Prediction of causative variants in cancer using NGS and GWAS

by

Sohaib Aslam

A dissertation submitted in partial fulfillment of the requirements

for the degree of Master of Science in Computational Science and Engineering

Supervised by

Dr. Shumaila Sayyab

Research Center for Modeling and Simulation

National University of Sciences and Technology

Islamabad, Pakistan

© Sohaib Aslam, 2016

Declaration

I declare that this thesis comprises of my own research work, all the parts of this thesis are original; nothing has been plagiarized. Any prior work is duly referenced and contributions made by other people are acknowledged.

Sohaib Aslam
NUST201361530MRCMS64213F

Abstract

The Cancer Genome Atlas (TCGA) data is used for analysis in this study for prediction of causative variants in cancer. We have used Next-Generation Sequencing (NGS) data containing copy number variants and Genome Wide Association Studies (GWAS) data containing single nucleotide variants in three different types of lung related cancers (Lung adenocarcinoma, Lung squamous cell carcinoma and Mesothelioma) using the customised pipelines. Annotation (gene based, transcription factor binding sites, conserved elements, microRNAs and snoRNAs), functional enrichment analysis and protein-protein interaction have been covered in this study. Variants lying in highly conserved regions or overlapping the highly conserved regions are identified. Our results show that three types of microRNAs (hsa-mir-3149, hsa-mir-933 and hsa-mir-4307) were common in all three types of Lung cancers. Genes containing zinc finger domain were identified in all three type of lung cancers. Transcription factor binding sites (nks39, cdc5 and foxo3) were common in all three types of cancers, suggesting their regulatory function.

Acknowledgement

All praises and thanks for Almighty Allah, the Creator, most Merciful, whose blessings enabled me to accomplish this work. My humblest gratitude to His Prophet Hazrat Muhammad (Peace be upon him) whose way of life is a continuous guidance for me.

My appreciation goes to my parents without their kind support and help it was impossible for me to pursue my master degree.

I also wish to express my appreciation to the principal of RCMS, National University of Sciences and Technology, Islamabad, Dr. Ahmed Ejaz Nadeem for his valuable support and efforts for the students of his department.

My profound gratitude is for my supervisor Dr. Shumaila Sayyab, Assistant Professor at RCMS for her kind support, encouragement and guidance.

I pay my special gratitude to GEC members: Dr. Zartasha Mustansar, Dr. Rehan Zafar Paracha and Dr. Zamir Hussain for enthusiastic guidance through out my thesis.

I pay thanks to all the faculty members at RCMS, especially Assistant Professor Mr. Tariq Saeed, Dr. Fouzia Malik and Dr. Salma Sherbaz for their kind motivation and moral support.

My dear friends and well wishers Hammad Ali Hassan and Usman Yousaf, thanks a lot.

My thanks also go to all of my friends. Shahid Mahmood, Kashif Zaheer, Tauseef Mushtaq, Imraan Nawaz, Ali bhai, Mureed bhai, Yasir Butt, Saad, Saqib ch, Mustafa Kamal Pasha, Muzammil and MA Khan for making my time memorable.

I pay special thanks to my uncle Mr. Muhammad Arshid for providing me moral and financial support during my studies.

Sohaib Aslam

To My Beloved Parents

Contents

List of Figures

List of Tables

Appendix

CCLE	Cancer Cell Line Encyclopedia
CGP	Cancer Genome Project
DNA	Deoxyribonucleic acid
EGS	External Guide Sequence
EDRN	Early Detection Research Network
FDR	False Discovery Rate
GWAS	Genome-Wide Association Studies
GB	Gigabyte
ICGC	International Cancer Genome Consortium
KEGG	Kyoto Encyclopedia of Genes and Genomes
LUAD	Lung Adenocarcinoma
LUSC	Lung Squamous Cell Carcinoma
MESO	Mesothelioma
miRNA	micro RNA
NGS	Next-Generation Sequencing
OMIM	Online Mendelian Inheritance in Man
PANTHER	Protein ANalysis THrough Evolutionary Relationships
RBM	Reducing Body Myopathy
PERL	Practical Extraction and Report Language
PPI	Protein-Protein Interaction
RNA-seq	RNA sequencing
SNPs	Single Nucleotide Polymorphisms
SNVs	Single Nucleotide Variations
TCGA	The Cancer Genome Atlas
UCSC	University of California, Santa Cruz
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Chapter 1

Introduction

1.1 Overview

Every organism retains a genome which carries genetic information for constructing and maintaining living example of organism. Mostly, the genomes of humans and other cellular organisms is made up of DNA (deoxyribonucleic acid) except in case of few viruses which contain RNA (ribonucleic acid) [1]. In a cell chromosomes represents genome. Chromosomes are thread like structures composed of nucleic acid and protein. Two strands of DNA lie in each chromosome that are wrapped in the form of double helix (Figure 1.1). DNA is composed of four nucleotides Adenine, Thymine, Guanine and Cytosine. Nucleotides in one strand are complement to other strand, as a rule, Adenine is complement to Thymine and Guanine is complement to Cytosine, such complementary behaviour of nucleotides results in form of base pairs [2]. Proteins play important role in all biological processes. Chain of amino acids form primary structure of proteins. Proteins are functionally diverse in nature due to different combinations of amino acids. [1]. Change in DNA or amino acid sequence may bring harmful effects to an organism.

Cancer is a disease, caused by changes in certain genes that control cell function, especially how cell grow and divide. These changes include mutations in DNA (Gene coding DNA) [4]. Cancer is among leading causes of mortality throughout the world. The most common organs affected by cancer are lung, prostate and colorectum, there are 1.59 million deaths due to lung cancer, 0.745 million deaths due to liver cancer and 0.723 million deaths due to colorectum cancer [5].

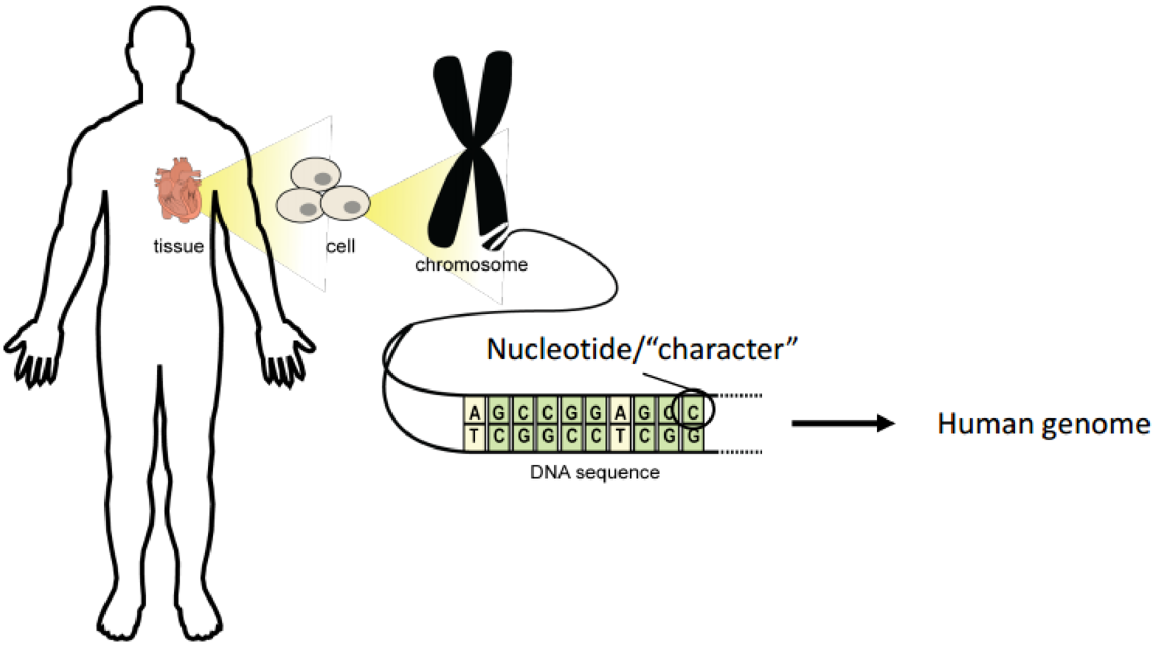


Figure 1.1: General overview of human genome

1.2 Genetic variations

Genetic variation describes naturally occurring genetic differences among individuals of the same species. Variation is essential for survival of population in environmental changes [6]. Genetic variation also plays important role in human diseases and rare Mendelian disorders [7]. Mutation results from DNA that does not corrected by DNA repair mechanism. Mutation can contribute towards normal or diseased phenotype. Inevitable mutations may have adverse effects on human [8].

Two types of mutation may occur i.e. somatic and germline. Somatic mutations are simultaneously occurring mutation that do not pass to next generation like cancer mutations while germline mutations are passed from parents to off springs Figure (1.2).

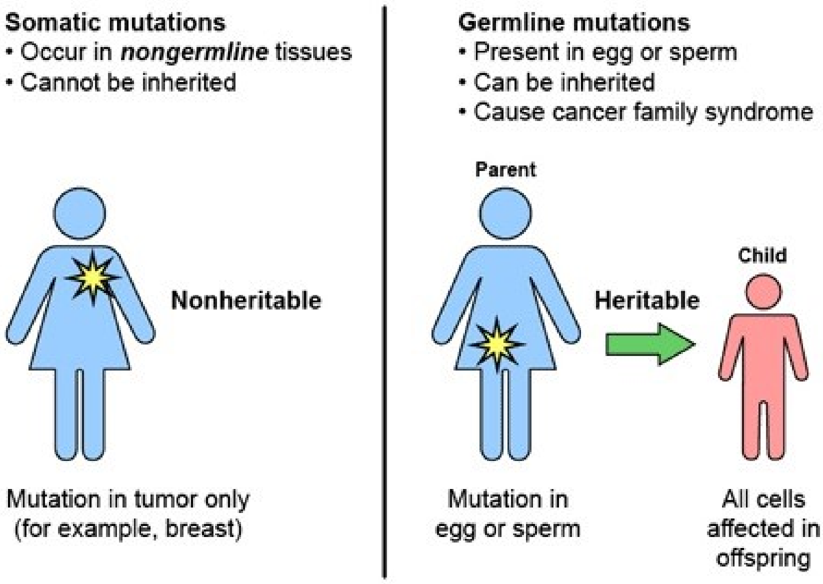

Somatic mutations
• Occur in *nongermline* tissues
• Cannot be inherited
Nonheritable
Mutation in tumor only (for example, breast)
Germline mutations
• Present in egg or sperm
• Can be inherited
• Cause cancer family syndrome
Parent
Heritable
Child
Mutation in egg or sperm
All cells affected in offspring

Figure 1.2: Comparison of somatic and germline mutations [9]

Some times DNA variant is different from others due to absence or presence of a nucleotide or small number of nucleotides at specific position that is termed as "insertion or deletion". Large insertions or deletions results from change in copy number of sequences, insertion or deletion of more than 100 nucleotides is termed as "copy number variation" [10].

Most common type of genetic variation in humans is single nucleotide substitution [8]. For example, variant G replaces with C at specific point. This is called SNV. If frequency of two or more alternative variants in a population exceeded from 0.01 then it is called SNP [8]. In this thesis, work on SNVs and CNVs is carried out.

1.2.1 Somatic mutations

Most of somatic mutations in our body do not contribute to any harmful activity, role of somatic mutations in cancer was supported by the discovery of mutagenic activity performed by carcinogenic chemicals [11].

1.2.2 Single nucleotide variants

SNVs (Single nucleotide variants) are most common form of intra-species variations. In humans number of SNVs range from 3-5 million. Common SNVs are main focus of medical and population genetics research [12].

1.2.3 Single nucleotide polymorphisms

SNP (pronounced as "snip") is difference of a nucleotides at the same position among individuals. If a population doest not carry same nucleotide more than 1%, at a specific position, then such variation is termed as SNP. SNP may involve in variation of amino acid sequence but it is not compulsory because it may occur in non-coding regions of DNA [13]. While comparing two DNA sequences there is a SNP after every 1000-2000 nucleotides. Among 3.2 billion nucleotides in human genome there are about 1.6 million to 3.2 million SNPs [14].

