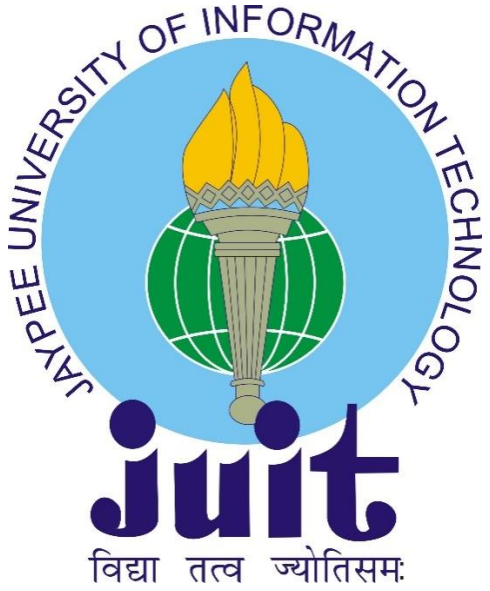
COMPUTATIONAL STUDIES TO INVESTIGATE DNA REPAIR MECHANISM AND MUTATIONS INVOLVED IN LUNG CANCER

Enrolment No- **171509**

Name of the student- **Aagam Mishra**

Name of Supervisor- **Dr. Tiratha Raj Singh**



May, 2021

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS,

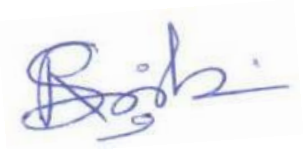
JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, SOLAN 173234, HIMACHAL PRADESH, INDIA

CERTIFICATE

This is to certify **Ms. Aagam Mishra**, B.Tech. Bioinformatics 8th semester student of Jaypee University of Information Technology, Solan, bearing a **Roll No. 171509** has worked on the project entitled **"Computational studies to investigate DNA repair mechanism and mutations involved in lung cancer"** at the department of biotechnology and bioinformatics, under my guidance from 15 July, 2020 to 19 May, 2021. She has successfully completed her training and her conduct is satisfactory.

I wish her a successful career.



Dr. Tiratha Raj Singh

Associate Professor

Department of biotechnology and bioinformatics

Jaypee University of Information Technology, Solan

Date: 19 May, 2021

Candidate's Declaration

I hereby declare that the work presented in this report entitled **"Computational studies to investigate DNA repair mechanism and mutations involved in lung cancer"** in the fulfilment of the requirements for the major project submitted in the department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat, Solan, is an record of my own work and carried out over a period from 15 July, 2020 to 19 May, 2021 under the supervision of Dr. Tiratha Raj Singh.

Aagam Mishra (171509)

ACKNOWLEDGEMENT

I would like to begin by thanking our supervisor Dr. Tiratha Raj Singh, Associate Professor at the Department of Biotechnology and Bioinformatics in Jaypee University of Information Technology, for giving us the opportunity to work with him. His immense understanding and curiosity in research allowed us to learn many things. His guidance and encouragement were indispensable in completing this dissertation.

Also, I would like to thank Arvind Yadav, Research Scholar for constant encouragement and help in understanding DNA Repair Mechanism.

My sincere thanks goes to all the staff, professors and PhD scholars at JUIT, whose Kindness and friendly nature provided a very enjoyable research environment and a platform of excellence to do my dissertation.

Aagam Mishra

LIST OF FIGURES

Figure-01: Statistics representing estimated number of new cases of lung cancer (11.4%) in 2020 for both the sexes and all ages.

Figure-02: Statistics showing the estimated age-standardized incidence and mortality rates in India as well as in the world for the year 2020 for both the sexes and all ages.

Figure-03: Somatic Mutations associated with the lung cancer

Figure-04: Frameshift Mutations in coding and non-coding strands

Figure-05: Representation of the SNP profile of all chromosomes

Figure-06: Clinical Applications of GWAS in Lung Cancer

Figure-07: Schematic representation of the methodology used in this study.

Figure-08: Automated Codon Usage Analysis

Figure-09: Workflow of Network analyst which works in three consecutive steps-Data Processing, Network Construction, Network Analysis.

Figure-10: Comparison of correlation and t-test values generated with the help of SHIFT webserver.

Figure-11: (a) 2D representation; (b) 3D representation of network analysis approach to understand gene expression patterns

Figure-12: List of the gene represented through Heatmaps.

Figure-13: Manhattan plot showing the Gene Ontology analysis

Figure-14: Detailed Description of molecular, biological and cellular processes associated with the protein coding genes

Figure-15: Representation of frameshift mutation at the specific position Chr1: 197073232-197073232

Figure-16: Overview of the Mutation, CNV Amplification, CNV Deletion and Expression

LIST OF TABLES

Table-01: Results obtained from the ACUA tool (AT percentage, GC percentage, AT1 skewness, GC1 skewness, CAI and ENc)

Table-02: Codon usage table with the frequency of amino acids and their fractional values

Table-03: Genes with their degree and betweenness centrality values

Table-04: Genes that were involved in various pathways with their p-values and false discovery rates

Table-05: Genes with their mutations and prediction of disease associated variations

Table-06: Gene Expression in percentiles present in different genomic resources with their cell names.

Table-07: Description of Position, CDS mutation, AA mutation, Zygosity and Mutation type for specific cell line.

Table-08: Copy number of a specified gene with their cell lines and tissue association information

TABLE OF CONTENTS

ABSTRACT

INTRODUCTION

LUNG CANCER

GENOMIC INSTABILITY AND DNA DAMAGE IN LUNG CANCER

THEORY OF SOMATIC MUTATIONS

SPECIFIC MUTATIONS RELATED TO LUNG CANCER

SMOKING

SCREENING

PROTEIN CODING SEQUENCES

CODON USAGE ANALYSIS

FRAMESHIFT MUTATION

SINGLE NUCLEOTIDE POLYMORPHISM ANALYSIS

GENOME WIDE ASSOCIATION STUDIES

METHODOLOGY

THE CANCER GENOME ATLAS PROGRAM

COMPUTATIONAL ANALYSIS THROUGH VARIOUS TOOLS AND SOFTWARES

ACUA

MEASURES OF CODON BIAS

SHIFT

NETWORK ANALYST

G:PROFILER

GWAS CATALOG

SNPs3D

PANTHER

SNPs&GO

GEMICCL

EXPECTED OUTCOMES

RESULTS AND DISCUSSION

CONCLUSION

REFERENCES

ABSTRACT

The most common malignancy in the western world is lung cancer which is caused by smoking known as the single greatest risk factor. A study was done in the northern part of India where it was proved that 90% people were diagnosed with lung cancer at an advanced stage. Apart from this, it was observed that the factors such as tobacco consumption and air pollution were solely responsible for lung cancer in India. Before an affected individual becomes symptomatic, the growth cycle of lung cancer has reached an advanced stage in their natural history. In the past 30 years, there were very modest improvements in the prognosis of lung cancer. Radiography and sputum cytology is a failure in the treatment of disease. In lung cancer, it is difficult to identify premalignant lesions. It has been found that certain somatic mutations are involved in the evolution of lung cancer. By insertion or deletion, frameshift mutations induce more immunogenic tumor-specific neoantigens which proved to be a better response for immune checkpoint inhibitors in NSCLC. In the present study, the investigation of mutations and their role in lung cancer were analyzed. Network analysis was done to study novel pathogenesis pathways that can be helpful in the determination and identification of potential therapeutic targets. Apart from that, analysis of SNP data will provide insight into the genome-wide association studies and their relation to potential drug data. All these processes were explored w.r.t. their association with DNA repair mechanisms. Expression level of genes were also performed for the prediction was disease associated variability and the therapeutic targets. Results obtained through computational meta-analysis will provide insights about the plausible biomarkers involved in the regulation of DNA repair mechanisms involved in lung cancer.

INTRODUCTION

Lung cancer:

The human body is made up of trillions of cells and cancer can start almost anywhere in the human body. Lung cancer is also known as the disease of modern man [1]. In the United States of America, lung cancer is the top killer for both men and women. There were 228,820 new cases however, 135,720 deaths were predicted for lung cancer in the USA during 2020. In India, the reported cases of male patients were 679,421 (94.1 per 100,000) whereas the reported cases of females patients were 712,758 (103.6 per 100,000) for the year 2020. In metropolitan cities and the southern part of India, lung cancer was observed to be a leading site [2]. People die more because of lung cancer than the other common four cancer types (Colon/rectal, breast, pancreas, and prostate). Its mortality is associated with 20 years of smoking history [3]. The susceptibility for lung cancer depends on competitive gene enzyme interactions which cause activation or detoxification of procarcinogens as well as the integrity of endogenous mechanism for repairing lesions in DNA. Depending on its anatomic location, lung cancer presents variable symptoms and is highly heterogeneous in many different sites of the bronchial tree. In the USA, 70% patients are diagnosed with stage (III) or stage (IV) lung cancer.

According to the recent publication, there are only 15% chances with a five-year survival rate for lung cancer [4]. Using cytotoxic chemotherapeutic agents as personalized medicine, lung cancer can be treated. These are also known as combinatorial therapies. Regimes like platinum-based DNA damaging agent Cisplatin and carboplatin are being used in the treatment of Lung cancer. The cytotoxic drugs divide cells where they lead to irreparable DNA damage.

Combined activities of DNA repair proteins are largely maintained through the stability of genomic code in noncancerous cells. These proteins not only detect damages but also initiate

cell cycle checkpoint signaling, repair the lesion and activate apoptosis pathways. In cancer cells, regulation of these processes are lost which leads to the accumulation of mutations. The initial transformation of cells is contributed by deregulation of DNA repair which in continuation may result in metastatic disease. As a result, high selective pressure is experienced by the cell, including the need for function in a foreign environment.

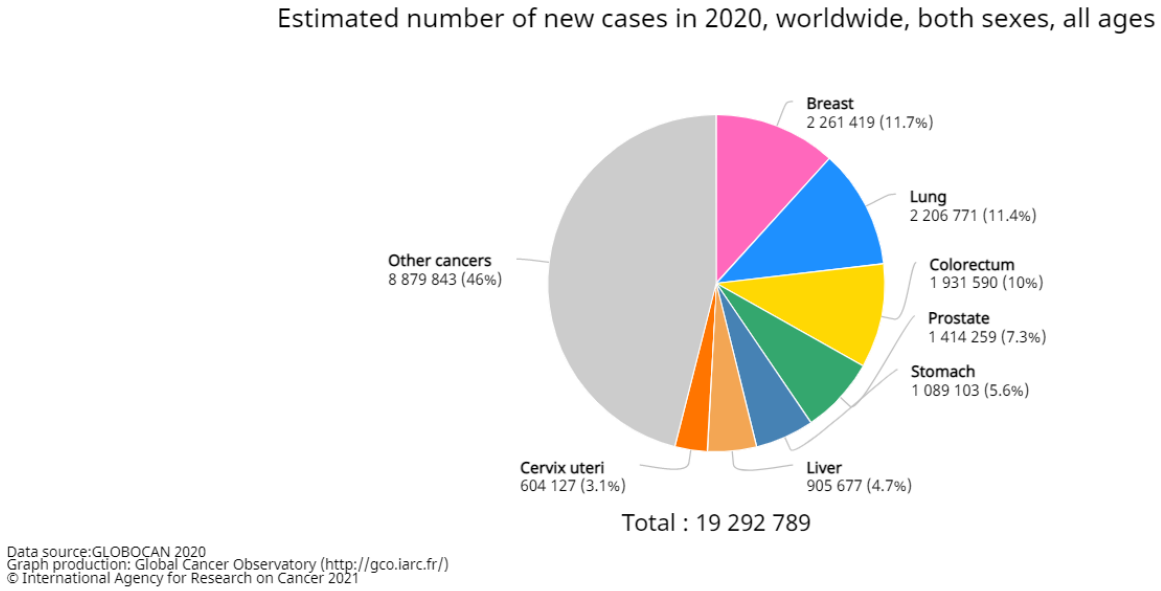


Figure-01: Statistics representing estimated number of new cases of lung cancer (11.4%) in 2020 for both the sexes and all ages.
(http://gco.iarc.fr/)

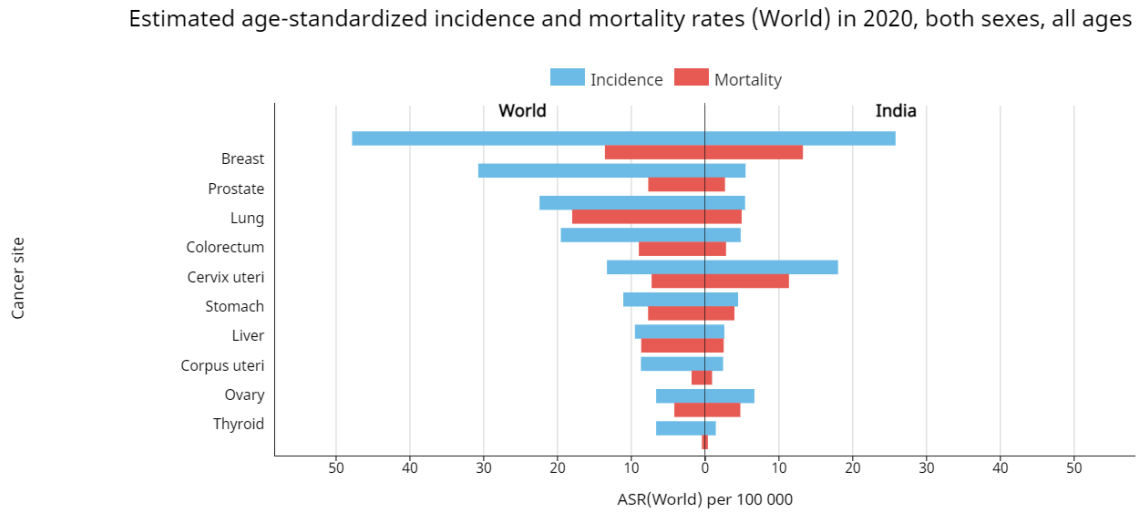


Figure-02: Statistics showing the estimated age-standardized incidence and mortality rates in India as well as in the world for the year 2020 for both the sexes and all ages.

(http://gco.iarc.fr/)

Genomic Instability and DNA Damage in Lung Cancer:

Each day, normal cells receive up to 30,000 DNA damage events which include oxidation or deamination of DNA bases, the formation of nucleotide photo adducts as well as generation of ssDNA and dsDNA breaks [5]. Endogenous processes such as somatic and meiotic recombination, reaction with reactive oxygen, the collapse of stalled DNA replication forks resulted from DNA breaks. These breaks can also occur through exogenous such as radiation exposure, chemotherapeutics and carcinogenic environmental compounds [6]. There are five thousand compounds that comprise cigarette smoke from which 73 are carcinogenic of which 20 specifically affect the lungs. Through direct interaction with DNA, many of these compounds appear mutagenic [7].

Theory of somatic mutation:

There is a myriad of cells in the body where genetic mutations occur at a slow rate. Such mutations that occur at a slow rate have no reproductive advantage [8]. Whenever the cells such as the mutant one try to grow then at the time it overcomes control mechanisms that are present within the body. One such example is apoptosis in which cells such as the cancer ones undergo systemized cell death. For a normal cell to evade the mechanisms, at least six genetic mutations are required. This process further helps the cell to become a cancerous one. Certain mutations surge in the probability of occurrence of the subsequent mutations. The example includes expanded target population or an increase in overall mutation rate [9].

There are several genes that are classified into three main groups such as proto-oncogenes, tumor suppressor genes, and DNA repair genes. Proto-oncogenes are those which after mutation change into oncogenes [10]. Oncogenes are further involved in the cell cycle control which inappropriately activates and stimulates cell division. To have a phenotypic effect, Oncogenes require an allele of one of the two chromosomes. Cell division is inhibited by tumor suppressor genes in response to DNA damage until and unless it is repaired again. To change the behavior in a particular cell, both alleles must be inactivated. Loss of

heterozygosity is a conversion of heterozygous inactivated state to a homozygous inactivated state where the loss of tumor suppressor requires a mutation and this can further be explained through the chromosomal regions that lie in blood cells present at the boundaries in conjunction with tumor cells. For the inactivation of the normal repair mechanism, both the alleles must mutate in a Mutator gene.

