Targeting Early Detection:

Whole-Exome Analysis for Lung Cancer Biomarkers

Introduction:

A major public health issue, lung cancer continues to be one of the most common and deadly cancers in the world. The majority of lung cancer cases are discovered at an advanced stage when there are few viable treatment options and the chance of survival is significantly decreased. Therefore, finding efficient early detection methods is crucial for enhancing patient outcomes.

In the effort to find and validate biomarkers that can aid in the early diagnosis of lung cancer, whole-exome analysis, a potent genomic tool, has become a promising direction. This novel method enables a thorough analysis of the human genome's coding regions, providing information on the genetic changes that underlie the onset and development of lung cancer.

Project Objective:

1. Identify and scrutinize potential biomarkers, focusing on single nucleotide polymorphisms (SNPs) and insertions/deletions (Indels) associated with early-stage lung cancer.

2. Investigate the genes linked to these biomarkers and explore their pathways to determine their relevance and significance in the context of lung cancer.

With improved management and early detection of lung cancer, this study hopes to lower the disease's high mortality rate. In order to accomplish this, in this project we will perform whole-exome sequencing and comparative analysis on DNA samples taken from stage 1 lung cancer tissue and nearby normal lung tissue, concentrating on somatic mutations specific to the tumor tissue that may act as early-stage lung cancer biomarkers.

Dataset Details:

The normal paired sample:

`N_231335_R1_chr5.fastq.gz`

`N_231335_R2_chr5.fastq.gz`

The Tumor Paired Sample:

`T_231336_R1_chr5.fastq.gz`

`T_231336_R2_chr5.fastq.gz`

`RG\tID:group1\tSM:sample1\tPL:illumina\tLB:lib1\tPU:unit1`

`Reference Genome: hg19.chr5_12_17.fa`

Methods:

There are several crucial steps involved in the analysis of whole-exome sequencing data to find potential biomarkers linked to early-stage lung cancer. Here is a strategy for approaching this analysis:

1. **Preprocessing:**

   We skipped this step, due to the following reasons.

   - The reference genome is well-established and suitable for the analysis at hand.
   - The reference genome is accurate and comprehensive.

2. **Mapping:**

   Mapping paired-end DNA sequences of normal pair and tumor pair against a human reference genome hg19.chr5_12_17.fa having Chr5, Chr12, and Chr17 data.

   Mapping is a crucial step in our project where we align paired-end DNA sequences from both normal and tumor samples to the human reference genome hg19.chr5_12_17.fa, specifically targeting genomic regions on Chromosomes 5, 12, and 17. This process allows us to pinpoint and understand variations, mutations, or structural alterations in these specific chromosomal regions associated with the normal and tumor samples. By aligning the sequences to the reference genome, we can accurately identify and analyze differences, aiding in the detection of potential genetic abnormalities or markers related to the tumor. Mapping is the foundation of our analysis, providing a comprehensive view of genomic changes and paving the way for further investigations into the molecular landscape of the samples under study. In essence, mapping serves as a critical bridge connecting raw sequencing data to meaningful insights about the genomic makeup of normal and tumor pairs.

   Algorithm:

   ## 1. Alignment with BWA-MEM

   ```
   bwa mem -R '@RG\tID:group1\tSM:sample1\tPL:illumina\tLB:lib1\tPU:unit1'
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/hg19_ref/hg19.chr5_12_17.fa
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/N_231335_R1_chr5.fastq.gz
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/N_231335_R2_chr5.fastq.gz > normal_sample.sam
   ```

   ```
   bwa mem -R '@RG\tID:group1\tSM:sample1\tPL:illumina\tLB:lib1\tPU:unit1'
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/hg19_ref/hg19.chr5_12_17.fa
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/T_231336_R1_chr5.fastq.gz
   /class/dsa8110_genomic_analytics/MR_DATA/test_t-
   n_somatic/T_231336_R2_chr5.fastq.gz > tumor_sample.sam
   ```

   The tool used: BWA MEM (V. 0.7.17-r1188)

   BWA(the Burrows-Wheeler Aligner) is a software package for mapping DNA sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the rest two for longer sequences

ranged from 70bp to a few megabases. BWA-MEM and BWA-SW share similar features such as the support of long reads and chimeric alignment, but BWA-MEM, which is the latest, is generally recommended as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

Input Files: Fasta samples of normal and tumor-paired end DNAs.

Output Files : .sam files for normal and tumor samples.

