LIST OF FIGURES

Figure 1: NGS overview

Figure 2: Data portal web face

Figure 3: maf file

Figure 4: convert2annovar.pl

Figure 5: .avinput script format

Figure 6: vcf file

Figure 7: .avinput file

Figure 8: .avinput file

Figure 9: matching vcf variants with DisGeNET

Figure 10: conserved region variations

Figure11: genomicSuperDups filteration

Figure12: polyphen2 scores

Figure13: Cosmic validation

Figure14: Exonic variants

Figure15: Transcription factor binding site

Figure16: cytoBand

Figure 17: Phenolyzer details

LIST OF TABLES

Table 1: Data center

Table 2: TCGA Data Level

Table 3: Data description

LIST OF ABBREVIATIONS

PRAD: Prostate Adenocarcinoma

TCGA: The Cancer Genome Atlas

DNA: Deoxyribonucleic acid

RNA-Seq: RNA Sequencing

SNPs: Single Nucleotide Polymorphisms

CNV: Copy Nnmber Variation

Maf: Mutation Annotated Format

Vcf: Variant Calling Files

Analysis and Annotation of exome sequencing data to identify and prioritize genes responsible for Prostate Adenocarcinoma

Ashish Chahal
Delhi Technological University

ABSTRACT

After skin cancer prostate cancer is the second most prevalent cancer in men. Somatic mutations in Prostate Adenocarcinoma are revealed by processing of the next-generation DNA sequencing data of the exome region. Mutation in exome region directly effects the expression of the genes and sometimes inhibits the expression which can lead to several diseases. High throughput technologies and NGS analysis enable us to find out variations in the exome region that are involved in complex pathways of cancers. Biomarkers can be identified using NGS and exome sequencing analysis pipelines which can help in diagnosis, treatment and prognosis of the cancer. Exome play a major role in protein profiling so any change in this region affect the individual. PRAD exome data was used to analyze the variations in the exome region. Data for PRAD was downloaded from the TCGA web portal for tumor matched with normal types 17 samples on which exome sequence analysis pipeline were applied to predict and prioritize the genes involved for PRAD pathway.

Perl programming language was used to prioritize and analyze the exome data. Perl script maf2vcf.pl, DisGeNET, Annovar software packages were used to find out 93 probable genes that were filtered from DisGeNET. Then 54 genes were found in conserved regions with phastconselements46way score > 400.

17 TCGA IDs samples showed sequence alignment errors which were filtered by matching with segmented duplications. Polyphen2 annotations were used to give scores about the deleterious effect of the variants. After these steps we got the most probable genes that might be responsible for the cause of Prostate Adenocarcinoma (PRAD). GSTT1, TP53, CYP19A1, BRAF genes were already involved in the pathway of occurrence of prostate cancer and these genes were also present in the filtered genes in this study. Using experimental validation methods on the filtered genes we may help in finding out the novel genes that are involved in the complex pathway of prostate cancer.

INTRODUCTION

Revolution in research is due to the wide range of applications and high throughput efficiency of the Next Generation Sequencing (NGS). NGS helps in studying the alternative splicing complexity (Martin *et al.*, 2011), landscape of mutations in cancer. The recent advancement in NGS following the first Human Genome draft (2003), massive data have been generated for various types of cancers. To analyze these data various computational and NGS pipelines approaches are used. The most widely used application is whole exome sequence analysis which is used to find the genetic basis of human diseases phenotype (Mamanova *et al.*, 2010). Many recent studies have been carrying out integrative analysis of epigenetic and exome sequences (Liao *et al.*, 2013).

Capture sequencing of the exome has been quite popular since such an approach showed its clinical utility to identify known and novel variants associated with Mendelian diseases. In a landmark report, exome capture followed by next generation sequencing was shown to have clinical relevance in Miller's syndrome. Furthermore, a number of researchers have used exome sequencing for diagnosis and identification of novel genetic variants and novels genes associated with Mendelian diseases. Additionally, it has also been used to finely map variations in complex diseases and used them in clinical diagnosis of cancers.