1.2.4 Copy number variations

CNVs (copy number variations) play important role in genetic susceptibility to many diseases [15]. Discovery of CNVs has dramatically changed view on structural variation of DNA and disease. It is thought that CNVs encompass more nucleotides and arise more frequently than SNPs [16]. Deletions, insertion duplications and translocations results in CNVs (Figure 1.3). CNVs formation is faster than any other type of mutation [18]. In humans, CNVs are important source of DNA polymorphism [19]. Contribution of CNVs to human genome is around 13% [16].

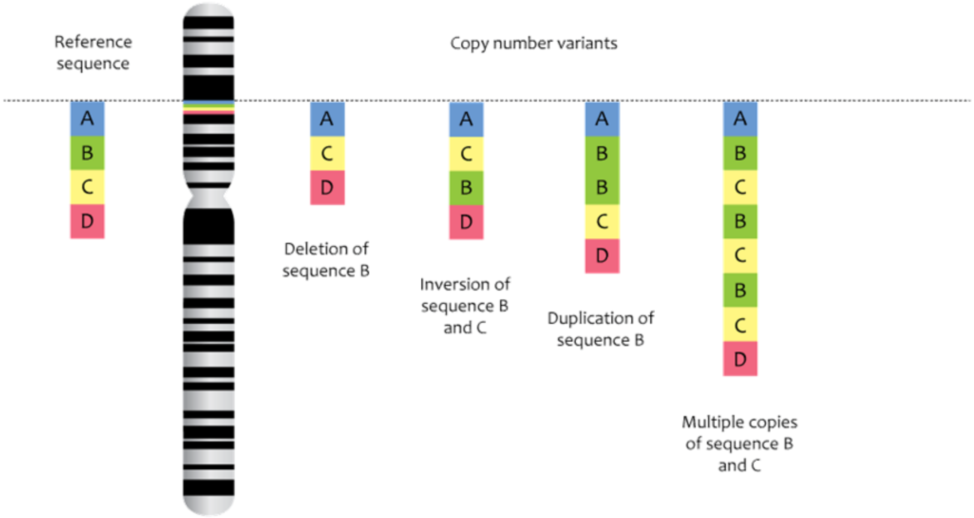

Reference sequence
Copy number variants
A
B
C
D
A
C
D
Deletion of sequence B
A
C
B
D
Inversion of sequence B and C
A
B
B
C
D
Duplication of sequence B
A
B
C
B
C
B
C
D
Multiple copies of sequence B and C

Figure 1.3: Description of Copy number variants [17]

1.3 Genome-wide association studies

Completion of human genome project, discovery of millions of SNPs in human genome, LD pattern characterization by Hap Map project and availability of high-throughput genotyping platforms at low cost, made possibility of genome-wide association studies. GWAS assay hundreds of thousands of SNPs by using high-throughput genotyping technologies and relate them with clinical conditions [20]. GWAS have been fruitful in studying complex diseases by identifying thousands of common variants associated with complex diseases [22].

GWAS have proved an important tool for discovering the regions of genome carrying genetic variant, confer risk for different types of cancers [23].

1.4 Next- generation sequencing

Need of Sequencing technologies has increased with the passage of time, nowadays it is use in number of research fields like genomics, evolution, applied medicine and forensic sciences. Old sequencing technologies like Sanger sequencing is expensive and time consuming. Several emerging technologies are promising for fast and cost effective genome sequencing [24].

For past 30 years, Sanger sequencing has been gold standard for DNA sequencing. It has achieved major accomplishments like completion of human genome project. Pyrosequencing was first commercially introduced in 2005 that laid foundation of new era, of high-throughput genomic analysis referred to NGS (Next Generation Sequencing) [25].

While discussing applications of NGS, there are lot of information related to cancer genomic alterations complexity are available like small insertion or deletion, point mutation, copy number alterations and structural variations. By comparing altered to matched normal samples (tumor and normal samples), researchers are able to distinguish between two categories of variants: germline and somatic. Meanwhile whole transcriptome approach (RNA-Seq) can help in detection of RNA editing, alternative splicing and gene expression profiles. Combination of these NGS technologies provides global view and high resolution of cancer genome. While using Bioinformatics researchers are able to understand biology of cancer that is leading to development of personalised treatment strategies [26].

1.5 Cancer

Cancer is genomic disease and arises from set of somatic alterations [27]. In cancer, body cells start dividing without any stop and spread into nearby tissues. Cancer can start in any part of the body. In normal body, old-cells die and new cells take place of them however in cancer, disturbance in such a mechanism is observed (figure 1.4) [28]. After cardiovascular diseases, cancer is the second leading cause of death [29]. Now a days millions of people extend their life by early diagnosis and treatment of cancer [30].

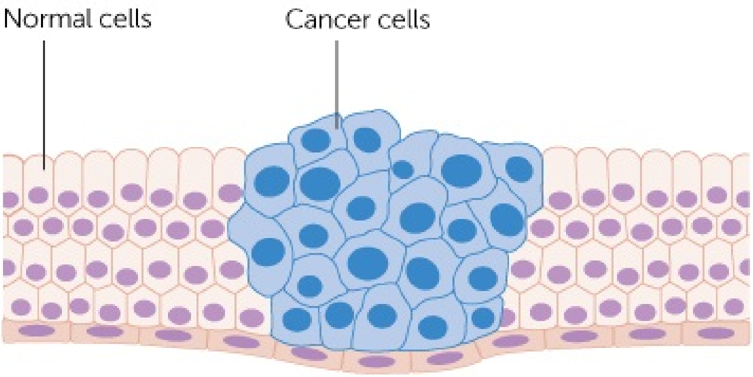


Figure 1.4: Difference between normal and cancer cells [31]

1.6 Cancer data collection

Publicly available resources has improved advancement in scientific discovery. Big data and genomics brought the concept of data sharing and collaboration to make research effective. Few on-line resources are mentioned below [65].

- **ICGC:** Established in 2008, it has produced terabytes of data from 12,232 donors and 50 cancer types. Somatic variant data is publicly accessible in ICGC (*icgc.org*).
- **CGP:** Collects genomic data from 50 different types of cancers. Major focus of CGP is discovery of frequently mutated genes in tumours and identify pattern of mutations in cancer cells (*www.sanger.ac.uk*).
- **TCGA:** Established in 2006, examining spectrum of genomic changes in more than 20 types of cancers. Main goal of TCGA is to improve our scientific ability to diagnose, treat and prevent cancer (*cancergenome.nih.gov*).

- **CCLE** Provides public access to about 1000 cell lines data includes genomic data, analysis and visualization. It focus on large human cancer models for genetic and pharmacologic characterization (*www.broadinstitute.org/ccle*).

1.7 Lung cancer

There are more than 200 forms of cancers. In TCGA data of more than 20 different types of cancers is available. Lung cancer is one of the leading cause of deaths in the world. Each year about 1.2 million cases of lung cancer are diagnosed. Cigarette smoking is major cause of lung cancer [33]. Exposure to harmful gases, air pollution, exposure to certain chemicals and lowered immunity can cause lung cancer [21]. Types of lung cancers discussed in this study are.

1.7.1 Lung adenocarcinoma

Lung adenocarcinoma is one of the most common form of lung cancer and have only 15% survival rate per 5 year. Lung adenocarcinoma in non-smoking patients contain mutations within tyrosine kinase domain of Epidermal Growth Factor Receptor (EFGR) gene [33].

1.7.2 Lung squamous cell carcinoma

Lung squamous cell carcinoma is form of lung cancer that is common in world with about 0.4 million deaths annually worldwide. No genomic alterations in Lung squamous cell carcinoma have been comprehensively characterized. No targeted molecular agents have been specified for treatment [34].

1.7.3 Mesothelioma

Malignant mesothelioma is asbestos-related rare cancer, forms on thin protective tissues that cover the abdomen and lungs [56]. Mesothelioma is aggressive cancer. After diagnosis death occurs within one year [35].

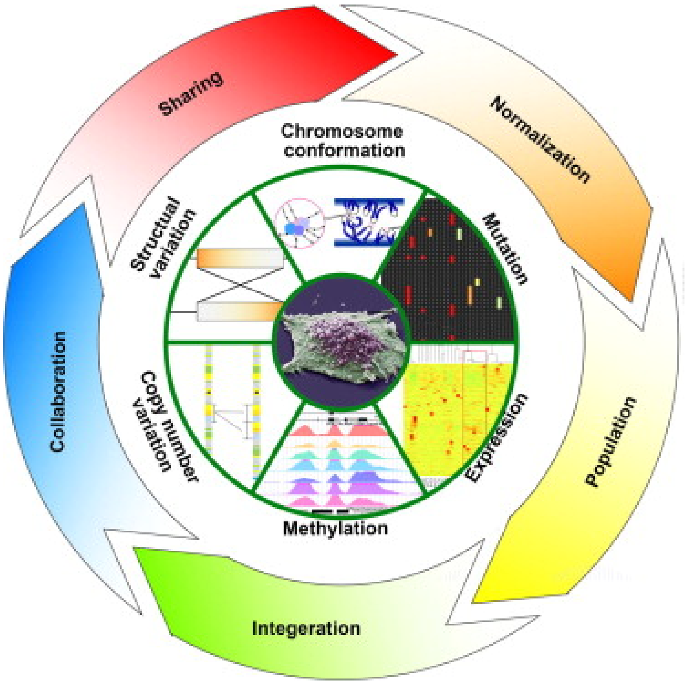


Figure 1.5: Future of cancer research [36]

1.8 Programming languages and tools

1.8.1 Perl

Perl provides excellent support for common application oriented based tasks [63]. Perl is used by the people, who want to analyse or convert lots of data quickly [38]. Few attributes of Perl that make it attractive [39]

- For matching and manipulating the strings, Perl provides powerful ways through usage of regular expressions.
- Writing programs as libraries is easy by modularity in Perl.
- System calls and pipes of Perl can be used for incorporating external programs.

- Perl is easy to code and a good prototyping language. Testing of new algorithms can be easily done in Perl before use of any rigorous language.
- Perl is excellent platform for writing CGI scripts for interfacing with Web.
- Perl provides excellent support for object oriented programming development.

1.8.2 R programming

R is a scripting language for manipulation and analysis of Statistical data. Some of R virtues are:

- It is available for Linux, Windows and Mac operating systems.
- It incorporate the features that are available in OOP and functional programming languages.
- It is superior to number of commercially available products, different operations available like graphics, programmability and, so on.
- R is open source therefore it is easy to take help from programmers community.
- System saves history of previous commands, different sessions can also be saved.

1.8.3 Annovar

Annovar is used for functionally annotating genetic variants belonging to the diverse genomes. Annovar is used for annotating SNVs, deletions/insertions, cytogenetic bands inferring, functional importance scores reporting, conserved region variant finding, identifying reported variants in 1000 Genome Project [40]. Annovar can be used for.

- Gene-based annotation

- Region-based annotation
- Filter-based annotation

1.8.4 DAVID

DAVID is a online web-based bioinformatics resource (https://david.ncifcrf.gov/), consists of analytic tools and integrated biological knowledge-base, aimed at extracting biological meaning from large gene or protein lists. DAVID can handle any type of gene list independent of any platform or software [41] [42]. Functional annotation suite web-based provides:

1. DAVID Gene Functional Classification Tool
2. DAVID Functional Annotation Tool
3. DAVID Gene ID Conversion Tool
4. DAVID Gene Name Viewer
5. DAVID NIAID Pathogen Genome Browser.

1.9 Problem statement

To understand the genetics of lung cancer, there is need to analyze and identify the genetic variants that are involved in causing lung cancer by making use of already available NGS and GWAS data.

1.10 Objectives

The objectives of the current study are as follows:

- Analysis of NGS and GWAS data

- Functional Annotation of variants (SNVs and CNVs)
- Analysis using conservation, Transcription factor binding site and micro RNA
- Pathway analysis of genes containing significant candidates
- Post GWAS Analysis using databases.