3. Sam to Bam and sorting:

The SAM file is a text-based, tab-separated file that provides detailed information about each read and its respective alignment to the reference genome. The compressed binary counterpart of a SAM file is known as a BAM file, which is used to minimize storage requirements and to enable quick indexing and data retrieval

SAM files from the tumor and normal samples are converted to sorted BAM files using Sam tools. SAM (Sequence Alignment Map): This is a human-readable, text-based format for storing biological sequences that are aligned to a reference genome. BAM (Binary Alignment Map): This is the binary version of a SAM file. It is more space-efficient and allows for faster data retrieval.

Algorithm:

```
samtools view -bS tumor_sample.sam | samtools sort -o tumor_sample_sorted.bam
samtools view -bS normal_sample.sam | samtools sort -o normal_sample_sorted.bam
```

input files: .sam files
output files: .bam files

The tool used: Sam tools (V. 1.10.2)

4. Marking Duplicate Reads:

The process of removing duplicates is an important step in the analysis of Next-Generation Sequencing (NGS) data. Duplicates can arise from various sources including PCR amplification and sequencing errors. These duplicates can skew downstream analyses such as variant calling and gene expression studies, leading to inaccurate results.

In this step, the duplicate reads, and the sorted BAM files of both tumor and normal samples are identified, ensuring data integrity and enhancing the accuracy of downstream genomic analyses. The generated metrics, encapsulated in files like `tumor_dedup_metrics.txt` and `normal_dedup_metrics.txt`, offer valuable

insights into the deduplication process, contributing to a comprehensive assessment of data quality.

Algorithm/Program:

gatk MarkDuplicates -I tumor_sample_sorted.bam -O tumor_sample_dedup.bam -M tumor_dedup_metrics.txt

gatk MarkDuplicates -I normal_sample_sorted.bam -O normal_sample_dedup.bam -M normal_dedup_metrics.txt

input files: .bam files
output files: dedup.bam and dedup_metrics.txt files

The tool used: GATK's Mark Duplicates (V. 4.1.0.0)

5. Generate recalibration report for tumor and normal pair:

Correct variant calling depends on the quality of the bases, and low quality bases are less trustworthy. Systematic errors arise in sequencing that can cause the base quality to be more or less than its correct value. GATK uses machine learning to model the errors and adjust the quality scores accordingly. This never changes the base itself, just the quality score. The first of two steps is the BaseRecalibrator. BaseRecalibrator takes a set of known variants from, e.g., dbSNP, to prevent modelling true variants as errors, and then produces a recalibration table based on various covariates. By default these are read group, reported quality score, machine cycle (position in the read), and the nucleotide context (represented by all possible dinucleotides). In the output are several tables of covariate values and other metrics.

Recalibration reports were generated for both tumor and normal samples, leveraging known variant sites from the provided dbSNP database (dbsnp.b147.chr5_12_17.vcf). These recalibration tables, such as tumor_recal_data.table and normal_recal_data.table plays a critical role in refining the accuracy of variant calls by correcting for systematic errors in the sequencing data.

Algorithm/Program:

gatk BaseRecalibrator -I tumor_sample_dedup.bam -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa --known-sites /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/dssnp/dbsnp.b147.chr5_12_17.vcf -O tumor_recal_data.table

gatk BaseRecalibrator -I normal_sample_dedup.bam -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa --known-sites

/class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/dssnp/dbsnp.b147.chr5_12_17.vcf -O normal_recal_data.table

input files: dedup.bam files

output files: normal_recal_data.table `files`

The tool used: GATK Recalibration (V. 4.1.0.0)

Recalibrating:

The recalibration data generated earlier is applied to refine the base quality scores of reads in both tumor and normal samples. This crucial step enhances the accuracy of base calls, correcting for systematic errors, and the resulting recalibrated BAM files, such as tumor_sample_recal.bam and normal_sample_recal.bam, are ready for subsequent genomic analyses.

Algorithm/Program:

gatk ApplyBQSR -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa -I tumor_sample_dedup.bam --bqsr-recal-file tumor_recal_data.table -O tumor_sample_recal.bam

gatk ApplyBQSR -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa -I normal_sample_dedup.bam --bqsr-recal-file normal_recal_data.table -O normal_sample_recal.bam

The tool used: GATK's Apply BQSR (V. 4.1.0.0)

6. Somatic Variant Calling:

Somatic variant calling compares the aligned and recalibrated BAM files of the tumor and normal samples to identify genomic variations specific to the tumor.

Algorithm/Program:

gatk Mutect2 -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa -I tumor_sample_recal.bam -I normal_sample_recal.bam -normal sample1 -O output.vcf

Input File: The recal.bam files of tumor and normal sample

Output File: output.vcf file

The tool used: GATK's Mutect2(V 2.1)

7. Further Filtering:

This step further filters out the variants with high confidence.