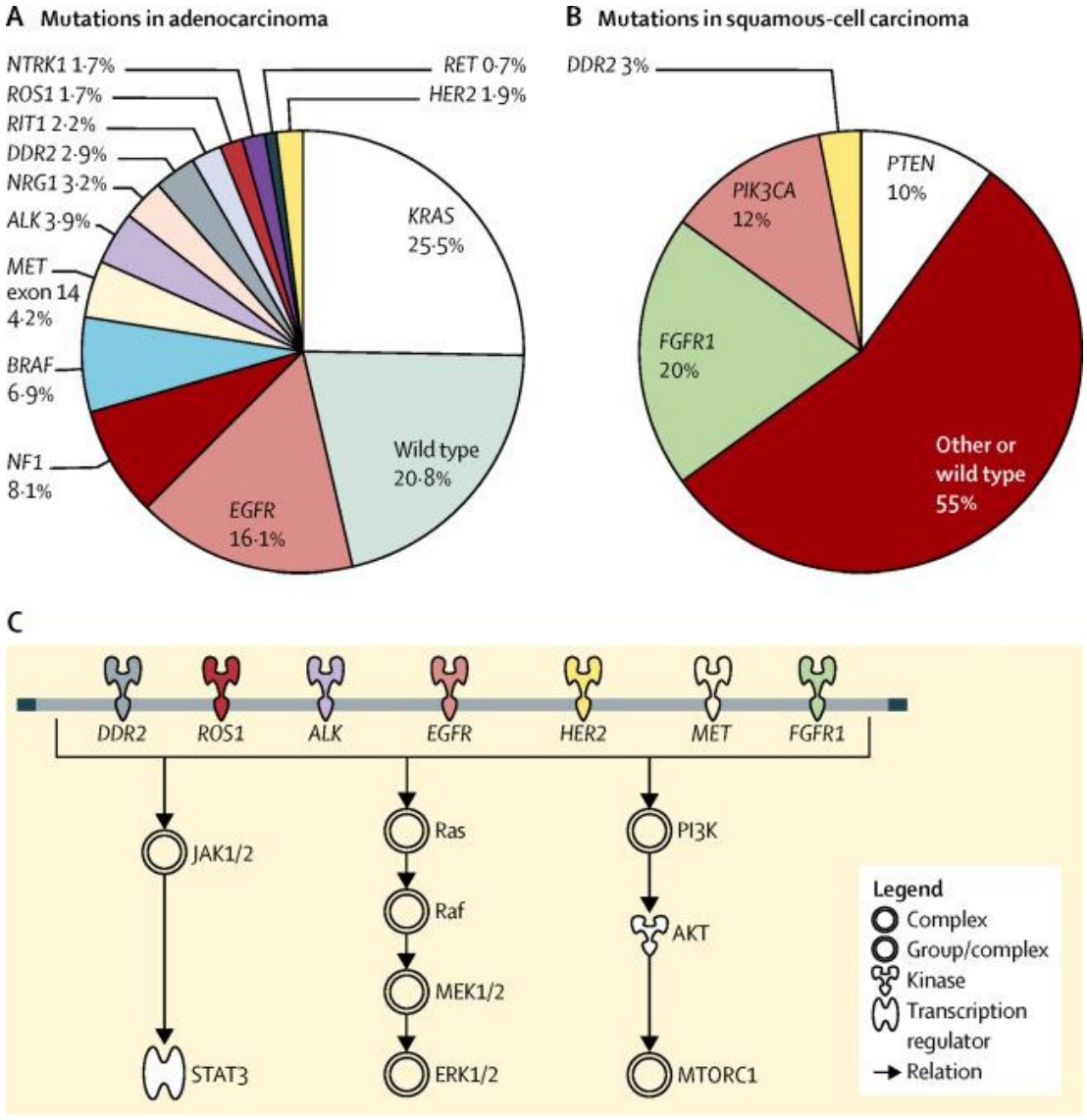


Figure-03: Somatic Mutations associated with the lung cancer (https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(15)01125-3/fulltext)

Specific mutations related to lung cancer:

A small number of lesions occur due to a series of evolvement in epithelial cancers that may arise into invasive cancers. These changes further result in a series of somatic mutations for a particular tumor type [11]. Earlier precursor lesions could not be identified in the lungs and

therefore, identifying a characteristic in somatic mutations becomes difficult. There were many studies performed to find dysplasia. In this bronchoscopic biopsy specimens were taken from the fixed site from different individuals that involve asymptomatic smokers, non-smokers and ex-smokers [12]. A comparison was made between longitudinal sampling and parallel sampling. A longitudinal sampling involves samples from the same patient taken at different time intervals whereas a parallel sampling involves specimens that were taken at the same time interval from different mucosal sites. These samples resulted that a higher rate of somatic mutations occur in mucosa as it is highly exposed to harmful carcinogens that may be present in cigarette smoke. Any chromosomal abnormalities can be explained as lung cancers consist of very complex karyotypes. Both small cell and non-small cell lung cancer involve mutation that is present on chromosome 3 which may be responsible for the deletion or loss of heterozygosity [13]. Premalignant changes take place in this chromosome and hence, its various regions are rapidly damaged. There is another common mutation found in the majority of small cell and a minute in non-small cell lung cancer and this mutation is of p53 gene that takes place at location 17p13 and it is probably responsible for the single genetic change that occurs in all human cancers. One of the p16 gene at 9p21 was represented in terms of gene locus found in leukemia and it is furthermore observed in the pathogenesis that takes place in non-small cell lung cancer.

Smoking:

One of the greatest risk factors for lung cancer is smoking as it consists of at least 43 known carcinogens that are quite harmful [14]. These cause a field effect with the enhanced somatic mutations in the respiratory mucosa. Such changes increase the extent as well as severity within a particular set of years. Somatic mutation remains that explains the further risks for the development of lung cancer in a former smoker. Some somatic mutations are considered as the prognostic markers in non-small cell lung cancer. Example include p53 mutation which is defined as the predictor of death. In resectable non-small cell lung cancer, a poor prognosis is predicted by K-Ras and Neu oncogenes. Apart from this, these genes further reduced the survival rate [15].

Screening:

Pathogenesis of lung cancer provides us a vast amount of knowledge related to the occurrence of mutations [16]. Few examples include: damage to the p53 gene done by allelic loss on chromosome 3, allelic loss seen in dysplasia occurs in a sequential pattern in chromosome 3. Common sequences of mutations need to be defined before the genetic analysis [17].

Protein coding sequences:

The DNA sequences that involve the translation of mRNA molecules into a polypeptide chain are referred to as protein-coding sequences. The combination of three nucleotides is referred to as codon which forms 1 amino acid in the polypeptide chain. Such sequences have a start (ATG) and the stop codon (TAA). These protein-coding sequences consist of several regions termed protein domains and are defined as the composite parts of such domains. The N-terminal domain is special as it contains a start codon that is situated at a distance from a ribosomal binding site [18]. Apart from this, the presence of a myriad of special features at different coding regions may include protein export tags, lipoprotein cleavage and attachment tags, respectively. These are referred to as head domains as they occur at the beginning of coding regions. A relatively independent fold in a sequence of amino acids defines a protein domain and the DNA sequences of such domains include a multiple of three bases and maintain in-frame translation. These are called Internal domains as they remain inside a protein-coding sequence. If such domains involve functions like splicing or protein cleavage then these are called special internal domains. The C-terminal domain consists of a stop codon with special features that includes degradation tags. These are called tail domains when they are unable to function being internally in the coding regions.

Codon Usage Analysis:

There has been defined an appropriate system or coding strategy in the genome hypothesis for each species that will choose their codons from among the synonymous ones. Since the repetition of the system takes place in the genome of each gene, it is defined as the characteristic of the genome. Codon usage bias is related to the number of times the synonymous codons occur in the coding DNA which could be different in different

organisms. In a particular organism, some codons are used more than others. The degree of codon usage bias is related to the level of gene expressions [19].

Frameshift Mutations:

Frameshift mutations are often referred to as reading frameshift that is resulted from the insertions and deletions (indels) of nucleotides in a DNA sequence. It is present in the triplet form and due to this nature, reading frames are often changed with the help of insertion or deletions. This results in a completely different translation which causes more alter protein. This mutation is associated with codonic reading whenever the mutation to code occurs for different amino acids. This will further lead to the alteration of the first stop codon and hence, the polypeptide chain would be abnormally short or long [20].

They occur in regions where there is a repetitive sequence. The mutation becomes more pathogenic in case whenever DNA mismatch repair is unable to fix the addition or deletion of bases. According to the experimental observations, it is found that longer microsatellites cause a high percentage of frameshift mutations.

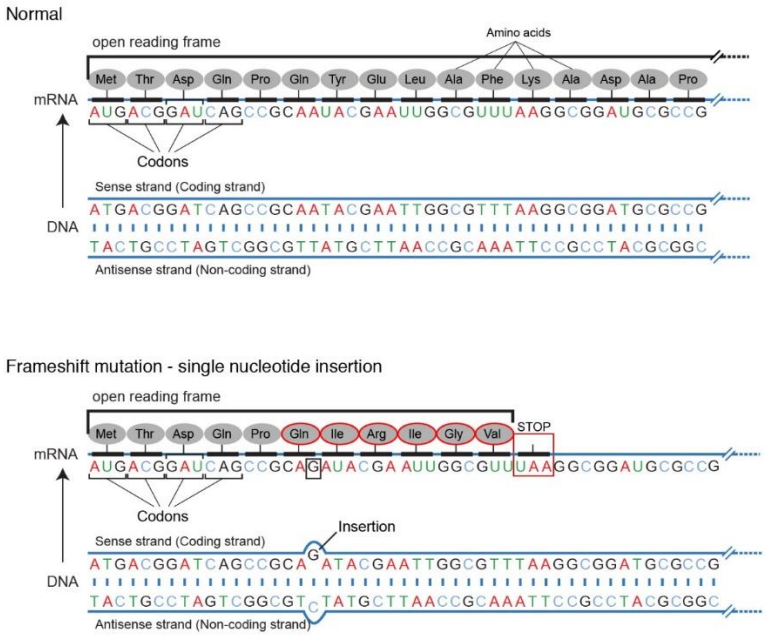


Figure-04: Frameshift Mutations in coding and non-coding strands

(_https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.genome.gov%2Fgenetics-glossary%2FFrameshift-Mutation&psig=AOvVaw3yVJS--Kdz7iWjOS-St_7d&ust=1621006607269000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCMj3p6n-xvACFQAAAAAdAAAAABAP)

Single Nucleotide Polymorphism Analysis:

When there is a change in the single nucleotide between paired chromosomes in humans or a biological species, then this variation in a DNA sequence is defined as SNP (Single Nucleotide Polymorphism) [21]. In today's world, there is a lot of research done in cancer genomic projects that are typically based on tumor-specific alterations such as somatic mutations. Structural variations can also result in genetic factors, the ones that are found in the tumor and normal matched DNA (SNPs). These factors can further detect the severity, kind of side effects and chemotherapy or targeted therapy response of a patient. There are various regions where we can find the SNPs, these are the intergenic regions, coding sequences of genes and non-coding regions of genes. When both alleles are the same, the taxonomy of SNPs is referred to as homozygous and when the alleles are different, it is termed as heterozygous. Oxidative stress is reduced by chemotherapy which leads to a fall in survival and proliferation rates of cancer cells. This process results in an objective treatment response that is radiographically quantified. To observe the toxicity and severity of chemotherapy, the potential predictive value has been examined. One such example includes non-small cell lung cancer (NSCLC).

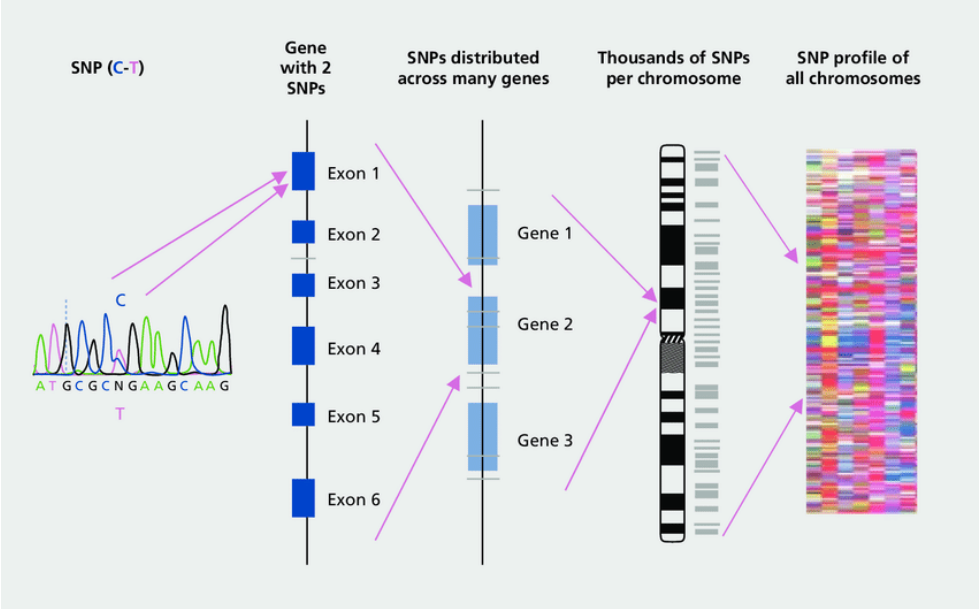


Figure-05: Representation of the SNP profile of all chromosomes (https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.researchgate.net%2Ffigure%2FSingle-nucleotide-polymorphisms-SNPs-from-a-single-SNP-to-an-SNP-profile_fig1_51751873&psig=AOvVaw3gqaEmyksqnonGByHKUC1X&ust=1620303855022000&source=images&cd=vfe&ved=0CAoQjRxqFwoTCOj127DEsvACFQAAAAAdAAAAABAT)

Genome-Wide Association Studies:

It is the method in which the multiple individual gene SNPs are detected which are based on linkage equilibrium. Genome-wide association study (GWAS) is used to distinguish gene characteristics as well as analyze genotypes that are associated with the diseases. GWAS method is applied to the study of deeper insights into tumors from the last few decades. A myriad of biological functions of variations loci along with genetic mechanisms could not be described appropriately besides the fact that many pleiotropic loci that have been identified by GWAS were associated with complex phenotypes [22]. With the help of GWAS, 450 genetic variants have been identified so far which were also associated with the increased risks. The main purpose of GWAS is that it provides an opportunity to analyse the process of repositioning, drug discovery and is quite helpful with the studies for cancer prevention. There are some of the key points that are associated with GWAS. These are as follows:

1.) The biological basis of the associations that are identified is quite challenging. However, it is necessary to decipher such a basis so that one can realize the potential of GWAS.

2.) There are several clinical relevances of GWAS. These may include drug prognostication and repositioning, identification of therapeutic targets, identification of nongenetic aetiological risk factors as well as optimizing population screening.

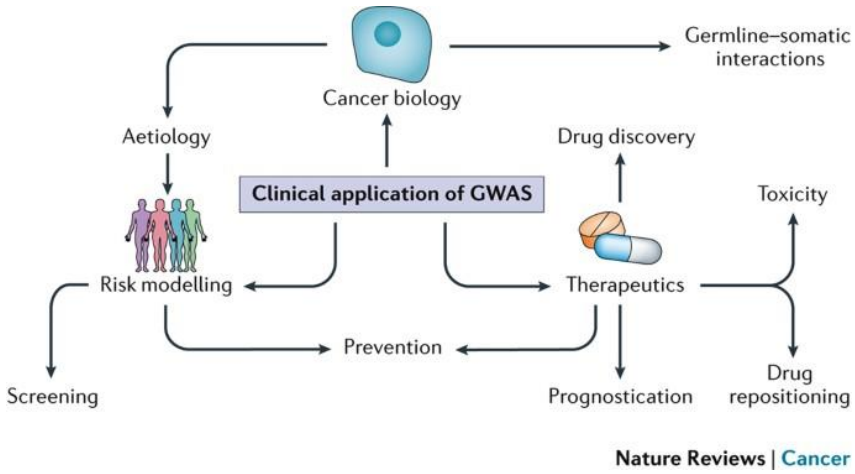


Figure-06: Clinical Applications of GWAS in Lung Cancer (https://www.nature.com/articles/nrc.2017.82)

METHODOLOGY

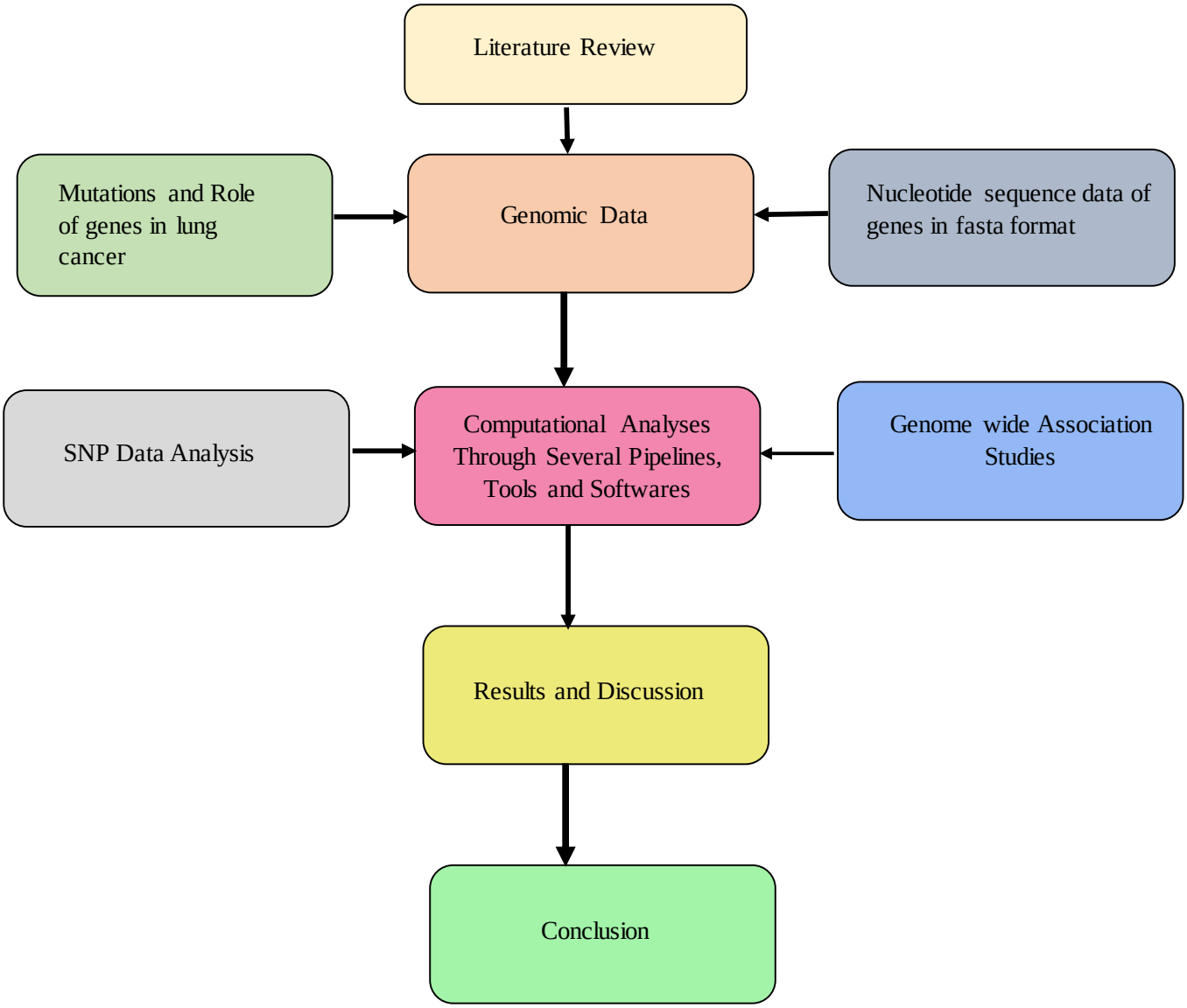


Figure-07: Schematic representation of the methodology used in this study.