Chapter 2

Literature review

2.1 Overview

Next-generation sequencing (NGS) technologies are creating petabytes of raw and scattered data that is decentralized in databases, archives and sometimes in isolated hard-disks that is not available for browsing and analysis. Curated secondary databases may help to organize some of Big data by letting users to navigate, search and compute better on it.

National and international collaborations, such as ICGC, GCGC, EDRN and TCGA are generating large amounts of data, majority of data from NGS technologies. TCGA data will exceed to 100 petabytes at the end of the project [46]. According to a recent survey, files generated by TCGA are doubling every seven months since 2010, having total count of more than 700,000 files [46]. It is desireable for cancer biologists to use this data for developing and testing hypothesis, few wet-laboratory researchers have knowledge about the scores of bioinformatics tools that are complex in nature having higher understandings. These types of challenges are leading for development of tools and secondary databases which are expected for Big Data use [47] [48].

According to Tsung-Jung Wu et al. 2014, genomic data is large, heterogeneous, varied and widely distributed. Extracting and converting this data into useful information and comparing results is becoming hurdle for personalized genomics. Due to complexity and size of NGS data there is need of methods to store, analyse and curate genomics data.

One major purpose of NGS is to identify human genetic variants, which can be used for improving understanding of human diseases [66]. Computational approaches are available for prediction of variants that are deleterious and potentially associated with disease [50].

2.2 LUSC molecular profiling

TCGA research group performed a comprehensive molecular profiling of Lung squamous cell carcinoma. They have studied profile of 178 LUSC and find mutation in 11 genes, interestingly tumor protein p53 mutation was found nearly in all specimens. This study has provided potential therapeutic targets, offering new opportunities for treatment of LUSC [51].

2.3 LUAD molecular profiling

TCGA research group reported molecular profiling of 230 Lung adeno carcinoma using mRNA, DNA methylation, copy number alterations and protein expression. They observed high rates of somatic mutations, eight genes were statistically significant mutated. EGFR mutations in female patients were more frequent, whereas in males RNA Binding Motif Protein 10 mutations were very common. Three benefits are achieved, first new knowledge is gained about genomic alteration by performing these studies, second previously unappreciated altered genes are highlighted and third personalized treatment for deadly disease is improved [52].

2.4 Proteomic perspectives of pan-cancer

Rehan Akbani et al. 2012, compared protein data from 11 diseases and 3467 samples and integrated resultant proteomic data with transcriptomic and genomic analyses of the same samples to identify differences, similarities, network biology and emergent pathways within and across tumor lineages. In addition, signals specific to tissue are computationally reduced to enhance biomarker and target discovery over multiple tumor lineages. This analysis, with emphasis on potentially key proteins and pathways, provides a basic structure for determining the relevance of functional proteome to prognostic, predictive and therapeutics [53].

2.5 Glioma susceptibility loci

For identifying risk variants causing glioma, Sanjay Shete et al. 2009, conducted a meta analysis of two GWAS by genotyping 550k tagging SNPs with 3670 controls and 1878 cases, with control of three additional independent series having 2953 controls and 2545 cases. They have identified 5 risk loci 5p15.33, 8q24.21, 9p21.3, 20q13.33 and 11q23.3. They used Haploview for inferring LD structure of genome in regions glioma risk associated loci. For examining relationship between mRNA expression and SNP genotype in normal human cortex and lymphocytes publicly available data set is used [54].

2.6 Somatic mutation in Glioblastoma

Cameron W. Brennan et al. 2013, studied genomic alteration based on comprehensive and multidimensional characterization of more than 500 GBMs tumors. Dataset consists of molecular and clinical data of about 543 patients. They have identified novel mutated genes and signature receptors rearrangements from data set of patients, including the PDGFRA and EGFR. They have observed mutations of TERT promoter are correlating with increased mRNA expression that plays an important role in reactivation of the telomerase. This type of data will provide facility to discover the target candidates(therapeutic and diagnostic) [55].

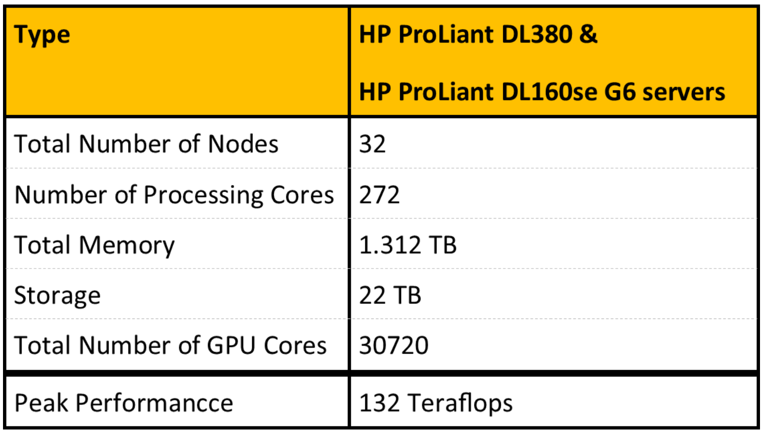
Chapter 3

Methodology

3.1 Experimental conditions

3.1.1 Hardware

This work was done by the use of resources provided by Research Center for Modeling and Simulation (RCMS). Most of the analysis were performed using Super Computing Research And Education Centre (ScREC) [1] allocated desktop computer. ScREC super computer specifications are described in Table (3.1).



Type	HP ProLiant DL380 & HP ProLiant DL160se G6 servers
Total Number of Nodes	32
Number of Processing Cores	272
Total Memory	1.312 TB
Storage	22 TB
Total Number of GPU Cores	30720
Peak Performancce	132 Teraflops

Table 3.1: ScREC super computer specifications

[1] ScREC is state of the art facility established by RCMS

3.1.2 Software

Annovar was provided by Kai Wang after providing institutional affiliation details on annovar web portal. Strawberry perl (version 5.24.0.1) was downloaded from strawberry perl website for running Annovar. R programming language (version 3.2.5) for data processing was downloaded from cran.r-project website.

3.2 Data collection

Different cancer databases were searched for collection of cancer data with the help of literature. During data mining step, TCGA was found more comprehensive and collective database rather than other cancer repositories. TCGA researchers have analysed more than 30 types of human cancers by genome sequencing and multi-dimensional analyses.

3.2.1 Data levels

Data level is a method of TCGA data characterization for researchers, to facilitate, locate and download data of interest. There are four levels of data (figure 3.1).

1. **Raw data:** It is raw data of single sample that is not normalized.

2. **Processed data:** It is normalized data of single sample.

3. **Segmented/Interpreted:** It is processed data aggregation of single sample.

4. **Summary/ROI:** It is quantified association across samples.

3.2.2 Data types

On TCGA, data matching to normal and tumor is available. Data types include

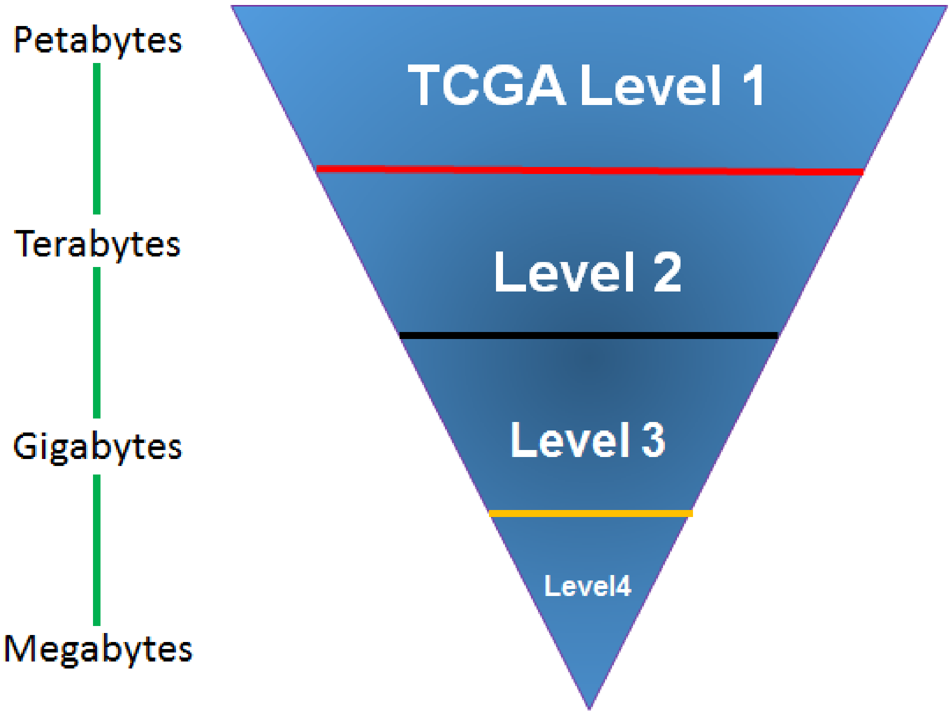


Figure 3.1: TCGA data level and its units of storage

- Exome(variant analysis)
- SNP
- mRNA
- Methylation
- miRNA
- Patient clinical information

3.2.3 Data access

Two types of data access is available

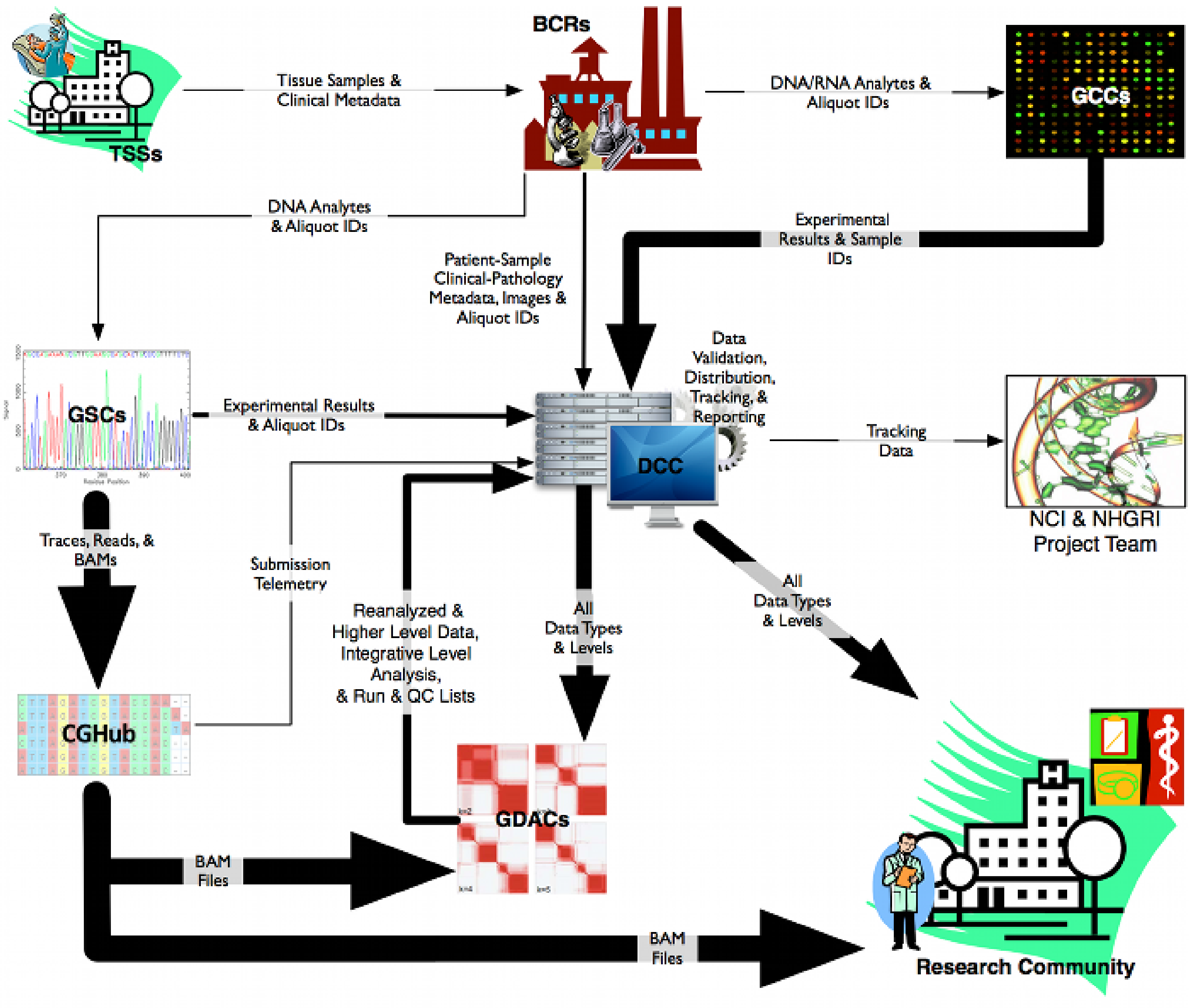


Figure 3.2: Illustration of TCGA data flow [57]

- **Controlled access-** data is not available publicly. Certification is required to be reviewed by data access committee of TCGA. Data is not unique to explore identity of a patient.