As the result obtained had no variants only the header part, 'Bcftools' was used to generate the VCF files separately for both normal and tumor samples.

Algorithm/Program

gatk FilterMutectCalls -R /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa -V output.vcf -O filtered.vcf

Input Files : output.vcf file

Output File : Filtered.vcf

The tool used: GATK's Filter Mutect Calls

(Note: Due to not getting the correct variants went a step back to create VCF files separately for Normal and Tumor Pair using BCF Tools below are the pieces of code used for that and then went for the next step which is annotation using the VEP website and the Filtered the variants after annotating using Python.

bcftools mpileup -Ou -f /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa normal_sample_dedup.bam | bcftools call -mv -Ov -o normal_aligned_variants.vcf

bcftools mpileup -Ou -f /class/dsa8110_genomic_analytics/MR_DATA/test_t-n_somatic/hg19_ref/hg19.chr5_12_17.fa tumor_sample_dedup.bam | bcftools call -mv -Ov -o tumor_aligned_variants.vcf)

8. Annotation:

In this step, the Annotation tool VEP assessed the functional impact of variants by predicting how they might affect genes, proteins, or regulatory regions.

The tool used: VEP Web interface (V. 110)

The output file: VCF files and text files

9. Filtering Data Using Python:

1. Filtered only the missense_variants.
2. Filtered the variants that are only present in the tumor sample.
3. Filtered the top 20 genes having maximum variants.
4. Checked for the genes responsible for lung cancer
5. Used all those 20 genes and also those 5 genes to see the differences in pathways, they are mostly the same.

Reference: Module 8/ Reports/ project_final_steps.ipynb for the programs.

Results:

Annotation Results:

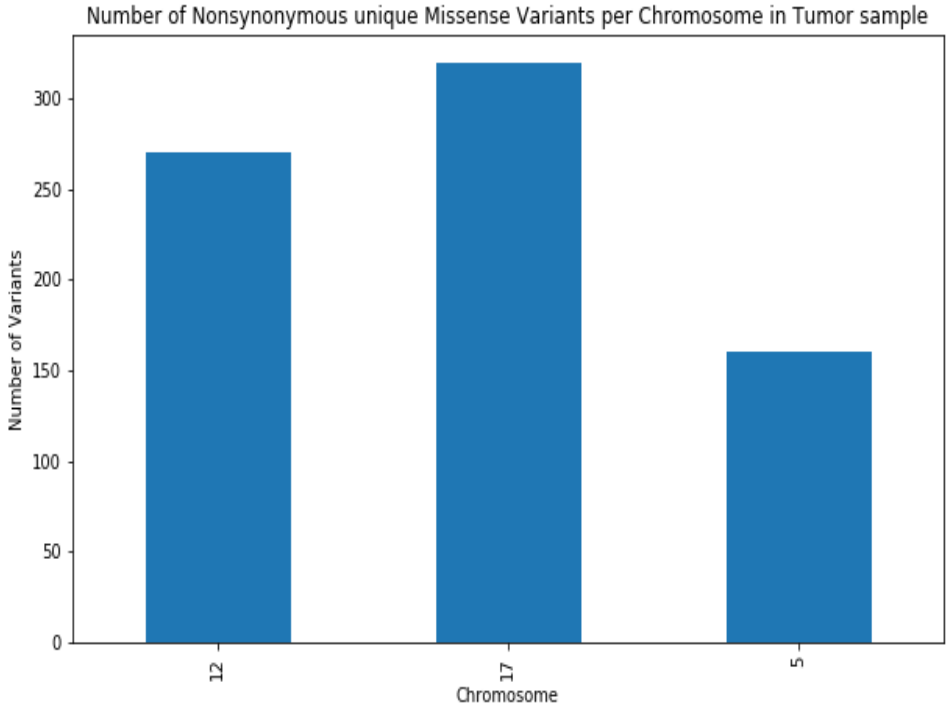
Tumor sample variants:
Coding consequences:
Missense variant: 61%
Synonymous variant: 26%
Frameshift variant: 6%
Stop lost: 3%
Stop gained: 2%
Inframe deletion: 1%
Stop retained variant: 0%
Inframe insertion: 0%
Start lost: 0%