In this study, we plan to implement computational genomics and bioinformatics approaches to investigate DNA repair mechanisms and mutations involved in lung cancer. To commence with, we studied different research papers and collected the different protein-coding genes data. Furthermore, we made use of different tools and databases such as Network Analyst, SHIFT, ACUA, g:Profiler, GWAS catalog, SNPs3D, Panther, SNP&GO and GemiCCL.

With the help of ACUA, we critically analyzed and determined certain values such as RSCU, GC percentage and codon adaptive index which potentially help in the prediction of the level of gene expression within the species as well as these values also help to identify protein-coding reading frames. After this, we kept our focus on the frameshift mutations that are involved in lung cancer. Through the SHIFT webserver, we calculated the correlation between codon usage frequencies and the contribution of codons to hidden stops in the respective sequences. Network Analyst tool was used to gain system-level understanding in different cell types and other biological/experimental conditions. We used a g:Profiler application associated with the gene ontology analysis to have gain a deeper insight about the molecular, biological and cellular processes. At last, we proceeded with the SNP analysis and the clinical applications of GWAS. Here, we were required to find out the genes that were disease associated or the ones that were neutral. This was done with the help of SNPs3D, Panther and SNPs&GO. This could further help in the drug development process with the quintessential targets. After the SNP analysis of genes, we were require to retrieve cell lines associated with genes as these cell lines were considered to be a potential step in the drug development process.

The Cancer Genome Atlas (TCGA) Program

It is a cancer genomics program that consists of matched and primary samples of at least 33 different cancer types. It is developed in 2006 to join hands with researchers from diverse fields. This program was created with the joint efforts of NCI and the National Human Genome Research Institute. There are approximately 2.5 petabytes of data is generated from the fields of epigenomic, genomics, proteomics and transcriptomics over the next dozen years. The data collected will be helpful in the treatment, diagnosis and prevention of cancer that could be available publicly for the use of anyone in the research community.

We collected the protein-coding genes data from TCGA. Genes such as TOP2A, CDC 20, BUB1, MADL1, JUN, FYN, CAV1, SFN, ASPM, CCNB2, CDC45, MELK and UBE2T were retrieved for the analysis of genomic data which can further provide help in finding potential biomarker that might be associated to disease. These genes were potential molecular targets in the treatment of lung adenocarcinomas. The fasta format of these gene sequences were downloaded from NCBI. .

Computational Analyses Through Various Tools and Softwares

❖ ACUA (A software tool for automated codon usage analysis)

It is the tool that is limited to the inbuilt calculations which perform high throughput sequence analysis as well as aids statistical profiling of codon usage. After the detailed studies of tools such as CodonW, GCUA, EMBOSS and Jemboss, this tool has been developed. It is used to calculate measures that may include codon usage table, RSCU, CAI, Nc Value, selective positional AT, GC content and its skewness. There is a package present within it that is unique to ACUA and also provides on-click access to the sequence through the results that are obtained [23].

This tool using several programming languages such as Visual Basic, PERL and C++. It is created as a standalone package for the codon usage analysis. The input required by this software is the fasta formatted nucleotide sequence in a text file. Apart from this, the user can select references from tables or provide a table in Emboss .cut format. After this, user can select their preferences as per their choice.

The output is generated in a MS Excel file with two worksheets. The first one contains all the measures that were calculated using the given inbuilt preferences and the second one consist of a table for statistical analysis. The present version of this tool works only on a single processor machine but the upcoming version is built Message Passing Interface with the help of R programming language that will probably enable cluster computing.

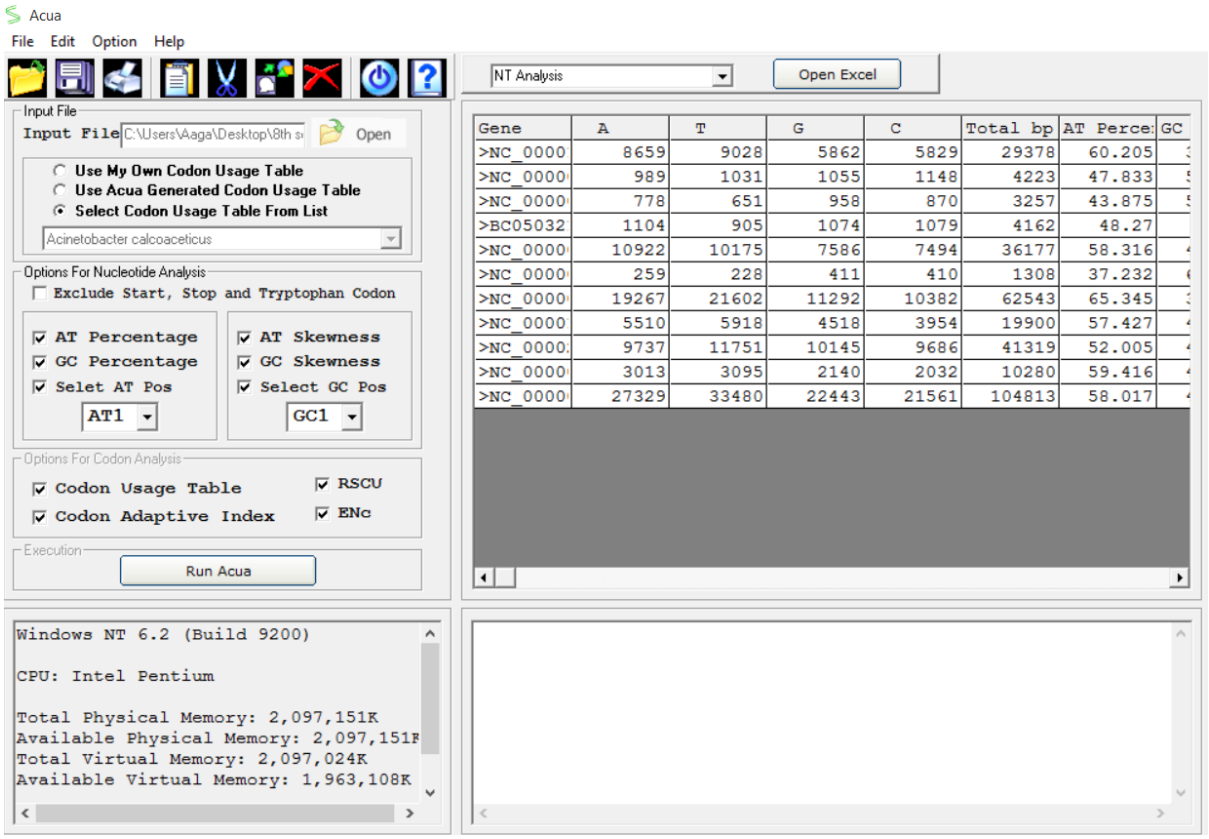


Figure-08: Automated Codon Usage Analysis

Measures of codon-usage bias:

There is a myriad of measures that may be quintessential while performing the codon usage analysis. These measures include:

1.) Frequency of optimal codons (Fop): It is referred to as the simplest measure which is species-specific. It is the frequency of optimal codons divided by the frequency of non-optimal codons. There are some exceptions associated with this measure. It excludes stop codons as well as the codons for methionine and tryptophan.

$$Fop= (Xop)/(Xop+Xnon)$$

2.) RSCU (Relative Synonymous Codon Usage): It determines the frequency that how many times a codon appears within a certain gene that is further divided by the frequency of expected occurrences of codon under equal codon usage.

$$RSCU(i)=X(i)/(1/n(\Sigma X(i))$$

Where n refers to the number of synonymous codons ($1 \leq n \leq 6$) for the amino acid under study, X(i) = number of occurrences of codon i.

3.) CAI (Codon Adaptive Index): It is the measurement of the degree through which the preferred codons are used by genes. It is the approach to calculate the level of

biasness of the codons that are favoured in highly expressed genes. The RSCU table helps in the identification of codons that are rapidly used for each amino acid. The relative adaptiveness of a codon (w(i)) is calculated as

$$W(i)=RSCU(i)/RSCU(max)$$

Where RSCU(max)= the RSCU value for the most frequently used codon for an amino acid.

4.) GC Percentage: It is basically used for the calculation of silent sites as well as replacement sites in the codon.

5.) Effective number of codon: It is the approach to calculate biasness from equal codon usage in a gene.

$$F(k)= (nS-1)/(n-1)$$

Where 'n' is the total number of codons for that amino acid.

6.) Scaled Chi-Square (SCS): This method is derived from the deviation from equal usage of degenerate codons. The calculation is done by taking uniform synonymous codon usage as the expected value which is further divided by twice the number of codon.

7.) Correlated Scaled Chi-Square (CSCS): It includes a correlation for GC percentage and works in the same manner as SCS.

8.) The Mutational Response Index (MRI): It provides the difference between SCS and CSCS.

9.) Intrinsic Codon Deviation Index (ICDI): ICDI is calculated based on S(k) which is as follows:

$$S(k)=\sum(RSCU(I)-1)^2/(k(k-1))$$

Where RSCU value is calculated for the ith codon and 'k' is either 2,3,4 or 6 that depends on the synonymous group degeneration.

$$ICDI= (\sum S_2+ S_3+\sum S_4+\sum S_6)/18$$

Hence, the S(k) values are combined to calculate ICDI.

❖ SHIFT: A Webserver for Hidden Stops analysis In Frameshifted Translation

In the process of translation from nucleotide sequences, the reading frames play an important role. The lethal results might be produced due to the wrong selection of reading frames and the wrong protein products [24]. During the translation process, the changes that occur in the reading frame are very rare. For example the frameshift event. It is a genetic mutation that is caused due to the insertion and deletion that takes place in the given nucleotides. Many stop codons appear off-frame. However, in some cases, stop codons are lacked in coding sequences. The hidden stop codons are present in the form of +1 and -1 shifted reading frames. There is a possibility that the whole sequence can be mutated using a single indel. This could bring a complete change of stop codons. The objective of the stop codon is to control the protein products and vigil over the process of translation. Frameshifts often lead to synthesization of peptides that cytotoxic in nature. Besides this, it causes waste of resources, energy and the activity of the biosynthetic pathway. There is a growth of protein in large amounts which could cause diseases such as cancer if these stop codons are not read. There is the connectivity and relationship seen between the read-through stop codon and the disorders.

SHIFT webserver is a tool that is developed to identify all such hidden stop codons that might be present in a genomic DNA sequence. With the respective genetic code systems, both +1 and -1 frameshift hidden stops will be checked by this webserver. The categories of codons as well as the correlation between codon usage frequencies and codonic contribution will be calculated through this server. For statistically significant correlations, t-values are generated using one-tailed t-test. Analysis of frameshifted translation and their evolutionary implications can be found with the help of this server which can further help computational and evolutionary biologists.

❖ Network Analyst: a comprehensive network visual analystics platform for gene expression analysis.

Network Analyst is a tool developed to overcome the challenges such as it can be used for a large amount of data and the complex analysis can be performed through the simple interface. This tool is used for a myriad of analysis analyses that may include meta-analysis, system-level understanding of gene expression data along with the comprehensive profiling. In the year 2015, the tool was updated with the new user interface version that also has an

enhanced workflow about the meta-analysis of multiple gene expression studies. There are five different approaches in which the data can be entered in the Network Analyst. These are common network files, raw RNAseq reads, one or multiple gene lists, a single gene expression data table and multiple gene expression data tables. The analysis involves data input that might corresponds to specific data processing steps and is represented in the highly interactive visual analytics methods that also includes functional enrichment analysis. To help understand complex molecular interactions, biological networks provides an intuitive framework. A Deeper mechanistic insights can be obtained only through the analysis of such network pathways. To facilitate novel hypothesis generation, gene co-expression networks can complement experimental evidence networks based on large large-scale gene expression studies. The three unique gaps present in the bioinformatics tools are addressed by the Network Analyst. Initially, it creates and visualise visualize biological networks to complement tools such as Cytoscape. Using R and Bioconductor packages, it enables web web-based meta meta-analysis of gene expression data [25].

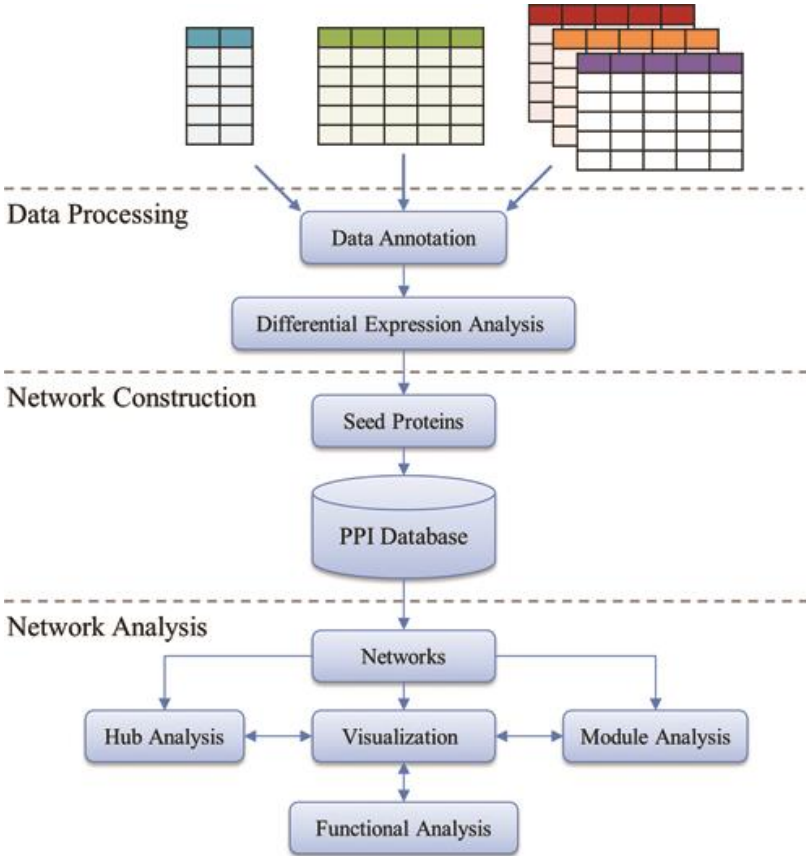


Figure-09: Workflow of Network analyst which works in three consecutive steps-Data Processing, Network Construction, Network Analysis.

❖ g:Profiler: a webserver for functional enrichment analysis

Several challenges have been observed regarding the utility of the data related with the application of genome-wide experimental approaches and the biological problems. With the help of a myriad of computational methods, a focus can be made on high throughput data. To filter down the complexity of –omics data, databases such as Gene Ontology can be used. A number of GO tools are also required for functional groups as well as the biological classification of genes. These analysis will help to uncover the biological effects [26]. The analysis can be visualised in a hierarchial manner with the help of a plot and this plot can be developed from the g:Profiler applications. This webserver manipulates gene lists and further help in the powerful visualisation. Apart from this, it also performs statistical enrichment analysis where data is provided from multiple sources for functional evidence. It maintains the focus of the user on a specific disease information or various biological, molecular and cellular components.

❖ GWAS Catalog

All the genome wide association studies of the unstructured data obtained from a myriad number of resources are compiled and kept within this online database. It consist of various information that may include SNP disease association information, study group information along with the identification of the genetic loci that could be analysed to know further about a specific disease and common traits. It is a program that was created with the collaboration of National Human Genome Research Institute and European Bioinformatics Institute in the year 2008 and 2010. It also has some additional features of a curation interface and graphical user interface [27].