- **Open access-** data is publicly available, it is not unique to explore identity of a patient.

3.2.4 Downloading data

There are three different ways of downloading data

1. **Data Matrix:** Subsets of data are downloaded by users based on specific criteria.

2. **Bulk Download:** Archives of data are searched and downloaded by users that are uploaded by TCGA centres.

3. **HTTP Directories:** HTTP directories are directly accessed by users where data archives are stored.

3.3 Data types selected for study

Major focus of this study was to explore NGS and GWAS data. CNVs data from GWAS by using Affymetix SNP 6.0 platform and SNVs data of somatic mutations from NGS by using Whole Exome Sequencing was available on TCGA. These two types of data were selected for this study.

3.4 Lung cancer data collection

More than twenty different types of cancers data is available on TCGA. Focus of this study was to explore variants of one specific organ and lung was selected. There were three different types of lung cancer data available on TCGA.

1. Lung adenocarcinoma

2. Lung squamous cell carcinoma

3. Mesothelioma

CNVs, level 3 data of LUAD, LUSC and Meso was downloaded from TCGA by using TCGA Data Matrix. Data comprises of a folder consisting different files.

SNVs, level 2 data of LUAD, LUSC and Meso was downloaded from TCGA by using TCGA Data Matrix. There was a .maf file for each type of cancer.

3.5 Data preprocessing

After downloading data it was challenging to make data suitable for pipelines. Scattered files were managed by writing R programming scripts. Different files were merged to a single file and Annovar accepted format was made (consisting of chromosome, position, alternative and reference allele).

3.6 Gene-based annotation

Gene-based annotation libraries were downloaded in Annovar. Ensembl genes and human genome reference hg19 build was used for gene based annotation in Annovar. To avoid excel auto gene id conversion to date, we used Ensemble gene id database. Gene-based annotation was used for getting two types of information, one location of SNVs and CNVs in genome (exonic, intronic or intergenic) and second Ensembl gene ids.

3.7 Conserved variants finding

After gene based annotation, there was a step of finding variants that lie in conserved regions (SNVs) or overlap conserved regions (CNVs). 46-way alignment was used for annotating variants that fall in conserved regions.

3.8 1000 Genome project

It is highly desirable to match subsets of variants to other existing variant databases, by doing this step we cross check our variants with information of 1092 individual variants available in 1000 genome project databases. For this purpose, database of variants annotated in 1000 Genome project (version August 2015) was used. Variants having same start and end positions, and same observed alleles were dropped out during this step .

3.9 microRNAs and snoRNAs

miRNAs and snoRNAs have distinct and central role of regulation in cells. There is functional and evolutionary relationship between subsets of miRNAs and snoRNAs. Regions overlapping microRNAs and snoRNAs were identified by using UCSC wgRna table for microRNAs and snoRNA.

3.10 Functional enrichment analysis

Gene symbols were extracted for further analysis using notepad++. Functional annotation was performed by using on-line tool DAVID.

3.10.1 Disease

Online Mendelian Inheritance in Man (OMIM) is a database of information related to genes and heritable traits of humans. Genes enrichment in diseases was studied by checking there overlapping with Online Mendelian Inheritance in Man knowledge base.

3.10.2 Pathways

Biological pathways control person inner world, pathways involved in gene regulation, metabolism and transmission of signal are important pathways, problem in pathway can trigger disease. In this step genes enriched in KEGG pathways were studied.

3.10.3 Protein domains

Protein domain is conserved part of protein sequence. For protein domain analysis Panther (Protein Analysis Through Evolutionary Relationships) database was used.

3.10.4 Protein interaction

Regulatory regions of eukaryotes are characterized by transcription factor binding site presence. Genes enriched in transcription factor binding sites were studied by using TFBS UCSC database.

3.11 Protein-protein interaction

Knowledge of direct and indirect interactions between proteins of a cell provide comprehensive view of cellular mechanism and function [44]. STRING data base provides assessment and integration of protein-protein interactions, it includes direct(physical) and indirect (functional) associations. STRING (version 10.0) covers more than 10000 organisms [45]. Proteins are product of genes, it is important to find out the interaction of proteins with each other. PPIs were studied by using online tool STRING (version 10.0).

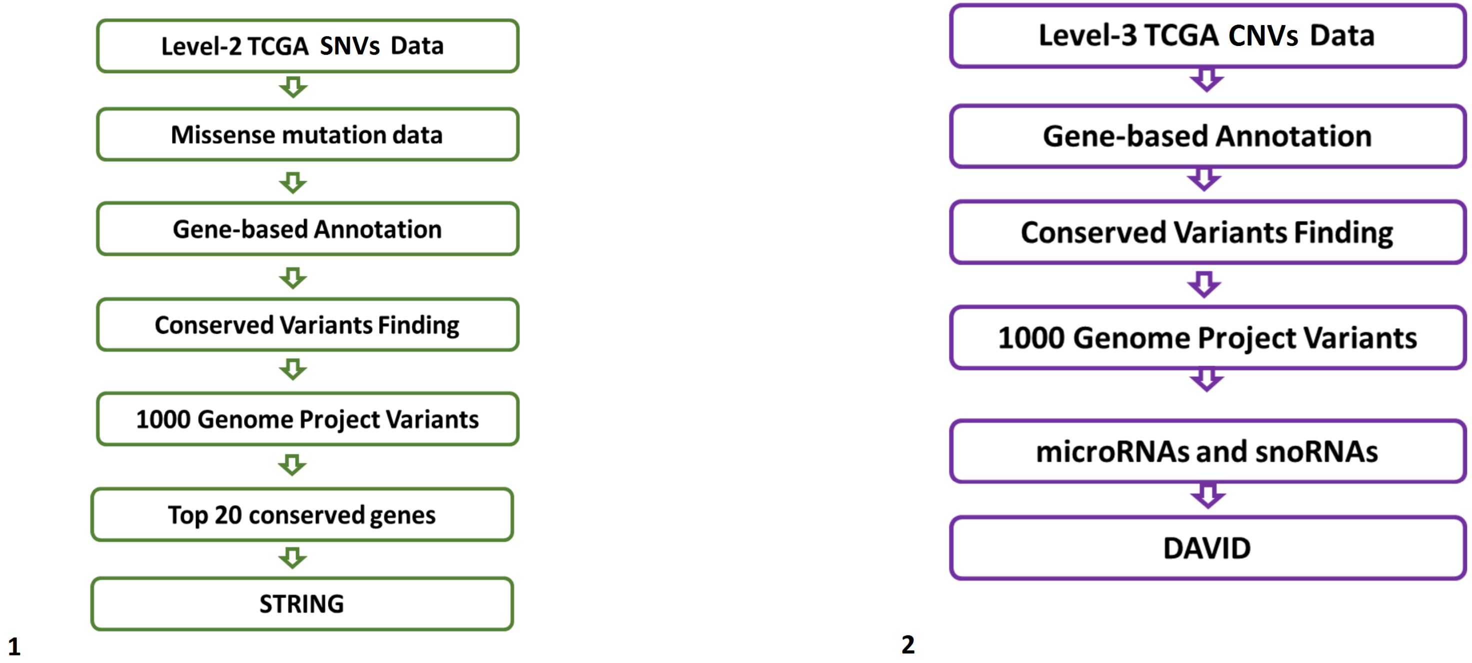


Figure 3.3: 1-SNVs Data analysis pipeline 2-CNVs Data analysis pipeline.

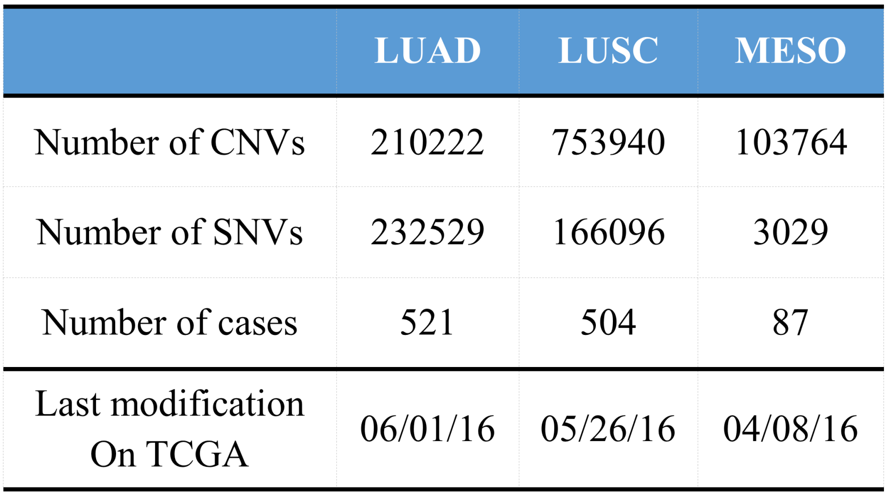
Chapter 4

Results and Discussion

In this chapter cancer data analysis is presented and discussed. Our project was divided into two parts i.e. NGS data analysis and GWAS data analysis to identify candidate risk factors for three types of factors i.e. LUAD, LUSC and MESO.

4.1 Overview of data

This overview of data was compiled during data download and organizing step. ”Last modification on TCGA” in Table (4.1) is representing the downloaded data version.

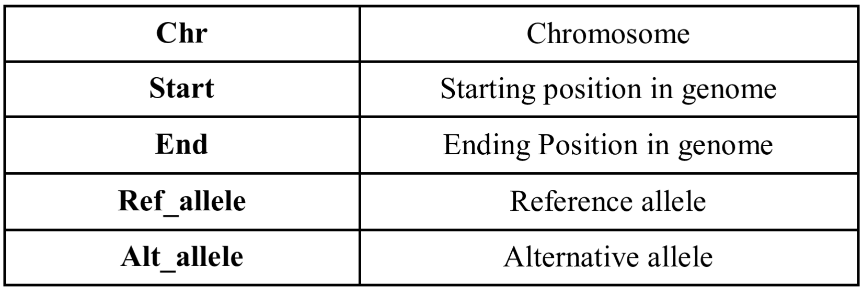


	LUAD	LUSC	MESO
Number of CNVs	210222	753940	103764
Number of SNVs	232529	166096	3029
Number of cases	521	504	87
Last modification On TCGA	06/01/16	05/26/16	04/08/16

Table 4.1: Overview of data

4.2 Annotation parameters

Annotation was done using Annovar. Annovar required input file in a specific format called ".AVINPUT". Parameters used in ".AVINPUT" format are described in Table(4.2). ".AVINPUT" is tab-deliminated file. If information of reference and alternative allele is not available then "0" is used in reference and alternative allele column for making Annovar acceptable file.