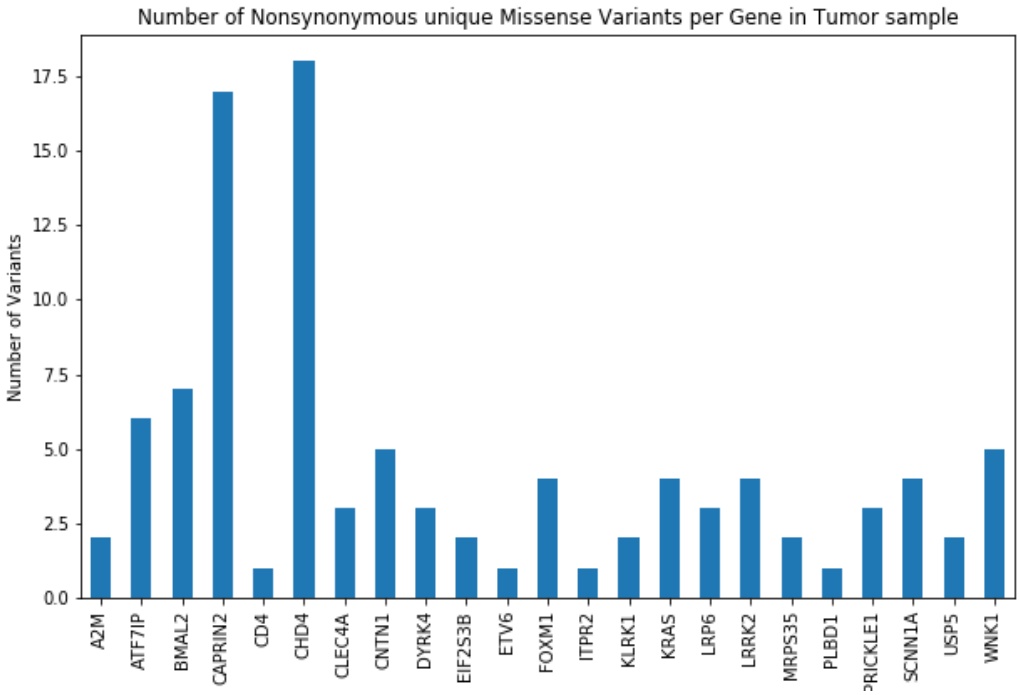
Category	Count
Variants processed	243405
Variants filtered out	0
Novel / existing variants	235438 (96.7) / 7967 (3.3)
Overlapped genes	8754
Overlapped transcripts	38313
Overlapped regulatory features	21858

Normal Sample Variants :
Coding consequences:
missense_variant: 59%
synonymous_variant: 25%
frameshift_variant: 7%
stop_lost: 3%
stop_gained: 2%
inframe_deletion: 1%
stop_retained_variant: 1%
protein_altering_variant: 0%
inframe_insertion: 0%Others

Category	Count
Variants processed	180468
Variants filtered out	0
Novel / existing variants	174727 (96.8) / 5741 (3.2)
Overlapped genes	8538
Overlapped transcripts	37718
Overlapped regulatory features	18914

The bar charts of the distribution of somatic nonsynonymous Missense variants found after filtration:





The top 20 Gene Mutations:

CHD4	18
CAPRIN2	17
BMAL2	7
ATF7IP	6
WNK1	5
CNTN1	5
LRRK2	4
FOXM1	4
KRAS	4
SCNN1A	4
CLEC4A	3
LRP6	3
PRICKLE1	3
DYRK4	3
KLRK1	2
EIF2S3B	2
A2M	2
USP5	2
MRPS35	2
PLBD1	1

Analysis:

Lung Cancer Genes Analysis:

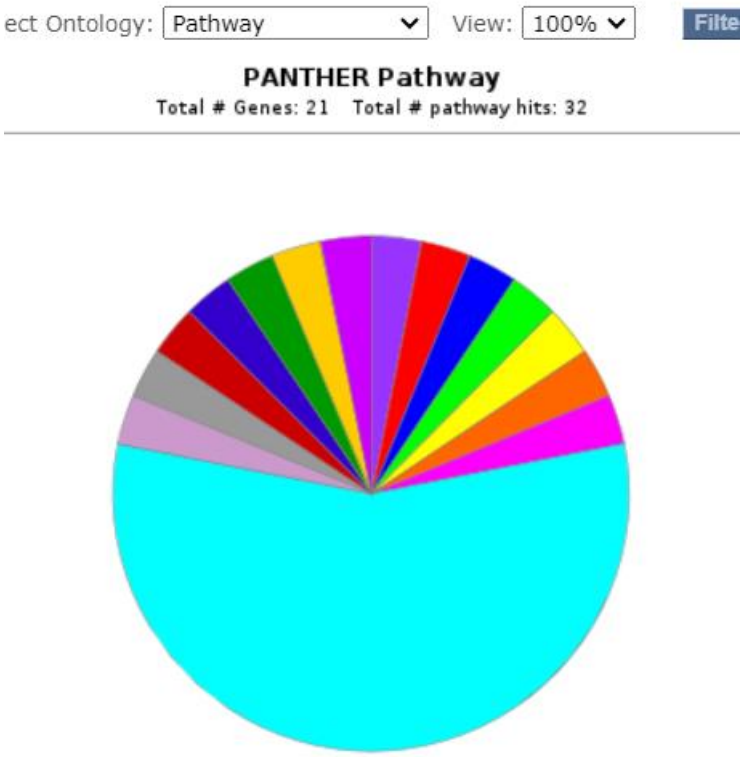
Among the top 20 genes, we tried to analyze if any of them are related to lung cancer and found that there are 5 such genes which are.