❖ SNPs3D

With the help of SNP data, one can gain insights about the relationship between the genotype and the statistical data of a specific disease that have been provided. There are diseases such as cancer and Alzheimer of which people have gained very less knowledge about it. All the association studies of SNPs identifies low signal which tell about the risk that might be develop from any single variant. These studies further include disease probability influence. SNPs3D serves many purposes. The first one is to find appropriate candidate genes with the help of appropriate path. The genes

which might be solely responsible to influence disease susceptibility. Besides this, it also helps in choosing most appropriate non-synonymous SNPs that are present within such genes. The second purpose is to combine the molecular level data as well as mechanisms associated with disease and genetic variations [28].

❖ PANTHER (Protein Analysis Through Evolutionary Relationships)

In order to facilitate the high throughput analysis, this classification system was developed. There are several tools present in the panther such as gene list analysis, sequence search and cSNP scoring. cSNP usually refers to the coding SNPs. To cause a functional impact on the protein, it calculates likelihood of the nonsynonymous coding SNP. It uses a specific method called PANTHER-PSEP (position-specifc evolutionary preservation) which calculates the time period in which amino acid was preserved and with that, it explains how much functional impact can be there for that SNP [29].

❖ SNPs&GO (Predicting disease associated variations using GO terms)

SNPs is mainly responsible for all the genetic basis that takes place within human variability. Most common ones are the missense mutations that have been identified to cause residue substitutions in the protein. The information collected from the 3D structure, protein sequence, protein function and protein sequence profile is all integrated within SNPs&GO. It helps in relating all the protein variations that are present in the database with diseases [30]. It is based on the SVM classifier and its output consist of the predictions that identifies whether or not mutation is disease-related. It captures all the complementary knowledge correlations that eist in the database.

❖ GEMiCCL (Gene Expression and Mutations in Cancer Cell Lines)

A variety of biomedical studies include cell lines due to some underlying factors. These factors are homogeneous characteristics, cost-effectiveness, convenience and unlimited supply. In the drug development process, cell lines are considered to be the most essential part. For a study design and for choosing the appropriate cell line, it is needed that both the molecular as well as clinical characteristics are taken into

account. Earlier, more options were provided for gathering information on cell lines in biological experiments and drug testing. This information consist of molecular characteristics including genetic abnormalities. Therefore, it was needed that the information should be obtained in a unified manner with the help of some comprehensive database. It is where GEMiCCL comes into play. This database is a composition of three genomic data resources through which much information can be retrieved. Information about gene expression, copy number alteration, cancer cell lines and their various comparisons, searches and browsing with the help of interactive as well as intuitive features [31].

EXPECTED OUTCOMES

We will analyze the protein-coding gene data which will further provide the pathway level information when we will use the systems biology approach for network analysis. Sequence and structural level analysis of genes for SNPs and their impact will provide biologically meaningful information.

RESULTS AND DISCUSSION

With the help of the ACUA tool, we calculated CAI, ENc, AT and GC content values of lung cancer protein-coding genes such as TOP2A, CDC 20, BUB1, MADL1, JUN, FYN, CAV1, SFN, ASPM, CCNB2, CDC45, MELK and UBE2T. We found that the CAI value of >NC_000001.11:c197146669-197084127 is 0.668 which is quite high when compared with others. This means that ASPM is a highly expressed gene as the CAI value is high for highly expressed genes and low for the lowly expressed gene. CAI range from 0 for no bias and 1 for the strongest bias. Apart from this, the RSCU value of ASPM came out to be 1 that means the synonymous codons of an amino acid are used with equal frequencies.

Table-01: Results obtained from the ACUA tool (AT percentage, GC percentage, AT1 skewness, GC1 skewness, CAI and ENc)

Gene Position on Chromosome	A	T	G	C	Total bp	A1	T1
>NC_000017.11:c40417902-40388525 Homo sapiens chromosome 17, GRCh38.p13 Primary Assembly	8659	9028	5862	5829	29378	2898	3032
>NC_000001.11:43358981-43363203 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly	989	1031	1055	1148	4223	347	324
>NC_000001.11:c58784047-58780791 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly	778	651	958	870	3257	224	185
>BC050321.1 Homo sapiens TBC1 (tre-2/USP6, BUB2, cdc16) domain family, member 1, mRNA (cDNA clone MGC:48435 IMAGE:5262043), complete cds	1104	905	1074	1079	4162	377	239
>NC_000007.14:116525009-116561185 Homo sapiens chromosome 7, GRCh38.p13 Primary Assembly	10922	10175	7586	7494	36177	3645	3393
>NC_000001.11:26863149-	259	228	411	410	1308	92	66

26864456 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly							
>NC_000001.11:c197146669-197084127 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly	19267	21602	11292	10382	62543	6386	7224
>NC_000015.10:59105146-59125045 Homo sapiens chromosome 15, GRCh38.p13 Primary Assembly	5510	5918	4518	3954	19900	1789	1930
>NC_000022.11:19479294-19520612 Homo sapiens chromosome 22, GRCh38.p13 Primary Assembly	9737	11751	10145	9686	41319	3227	3897
>NC_000001.11:c202341936-202331657 Homo sapiens chromosome 1, GRCh38.p13 Primary Assembly	3013	3095	2140	2032	10280	1045	1030
>NC_000009.12:36572870-36677682 Homo sapiens chromosome 9, GRCh38.p13 Primary Assembly	27329	33480	22443	21561	104813	9110	11212

Table-01 (Continued)

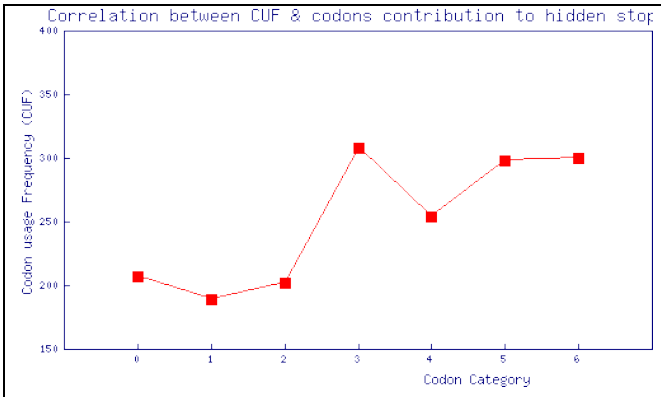
G1	C1	AT1 Percent	GC1 Percent	AT1 Skewness	GC1 Skewness	CAI	ENc
1958	1905	20.185	13.149	-0.023	0.014	0.686	52.871
382	355	15.889	17.452	0.034	0.037	0.739	55.174
354	323	12.558	20.786	0.095	0.046	0.709	57.745
379	392	14.801	18.525	0.224	-0.017	0.714	54.531
2551	2470	19.454	13.879	0.036	0.016	0.7	52.73
159	119	12.08	21.254	0.165	0.144	0.71	58.075
3735	3503	21.761	11.573	-0.062	0.032	0.668	48.836
1613	1301	18.688	14.643	-0.038	0.107	0.699	52.942
3380	3269	17.241	16.092	-0.094	0.017	0.718	52.999
715	637	20.185	13.152	0.007	0.058	0.703	52.404
7471	7145	19.389	13.945	-0.103	0.022	0.692	52.858

The second table lists the RSCU values as well as the codon usage table which can be used for further statistical analysis. The complete information retrieved can be helpful in gene ontology classifications. RSCU is 0 for low preferences and 1 for high preferences.

Table-02: Codon usage table with the frequency of amino acids and their fractional values

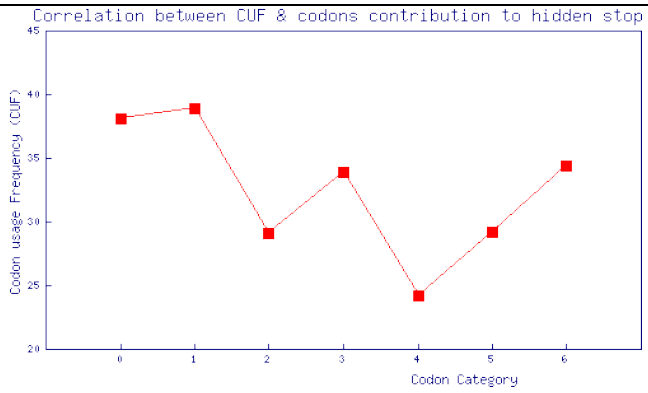
Codon	Amino Acid	Fraction	/1000	Number
GCA	A	0.29	13.763	1456
GCC	A	0.29	13.754	1455
GCG	A	0.075	3.564	377
GCT	A	0.345	16.363	1731
TGC	C	0.408	14.926	1579
TGT	C	0.592	21.694	2295
GAC	D	0.391	8.659	916
GAT	D	0.609	13.461	1424
GAA	E	0.5	16.646	1761
GAG	E	0.5	16.646	1761
TTC	F	0.297	19.236	2035
TTT	F	0.703	45.571	4821
GGA	G	0.268	14.68	1553
GGC	G	0.251	13.763	1456
GGG	G	0.237	13.007	1376
GGT	G	0.243	13.328	1410
CAC	H	0.447	13.716	1451
CAT	H	0.553	16.958	1794
ATA	I	0.335	19.151	2026

After the calculation of measures, we proceeded towards finding out the frameshifted translation in the nucleotide sequences of various genes. For this, we made use of the SHIFT database as we need to find out the hidden stop codon in the genomic DNA sequence. We obtained different graphs for genes with different correlation and t-test values.



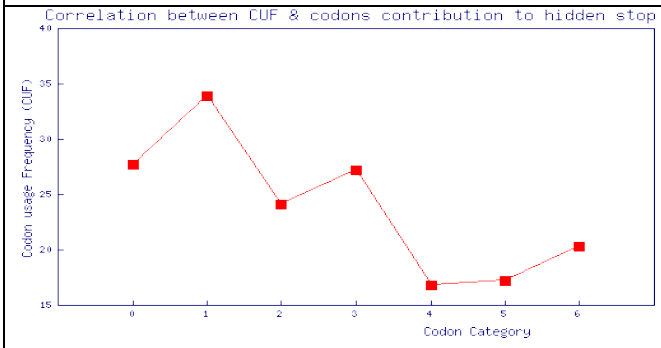
Correlation Coeffcient(*r*): 0.817 t-value: 3.17

(a)



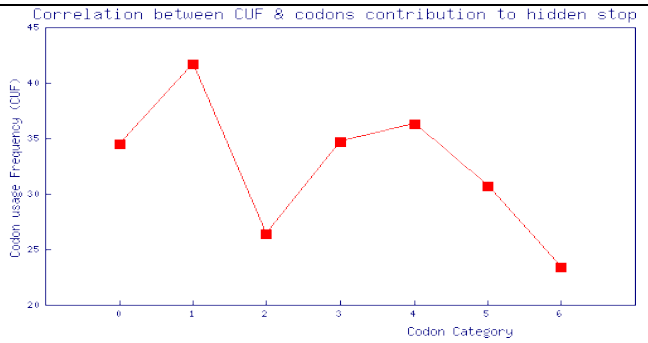
Correlation Coeffcient(*r*): -0.516 t-value: -ve value

(b)



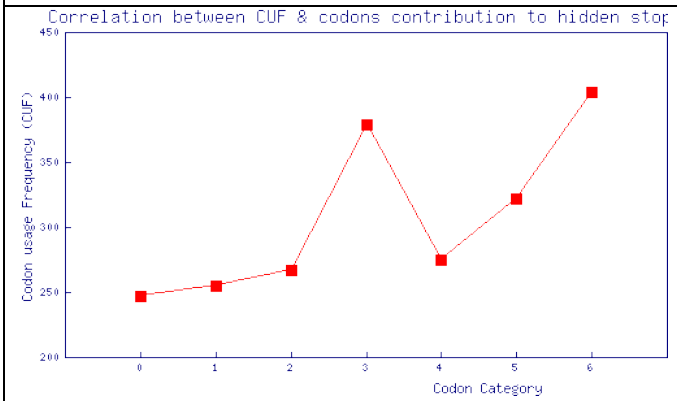
Correlation Coeffcient(*r*): -0.778 t-value: -ve value

(c)



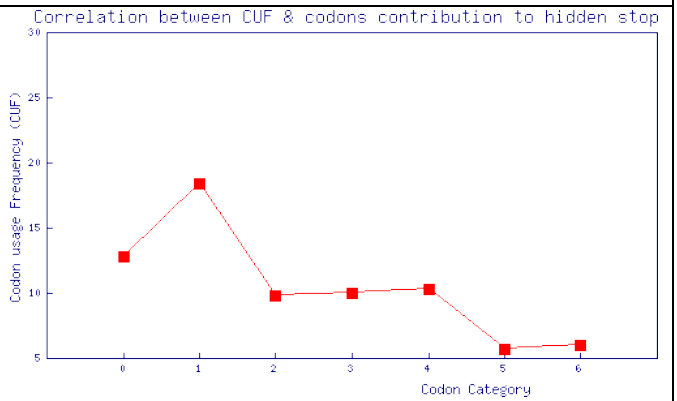
Correlation Coeffcient(*r*): -0.561 t-value: -ve value

(d)



Correlation Coeffcient(*r*): 0.752 t-value: 2.55

(e)



Correlation Coeffcient(*r*): -0.811 t-value: -ve value

(f)

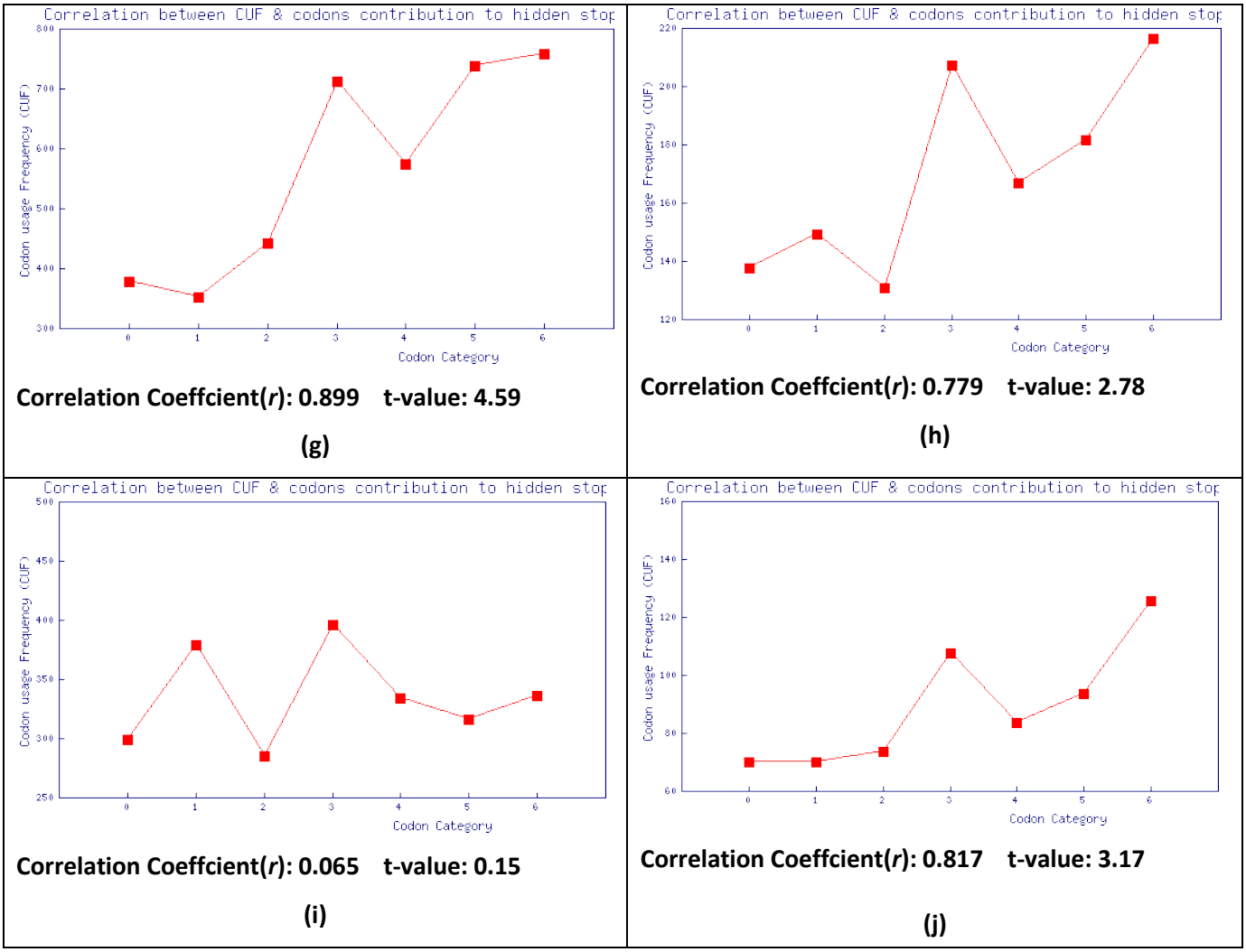
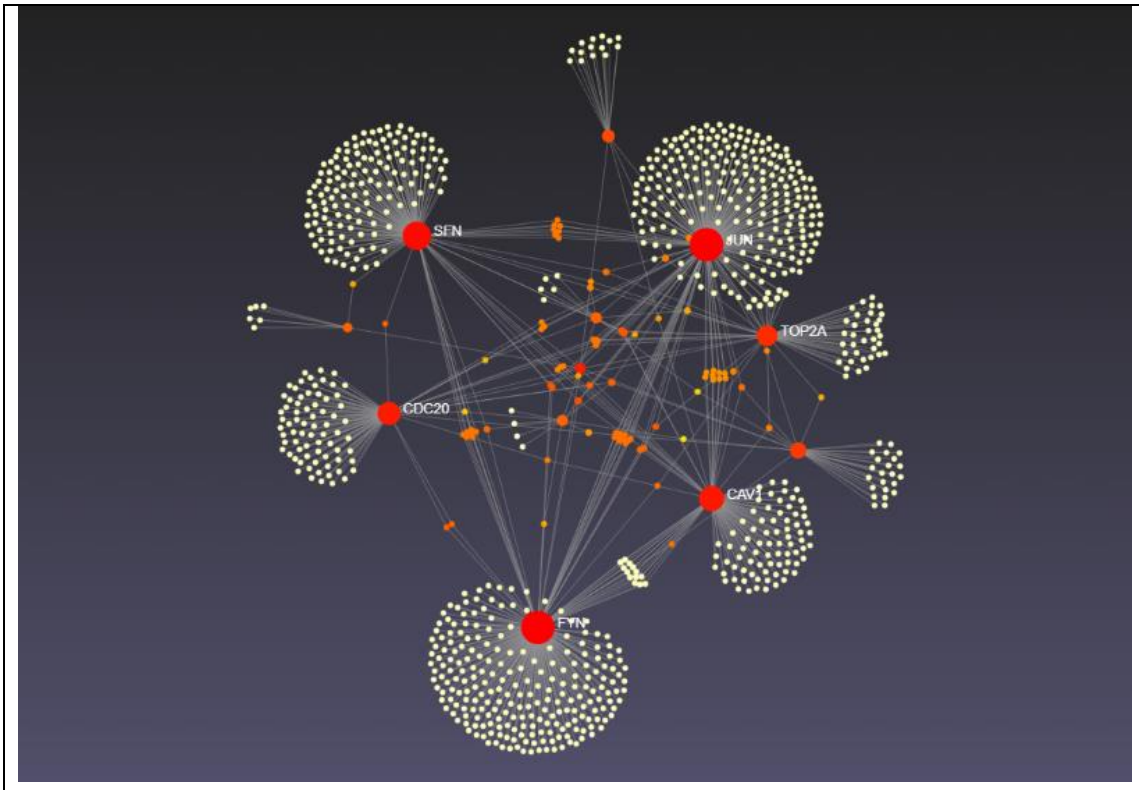