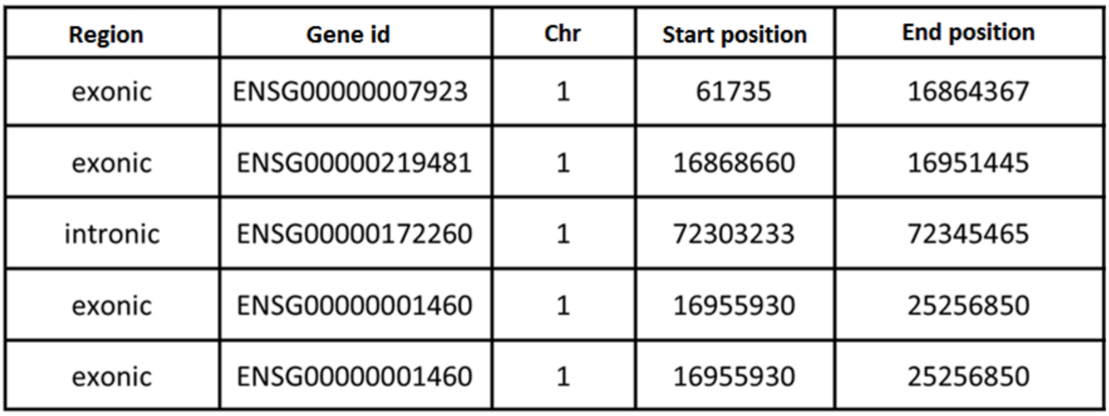


Chr	Chromosome
Start	Starting position in genome
End	Ending Position in genome
Ref_allele	Reference allele
Alt_allele	Alternative allele

Table 4.2: Description of .AVINPUT format

4.3 Gene-based annotation

Gene based annotation was performed after downloading and manipulation of LUAD, LUSC and MESO level-3 data from TCGA. 210222 variants of LUAD, 753940 variants of MESO and 103764 vaiants of LUSC were annotated by using Annovar. First column in (Table 4.3) is explaining the region of genome where specific variant is located, it can be intronic, intergenic or exonic etc. Second column of Table (4.3) is explaining Ensemble gene ids, genes in which variant lies or distance between two genes with gene ids.



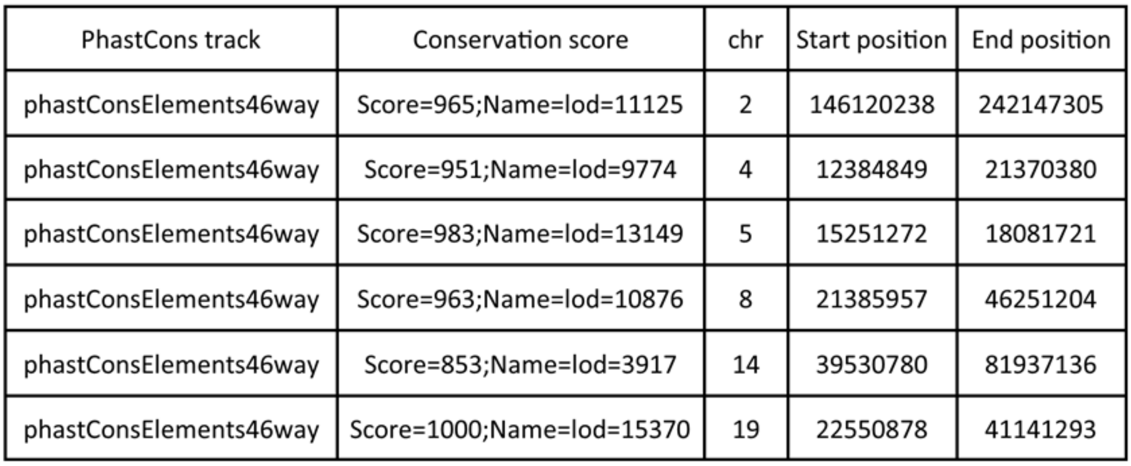
Region	Gene id	Chr	Start position	End position
exonic	ENSG00000007923	1	61735	16864367
exonic	ENSG00000219481	1	16868660	16951445
intronic	ENSG00000172260	1	72303233	72345465
exonic	ENSG00000001460	1	16955930	25256850
exonic	ENSG00000001460	1	16955930	25256850

Table 4.3: Results of gene based annotation using annovar

Annotation is an important genome analysis and essential for genome exploration. It required one year per person for manually annotating 1 mb genome [58]. There are chances of error in automated annotation methods but we have to rely on automate based options because it is impossible to manually annotate such a big sequencing data [62].

4.4 Conserved variants search

Norm score threshold value of 950 was used for getting variants lie in highly conserved regions of genome. Three different analysis were performed on three different type of lung cancer related data. In Table (4.5) first column is representing 46way alignment, second column is showing UCSC phast cons score.

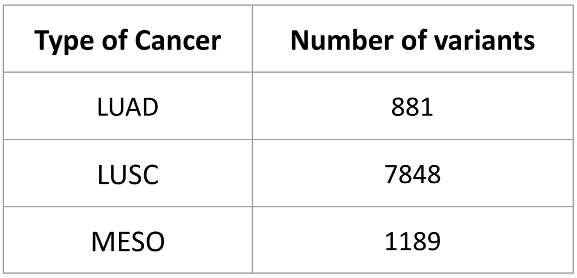


PhastCons track	Conservation score	chr	Start position	End position
phastConsElements46way	Score=965;Name=lod=11125	2	146120238	242147305
phastConsElements46way	Score=951;Name=lod=9774	4	12384849	21370380
phastConsElements46way	Score=983;Name=lod=13149	5	15251272	18081721
phastConsElements46way	Score=963;Name=lod=10876	8	21385957	46251204
phastConsElements46way	Score=853;Name=lod=3917	14	39530780	81937136
phastConsElements46way	Score=1000;Name=lod=15370	19	22550878	41141293

Table 4.4: Results of variants lying in conserved regions using Annovar.

It is interesting to find variants at conserved genomic regions, as they are regions in the genome that are conserved in multiple species and any change in these sites might contribute to a disease or functionally important phenotype. In case of SNVs, variants lie in conserved region while when dealing with CNVs, variants overlap conserved regions.

Variants are potentially causative that lie in genomic sequences that are conserved across species [60]. Table (4.5) shows that 881 variants in LUAD, 7848 in LUSC and 1189 in MESO lie in conserved regions.



Type of Cancer	Number of variants
LUAD	881
LUSC	7848
MESO	1189

Table 4.5: Number of variants lying in conserved regions.

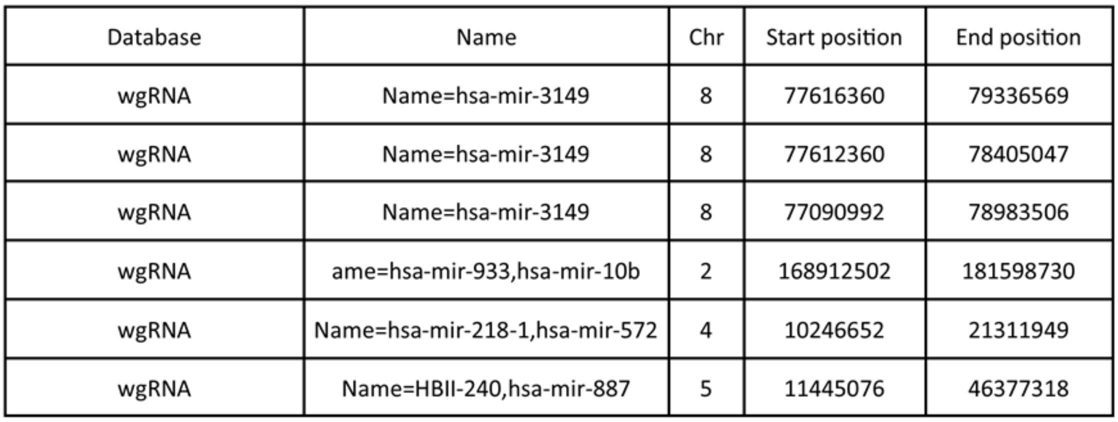
4.5 1000-genome project

No variant was dropped out from three lung cancer data sets, while searching a list of variants from 1000 genome project (Aug 2015 version) database, it shows that no variant is from 1000 genome project.

The aim of using 1000 genome project was to identify those variants that may be common or present in 1000 genomes. As the individuals selected for 1000 genomes were normal, so any variant overlaping 1000 genomes individuals will suggest that it may not be contributing in cancer disease. However, in our case, none of our variants overlapped 1000 genomes variants, suggesting that our variants are important in causing disease.

4.6 microRNAs and snoRNAs

Variants from LUAD, LUSC and MESO were found that overlaps with microRNAs and snoRNAs. Table (??) shows the results of mcroRNAs and snoRNAs. In LUAD 631, LUSC 6789 and MESO 1049 variants were overlapping microRNAs and snoRNAs.

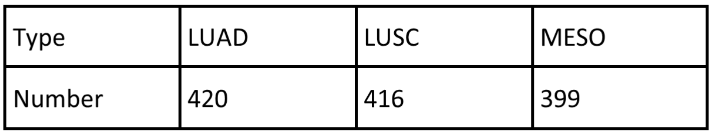


Database	Name	Chr	Start position	End position
wgRNA	Name=hsa-mir-3149	8	77616360	79336569
wgRNA	Name=hsa-mir-3149	8	77612360	78405047
wgRNA	Name=hsa-mir-3149	8	77090992	78983506
wgRNA	ame=hsa-mir-933,hsa-mir-10b	2	168912502	181598730
wgRNA	Name=hsa-mir-218-1,hsa-mir-572	4	10246652	21311949
wgRNA	Name=HBII-240,hsa-mir-887	5	11445076	46377318

Table 4.6: Results of microRNAs and snoRNAs using Annovar.

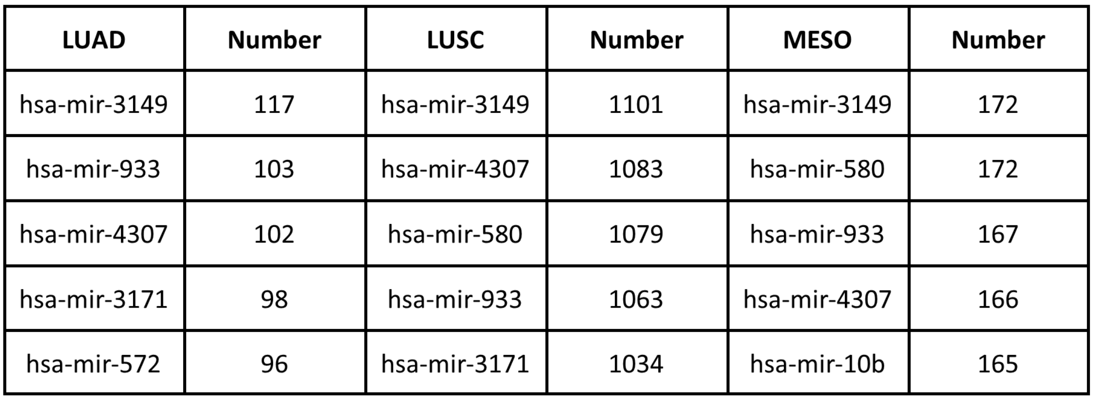
We have found 420, 416, 399 microRNAS and snoRNAs in LUAD, LUSC and MESO, respectively, (Table (4.7)). hsa-mir-3149, hsa-mir-933 and hsa-mir-

4307 are most occuring and common miRNA in all three types of cancers. It implies that may they have important role in progression of lung cancer Table (4.8).



Type	LUAD	LUSC	MESO
Number	420	416	399

Table 4.7: Number of micro RNAs and snoRNAs found in overlapping regions of CNVs .