CNTN1, KRAS, SCNN1A, KLRK1, and USP5 are there among the 2539 proteins co-occurring, with the tissue **lung cancer cell** in abstracts of biomedical publications from the TISSUES Text-mining Tissue Protein Expression Evidence Scores dataset. **CHD4** has a role in all types of cancers, including Lung Cancer.

Ref:
https://maayanlab.cloud/Harmonizome/gene_set/lung+cancer+cell/TISSUES+Text-mining+Tissue+Protein+Expression+Evidence+Scores

Panther Pathway Analysis:

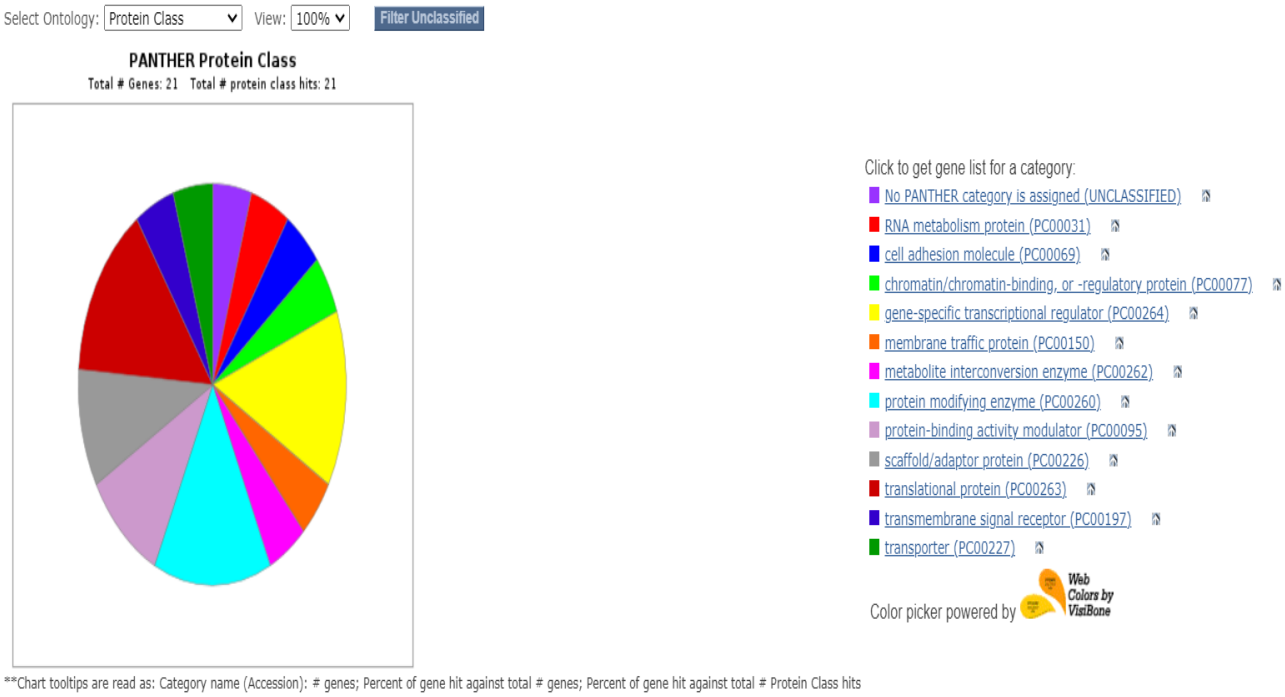
Used the Panther Pathway Analysis Database to find out which pathways are getting affected when those genes are muted. And here is the overall result. Out of all the pathways it generated on further investigation, we realized few of them are related to cancer, which has been explained explicitly below and has shared the pathway maps. Generated the pathway maps separately as well, to study in detail about each gene we found in the top 20 genes, that has been uploaded in Module Report folder under Pathways folder.