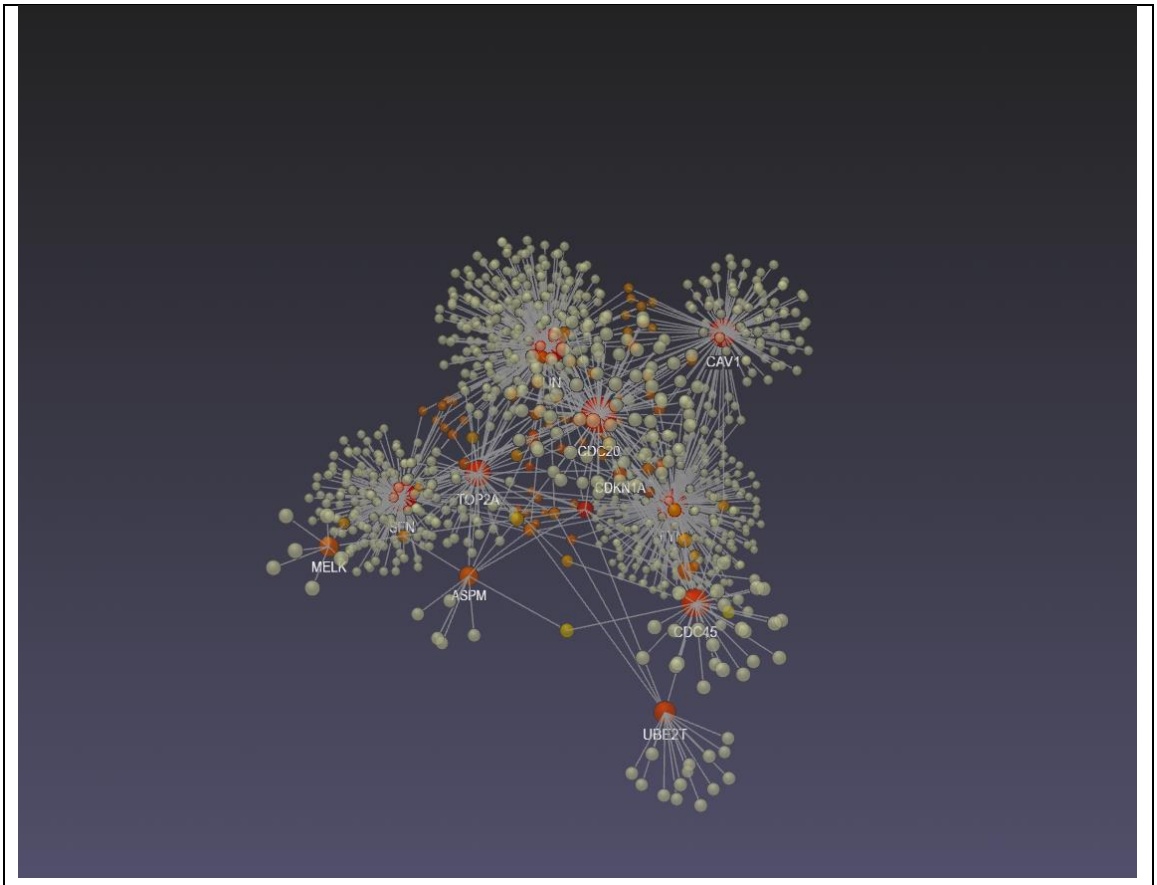
Figure-10: Comparison of correlation and t-test values generated with the help of SHIFT webserver for different protein coding genes: (a) TOP2A, (b) CDC20, (c) BUB1, (d) JUN (e) FYN (f) CAV1, (g) ASPM, (h) SFN, (i) CCNB2 (j) MELK. The graph was generated between codon usage frequency and condonic contribution to hidden stops.

With the identification of hidden stop codons, the computational and evolutionary biologists will be able to analyze implications of frameshifted translations which could help to stop the massive growth of protein products and further helps in the control of diseases such as lung cancer. Furthermore, we did the network analysis of the genes with the help of the Network Analyst tool. This tool provides quintessential information to gain a system-level understanding of the gene expression patterns that are present in different cell types, disease states as well as other experimental conditions. Diseases such as lung cancer that are primarily driven by mutations, novel pathogenesis pathways

revealed from the network-based approaches can be helpful in the determination and identification of potential therapeutic targets.



(a)



(b)

Figure-11: (a) 2D representation; (b) 3D representation of network analysis approach to understand gene expression patterns

For the proteins that directly interact with the seed proteins, a search algorithm is performed for every individual (seed) protein. The approach returns two types of sub-networks. Large sub-networks is regarded as the continents and smaller one is defined as the island. There are two widely used measures provided by the Network Analyst. These are-degree and betweenness centrality. The number of nodes connected with the other nodes is probably determined by its degree of centrality whereas the shortest path passing through any node is measured in terms of betweenness centrality. Potentially important hubs that are required for cellular signal trafficking are the nodes that have a high degree of betweenness values. If there is a seed protein then the log fold change value will be retrieved. If the given node is not a seed protein, it would be determined with the symbol "-". The p-value mentioned in the given table describes the significance of each module. This significance is calculated based on the difference that occurs between the edges that connect the node with the network and the number of edges.

Table-03: Genes with their degree and betweenness centrality values

Node Id	Label	Degree	Betweenness
3725	JUN	279	198540.1
2534	FYN	273	198739.9
2810	SFN	165	122908.4
857	CAV1	115	80562.56
991	CDC20	89	67556.36
7153	TOP2A	64	38271.59
8318	CDC45	33	22857.34
29089	UBE2T	18	13135.99
259266	ASPM	12	5592.64
9133	CCNB2	12	5109.42
7316	UBC	11	66010.12
9833	MELK	8	5680.76
7157	TP53	5	8546.27
83737	ITCH	4	11938.66
2099	ESR1	4	7338.53
1026	CDKN1A	4	5154.72
6613	SUMO2	4	2086.58
2908	NR3C1	3	8956.23
25	ABL1	3	8956.23
1499	CTNNB1	3	6704.69
351	APP	3	5638.71
6667	SP1	3	4625.25
1445	CSK	3	4625.25
983	CDK1	3	4041.54
3172	HNF4A	3	3611.64
1457	CSNK2A1	3	3607.52
10987	COPS5	3	2853.66
3066	HDAC2	3	2853.66
3065	HDAC1	3	2853.66
4738	NEDD8	3	1985.58
5594	MAPK1	3	1924.3
672	BRCA1	3	1526.86
8626	TP63	3	1514.4
1017	CDK2	3	890.44
4609	MYC	3	496.52
10013	HDAC6	2	6253.27
996	CDC27	2	6253.27
11186	RASSF1	2	5478.78
25920	COBRA1	2	3449.54
3320	HSP90AA1	2	3449.54
602	BCL3	2	3449.54
6774	STAT3	2	3449.54

2516	NR5A1	2	3449.54
10399	GNB2L1	2	3449.54
7409	VAV1	2	3449.54
1601	DAB2	2	3449.54
2571	GAD1	2	3449.54
4734	NEDD4	2	3449.54
326	AIRE	2	3255.16
4998	ORC1	2	3189.5
9564	BCAR1	2	2964.18
867	CBL	2	2964.18
5062	PAK2	2	2964.18
5509	PPP1R3D	2	2964.18
5894	RAF1	2	2964.18
9743	ARHGAP32	2	2964.18
8871	SYNJ2	2	2964.18
9759	HDAC4	2	2542.52
1386	ATF2	2	2542.52
9734	HDAC9	2	2542.52
10980	COPS6	2	2542.52
4193	MDM2	2	2542.52
334	APLP2	2	2542.52
64326	RFWD2	2	2542.52
3397	ID1	2	2309.31
23411	SIRT1	2	2153.68
2033	EP300	2	2153.68
1387	CREBBP	2	2153.68
1956	EGFR	2	1806.72
7249	TSC2	2	1806.72
5576	PRKAR2A	2	1806.72
5531	PPP4C	2	1600.77
367	AR	2	1175.71
5595	MAPK3	2	1175.71
4843	NOS2	2	1175.71
7186	TRAF2	2	1175.71
7852	CXCR4	2	1175.71
4087	SMAD2	2	1175.71
2697	GJA1	2	1175.71
7124	TNF	2	1175.71
5743	PTGS2	2	1175.71
9636	ISG15	2	1065
9040	UBE2M	2	723.28
81620	CDT1	2	699.98
2316	FLNA	2	565.94
994	CDC25B	2	536.55
7048	TGFBR2	2	509.82
8358	HIST1H3E	2	401.83
7161	TP73	2	272.59

1163	CKS1B	2	137.61
8452	CUL3	2	97.81

Table-04: Genes that were involved in various pathways with their p-values and false discovery rates

Pathway	Total	Expected	Hits	P value	FDR
Cell cycle	124	9.68	73	1.24E-49	3.93E-47
HTLV-I infection	219	17.1	83	2.57E-37	4.08E-35
Hepatitis B	163	12.7	68	1.11E-33	1.18E-31
Pathways in cancer	530	41.4	124	2.14E-31	1.70E-29
Chronic myeloid leukemia	76	5.93	40	6.32E-25	4.02E-23
T cell receptor signaling pathway	101	7.88	45	5.41E-24	2.87E-22
Prostate cancer	97	7.57	44	7.06E-24	3.21E-22
Viral carcinogenesis	201	15.7	63	3.97E-23	1.58E-21
MAPK signaling pathway	295	23	77	1.65E-22	5.82E-21
Progesterone-mediated oocyte maturation	99	7.73	43	1.98E-22	6.29E-21
Neurotrophinsignaling pathway	119	9.29	47	2.80E-22	8.09E-21
Osteoclast differentiation	128	9.99	48	1.39E-21	3.69E-20
AGE-RAGE signaling pathway in diabetic complications	100	7.81	42	3.02E-21	7.40E-20
Chagas disease (American trypanosomiasis)	103	8.04	42	1.19E-20	2.70E-19
Ubiquitin mediated proteolysis	137	10.7	48	4.03E-20	8.54E-19
Hepatitis C	155	12.1	51	6.29E-20	1.25E-18
ErbBsignaling pathway	85	6.64	37	1.71E-19	3.20E-18
TNF signaling pathway	110	8.59	42	2.32E-19	4.09E-18
Fc epsilon RI signaling pathway	68	5.31	33	2.51E-19	4.20E-18
FoxOsignaling pathway	132	10.3	46	3.23E-19	5.14E-18
Kaposi's sarcoma-associated herpesvirus infection	186	14.5	55	5.65E-19	8.56E-18

Colorectal cancer	86	6.71	36	2.59E-18	3.74E-17
Epstein-Barr virus infection	201	15.7	56	5.66E-18	7.51E-17
Proteoglycans in cancer	201	15.7	56	5.66E-18	7.51E-17
Oocyte meiosis	125	9.76	43	8.64E-18	1.10E-16
Pancreatic cancer	75	5.85	33	1.09E-17	1.34E-16
Fluid shear stress and atherosclerosis	139	10.9	45	2.15E-17	2.53E-16
Apoptosis	136	10.6	44	5.09E-17	5.78E-16
Acute myeloid leukemia	66	5.15	30	1.13E-16	1.24E-15
Th17 cell differentiation	107	8.35	38	2.19E-16	2.32E-15

If we need to visualize the gene expression data, there is a popular method of visualizing it. This method is called Heatmaps. Besides the fold change patterns, it tells whether the gene is present or absent in a particular gene list whenever we used this method for multiple genes lists visualization. Apart from this, we get to know about the overall view of DEGs that are shared across multiple lists.

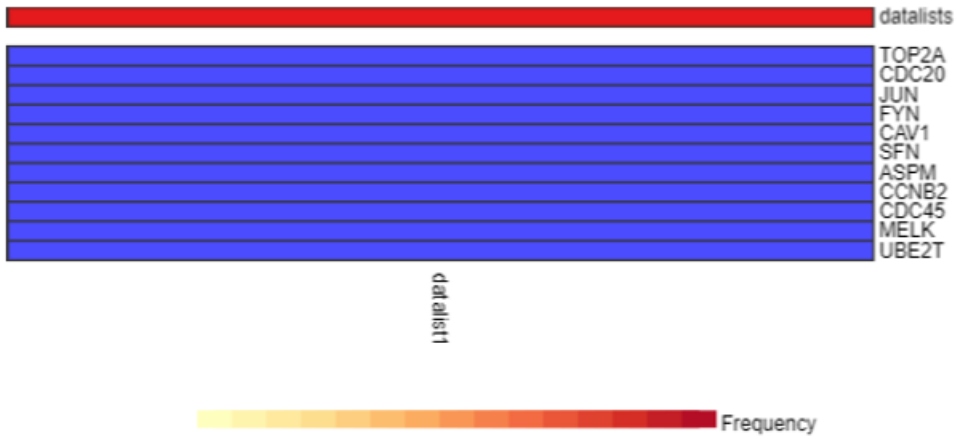


Figure-12: List of the gene represented through Heatmaps.

With this, we also did the gene ontology analysis using g:Profiler application. The Manhattan plot generated provides over-representation of information that have been analysed using Gene Ontology terms, biological pathways, human disease gene

annotations and protein-protein network. The x-axis shows the group of functional terms whereas the y-axis consist of the adjusted enrichment p-values in negative log10 scale. It provides a large set of genes that identifies all those functional terms that are associated with the experimental setup and dramatic changes within it.