LUAD	Number	LUSC	Number	MESO	Number
hsa-mir-3149	117	hsa-mir-3149	1101	hsa-mir-3149	172
hsa-mir-933	103	hsa-mir-4307	1083	hsa-mir-580	172
hsa-mir-4307	102	hsa-mir-580	1079	hsa-mir-933	167
hsa-mir-3171	98	hsa-mir-933	1063	hsa-mir-4307	166
hsa-mir-572	96	hsa-mir-3171	1034	hsa-mir-10b	165

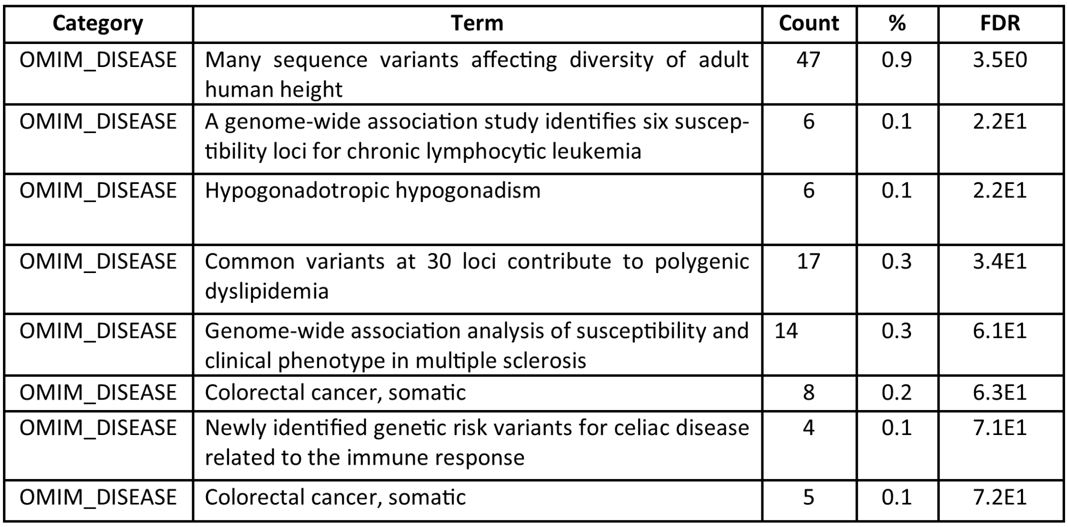
Table 4.8: Most common variants in all three types of cancers.

4.7 Functional enrichment analysis

4.7.1 Disease

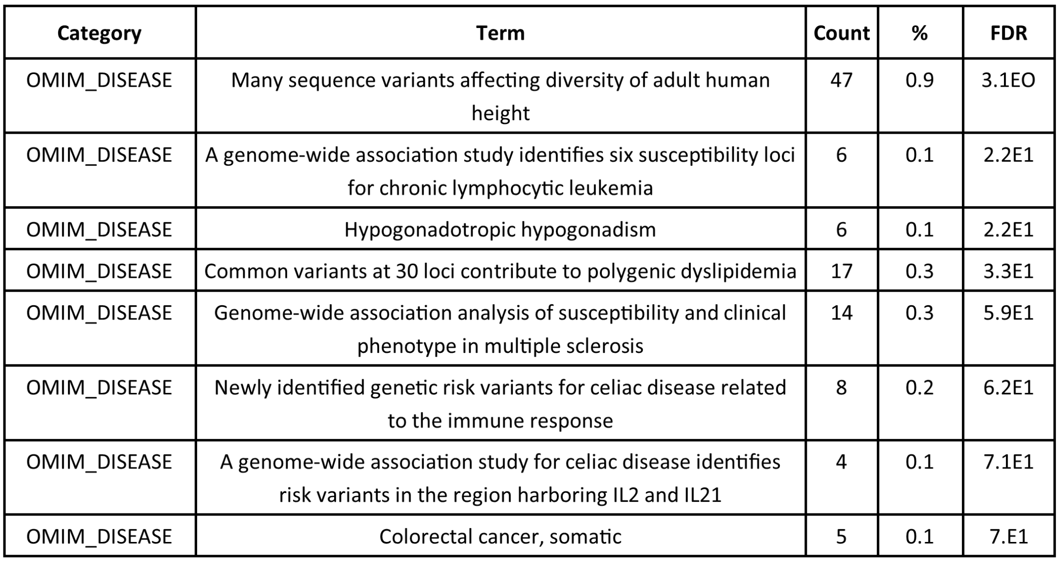
Gene list was checked for its enrichment in already published disease knowledge base OMIM (Online Mendelian Inheritance in Man) by using DAVID. OMIM is a comprehensive and timely knowledge base of human genes and genetic disorders for supporting research and education on human genetics [59]. There were 15.3% genes from our list that overlapped with OMIM_disease knowledge base. Value of FDR<0.1 was set for finding genes enriched in OMIM_Disease. Count in Table

(4.9) is explaining number of genes, % column is explaining percentage of genes from total genes list, FDR column is presenting values of FDR.



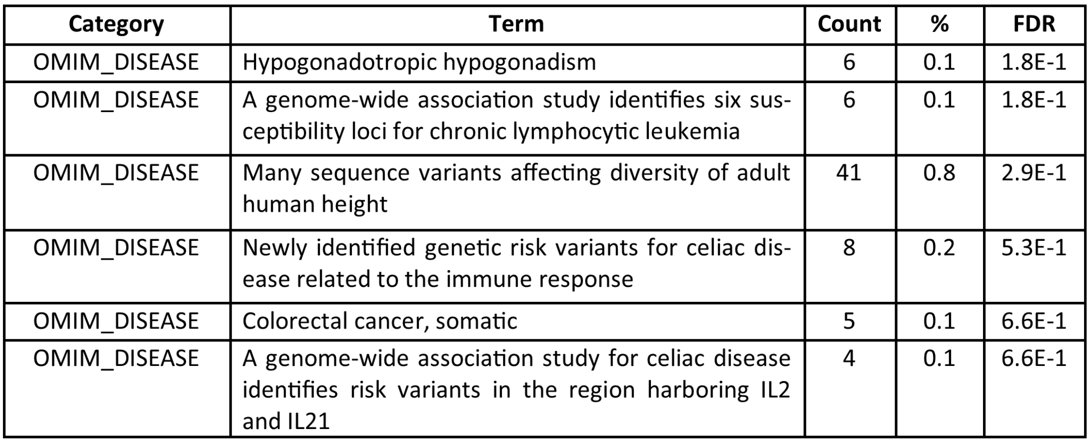
Category	Term	Count	%	FDR
OMIM_DISEASE	Many sequence variants affecting diversity of adult human height	47	0.9	3.5E0
OMIM_DISEASE	A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia	6	0.1	2.2E1
OMIM_DISEASE	Hypogonadotropic hypogonadism	6	0.1	2.2E1
OMIM_DISEASE	Common variants at 30 loci contribute to polygenic dyslipidemia	17	0.3	3.4E1
OMIM_DISEASE	Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis	14	0.3	6.1E1
OMIM_DISEASE	Colorectal cancer, somatic	8	0.2	6.3E1
OMIM_DISEASE	Newly identified genetic risk variants for celiac disease related to the immune response	4	0.1	7.1E1
OMIM_DISEASE	Colorectal cancer, somatic	5	0.1	7.2E1

Table 4.9: OMIM_DISEASE results of Lung Adenocarcinoma.



Category	Term	Count	%	FDR
OMIM_DISEASE	Many sequence variants affecting diversity of adult human height	47	0.9	3.1EO
OMIM_DISEASE	A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia	6	0.1	2.2E1
OMIM_DISEASE	Hypogonadotropic hypogonadism	6	0.1	2.2E1
OMIM_DISEASE	Common variants at 30 loci contribute to polygenic dyslipidemia	17	0.3	3.3E1
OMIM_DISEASE	Genome-wide association analysis of susceptibility and clinical phenotype in multiple sclerosis	14	0.3	5.9E1
OMIM_DISEASE	Newly identified genetic risk variants for celiac disease related to the immune response	8	0.2	6.2E1
OMIM_DISEASE	A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21	4	0.1	7.1E1
OMIM_DISEASE	Colorectal cancer, somatic	5	0.1	7.E1

Table 4.10: OMIM_DISEASE results of Lung Squamous Cell Carcinoma.

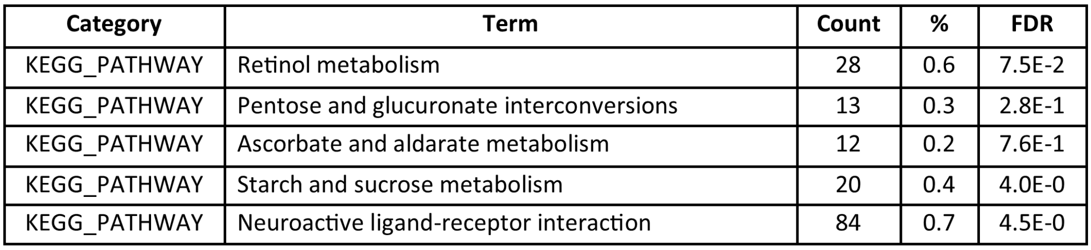


Category	Term	Count	%	FDR
OMIM_DISEASE	Hypogonadotropic hypogonadism	6	0.1	1.8E-1
OMIM_DISEASE	A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia	6	0.1	1.8E-1
OMIM_DISEASE	Many sequence variants affecting diversity of adult human height	41	0.8	2.9E-1
OMIM_DISEASE	Newly identified genetic risk variants for celiac disease related to the immune response	8	0.2	5.3E-1
OMIM_DISEASE	Colorectal cancer, somatic	5	0.1	6.6E-1
OMIM_DISEASE	A genome-wide association study for celiac disease identifies risk variants in the region harboring IL2 and IL21	4	0.1	6.6E-1

Table 4.11: OMIM_DISEASE results of Mesothelioma

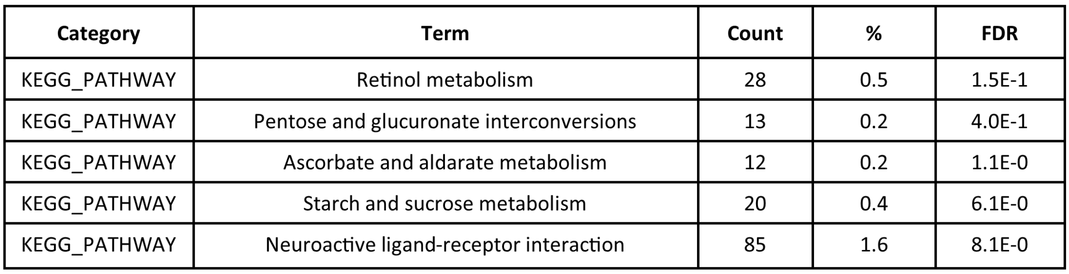
4.7.2 Pathways

25.4% genes were enriched in KEGG pathway. FDR<0.01 was selected for significant enrichment. Table (4.12), Table (4.13) and Table (4.14) shows that three lung cancers have same significant enriched genes.