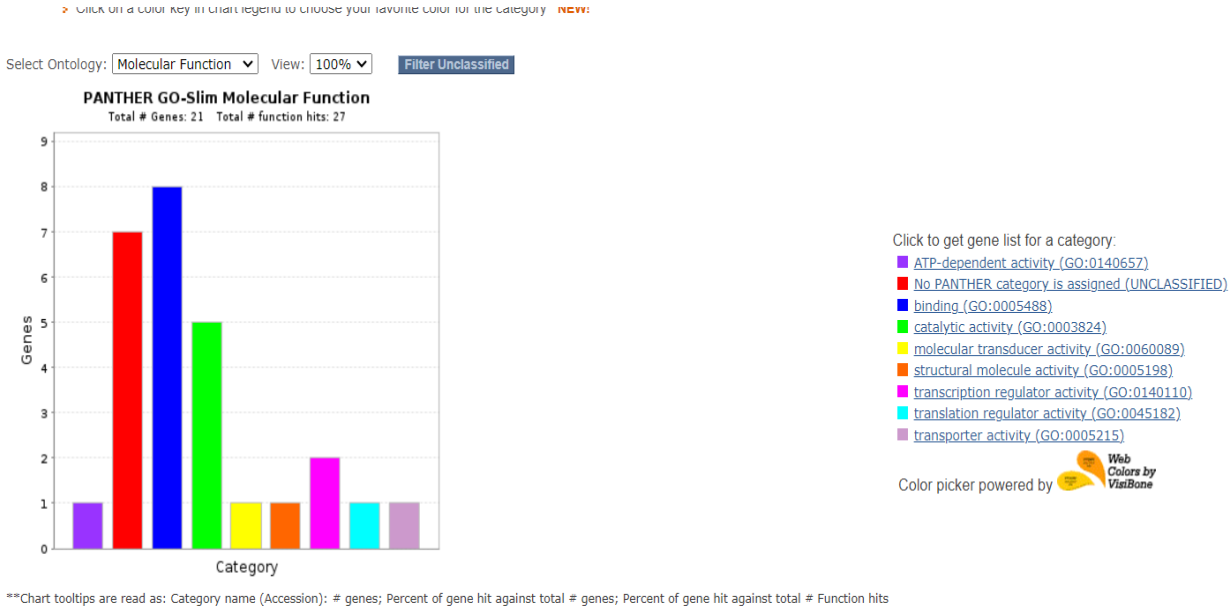
	Blood coagulation (P00011)	1	4.8%	3.1%		
2	No PANTHER category is assigned (UNCLASSIFIED)	18	85.7%	56.3%		
3	Angiogenesis (P00005)	1	4.8%	3.1%		
4	Integrin signaling pathway (P00034)	1	4.8%	3.1%		
5	Alzheimer disease-presenilin pathway (P00004)	1	4.8%	3.1%		
6	Inflammation mediated by chemokine and cytokine signalling pathway (P00031)	1	4.8%	3.1%		
7	Wnt signaling pathway (P00057)	1	4.8%	3.1%		
8	VEGF signaling pathway (P00056)	1	4.8%	3.1%		
9	p53 pathway feedback loops 2 (P04398)	1	4.8%	3.1%		
10	TGF-beta signaling pathway (P00052)	1	4.8%	3.1%		
11	PI3 kinase pathway (P00048)	1	4.8%	3.1%		
12	FGF signaling pathway (P00021)	1	4.8%	3.1%		
13	PDGF signaling pathway (P00047)	1	4.8%	3.1%		
14	EGF receptor signaling pathway (P00018)	1	4.8%	3.1%		
15	Ras Pathway (P04393)	1	4.8%	3.1%		

Panther Protein Analysis:

1	protein modifying enzyme (PC00260)	3	14.3%	14.3%		
2	transporter (PC00227)	1	4.8%	4.8%		
3	scaffold/adaptor protein (PC00226)	2	9.5%	9.5%		
4	membrane traffic protein (PC00150)	1	4.8%	4.8%		
5	No PANTHER category is assigned (UNCLASSIFIED)	1	4.8%	4.8%		
6	cell adhesion molecule (PC00069)	1	4.8%	4.8%		
7	protein-binding activity modulator (PC00095)	2	9.5%	9.5%		
8	transmembrane signal receptor (PC00197)	1	4.8%	4.8%		
9	RNA metabolism protein (PC00031)	1	4.8%	4.8%		
10	gene-specific transcriptional regulator (PC00264)	3	14.3%	14.3%		
11	translational protein (PC00263)	3	14.3%	14.3%		
12	metabolite interconversion enzyme (PC00262)	1	4.8%	4.8%		
13	chromatin/chromatin-binding, or -regulatory protein (PC00077)	1	4.8%	4.8%		