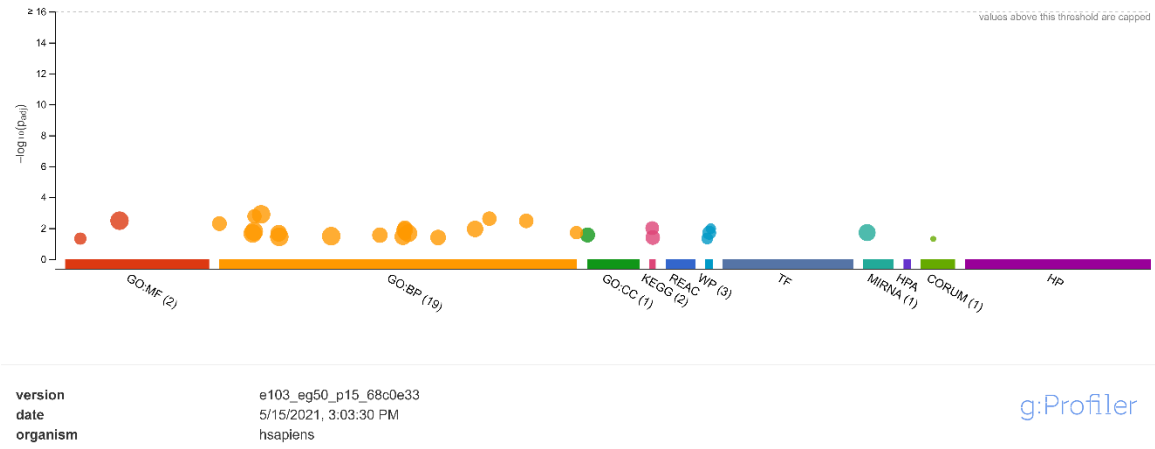


Figure-13: Manhattan plot showing the Gene Ontology analysis

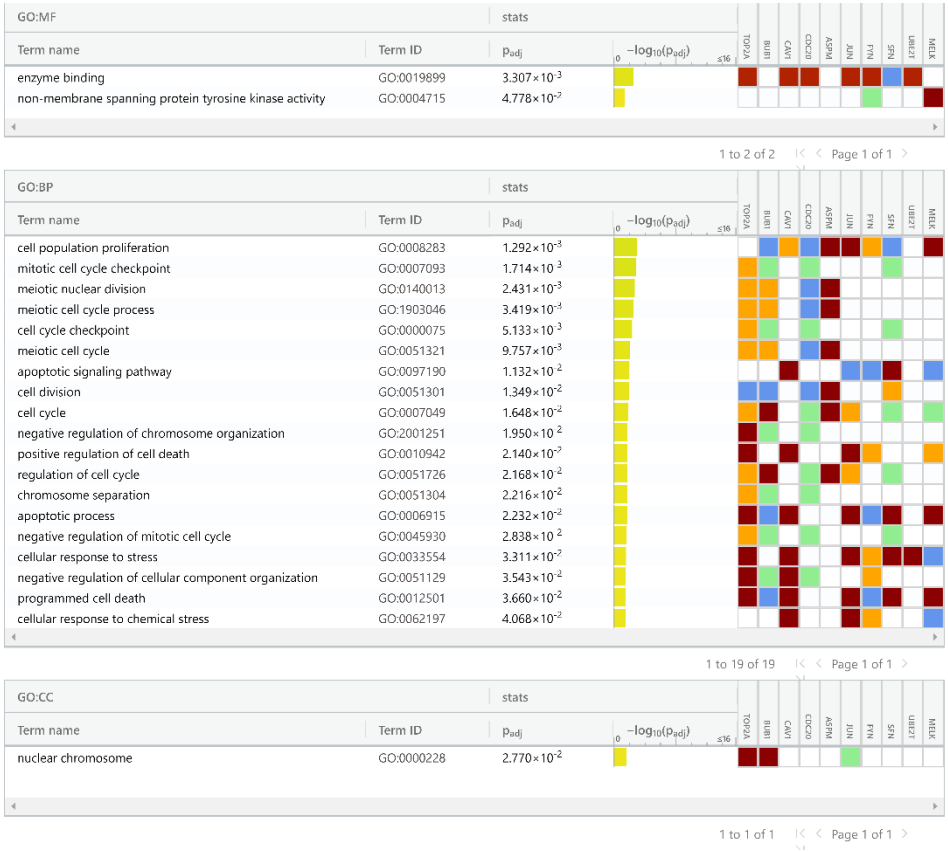


Figure-14: Detailed Description of molecular, biological and cellular processes associated with the protein coding genes

After the gene ontology analysis, we were required to obtain SNP analysis and find out which genes are disease associated and which are neutral. We firstly retrieved all the genome wide association studies involved with the genes. We got a lot of information regarding the risk alleles, p values, RAF values, CI values and the location of the mapped genes. From there, we got reference SNP cluster id which were required so that we might refer to a specific SNP. We further used SNPs3D to obtain non-synonymous SNPs with their functional impact. Deleterious SNPs are classified with the help of negative SVM score. We took the SNP ids from here and downloaded the protein sequence of these genes from uniprot. We used panther database for the evolutionary analysis of coding SNPs. Here, on the basis of PSEP, we can calculate preservation time of an amino acid which will further provide the probability of deleterious effect based on Pdel value. If the time > 450 my and has the false positive rate of 0.2, it means that the substitution is probably damaging. If the time is less than 450 my and greater than 200 my with a false positive rate of 0.4, it means that the substitution is possibly damaging. If the time < 200 my, it means that the substitution is probably benign. On the basis of this analysis, we further proceeded for the prediction of disease associated variation. We used SNPs&GO for this. Based on the mutation and the protein sequence, we retrieved the RI values and the probability. RI (Reliability Index) is evaluated with the help of SVM only on places where the sign of stability change is predicted. If the probability is greater than 0.5 then that mutation is referred to be disease associated. In this study, we found that genes such as TOP2A, CDC20, JUN, FYN and BUB1 were disease associated and genes such as ASPM and SFN were found to neutral or benign.

Table-05: Genes with their mutations and prediction of disease associated variations

Gene Name	**RS_ID**	**SNP**	**Preservation Time**	**Pdel values**	**RI Values**	**Probability**	**Prediction**
TOP2A	rs13695, rs16965748, rs11867902	S654I	1189	0.85	2	0.585	Probably damaging and disease associated
ASPM	rs10922162	R430G	30	0.13	8	0.111	Probably benign and neutral

							association
CDC20	rs45461499	V402M	1629	0.89	4	0.693	Probably damaging and disease associated
SFN	rs72307178, rs546525796	M155I	325	0.5	9	0.057	Possibly damaging and Neutral association
FYN	rs2148710, rs7757969, rs28763977	A438D	911	0.85	8	0.886	Probably damaging and disease associated
BUB1	rs140983998, rs185949673	G20D	1628	0.89	4	0.680	Probably damaging and disease associated
JUN	rs17073012	T297M	797	0.74	3	0.674	Probably damaging and disease associated

With this step, we proceeded to determine the molecular and clinical information about the genes as these two are considered to be the most important factors in functional and screening experiments whenever a cancer cell line is selected. There is a table displayed with the gene expression values in percentiles. These values will lead our focus towards the pathways or activities of genes. The second table describes the mutations present in a specified cell line. There are thousands of mutations identified by exome sequencing that makes hard for us to locate specific gene of interest. We have also used the browsing function to determine copy number alterations, expression levels along the presence of mutations. This information will help the potential experimental biologists to look for the specific molecular characteristics. This will further help them in quality control and authentication of cell lines for drug testing. It could also be an essential step in compound toxicity testing with the development of drugs. In our study, we found that the expression level of ASPM was significantly increased when we observed it for the lung cancer. This could be advantageous for predicting the lung cancer in advance. Due to clinical manifestations, there has been no therapeutics advancement made in the field of lung cancer. However, ASPM shows co-expression with CDK4 which indicates its important role as a therapeutic target. Therefore, we kept our focus on ASPM to retrieve its cell line names with other information as well.

Table-06: Gene Expression in percentiles present in different genomic resources with their cell names.

GeneName	Cell line Name	Entrez	CCLE	COSMIC	NCI60
ASPM	CHP-126	259266	51.58%	61.07%	-
ASPM	CHP-126	259266	51.58%	61.07%	-
ASPM	CHP-212	259266	64.77%	57.18%	-
ASPM	IMR-32	259266	64.46%	-	-
ASPM	Kelly	259266	66.06%	55.58%	-
ASPM	KP-N-RT-BM-1	259266	59.81%	-	-
ASPM	KP-N-RT-BM-1	259266	59.81%	-	-
ASPM	KP-N-SI9s	259266	62.53%	-	-
ASPM	KP-N-YN	259266	67.88%	62.32%	-
ASPM	MHH-NB-11	259266	55.72%	56.26%	-
ASPM	SJNB-1	259266	67.56%	63.16%	-
ASPM	SK-N-DZ	259266	63.22%	58.59%	-
ASPM	SK-N-FI	259266	56.23%	46.72%	-
ASPM	HuCC-T1	259266	64.32%	51.20%	-
ASPM	HuH-28	259266	66.91%	-	-
ASPM	Hs 706.T	259266	64.66%	-	-
ASPM	Hs 737.T	259266	45.91%	-	-
ASPM	Hs 821.T	259266	51.31%	-	-
ASPM	CAL-78	259266	60.02%	59.21%	-
ASPM	Hs 819.T	259266	46.02%	-	-
ASPM	A-673	259266	63.31%	58.88%	-
ASPM	CADO-ES1	259266	60.08%	53.22%	-
ASPM	Hs 822.T	259266	40.63%	-	-
ASPM	Hs 863.T	259266	52.25%	-	-
ASPM	MHH-ES-1	259266	70.58%	-	-
ASPM	143B	259266	59.08%	-	-
ASPM	G-292 clone A141B1	259266	65.56%	56.34%	-
ASPM	HOS	259266	64.94%	56.70%	-
ASPM	Hs 870.T	259266	65.26%	-	-
ASPM	Hs 888.T	259266	59.36%	-	-
ASPM	MG-63	259266	64.94%	56.20%	-
ASPM	AU565	259266	58.74%	50.60%	-
ASPM	CAMA-1	259266	57.12%	34.90%	-
ASPM	HCC1428	259266	61.90%	55.08%	-
ASPM	HCC1428	259266	61.90%	55.08%	-
ASPM	MDA-MB-231	259266	66.60%	58.67%	46.05%

ASPM	MDA-MB-231	259266	66.60%	58.67%	46.05%
ASPM	MDA-MB-361	259266	55.09%	50.20%	-
ASPM	MDA-MB-361	259266	55.09%	50.20%	-
ASPM	MDA-MB-415	259266	54.30%	55.14%	-
ASPM	MDA-MB-453	259266	69.84%	52.21%	-
ASPM	MDA-MB-453	259266	69.84%	52.21%	-
ASPM	MDA-MB-468	259266	61.05%	57.30%	-
ASPM	CAL-120	259266	61.34%	59.45%	-
ASPM	CAL-120	259266	61.34%	59.45%	-
ASPM	CAL-148	259266	58.85%	50.10%	-
ASPM	CAL-148	259266	58.85%	50.10%	-
ASPM	CAL-51	259266	67.20%	53.96%	-
ASPM	CAL-85-1	259266	61.80%	54.19%	-
ASPM	DU4475	259266	66.98%	57.25%	-
ASPM	EFM-192A	259266	60.96%	56.54%	-
ASPM	HCC1569	259266	62.72%	56.39%	-
ASPM	HCC2218	259266	55.81%	46.66%	-
ASPM	HDQ-P1	259266	54.82%	43.10%	-
ASPM	Hs 274.T	259266	61.77%	-	-
ASPM	Hs 281.T	259266	57.17%	-	-
ASPM	Hs 343.T	259266	61.68%	-	-
ASPM	Hs 606.T	259266	53.94%	-	-
ASPM	Hs 739.T	259266	53.14%	-	-
ASPM	Hs 742.T	259266	36.38%	-	-
ASPM	JIMT-1	259266	61.09%	57.40%	-
ASPM	MDA-MB-157	259266	64.53%	50.58%	-
ASPM	EFM-19	259266	63.08%	49.71%	-
ASPM	HCC1143	259266	59.79%	58.14%	-
ASPM	HCC1187	259266	64.06%	59.34%	-
ASPM	HCC1395	259266	63.58%	60.23%	-
ASPM	HCC1419	259266	54.16%	44.15%	-
ASPM	HCC1500	259266	51.02%	27.17%	-
ASPM	HCC1937	259266	66.53%	58.80%	-
ASPM	HCC1954	259266	63.33%	52.47%	-
ASPM	HCC202	259266	59.15%	51.12%	-
ASPM	HCC202	259266	59.15%	51.12%	-
ASPM	HCC2157	259266	60.33%	54.41%	-
ASPM	HCC38	259266	63.20%	58.12%	-
ASPM	HCC70	259266	63.61%	53.66%	-
ASPM	BT-20	259266	63.13%	56.24%	-

ASPM	BT-474	259266	68.65%	46.36%	-
ASPM	BT-483	259266	54.90%	48.77%	-
ASPM	BT-483	259266	54.90%	48.77%	-
ASPM	BT-549	259266	61.56%	59.33%	45.33%
ASPM	Hs 578T	259266	55.02%	49.55%	33.45%
ASPM	KPL-1	259266	57.91%	-	-
ASPM	MCF-7	259266	55.06%	52.31%	40.05%
ASPM	MCF-7	259266	55.06%	52.31%	40.05%
ASPM	MDA-MB-175-VII	259266	56.43%	-	-
ASPM	MDA-MB-436	259266	58.62%	66.15%	-
ASPM	MDA-MB-134-VI	259266	51.08%	52.97%	-
ASPM	HCC1806	259266	67.59%	51.80%	-
ASPM	1321N1	259266	65.64%	-	-
ASPM	Becker	259266	60.48%	23.42%	-
ASPM	CCF-STTG1	259266	59.55%	55.07%	-
ASPM	H4	259266	56.95%	52.80%	-
ASPM	42-MG-BA	259266	64.29%	45.99%	-
ASPM	8-MG-BA	259266	54.43%	53.57%	-
ASPM	A-172	259266	65.66%	53.88%	-
ASPM	AM-38	259266	65.28%	58.07%	-
ASPM	CAS-1	259266	56.30%	45.30%	-
ASPM	DBTRG-05MG	259266	58.02%	51.66%	-
ASPM	DK-MG	259266	56.19%	48.45%	-
ASPM	GaMG	259266	59.05%	25.97%	-
ASPM	GMS-10	259266	62.96%	51.95%	-
ASPM	GOS-3	259266	58.49%	-	-
ASPM	KALS-1	259266	65.67%	56.48%	-
ASPM	KNS-42	259266	64.33%	59.83%	-
ASPM	KNS-60	259266	62.63%	-	-
ASPM	KNS-81	259266	64.18%	-	-
ASPM	KS-1 [Human glioblastoma]	259266	65.84%	57.61%	-
ASPM	LN-18	259266	61.17%	41.50%	-
ASPM	LN-229	259266	67.17%	64.23%	-
ASPM	M059K	259266	65.32%	57.53%	-
ASPM	GI-1	259266	64.55%	55.22%	-
ASPM	D283 Med	259266	60.53%	42.28%	-
ASPM	D341 Med	259266	42.67%	-	-
ASPM	Daoy	259266	64.61%	58.59%	-
ASPM	Hs 683	259266	61.89%	56.54%	-
ASPM	KG-1-C	259266	64.40%	-	-
ASPM	AN3-CA	259266	59.59%	57.99%	-
ASPM	AN3-CA	259266	59.59%	57.99%	-

ASPM	HEC-1-A	259266	62.54%	-	-
ASPM	HEC-1-B	259266	63.20%	-	-
ASPM	HEC-108	259266	59.48%	-	-
ASPM	HEC-151	259266	61.66%	-	-
ASPM	HEC-265	259266	59.12%	-	-
ASPM	HEC-50B	259266	62.90%	-	-
ASPM	Ishikawa (Heraklio) 02 ER-	259266	66.46%	55.50%	-
ASPM	JHUEM-1	259266	48.31%	-	-
ASPM	JHUEM-2	259266	66.11%	-	-
ASPM	JHUEM-3	259266	58.15%	-	-
ASPM	KLE	259266	59.33%	62.39%	-
ASPM	MFE-280	259266	62.51%	57.10%	-
ASPM	MFE-296	259266	63.19%	57.71%	-
ASPM	MFE-319	259266	53.54%	46.74%	-
ASPM	MFE-319	259266	53.54%	46.74%	-
ASPM	RL95-2	259266	66.98%	63.76%	-
ASPM	EFE-184	259266	59.24%	-	-
ASPM	EN	259266	63.84%	57.43%	-
ASPM	EN	259266	63.84%	57.43%	-
ASPM	HEC-251	259266	60.37%	-	-
ASPM	HEC-59	259266	64.06%	-	-
ASPM	HEC-6	259266	62.80%	-	-
ASPM	ESS-1	259266	66.03%	50.73%	-
ASPM	F-36P	259266	49.85%	-	-
ASPM	KE-97	259266	59.67%	-	-
ASPM	BDCM	259266	62.56%	-	-
ASPM	P31/FUJ	259266	58.94%	48.51%	-
ASPM	Mono-Mac-1	259266	52.54%	45.77%	-
ASPM	Mono-Mac-6	259266	54.66%	52.77%	-
ASPM	GDM-1	259266	62.40%	52.27%	-
ASPM	HL-60	259266	62.94%	50.39%	43.58%
ASPM	Kasumi-6	259266	53.32%	-	-
ASPM	KG-1	259266	53.66%	56.45%	-
ASPM	KG-1	259266	53.66%	56.45%	-
ASPM	KO52	259266	61.76%	51.45%	-
ASPM	MOLM-13	259266	57.19%	50.28%	-
ASPM	MOLM-16	259266	58.19%	54.67%	-
ASPM	MOLM-16	259266	58.19%	54.67%	-
ASPM	NALM-6	259266	64.60%	56.36%	-
ASPM	RS4;11	259266	63.53%	54.29%	-
ASPM	KE-37	259266	67.05%	59.65%	-
ASPM	Loucy	259266	66.95%	61.31%	-
ASPM	MOLT-4	259266	65.57%	61.52%	44.93%
ASPM	A4/Fukuda	259266	60.85%	38.34%	-
ASPM	Kasumi-2	259266	63.80%	-	-

ASPM	MUTZ-5	259266	56.89%	-	-
ASPM	NALM-19	259266	56.18%	-	-
ASPM	RCH-ACV	259266	62.23%	60.36%	-
ASPM	CMK	259266	60.78%	57.28%	-
ASPM	CMK-11-5	259266	65.16%	-	-
ASPM	M-07e	259266	59.67%	-	-
ASPM	MV4-11	259266	58.17%	45.24%	-
ASPM	AML-193	259266	59.78%	-	-
ASPM	Kasumi-1	259266	61.58%	53.24%	-
ASPM	KOPN-8	259266	55.59%	62.48%	-
ASPM	MHH-CALL-2	259266	66.13%	58.82%	-
ASPM	MHH-CALL-3	259266	68.84%	-	-
ASPM	MHH-CALL-4	259266	63.39%	58.84%	-
ASPM	DND-41	259266	60.07%	62.08%	-
ASPM	HPB-ALL	259266	62.32%	-	-
ASPM	Jurkat	259266	63.00%	58.85%	-
ASPM	MOLT-13	259266	65.41%	60.57%	-
ASPM	MOLT-16	259266	62.88%	58.91%	-
ASPM	MEC-1	259266	61.04%	49.56%	-
ASPM	MEC-2	259266	59.66%	57.66%	-
ASPM	BV-173	259266	56.94%	56.83%	-
ASPM	CML-T1	259266	58.78%	55.40%	-
ASPM	EM-2	259266	60.12%	51.98%	-
ASPM	JK-1	259266	57.25%	-	-
ASPM	JURL-MK1	259266	51.17%	48.07%	-
ASPM	JURL-MK1	259266	51.17%	48.07%	-
ASPM	K-562	259266	56.80%	58.62%	45.60%
ASPM	K-562	259266	56.80%	58.62%	45.60%
ASPM	KCL-22	259266	62.50%	56.05%	-
ASPM	KCL-22	259266	62.50%	56.05%	-
ASPM	Ku812	259266	62.92%	61.23%	-
ASPM	KYO-1	259266	53.68%	-	-
ASPM	LAMA-84	259266	53.56%	54.24%	-
ASPM	MEG-01	259266	60.46%	48.76%	-
ASPM	MOLM-6	259266	53.10%	-	-
ASPM	NALM-1	259266	56.94%	-	-

Table-07: Description of Position, CDS mutation, AA mutation, Zygosity and Mutation type for specific cell line.