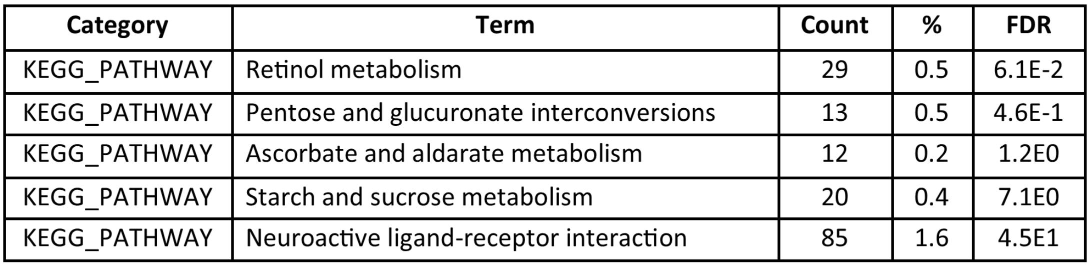
Category	Term	Count	%	FDR
KEGG_PATHWAY	Retinol metabolism	28	0.6	7.5E-2
KEGG_PATHWAY	Pentose and glucuronate interconversions	13	0.3	2.8E-1
KEGG_PATHWAY	Ascorbate and aldarate metabolism	12	0.2	7.6E-1
KEGG_PATHWAY	Starch and sucrose metabolism	20	0.4	4.0E-0
KEGG_PATHWAY	Neuroactive ligand-receptor interaction	84	0.7	4.5E-0

Table 4.12: KEGG_Pathways results of LUAD



Category	Term	Count	%	FDR
KEGG_PATHWAY	Retinol metabolism	28	0.5	1.5E-1
KEGG_PATHWAY	Pentose and glucuronate interconversions	13	0.2	4.0E-1
KEGG_PATHWAY	Ascorbate and aldarate metabolism	12	0.2	1.1E-0
KEGG_PATHWAY	Starch and sucrose metabolism	20	0.4	6.1E-0
KEGG_PATHWAY	Neuroactive ligand-receptor interaction	85	1.6	8.1E-0

Table 4.13: KEGG_Pathways results of LUSC

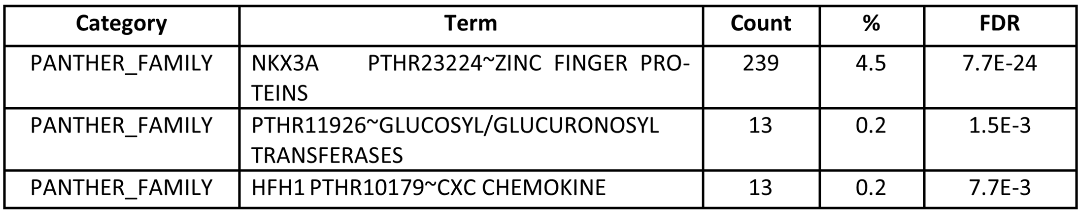


Category	Term	Count	%	FDR
KEGG_PATHWAY	Retinol metabolism	29	0.5	6.1E-2
KEGG_PATHWAY	Pentose and glucuronate interconversions	13	0.5	4.6E-1
KEGG_PATHWAY	Ascorbate and aldarate metabolism	12	0.2	1.2E0
KEGG_PATHWAY	Starch and sucrose metabolism	20	0.4	7.1E0
KEGG_PATHWAY	Neuroactive ligand-receptor interaction	85	1.6	4.5E1

Table 4.14: KEGG_pathways results of MESO

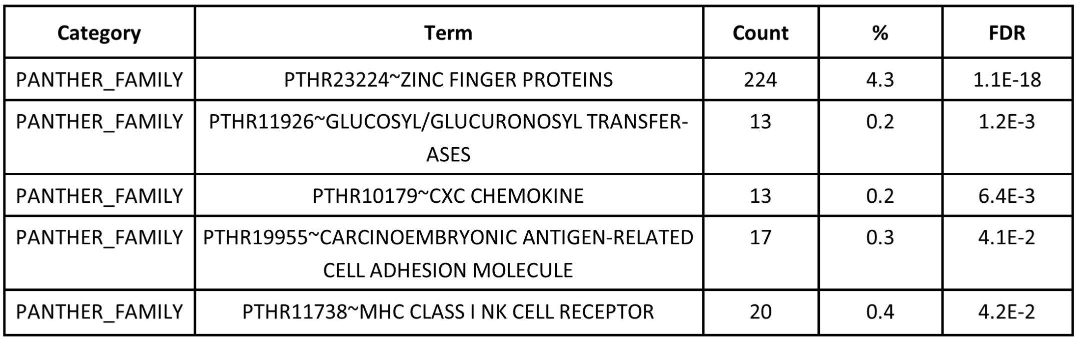
4.7.3 Protein Domains

Enrichment in protein domain was checked by using PANTHER databases. All genes from three different cancers were most significantly enriched in Zinc FINGER containing domains protein. Table(4.15), Table(4.16) and Table(4.17).



Category	Term	Count	%	FDR
PANTHER_FAMILY	NKX3A PTHR23224~ZINC FINGER PROTEINS	239	4.5	7.7E-24
PANTHER_FAMILY	PTHR11926~GLUCOSYL/GLUCURONOSYL TRANSFERASES	13	0.2	1.5E-3
PANTHER_FAMILY	HFH1 PTHR10179~CXC CHEMOKINE	13	0.2	7.7E-3

Table 4.15: Protein Domain analysis of LUAD genes using PANTHER database.



Category	Term	Count	%	FDR
PANTHER_FAMILY	PTHR23224~ZINC FINGER PROTEINS	224	4.3	1.1E-18
PANTHER_FAMILY	PTHR11926~GLUCOSYL/GLUCURONOSYL TRANSFERASES	13	0.2	1.2E-3
PANTHER_FAMILY	PTHR10179~CXC CHEMOKINE	13	0.2	6.4E-3
PANTHER_FAMILY	PTHR19955~CARCINOEMBRYONIC ANTIGEN-RELATED CELL ADHESION MOLECULE	17	0.3	4.1E-2
PANTHER_FAMILY	PTHR11738~MHC CLASS I NK CELL RECEPTOR	20	0.4	4.2E-2

Table 4.16: Protein Domain analysis of LUSC genes using PANTHER database.

Category	Term	Count	%	FDR
PANTHER_FAMILY	NKX3A PTHR23224~ZINC FINGER PROTEINS	221	4.4	6.0E-20
PANTHER_FAMILY	PTHR11926~GLUCOSYL/GLUCURONOSYL TRANSFERASES	13	0.3	7.3E-4
PANTHER_FAMILY	HFH1 PTHR10179~CXC CHEMOKINE	13	0.3	3.9E-3

Table 4.17: Protein Domain analysis of MESO genes using PANTHER database.

4.7.4 Protein Interaction

Genes were checked for protein interaction against transcription factor binding site database of UCSC by using DAVID protein interaction interface.

Category	Term	Count	%	FDR
UCSC_TFBS	NKX3A	2037	38.3	1.8E-2
UCSC_TFBS	CDC5	1959	36.8	3.3E-2
UCSC_TFBS	FOXO3	1216	22.9	8.1E-2
UCSC_TFBS	EGR3	757	14.2	8.5E-2
UCSC_TFBS	AP2	1110	20.9	1.3E-1

Table 4.18: Transcription Factor Binding Site analysis results of LUAD

Category	Term	Count	%	FDR
UCSC_TFBS	NKX3A	2026	38.7	8.7E-4
UCSC_TFBS	CDC5	1948	37.3	2.3E-3
UCSC_TFBS	FOXO3	1208	23.1	1.8E-2
UCSC_TFBS	HFH1	1945	37.2	2.5E-2
UCSC_TFBS	E4BP4	1872	35.8	3.5E-2

Table 4.19: Transcription Factor Binding Site analysis results of LUSC

Category	Term	Count	%	FDR
UCSC_TFBS	NKX3A	1933	38.7	2.8E-3
UCSC_TFBS	CDC5	1862	37.2	4.1E-3
UCSC_TFBS	HFH1	1871	37.4	6.2E-3
UCSC_TFBS	FOXO3	1159	23.2	1.3E-2
UCSC_TFBS	LHX3	1526	30.5	8.8E-2

Table 4.20: Transcription Factor Binding Site analysis results of MESO

Comparing results of all three different types of cancer it can be concluded that they share some common genetics because there are many common genes that are significantly enriched in all three types of cancers.

4.8 SNVs data analysis

In second part of project, TCGA SNVs level-2 was used. Level-2 data was annotated data of somatic mutations from three different types of lung cancers. we were interested in non synonymous substitution so we separated missense mutation data from other types of substitution data like Silent, Frame_Shift_Ins, Nonsense_Mutation and RNA data. R programming scripts were written for data preprocessing. Missense mutation data was re-annotated by using Annovar as mentioned in first part of work. Gene based annotation was done on data, list of genes associated to variants was obtained. In next step variants lying in conserved regions were identified on norm score of 600, normscore 600 was applied because no variant was identified in Mesothelimia at high norm scores. After conservation step, variants were passed through 1000 genome project database, few variants from all three types of cancers were dropped at this step.

4.8.1 Protein-Protein Interaction

We were interested in non synonymous variants lie in high conserved regions from list of highly conserved variant. We selected 30 genes associated with top normscore variants for checking their protein-protein interaction. We used STRING online PPIs tool. Figure (4.1), Figure (4.2) and Figure (4.3) are showing PPIs found in 30 proteins of LUAD, LUSC and MESO respectively.

Prediction of SNVs altering protein product is becoming popular in genomics and bioinformatics. PPIs present strong functional relationship among genes. Miguel Vázquez et al. 2015, studied protein-protein interactions in cancer related SNVs. They built a tool for PPIs named Structure-PPi. They performed analysis by this tool and and identified strong relationship of SNVs with cancer [64]. In our analysis, we identified strong relationship between proteins in all three types of cancers. Further analysis of involvement of these PPIs in relevant cancer can be carried out by mining already published.

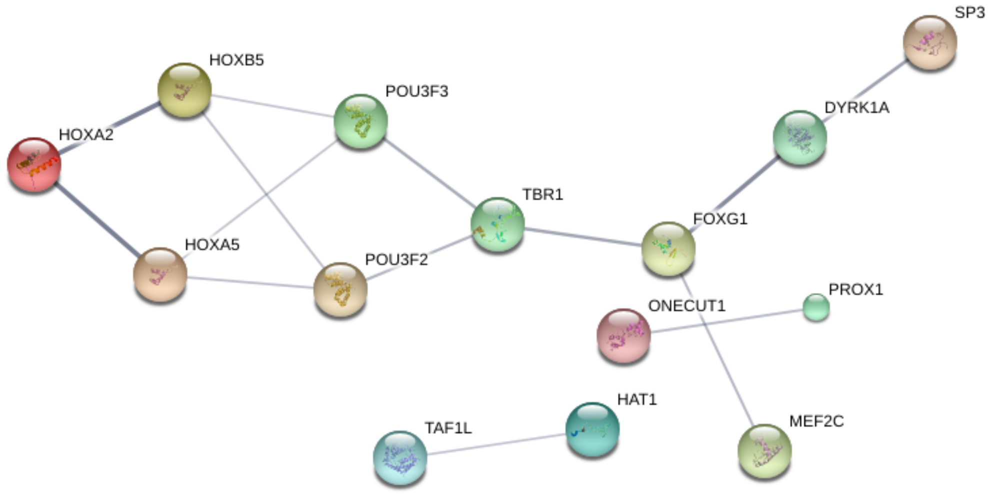


Figure 4.1: Protein-Protein interaction in LUAD. Node is presenting a single protein and network edges are presenting confidence. Disconnected nodes in the network are hidden.

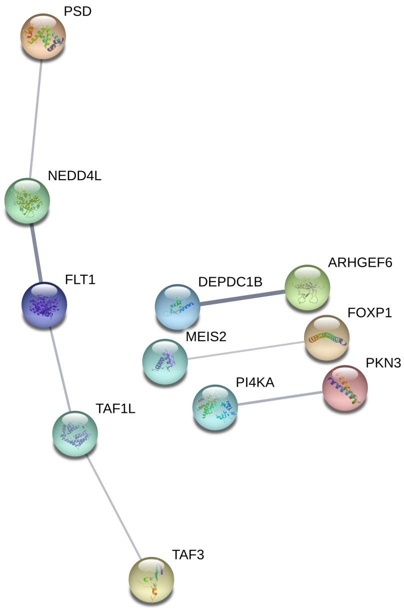


Figure 4.2: Protein-Protein interaction in LUSC. Node is presenting a single protein and network edges are presenting confidence. Disconnected nodes in the network are hidden

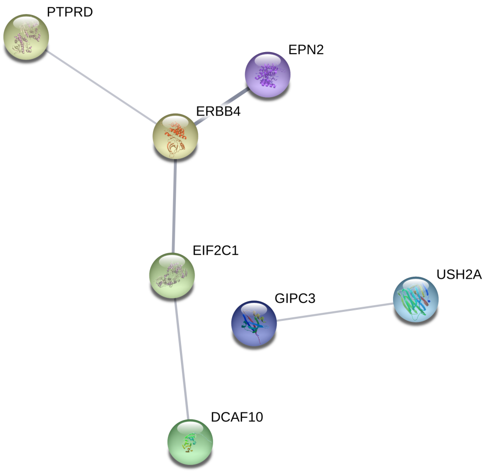


Figure 4.3: Protein-Protein interaction in LUSC. Node is presenting a single protein and network edges are presenting confidence. Disconnected nodes in the network are hidden.

