contain differentially expressed introns. The functions the genes are related to are shown and their p values and

Bonferri adjusted FDR values (adjusted due to multiple testing, or inferences made due to current observations).



Category	Term	Count	%	PValue	Genes	List Total	Pop Hits	Pop Total	Fold Enrichment	Bonferroni	Benjamini	FDR
UP_KEYWORDS	Phosphoprotein	1878	52.6936	4.53E-69	MEF2C, GF	3519	8246	20581	1.331985396	2.75E-66	2.75E-66	6.70E-66
UP_KEYWORDS	Cytoplasm	1132	31.76207	1.17E-39	RBPMS2, L	3519	4816	20581	1.374697537	7.10E-37	2.37E-37	1.73E-36
GOTERM_CC_DIRECT	extracellular exosome	745	20.90348	3.83E-31	LDHB, A2M	3360	2811	18224	1.437473531	3.88E-28	3.88E-28	6.06E-28
UP_SEQ_FEATURE	splice variant	1647	46.21212	1.23E-30	MEF2C, ZN	3482	7760	20063	1.222922637	9.19E-27	9.19E-27	2.44E-27
GOTERM_MF_DIRECT	protein binding	1925	54.01235	8.92E-29	RBPMS2, M	3165	8785	16881	1.168727932	1.85E-25	1.85E-25	1.54E-25
GOTERM_CC_DIRECT	cytosol	831	23.3165	1.77E-26	MEF2C, LD	3360	3315	18224	1.3596337	1.79E-23	8.97E-24	2.80E-23
UP_KEYWORDS	Cytoskeleton	326	9.147026	1.54E-23	KIFC2, SNC	3519	1138	20581	1.675415557	9.35E-21	2.34E-21	2.28E-20
UP_KEYWORDS	Polymorphism	2318	65.03928	2.24E-23	BTD, SCN3	3519	12043	20581	1.125708515	1.36E-20	2.72E-21	3.32E-20
UP_KEYWORDS	Cell cycle	210	5.892256	6.91E-22	KIFC1, MRF	3519	650	20581	1.889527182	4.19E-19	6.99E-20	1.02E-18
UP_KEYWORDS	Disease mutation	612	17.17172	7.17E-22	LDHB, XRC	3519	2550	20581	1.403648764	4.35E-19	6.22E-20	1.06E-18
UP_KEYWORDS	Extracellular matrix	109	3.058361	3.21E-21	ASPN, CTH	3519	258	20581	2.470893334	1.95E-18	2.44E-19	4.75E-18
UP_KEYWORDS	Secreted	486	13.63636	6.54E-20	CTHRC1, A	3519	1965	20581	1.446508268	3.97E-17	4.41E-18	9.67E-17
UP_SEQ_FEATURE	signal peptide	764	21.43659	1.60E-19	CTHRC1, A	3482	3346	20063	1.315632303	1.20E-15	6.00E-16	3.18E-16
GOTERM_BP_DIRECT	extracellular matrix organization	93	2.609428	2.05E-19	GFAP, PDG	3174	196	16792	2.510281239	1.38E-15	1.38E-15	4.02E-16
UP_KEYWORDS	Cell division	135	3.787879	2.60E-17	SEPT5, KIF	3519	388	20581	2.034928942	1.58E-14	1.58E-15	3.84E-14
INTERPRO	Epidermal growth factor-like domain	96	2.693603	2.79E-16	PEAR1, LTE	3370	231	18559	2.288673937	1.21E-12	1.21E-12	6.11E-13
UP_SEQ_FEATURE	sequence variant	2369	66.47026	3.06E-16	BTD, SCN3	3482	12443	20063	1.097001107	2.49E-12	8.30E-13	6.66E-13
GOTERM_CC_DIRECT	cytoplasm	1158	32.49158	3.11E-16	RBPMS2, M	3360	5222	18224	1.202752093	3.37E-13	1.12E-13	5.22E-13
GOTERM_CC_DIRECT	proteinaceous extracellular matrix	107	3.002245	3.49E-16	ASPN, CTH	3360	268	18224	2.16547619	3.37E-13	8.44E-14	5.22E-13
UP_KEYWORDS	Nucleotide-binding	435	12.20539	3.73E-16	KIFC2, KIFC	3519	1788	20581	1.422882206	2.02E-13	1.84E-14	4.88E-13
INTERPRO	EGF-like, conserved site	86	2.413019	6.66E-16	PEAR1, LTE	3370	199	18559	2.379962125	2.43E-12	1.21E-12	1.22E-12
GOTERM_CC_DIRECT	extracellular space	364	10.21324	7.73E-16	S100A4, ED	3360	1347	18224	1.465676813	7.87E-13	1.57E-13	1.23E-12
UP_KEYWORDS	EGF-like domain	93	2.609428	1.51E-15	PEAR1, LTE	3519	238	20581	2.285352504	9.43E-13	7.86E-14	2.30E-12
UP_KEYWORDS	Acetylation	749	21.01571	1.82E-15	MEF2C, RB	3519	3424	20581	1.279367363	1.08E-12	8.29E-14	2.63E-12
UP_KEYWORDS	ATP-binding	351	9.848485	2.05E-15	KIFC2, ADC	3519	1391	20581	1.475798934	1.21E-12	8.66E-14	2.95E-12
INTERPRO	EGF-like calcium-binding	63	1.767677	2.89E-15	LTBP1, LTB	3370	127	18559	2.73187925	1.05E-11	3.50E-12	5.31E-12
GOTERM_CC_DIRECT	focal adhesion	137	3.843996	6.26E-15	DLC1, ENA	3360	391	18224	1.900414079	6.30E-12	1.05E-12	9.84E-12