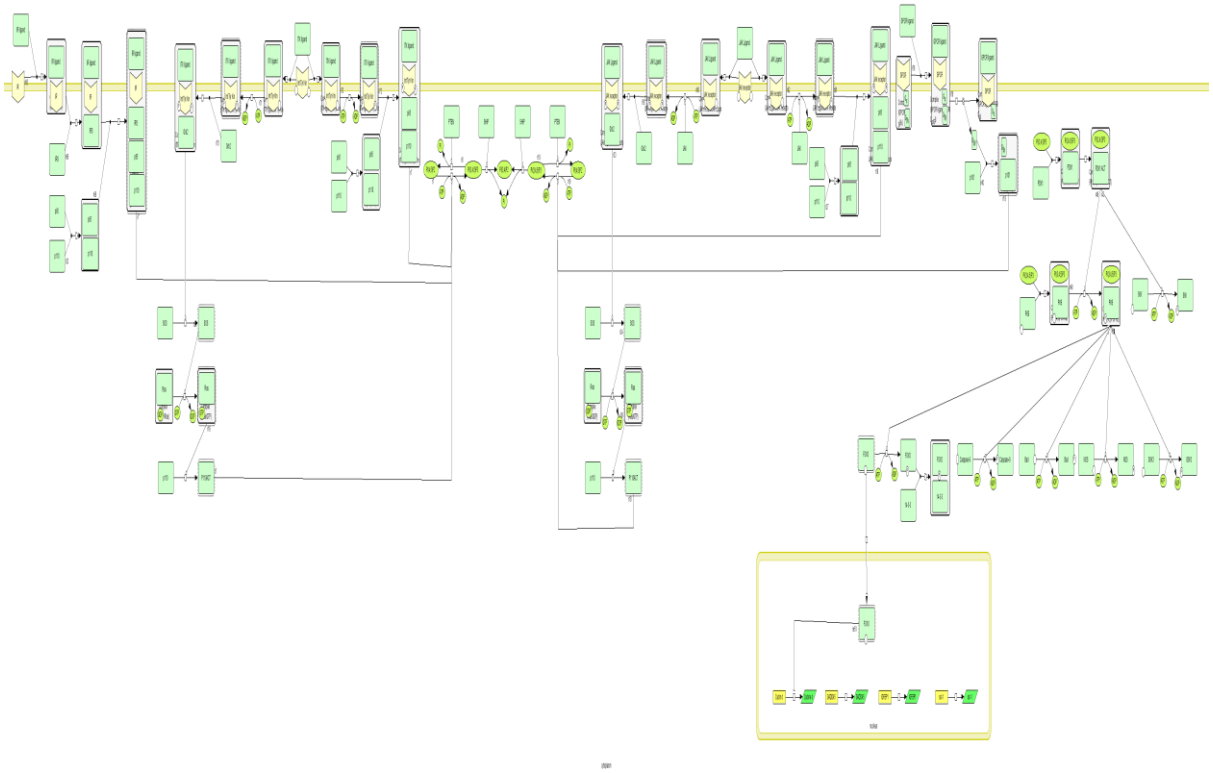
Panther Molecular Functions Analysis:



Pathways related to cancer (Phosphoinositide 3-kinase (PI3K):

Phosphoinositide 3-kinase (PI3K) activity is stimulated by diverse oncogenes and growth factor receptors, and elevated PI3K signaling is considered a hallmark of cancer. Many PI3K pathway-targeted therapies have been tested in oncology trials, resulting in regulatory approval of one isoform-selective inhibitor (idelalisib) for treatment of certain blood cancers, and a variety of other agents at different stages of development. In parallel to PI3K research by cancer biologists, investigations in other fields have uncovered exciting and often unpredictable roles for PI3K catalytic and regulatory subunits in normal cell function and in disease. Many of these functions impinge upon oncology by influencing the efficacy and toxicity of PI3K-targeted therapies. Here we provide a perspective on the roles of class I PI3Ks in the regulation of cellular metabolism and in immune system functions, two topics closely intertwined with cancer biology. We also discuss recent progress in developing PI3K-targeted therapies for the treatment of cancer and other diseases.