4.9 Limitations

Few limitations faced in this project are:

- Limited access to data is available, Only level 2 and level 3 data is available. Level 1 data can improve accuracy of results.
- There are chances of errors in automated annotation pipelines.

Chapter 5

Conclusion and Future directions

5.1 Conclusion

Information in our genome is like English alphabet. There are 26 alphabets in English but our genome has only 4 alphabet A, C, G and T, also termed as genetic letters. Genetic letter combine in a sequence and produce a story but cancer disrupt this story by causing small changes in these letters resulting in change of genomic word or sentence. Proteins formed by this procedure are inefficient, incomplete or lethal for human body. To study cancer, collection of different samples from different patients give deep insight, to researchers, for finding that what makes cancer different from one patient to other patient. For that there is need of exploring pipelines that are easy to use and accurate result wise. This project is focused on SNVs and CNVs data of three different types of lung cancers.

Our results suggest that several genes containing the copy number variants and single nucleotide variants are highly conserved. MicroRNAs and snoRNAs are very important in the regulation of genes and their functional affect. Any change in these may cause the disease condition. Three of the microRNAs (hsa-mir-3149, hsa-mir-933 and hsa-mir-4307) were identified as overrepresented in all three lung cancers, suggesting their role in the disease condition. Furthermore, genes containing zinc finger domain was identified to be common in these cancers.

Transcription factors (nks39, cdc5 and foxo3) binding sites were also common in the genes containing the CNVs, suggesting their regulatory function.

5.2 Future directions

In present stage this study can be further enhanced by diverting its scope across tumors like Skin cancer, Cervical cancer and Uterine carcinosarcoma. Collective study across tumors will be more powerful in nature and helpful in un covering two aspects, one finding molecular basis within tumor and second across tumor.

Different other types of data like DNA methylation ,gene expression is also available on TCGA. Pipelines can be designed for analysis of data for making this study more comprehensive.

Chapter 6

Resources

Annovar:	$http://annovar.openbioinformatics.org/en/latest/$
DAVID:	$https://david.ncifcrf.gov/$
Notepad++:	$https://notepad-plus-plus.org/download/v6.9.html$
Perl:	$http://strawberryperl.com/$
R-programming:	$https://cran.r-project.org/bin/windows/base/$
SIFT:	$http://sift.bii.a-star.edu.sg/$
STRING:	$http://string-db.org/$
TCGA:	$http://cancergenome.nih.gov/$

Bibliography

[1] Brown, Terence A. ”Genomes, Transcriptomes and Proteomes.” 2002.

[2] Noble, Denis. The music of life: biology beyond genes. Oxford University Press, 2008.

[3] Alberts, Bruce, et al. ”Molecular Biology of the Cell (3rd edn).” Trends in Biochemical Sciences 20.5 (1995): 210-210.

[4] http://www.cancer.gov/about-cancer/causes-prevention/genetics

[5] http://www.who.int/mediacentre/factsheets/fs297/en/

[6] Khanna, Pragya. Essentials of Genetics. IK International Pvt Ltd, 2010.

[7] Shigemizu, Daichi, et al. ”A practical method to detect SNVs and indels from whole genome and exome sequencing data.” Scientific reports 3 (2013).

[8] Strachan, Tom, Judith Goodship, and Patrick Chinnery. Genetics and Genomics in Medicine. Taylor & Francis, 2014.

[9] https://autismsciencefoundation.files.wordpress.com/2015/12/germlinesomatic1.gif

[10] Zhang, Feng, et al. ”Copy number variation in human health, disease, and evolution.” Annual review of genomics and human genetics 10 (2009): 451.

[11] Martincorena, Iñigo, and Peter J. Campbell. ”Somatic mutation in cancer and normal cells.” Science 349.6255 (2015): 1483-1489.

[12] Cline, Melissa S., and Rachel Karchin. ”Using bioinformatics to predict the functional impact of SNVs.” Bioinformatics 27.4 (2011): 441-448.

[13] http://www.nature.com/scitable/definition/single-nucleotide-polymorphism-snp-295

[14] Stoneking, Mark. "Single nucleotide polymorphisms: From the evolutionary past" Nature 409.6822 (2001): 821-822

[15] Craddock, Nick, et al. "Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls." Nature 464.7289 (2010): 713-720.

[16] Stankiewicz, Pawel, and James R. Lupski. "Structural variation in the human genome and its role in disease." Annual review of medicine 61 (2010): 437-455.

[17] http://www.emedmd.com/content/genomics-introduction

[18] Hastings, P. J., et al. "Mechanisms of change in gene copy number." Nature Reviews Genetics 10.8 (2009): 551-564.

[19] O'Donovan, Michael C., George Kirov, and Michael J. Owen. "Phenotypic variations on the theme of CNVs." Nature genetics 40.12 (2008): 1392-1393.

[20] Ding, Keyue, and Iftikhar J. Kullo. "Genome-wide association studies for atherosclerotic vascular disease and its risk factors." Circulation: Cardiovascular Genetics 2.1 (2009): 63-72.

[21] http://www.cancerresearchuk.org/about-cancer/type/lung-cancer/about/lung-cancer-risks-and-causes

[22] Machiela, Mitchell J., and Stephen J. Chanock. "GWAS is going to the dogs." Genome biology 15.3 (2014): 1.

[23] Chung, Charles C., et al. "Genome-wide association studies in cancer-current and future directions." Carcinogenesis (2009): bgp273.

[24] Metzker, Michael L. "Emerging technologies in DNA sequencing." Genome research 15.12 (2005): 1767-1776.

[25] Voelkerding, Karl V., Shale A. Dames, and Jacob D. Durtschi. ”Next-generation sequencing: from basic research to diagnostics.” Clinical chemistry 55.4 (2009): 641-658.

[26] Shyr, Derek, and Qi Liu. ”Next generation sequencing in cancer research and clinical application.” Biological procedures online 15.1 (2013): 1.

[27] Cline, Melissa S., et al. ”Exploring TCGA pan-cancer data at the UCSC cancer genomics browser.” Scientific reports 3 (2013): 2652.

[28] http://www.cancer.gov/about-cancer/understanding/what-is-cancer#related-diseases

[29] Ma, Xiaomei, and Herbert Yu. ”Global burden of cancer.” Yale J Biol Med 79.3-4 (2006): 85-94.

[30] Sudhakar, Akulapalli. ”History of cancer, ancient and modern treatment methods.” Journal of cancer science & therapy, 1.2 (2009): 1.

[31] http://www.cancerresearchuk.org/about-cancer/what-is-cancer

[32] Yang, Yadong, et al. ”Databases and web tools for cancer genomics study.” Genomics, proteomics & bioinformatics 13.1 (2015): 46-50.

[33] Ding, Li, et al. ”Somatic mutations affect key pathways in lung adenocarcinoma.” Nature 455.7216 (2008): 1069-1075.

[34] Cancer Genome Atlas Research Network. ”Comprehensive genomic characterization of squamous cell lung cancers.” Nature 489.7417 (2012): 519-525.

[35] Tsou, Jeffrey A., et al. ”Distinct DNA methylation profiles in malignant mesothelioma, lung adenocarcinoma, and non-tumor lung.” Lung Cancer 47.2 (2005): 193-204.

[36] http://en.cnki.com.cn/Journal_en/A-A006-GPBI-2015-01.htm

[37] The future of cancer research In the era of big data, one of the major... - Scientific Figure on ResearchGate. Available

from:https://www.researchgate.net/figure/272568145_fig1_The-future-of-cancer-research-In-the-era-of-big-data-one-of-the-major-challenges-is-to [accessed Aug 15, 2016]

[38] Wall, Larry, Tom Christiansen, and Jon Orwant. "Programming perl." O'Reilly Media, Inc., 2000.

[39] Niedner, R. Hannes, T. Murlidharan Nair, and Michael Gribskov. "Perl in bioinformatics." Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics, 2005.

[40] Wang, Kai, Mingyao Li, and Hakon Hakonarson. "ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data." Nucleic acids research 38.16 (2010): e164-e164.

[41] Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." Nature protocols 4.1 (2009): 44-57.

[42] Jiao, Xiaoli, et al. "DAVID-WS: a stateful web service to facilitate gene/protein list analysis." Bioinformatics 28.13 (2012): 1805-1806.

[43] Ng, Pauline C., and Steven Henikoff. "SIFT: Predicting amino acid changes that affect protein function." Nucleic acids research 31.13 (2003): 3812-3814.

[44] Franceschini, Andrea, et al. "STRING v9. 1: protein-protein interaction networks, with increased coverage and integration." Nucleic acids research 41.D1 (2013): D808-D815.

[45] Szklarczyk, Damian, et al. "STRING v10: protein–protein interaction networks, integrated over the tree of life." Nucleic acids research (2014): gku1003.

[46] Cole, Charles, et al. "Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data." BMC bioinformatics 15.1 (2014): 1.

[47] Robbins, David E., et al. "A self-updating road map of The Cancer Genome Atlas." Bioinformatics (2013): btt141.

[48] Deus, Helena F., et al. "Exposing the cancer genome atlas as a SPARQL endpoint." Journal of biomedical informatics 43.6 (2010): 998-1008.

[49] 1000 Genomes Project Consortium. "An integrated map of genetic variation from 1,092 human genomes." Nature 491.7422 (2012): 56-65.

[50] Karagiannis, Konstantinos, Vahan Simonyan, and Raja Mazumder. "SNVDis: a proteome-wide analysis service for evaluating nsSNVs in protein functional sites and pathways." Genomics, proteomics & 11.2 (2013): 122-126.

[51] Cancer Genome Atlas Research Network. "Comprehensive genomic characterization of squamous cell lung cancers." Nature 489.7417 (2012): 519-525.

[52] Cancer Genome Atlas Research Network. "Comprehensive molecular profiling of lung adenocarcinoma." Nature 511.7511 (2014): 543-550..

[53] Akbani, Rehan, et al. "A pan-cancer proteomic perspective on The Cancer Genome Atlas." Nature communications 5 (2014).

[54] Shete, Sanjay, et al. "Genome-wide association study identifies five susceptibility loci for glioma." Nature genetics 41.8 (2009): 899-904.

[55] Brennan, Cameron W., et al. "The somatic genomic landscape of glioblastoma." Cell 155.2 (2013): 462-477.

[56] https://www.asbestos.com/mesothelioma/

[57] https://wiki.nci.nih.gov/display/TCGA/Introduction+to+TCGA

[58] Eugene V Koonin and Michael Y Galperin. "Sequence - Evolution - Function Computational Approaches in Comparative Genomics." Kluwer Academic, 2003.

[59] Hamosh, Ada, et al. ”Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.” Nucleic acids research 33.suppl 1 (2005): D514-D517.

[60] Koufariotis, Lambros, et al. ”Regulatory and coding genome regions are enriched for trait associated variants in dairy and beef cattle.” BMC genomics 15.1 (2014): 1.

[61] Matloff, Norman. The art of R programming: A tour of statistical software design. No Starch Press, 2011.

[62] Poptsova, Maria S., and J. Peter Gogarten. ”Using comparative genome analysis to identify problems in annotated microbial genomes.” Microbiology 156.7 (2010): 1909-1917.

[63] Sanner, Michel F. ”Python: a programming language for software integration and development.” J Mol Graph Model 17.1 (1999): 57-61.

[64] Vazquez, Miguel, Alfonso Valencia, and Tirso Pons. ”StructurePPi: a module for the annotation of cancerrelated singlenucleotide variants at proteinprotein interfaces.” Bioinformatics, 2015.

[65] Yang, Yadong, et al. ”Databases and web tools for cancer genomics study.” Genomics, proteomics & bioinformatics 13.1 (2015): 46-50.

[66] 1000 Genomes Project Consortium. ”An integrated map of genetic variation from 1,092 human genomes.” Nature 491.7422 (2012): 56-65.