Figure 8: Gene Ontology Analysis with Bonferroni, Benjamini adjusted FDR[8]

5 Discussion and Conclusions

Differentially expressed introns play an important role in both tumorigenesis and tumor suppression, concurring with the stated hypothesis. The highest differentially expressed introns came from genes associated with phosphoproteins,the development of the Cytoplasm, and the production of extracellular exosomes. With p and adjusted FDR values of 4.53e-69, 2.75e-66, and 1.17e-39, 7.1e-7 respectively, the differentially expressed introns found are statistically significant.

Genes that participate in phosphoprotein production include MEF2C, GFER, AMOTL1.

1. MEF2c plays a role in myogenesis, the formation of muscular tissue.
2. GFER is one of the factors responsible for the capacity of the mammilian liver.
3. AMOTL1 encodes for a peripheral protein that is a component for tight junctions.

Genes that participate in cytoplasm development include RBPMS2 and LDHB.

1. RBPMS2 contributes to cell differentiation and proliferation in gastrointestinal levels.
2. LDHB catalyzes the interconversion of pyruvate and lactate.

The results above show how oncogenesis is not exclusively caused by genetics. There are substances inside the DNA, that cause tumor development.

6 Future Applications and Future Research

This experiment has laid the foundation to a new form of treatment towards many types of cancers. Through the use of this data, doctors and oncologists can diagnose the danger before the symptoms show. Currently, diagnosis of breast cancer is long, and tools such as mammograms are often costly and ineffective. Now that this experiment has effectively found a relationship between alternative splicing, introns, and tumorogenesis, sequencing can be a new, more effective method of diagnosing a tumor.

Additionally, for a long time, scientists have had no answer to the question, "what is an intron's importance?". This project has enforced the notion that introns have an importance in tumor suppression and tumor development. This project has effectively paved the way for additional research on the properties of the intron.

The next step to this research would be to implement machine learning with the data I have collected. With machine Learning, detection of potential tumor development can be easily established. Childhood genetic tests can be taken prior to determine the risk an individual has in contracting a tumor.

References

[1] Nancy Martinez-Montiel *Alternative Splicing in Breast Cancer and the Potential Development of Therapeutic Tools* (2017).

[2] Robinson, M. D. *EdgeR: a Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.* Bioinformatics, vol. 26, no. 1, (2009).

[3] Hernaez Long *GeneComp, a New Reference-Based Compressor for SAM Files. 2017 Data Compression Conference (DCC)* (2017).

[4] Competition Entrant. *Figure 5: The BCV CPM relationship found in each of the samples follows a negative binomial distribution.*

[5] Competition Entrant. *Figure 6: Multi-Dimensional Scaling Plot on Differentially Expressed Introns.*

[6] Competition Entrant. *Figure 7: QLFTest results.*

[7] Competition Entrant. *Figure 8: Log-fold box plot visualization.*

[8] Competition Entrant. *Figure 9: Gene Ontology Analysis with Bonferroni and Benjamini adjusted FDR.*