Cell line Name	GeneName	Entrez	Position	CDS Mutation	AA Mutation	Mutation Type	Zygosity
KCL-22	ASPM	259266	Chr1:197073232-	c.5149delA	p.I1717fs*1	Frame_Shift	Heterozygous

			197073232				
LoVo	ASPM	259266	Chr1:197073232-197073232	c.5149delA	p.I1717fs*1	Frame_Shift	Heterozygous
SNU-1040	ASPM	259266	Chr1:197073232-197073232	c.5149delA	p.I1717fs*1	Frame_Shift	Heterozygous
NCI-H630	ASPM	259266	Chr1:197073232-197073232	c.5149delA	p.I1717fs*1	Frame_Shift	Heterozygous
GP5d	ASPM	259266	Chr1:197111651-197111651	c.1731C>T	p.S577S	Silent	Heterozygous
SUP-T1	ASPM	259266	Chr1:197097732-197097732	c.2824C>T	p.R942C	Missense	Heterozygous
GR-ST	ASPM	259266	Chr1:197059121-197059121	c.9923G>A	p.R3308H	Missense	Heterozygous
HT115	ASPM	259266	Chr1:197073886-197073886	c.4495C>T	p.R1499W	Missense	Heterozygous
NCI-H2009	ASPM	259266	Chr1:197071152-197071152	c.7229C>T	p.S2410F	Missense	Heterozygous
NCI-H1703	ASPM	259266	Chr1:197073195-197073195	c.5186G>T	p.R1729L	Missense	Heterozygous
AMO1	ASPM	259266	Chr1:197070118-197070118	c.8263A>G	p.R2755G	Missense	Heterozygous
OPM-2	ASPM	259266	Chr1:197072622-197072622	c.5759T>C	p.F1920S	Missense	Heterozygous
MOLM-16	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
KYAE-1	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
RKN	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
SNU-175	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
IHH-4	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
KYSE-70	ASPM	259266	Chr1:197073406-	c.4975G>T	p.V1659F	Missense	Heterozygous

			197073406				
SBC-5	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
Lu-135	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
SJNB-1	ASPM	259266	Chr1:197073406-197073406	c.4975G>T	p.V1659F	Missense	Heterozygous
NCI-H1299	ASPM	259266	Chr1:197057419-197057419	c.10128G>T	p.L3376F	Missense	Heterozygous
NCI-H441	ASPM	259266	Chr1:197074260-197074260	c.4121C>T	p.S1374L	Missense	Heterozygous
CCK-81	ASPM	259266	Chr1:197070837-197070837	c.7544G>A	p.R2515Q	Missense	Heterozygous
OC 314	ASPM	259266	Chr1:197072215-197072215	c.6166G>A	p.A2056T	Missense	Heterozygous
IM95	ASPM	259266	Chr1:197073342-197073342	c.5039delA	p.N1680fs*4	Frame_Shift	Heterozygous
RL95-2	ASPM	259266	Chr1:197073342-197073342	c.5039delA	p.N1680fs*4	Frame_Shift	Heterozygous
IGROV-1	ASPM	259266	Chr1:197061057-197061057	c.9424A>T	p.N3142Y	Missense	Heterozygous
HCC2998	ASPM	259266	Chr1:197069984-197069984	c.8397A>C	p.K2799N	Missense	Heterozygous
HCT 15	ASPM	259266	Chr1:197072967-197072967	c.5414C>A	p.A1805D	Missense	Heterozygous
MDA-MB-231	ASPM	259266	Chr1:197074117-197074117	c.4264T>C	p.S1422P	Missense	Heterozygous
MCF-7	ASPM	259266	Chr1:197074298-197074298	c.4083T>A	p.Y1361*	Nonsense	Heterozygous
HCT 15	ASPM	259266	Chr1:197091653-197091653	c.3463T>G	p.Y1155D	Missense	Heterozygous
EBC-1	ASPM	259266	Chr1:197108957-197108957	c.1966A>G	p.T656A	Missense	Heterozygous
JeKo-1	ASPM	259266	Chr1:197108957-	c.1966A>G	p.T656A	Missense	Heterozygous

			197108957				
ES1	ASPM	259266	Chr1:197055969-197055969	c.10295T>A	p.I3432N	Missense	Heterozygous
TE-1	ASPM	259266	Chr1:197056003-197056003	c.10261C>A	p.Q3421K	Missense	Heterozygous
NUGC-4	ASPM	259266	Chr1:197056003-197056003	c.10261C>A	p.Q3421K	Missense	Heterozygous
5637	ASPM	259266	Chr1:197056028-197056028	c.10236G>A	p.M3412I	Missense	Heterozygous
ES4	ASPM	259266	Chr1:197056057-197056058	c.10206_10207delCT	p.Y3403fs*6	Frame_Shift	Heterozygous
EMC-BAC-1	ASPM	259266	Chr1:197056075-197056075	c.10189G>T	p.D3397Y	Missense	Heterozygous
MCF-7	ASPM	259266	Chr1:197056089-197056089	c.10175G>A	p.R3392K	Missense	Heterozygous
HCC202	ASPM	259266	Chr1:197057406-197057406	c.10141A>G	p.K3381E	Missense	Heterozygous
SNU-1040	ASPM	259266	Chr1:197059084-197059084	c.9960G>A	p.V3320V	Silent	Heterozygous
BL-41	ASPM	259266	Chr1:197059120-197059120	c.9924C>T	p.R3308R	Silent	Heterozygous
SNU-81	ASPM	259266	Chr1:197059154-197059154	c.9890C>A	p.S3297Y	Missense	Heterozygous
NUGC-4	ASPM	259266	Chr1:197059211-197059211	c.9833T>C	p.V3278A	Missense	Heterozygous
AN3-CA	ASPM	259266	Chr1:197059320-197059320	c.9829+6T>C	p.?	Unknown	Heterozygous
DG-75	ASPM	259266	Chr1:197059411-197059411	c.9744A>G	p.K3248K	Silent	Heterozygous
EHEB	ASPM	259266	Chr1:197060083-197060083	c.9533T>C	p.L3178P	Missense	Heterozygous
LS180	ASPM	259266	Chr1:197061035-197061035	c.9444+2T>C	p.?	Unknown	Heterozygous
JEG-3	ASPM	259266	Chr1:197061098-	c.9383A>G	p.Y3128C	Missense	Heterozygous

			197061098				
WM793	ASPM	259266	Chr1:197061104-197061104	c.9377G>A	p.R3126K	Missense	Heterozygous
TE-4	ASPM	259266	Chr1:197061171-197061171	c.9310G>T	p.A3104S	Missense	Heterozygous
RD	ASPM	259266	Chr1:197062189-197062189	c.9287G>A	p.R3096Q	Missense	Heterozygous
KM-H2	ASPM	259266	Chr1:197062305-197062305	c.9171G>C	p.L3057F	Missense	Heterozygous
MOLP-8	ASPM	259266	Chr1:197062375-197062375	c.9101G>A	p.C3034Y	Missense	Heterozygous
JJN-3	ASPM	259266	Chr1:197062394-197062394	c.9085-3T>G	p.?	Unknown	Heterozygous
GR-ST	ASPM	259266	Chr1:197063275-197063275	c.9023A>T	p.Q3008L	Missense	Heterozygous
P30/OHK	ASPM	259266	Chr1:197063305-197063305	c.8993delT	p.L2998fs*1	Frame_Shift	Heterozygous
HuO9	ASPM	259266	Chr1:197069698-197069698	c.8683G>A	p.A2895T	Missense	Heterozygous
HT	ASPM	259266	Chr1:197069715-197069715	c.8666delT	p.L2889fs*49	Frame_Shift	Heterozygous
RKO	ASPM	259266	Chr1:197069746-197069747	c.8634_8635insT	p.L2880fs*4	Frame_Shift	Heterozygous
VMRC-LCD	ASPM	259266	Chr1:197069824-197069824	c.8557C>T	p.R2853W	Missense	Heterozygous
EN	ASPM	259266	Chr1:197069927-197069927	c.8454A>G	p.A2818A	Silent	Heterozygous
SNU-407	ASPM	259266	Chr1:197069929-197069929	c.8452G>T	p.A2818S	Missense	Heterozygous
H-EMC-SS	ASPM	259266	Chr1:197069929-197069929	c.8452G>T	p.A2818S	Missense	Heterozygous
CW-2	ASPM	259266	Chr1:197069929-197069929	c.8452G>T	p.A2818S	Missense	Heterozygous
ASH-3	ASPM	259266	Chr1:197069929-	c.8452G>T	p.A2818S	Missense	Heterozygous

			197069929				
NCI-H2023	ASPM	259266	Chr1:197069938-197069938	c.8443A>T	p.S2815C	Missense	Heterozygous
A-704	ASPM	259266	Chr1:197070037-197070037	c.8344A>C	p.T2782P	Missense	Heterozygous
NCI-H2170	ASPM	259266	Chr1:197070094-197070094	c.8287G>C	p.E2763Q	Missense	Heterozygous
SK-N-DZ	ASPM	259266	Chr1:197070159-197070159	c.8222C>A	p.S2741Y	Missense	Heterozygous
EN	ASPM	259266	Chr1:197070185-197070186	c.8195_8196insA	p.N2734fs*17	Frame_Shift	Heterozygous
GP5d	ASPM	259266	Chr1:197070259-197070259	c.8122T>C	p.Y2708H	Missense	Heterozygous
OVISE	ASPM	259266	Chr1:197070312-197070312	c.8069G>A	p.R2690Q	Missense	Heterozygous
HCT 15	ASPM	259266	Chr1:197070313-197070313	c.8068C>T	p.R2690W	Missense	Heterozygous
TE-4	ASPM	259266	Chr1:197070331-197070331	c.8050G>A	p.D2684N	Missense	Heterozygous
HCC33	ASPM	259266	Chr1:197070372-197070372	c.8009T>C	p.I2670T	Missense	Heterozygous
NCI-H2347	ASPM	259266	Chr1:197070391-197070391	c.7990G>T	p.V2664L	Missense	Heterozygous
AN3-CA	ASPM	259266	Chr1:197070435-197070435	c.7946C>T	p.A2649V	Missense	Heterozygous
KYSE-150	ASPM	259266	Chr1:197070438-197070438	c.7943G>A	p.R2648K	Missense	Heterozygous
LNCaP clone FGC	ASPM	259266	Chr1:197070525-197070525	c.7856delA	p.K2619fs*22	Frame_Shift	Heterozygous
SNU-1040	ASPM	259266	Chr1:197070553-197070553	c.7828G>A	p.A2610T	Missense	Heterozygous
ASH-3	ASPM	259266	Chr1:197070604-197070604	c.7777A>G	p.K2593E	Missense	Heterozygous
SNU-81	ASPM	259266	Chr1:197070621-	c.7760A>G	p.Y2587C	Missense	Heterozygous

			197070621				
EFO-27	ASPM	259266	Chr1:197070706-197070706	c.7675G>A	p.V2559I	Missense	Heterozygous
P31/FUJ	ASPM	259266	Chr1:197070706-197070706	c.7675G>A	p.V2559I	Missense	Heterozygous
Karpas-45	ASPM	259266	Chr1:197070706-197070706	c.7675G>A	p.V2559I	Missense	Heterozygous
CRO-AP2	ASPM	259266	Chr1:197070721-197070721	c.7660C>T	p.H2554Y	Missense	Heterozygous
SAT [Human HNSCC]	ASPM	259266	Chr1:197070788-197070788	c.7593delG	p.W2531fs*14	Frame_Shift	Heterozygous
NCI-H2803	ASPM	259266	Chr1:197070812-197070812	c.7569A>G	p.Q2523Q	Silent	Heterozygous
MDA-MB-453	ASPM	259266	Chr1:197070828-197070828	c.7553G>A	p.R2518K	Missense	Heterozygous
HCC33	ASPM	259266	Chr1:197070884-197070884	c.7497A>G	p.T2499T	Silent	Heterozygous
KP-N-RT-BM-1	ASPM	259266	Chr1:197070906-197070906	c.7475G>A	p.R2492K	Missense	Heterozygous
KTCTL-195	ASPM	259266	Chr1:197070929-197070929	c.7452A>G	p.A2484A	Silent	Heterozygous
BxPC-3	ASPM	259266	Chr1:197071010-197071010	c.7371A>C	p.L2457F	Missense	Heterozygous
D-566MG	ASPM	259266	Chr1:197071048-197071049	c.7332_7333insGAGCCACC	p.I2445fs*16	Frame_Shift	Heterozygous
CL-34	ASPM	259266	Chr1:197071084-197071085	c.7296_7297insA	p.A2433fs*18	Frame_Shift	Heterozygous

Source	Cell line Name	GeneName	Entrez	Position	CDS Mutation	AA Mutation	Mutation Type	Zygosity
COSMIC	KCL-22	ASPM	259266	Chr1:197073232-197073232	c.5149delA	p.I1717fs*1	Frame_Shift	Heterozygous

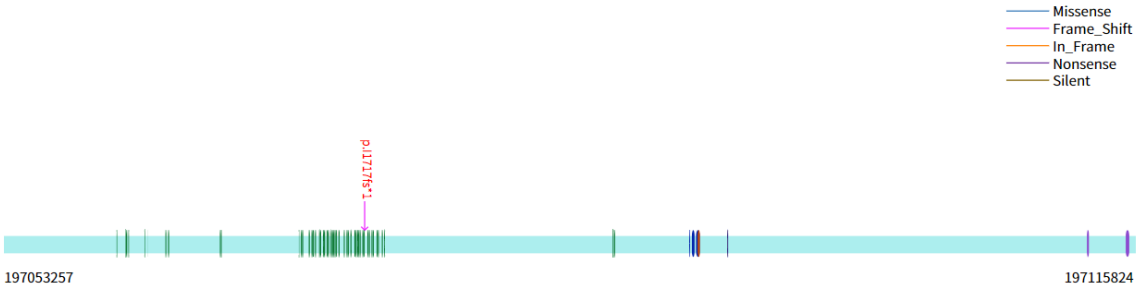


Figure-15: Representation of frameshift mutation at the specific position Chr1: 197073232-197073232

Table-08: Copy number of a specified gene with their cell lines and tissue association information

GeneName	Cell Line Name	Tissue	CCLE	COSMIC	NCI60
ASPM	HCC1195	LUNG	-0.04	-	-
ASPM	HCC2279	LUNG	-0.27	-	-
ASPM	HCC366	LUNG	-0.25	3	-
ASPM	NCI-H1650	LUNG	0.06	2	-
ASPM	NCI-H1666	LUNG	0.2	4	-
ASPM	NCI-H1781	LUNG	0.32	4	-
ASPM	NCI-H322	LUNG	0.13	-	-
ASPM	NCI-H358	LUNG	-0.07	3	-
ASPM	ChaGo-K-1	LUNG	-0.01	2	-
ASPM	COR-L23	LUNG	-0.14	2	-
ASPM	IA-LM	LUNG	-0.18	2	-
ASPM	LCLC-103H	LUNG	-0.37	3	-
ASPM	LCLC-97TM1	LUNG	-0.08	3	-
ASPM	LCLC-97TM1	LUNG	-0.08	3	-
ASPM	NCI-H1155	LUNG	-0.16	2	-
ASPM	NCI-H1299	LUNG	-0.31	3	-
ASPM	NCI-H1299	LUNG	-0.31	3	-
ASPM	NCI-H1581	LUNG	0.18	4	-
ASPM	NCI-H2106	LUNG	0.2	-	-
ASPM	NCI-H2126	LUNG	-0.02	3	-
ASPM	NCI-H2126	LUNG	-0.02	3	-
ASPM	NCI-H460	LUNG	0.28	4	0.29
ASPM	A-549	LUNG	0.01	3	-0.01
ASPM	ABC-1	LUNG	0.1	3	-