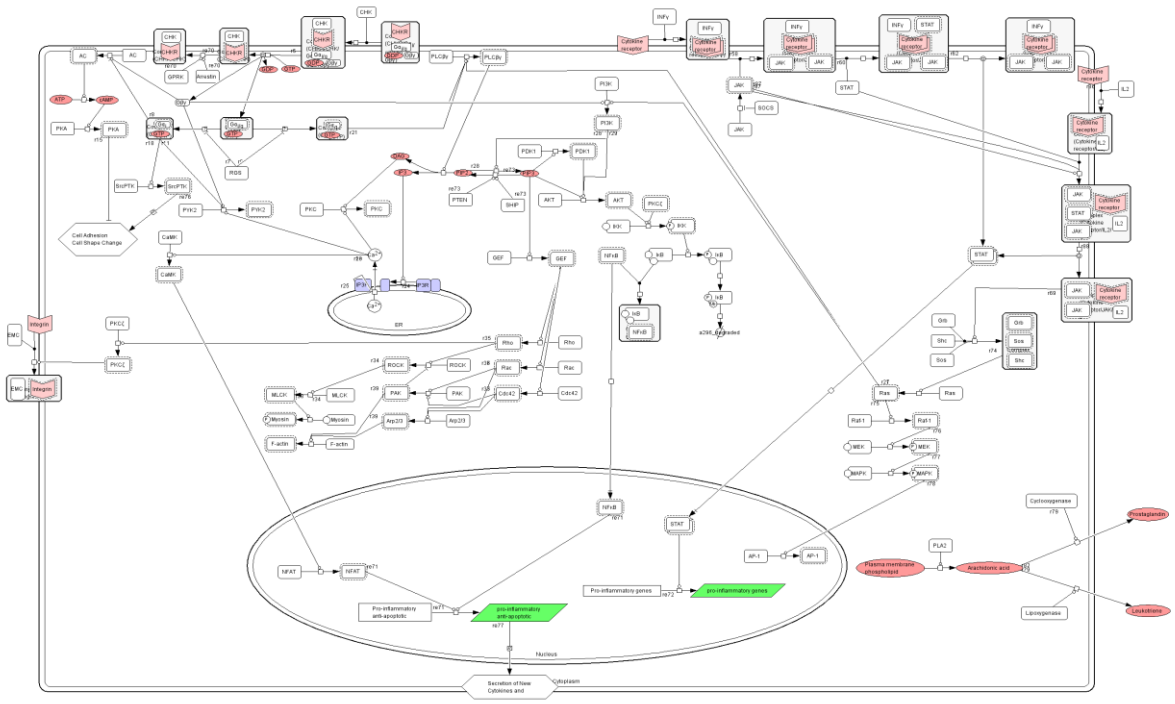
Ref: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5726441/



Pathways related to cancer (Inflammation mediated by chemokine and cytokine signaling pathway)

The relationship between chronic inflammation and neoplastic diseases is not fully understood. The inflammatory microenvironment of a tumor is an intricate network that consists of numerous types of cells, cytokines, enzymes, and signaling pathways. Recent evidence shows that the crucial components of cancer-related inflammation are involved in a coordinated system to influence the development of cancer, which may shed light on the development of potential anticancer therapies. Since the last century, considerable effort has been devoted to developing gene therapies for life-threatening diseases. When it comes to modulating the inflammatory microenvironment for cancer therapy, inflammatory cytokines are the most efficient targets. In this manuscript, we provide a comprehensive review of the relationship between inflammation and cancer development, especially focusing on inflammatory cytokines. We also summarize the clinical trials for gene therapy targeting inflammatory cytokines for cancer treatment. Future perspectives concerned with new gene-editing technology and novel gene delivery systems are finally provided.

Ref:https://pubmed.ncbi.nlm.nih.gov/33429846/#:~:text=The%20relationship%20between%20chronic%20inflammation,cytokines%2C%20enzymes%20and%20signaling%20pathways.



Conclusions and Future Work:

1. Limited Biomarker Detection: Despite identifying key biomarkers associated with lung cancer, the observed mutation counts were relatively low in our study.

2. Small Sample Size Impact: The restricted number of samples used in our research may have contributed to the lower mutation counts, potentially masking the true prevalence of certain genetic alterations.

3. Call for Further Investigation: To gain a more comprehensive understanding of early-stage lung cancer, it is imperative to expand the sample size in future studies.

4. Enhanced Precision with Larger Samples: Increasing the size of our sample pool will likely unveil a more accurate picture of the genes and pathways affected during the early stages of lung cancer.

5. Clinical Implications: A larger dataset holds the promise of identifying precise genetic markers, aiding in early detection and intervention, thereby improving patient outcomes and potentially saving lives.

Note:

The Trial Program that has not been used but written for the Program Pipeline can be found under Module 8 Reports Folder, Final Project Steps python file.