ASPM	Calu-3	LUNG	0.13	4	-
ASPM	COR-L105	LUNG	0.11	2	-
ASPM	DV-90	LUNG	-0.15	-	-
ASPM	HCC1171	LUNG	-0.19	-	-
ASPM	HCC2935	LUNG	0.05	-	-
ASPM	HCC4006	LUNG	0.14	-	-
ASPM	HCC44	LUNG	0.01	3	-
ASPM	HCC44	LUNG	0.01	3	-
ASPM	HCC78	LUNG	0.39	6	-
ASPM	HCC827	LUNG	-0.01	3	-
ASPM	Hs 229.T	LUNG	-0.11	-	-
ASPM	Hs 618.T	LUNG	-0.11	-	-
ASPM	LXF 289	LUNG	-0.2	2	-
ASPM	LXF 289	LUNG	-0.2	2	-
ASPM	MOR/CPR	LUNG	0.11	-	-
ASPM	NCI-H1355	LUNG	0.14	3	-
ASPM	NCI-H1373	LUNG	-0.02	-	-
ASPM	NCI-H1395	LUNG	0.11	4	-
ASPM	NCI-H1437	LUNG	-0.03	3	-
ASPM	NCI-H1563	LUNG	-0.09	4	-
ASPM	NCI-H1568	LUNG	-0.33	2	-
ASPM	NCI-H1573	LUNG	0.14	2	-
ASPM	NCI-H1573	LUNG	0.14	2	-
ASPM	NCI-H1623	LUNG	-0.02	3	-
ASPM	NCI-H1623	LUNG	-0.02	3	-
ASPM	NCI-H1648	LUNG	-0.26	2	-
ASPM	NCI-H1651	LUNG	-0.34	2	-
ASPM	NCI-H1693	LUNG	-0.31	2	-
ASPM	NCI-H1734	LUNG	-0.28	2	-
ASPM	NCI-H1755	LUNG	-0.01	4	-
ASPM	NCI-H1792	LUNG	-0.03	2	-
ASPM	NCI-H1793	LUNG	-0.11	3	-
ASPM	NCI-H1793	LUNG	-0.11	3	-
ASPM	NCI-H1838	LUNG	0.02	3	-
ASPM	NCI-H1838	LUNG	0.02	3	-
ASPM	NCI-H1944	LUNG	0.16	3	-
ASPM	NCI-H1975	LUNG	-0.06	2	-
ASPM	NCI-H2009	LUNG	-0.17	4	-
ASPM	NCI-H2009	LUNG	-0.17	4	-
ASPM	NCI-H2030	LUNG	-0.19	2	-
ASPM	NCI-H2087	LUNG	-0.01	3	-
ASPM	NCI-H2087	LUNG	-0.01	3	-
ASPM	NCI-H2122	LUNG	0.11	3	-
ASPM	NCI-H2228	LUNG	0.14	5	-
ASPM	NCI-H2291	LUNG	-0.04	3	-
ASPM	NCI-H23	LUNG	-0.24	3	-0.04

ASPM	NCI-H2342	LUNG	-0.19	3	-
ASPM	NCI-H2347	LUNG	0.16	5	-
ASPM	NCI-H2347	LUNG	0.16	5	-
ASPM	NCI-H2405	LUNG	-0.09	4	-
ASPM	NCI-H3255	LUNG	-0.01	3	-
ASPM	NCI-H3255	LUNG	-0.01	3	-
ASPM	Lu-65	LUNG	0.11	6	-
ASPM	Lu-99	LUNG	0.03	-	-
ASPM	EPLC-272H	LUNG	0.02	4	-
ASPM	NCI-H292	LUNG	0.17	3	-
ASPM	BEN	LUNG	-0.1	4	-
ASPM	CAL-12T	LUNG	-0.28	2	-
ASPM	Calu-6	LUNG	-0.26	2	-
ASPM	NCI-H1385	LUNG	0.59	-	-
ASPM	NCI-H1435	LUNG	0.06	2	-
ASPM	NCI-H1435	LUNG	0.06	2	-
ASPM	NCI-H1869	LUNG	0.26	6	-
ASPM	NCI-H1869	LUNG	0.26	6	-
ASPM	NCI-H1915	LUNG	0.05	3	-
ASPM	NCI-H2023	LUNG	0.1	3	-
ASPM	NCI-H2023	LUNG	0.1	3	-
ASPM	NCI-H2085	LUNG	0.11	4	-
ASPM	NCI-H2110	LUNG	0.08	3	-
ASPM	NCI-H2172	LUNG	-0.01	5	-
ASPM	NCI-H2172	LUNG	-0.01	5	-
ASPM	NCI-H2444	LUNG	-0.01	2	-
ASPM	LK2	LUNG	0.12	-	-
ASPM	NCI-H441	LUNG	0.06	3	-
ASPM	NCI-H441	LUNG	0.06	3	-
ASPM	COLO 668	LUNG	-0.07	3	-
ASPM	COLO 668	LUNG	-0.07	3	-
ASPM	COR-L24	LUNG	0.05	-	-
ASPM	COR-L279	LUNG	0.04	4	-
ASPM	COR-L311	LUNG	0.1	5	-
ASPM	COR-L47	LUNG	-0.07	-	-
ASPM	COR-L88	LUNG	-0.18	3	-
ASPM	COR-L95	LUNG	-0.1	3	-
ASPM	CPC-N	LUNG	-0.08	2	-
ASPM	DMS 114	LUNG	-0.07	3	-
ASPM	DMS 153	LUNG	0.47	-	-
ASPM	DMS 273	LUNG	-0.02	3	-
ASPM	DMS 454	LUNG	-0.03	-	-
ASPM	DMS 53	LUNG	-0.04	3	-
ASPM	DMS 79	LUNG	0.07	4	-
ASPM	HCC33	LUNG	0.04	3	-
ASPM	HCC33	LUNG	0.04	3	-

ASPM	NCI-H1048	LUNG	0.31	4	-
ASPM	NCI-H1092	LUNG	0.3	4	-
ASPM	NCI-H1105	LUNG	0.17	4	-
ASPM	NCI-H1184	LUNG	0.06	-	-
ASPM	NCI-H1339	LUNG	-0.12	-	-
ASPM	NCI-H1341	LUNG	0.08	5	-
ASPM	NCI-H1436	LUNG	0.33	3	-
ASPM	NCI-H146	LUNG	-0.11	3	-
ASPM	NCI-H1618	LUNG	0.09	-	-
ASPM	NCI-H1694	LUNG	0	3	-
ASPM	NCI-H1836	LUNG	-0.14	3	-
ASPM	NCI-H1876	LUNG	0.09	2	-
ASPM	NCI-H1930	LUNG	0.13	-	-
ASPM	NCI-H196	LUNG	0.31	4	-
ASPM	NCI-H1963	LUNG	0.22	3	-
ASPM	NCI-H2029	LUNG	-0.3	2	-
ASPM	NCI-H2081	LUNG	0.39	5	-
ASPM	NCI-H209	LUNG	-0.05	5	-
ASPM	NCI-H211	LUNG	0.25	3	-
ASPM	NCI-H2141	LUNG	0.15	4	-
ASPM	NCI-H2171	LUNG	-0.04	2	-
ASPM	NCI-H2196	LUNG	0.06	2	-
ASPM	NCI-H2227	LUNG	0.15	3	-
ASPM	NCI-H2227	LUNG	0.15	3	-
ASPM	NCI-H446	LUNG	0.2	4	-
ASPM	Calu-1	LUNG	-0.03	-	-
ASPM	HARA [Human squamous cell lung carcinoma]	LUNG	-0.32	2	-
ASPM	HCC15	LUNG	-0.07	2	-
ASPM	HCC95	LUNG	0.61	-	-
ASPM	HLF-a	LUNG	-0.17	-	-
ASPM	KNS-62	LUNG	0.02	3	-
ASPM	LC-1/sq-SF	LUNG	0.28	-	-
ASPM	LC-1F	LUNG	0.1	-	-
ASPM	LOU-NH91	LUNG	-0.12	2	-
ASPM	LUDLU-1	LUNG	0.19	-	-
ASPM	NCI-H1703	LUNG	-0.01	3	-
ASPM	NCI-H1703	LUNG	-0.01	3	-
ASPM	NCI-H2066	LUNG	0.14	3	-
ASPM	NCI-H2066	LUNG	0.14	3	-
ASPM	NCI-H2170	LUNG	0.19	8	-
ASPM	NCI-H2170	LUNG	0.19	8	-
ASPM	NCI-H226	LUNG	-0.01	2	-0.04
ASPM	NCI-H2286	LUNG	0.05	-	-

ASPM	NB-EbC1	LUNG	-0.35	-	-

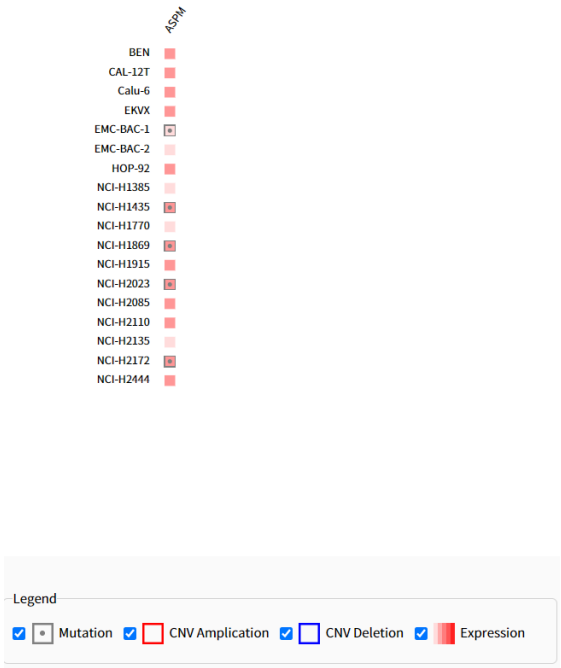


Figure-16: Overview of the Mutation, CNV Amplification, CNV Deletion and Expression

Above information reveals the frequent alterations of ASPM by frameshift mutations. Our study suggests how ASPM is involved in cell proliferation along with the growth of the tumor. This proliferation is the result of a number of mechanism that takes place. It was found slightly benign and is thus, associated with poor prognosis. It could be a novel therapeutic target for the treatment of lung cancer. The cell lines mentioned above can be further analyzed for the drug testing.

CONCLUSION

In this study, we performed computational analysis to investigate DNA repair mechanism and mutations involved in lung cancer. Genomic data analysis for non-small cell lung cancer was performed. A total of 11 putative genes were predicted based upon various kinds of analyses. Codon usage analysis was performed to identify the number of times synonymous codons that occur in the DNA. Frameshift mutations were also studied in these genes to look for the possible regulatory role of these mutations. Analysis of frameshifted translations can help to stop massive growth of protein products along with the control of diseases such as lung cancer. Network analysis was done to study novel pathogenesis pathways that can be helpful in the determination and identification of potential therapeutic targets. Gene ontology was also performed to gain the deeper insight about the molecular, biological and cellular processes of genes. Furthermore, SNP analysis was done for the prediction of disease associated variations. We predicted genes such as TOP2A, CDC20, FYN, BUB1 and JUN have a damaging effect on lung cancer tissues and were disease associated. On the other hand, genes such as ASPM and SFN showed neutral association and were slightly benign. Cell lines are a quintessential step towards drug development process and hence, we found in our study that the expression level of ASPM is significantly increased. ASPM co-expression with CDK4 was also observed which clearly explains its role as therapeutic target. This will be helpful for improved treatment efficacies and for avoiding unnecessary toxicities. It will help experimental biologists to look for possible direct targets conditioned to their experimental validations.

REFERENCES

[1] A. J. Alberg and J. M. Samet, "Epidemiology of lung cancer," *Chest*, vol. 123, no. 1, pp. 21S–49S, 2003.

[2] G. C. Kabat and E. L. Wynder, "Lung cancer in nonsmokers," *Cancer*, vol. 53, no. 5, pp. 1214–1221, 1984.

[3] S. G. Spiro and G. A. Silvestri, "One hundred years of lung cancer," *American journal of respiratory and critical care medicine*, vol. 172, no. 5, pp. 523–529, 2005.

[4] W. D. Travis, "Pathology of lung cancer," *Clinics in chest medicine*, vol. 32, no. 4, pp. 669–692, 2011.

[5] S. Behl, A. Sharma, P. Suravajhala, and T. R. Singh, "Computational studies to explore the role of MSI associated DNA mismatch repair mechanisms in HNPCC through expression and interaction data," *International Journal of Bioinformatics Research and Applications*, vol. 16, no. 4, pp. 408–416, 2020.

[6] A. Shukla, M. Sehgal, and T. R. Singh, "Hydroxymethylation and its potential implication in DNA repair system: a review and future perspectives," *Gene*, vol. 564, no. 2, pp. 109–118, 2015.

[7] A. Shukla, A. Moussa, and T. R. Singh, "DREMECELS: a curated database for base excision and mismatch repair mechanisms associated human malignancies," *PloS one*, vol. 11, no. 6, p. e0157031, 2016.

[8] H. Davies *et al.*, "Somatic mutations of the protein kinase gene family in human lung cancer," *Cancer research*, vol. 65, no. 17, pp. 7591–7595, 2005.

[9] A. Gemma *et al.*, "Somatic mutation of the hBUB1 mitotic checkpoint gene in primary lung cancer," *Genes, Chromosomes and Cancer*, vol. 29, no. 3, pp. 213–218, 2000.

[10] E. R. Velazquez *et al.*, "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer research*, vol. 77, no. 14, pp. 3922–3930, 2017.

[11] "Somatic Mutations Lead to an Oncogenic Deletion of Met in Lung Cancer | Cancer Research." https://cancerres.aacrjournals.org/content/66/1/283.short (accessed May 16, 2021).

[12] "Somatic mutations of epidermal growth factor receptor signaling pathway in lung

cancers - Shigematsu - 2006 - International Journal of Cancer - Wiley Online Library." https://onlinelibrary.wiley.com/doi/full/10.1002/ijc.21496 (accessed May 16, 2021).

[13] "Associations Between Somatic Mutations and Metabolic Imaging Phenotypes in Non–Small Cell Lung Cancer | Journal of Nuclear Medicine."

[14] C. J. Etzel, C. I. Amos, and M. R. Spitz, "Risk for smoking-related cancer among relatives of lung cancer patients," *Cancer research*, vol. 63, no. 23, pp. 8531–8535, 2003.

[15] D. Carbone, "Smoking and cancer," *The American journal of medicine*, vol. 93, no. 1, pp. S13–S17, 1992.

[16] R. M. Hoffman and R. Sanchez, "Lung cancer screening," *Medical Clinics*, vol. 101, no. 4, pp. 769–785, 2017.

[17] P. Nanavaty, M. S. Alvarez, and W. M. Alberts, "Lung cancer screening: advantages, controversies, and applications," *Cancer control*, vol. 21, no. 1, pp. 9–14, 2014.

[18] W. Lee *et al.*, "The mutation spectrum revealed by paired genome sequences from a lung cancer patient," *Nature*, vol. 465, no. 7297, pp. 473–477, 2010.

[19] A. Roth, M. Anisimova, and G. M. Cannarozzi, "Measuring codon usage bias," *Codon evolution: mechanisms and models*, pp. 189–217, 2012.

[20] J. R. Roth, "Frameshift mutations," *Annual review of genetics*, vol. 8, no. 1, pp. 319–346, 1974.

[21] L. J. Chin *et al.*, "A SNP in a let-7 microRNA complementary site in the KRAS 3′ untranslated region increases non–small cell lung cancer risk," *Cancer research*, vol. 68, no. 20, pp. 8535–8540, 2008.

[22] C. I. Amos *et al.*, "Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25. 1," *Nature genetics*, vol. 40, no. 5, pp. 616–622, 2008.

[23] U. Vetrivel, V. Arunkumar, and S. Dorairaj, "ACUA: a software tool for automated codon usage analysis," *Bioinformation*, vol. 2, no. 2, p. 62, 2007.

[24] T. R. Singh and K. R. Pardasani, "Ambush hypothesis revisited: Evidences for phylogenetic trends," *Computational biology and chemistry*, vol. 33, no. 3, pp. 239–244, 2009.

[25] J. Xia, E. E. Gill, and R. E. Hancock, "NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data," *Nature protocols*, vol. 10, no. 6, pp. 823–844, 2015.

[26] "g:Profiler—a web server for functional interpretation of gene lists (2016 update)| NucleicAcidsResearch|OxfordAcademic."

[27] “new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog) | Nucleic Acids Research | Oxford Academic.”

[28] “SNPs3D: Candidate gene and SNP selection for association studies | BMC Bioinformatics | Full Text.”

[29] “PANTHER: A Library of Protein Families and Subfamilies Indexed by Function.”

[30] “Functional annotations improve the predictive score of human disease-related mutations in proteins - Calabrese - 2009 - Human Mutation - Wiley Online Library.”

[31] “GEMiCCL: mining genotype and expression data of cancer cell lines with elaborate visualization | Database | Oxford Academic.”