The Cancer Genome Atlas (TCGA) is a common platform designed to distribute and handle large volumes of research data for 34 types of cancer. Prostate Adenocarcinoma (PRAD) is a common type of cancer prevalent in North America, Australia, Europe and New Zealand and is a major cause of death in men (Hsing and Chokkalingam, 2006). TCGA provides information about the variations in exome data observed in the tumor samples in which somatic changes and variations are observed. The smart architecture of TCGA enables a researcher to download raw or processed data wherever applicable and available. Independent studies can be carried out to compare, analyze and interpret the information from various platforms on a particular sample. As is the mission of TCGA, the atlas of variations can be analyzed and stored to reduce the gap in the cancer and its molecular biology. Biomarkers can be found by using this study which helps in prognosis, diagnosis and treatment of the PRAD. These markers will be useful for personalized oncology treatment of the cancer.

Data analysis in this study, first involves the conversion of maf files into vcf files using perl language programming script and then several databases and packages were used to filter and analyze the data. In this study, the data was downloaded for PRAD from batch 184 for 17 tumor matched with normal samples.

In this study, we implemented ANNOVAR package (which uses the perl programming language), twobittofa tool to obtain the reference genome in correct format, DisGeNET database, maftovcf.pl perl scripts and MS-Excel were used to filter and analyze the variations in exome sequencing data. Finally the results were viewed on Integrative Genomics Viewer (IGV, Broad Institute).

Review of literature

3.1 Next Generation Sequencing

The genome sequence is largely the same between individuals, all of us differ from each other in certain positions in the genome sequence. These changes are called genetic variations. These are associated with specific features which are shared with parents, so these variations are inherited one. The association of genetic variations and traits form the basis of human genetics. Now, not all genetic variations are associated with a human trait, but only a handful of them, mostly which fall in and around regions of protein coding genes (Conrad *et al.,* 2011). In many cases, the genetic variation is common in the population of individuals and these common variations are otherwise called polymorphisms.

Now sequencing individual genomes to understand the variations and arrive at the implications using the conventional Sanger sequencing methodology would have been extremely expensive. So the years which followed the human genome sequencing saw extensive investments into making nucleotide sequencing cheaper, fast and applicable in clinical settings. Consequent to these efforts, the field saw tremendous improvement in the throughput and speed and consequent drastic reduction in the costs of nucleotide sequencing. So much that it is now possible to sequence complete human genomes at a minuscule fraction of the cost incurred in the international human genome project. These technologies have been generally called next-generation sequencing technologies (NGS). As you would have imagined, NGS is not just one technology, but a generic name for a set of technologies which has enabled high-throughput sequencing of nucleotides. Each of the individual technologies significantly differ in the chemical reactions and readouts, but generally similar in the fact that they can sequence millions if not billions of sequences in one go (Koboldt *et al.,* 2010).

3.2 Exome Sequencing

The genome as we know is composed of over 3 billion bases. Not all of these three billion bases code for genes. Actually only a minuscule proportion of these 3 billion bases has protein coding potential. These regions are not contiguous in most cases of genes, but rather interspersed with regions which do not have a potential to encode for proteins. These are called exons. The exons encompass approximately over 50 million bases in the human genome. So one would naturally argue that it would be worthwhile to just sequence the 50 million odd bases in the human genome to understand genetic diseases. This is true in most cases with some exceptions. A number of mutations could also potentially affect the regulation and biogenesis of transcripts and could also cause diseases. Genetic mutations analysis show that these all fall in or near protein coding genes and exons.

Now sequencing the 50 million odd bases across the protein coding regions or exons and neighboring regulatory regions has been attempted. There are two popular approaches to 'capture' these regions. These are called capture methodologies. The principle is based on the fact that complementary oligonucleotides would hybridize the region of interest and could be separated. Two popular capture methodologies have been developed; in first method we capture fragments on a glass surface, while the other captures it on magnetic beads. Fragments on capturing regions denature and then we do sequencing of these regions.

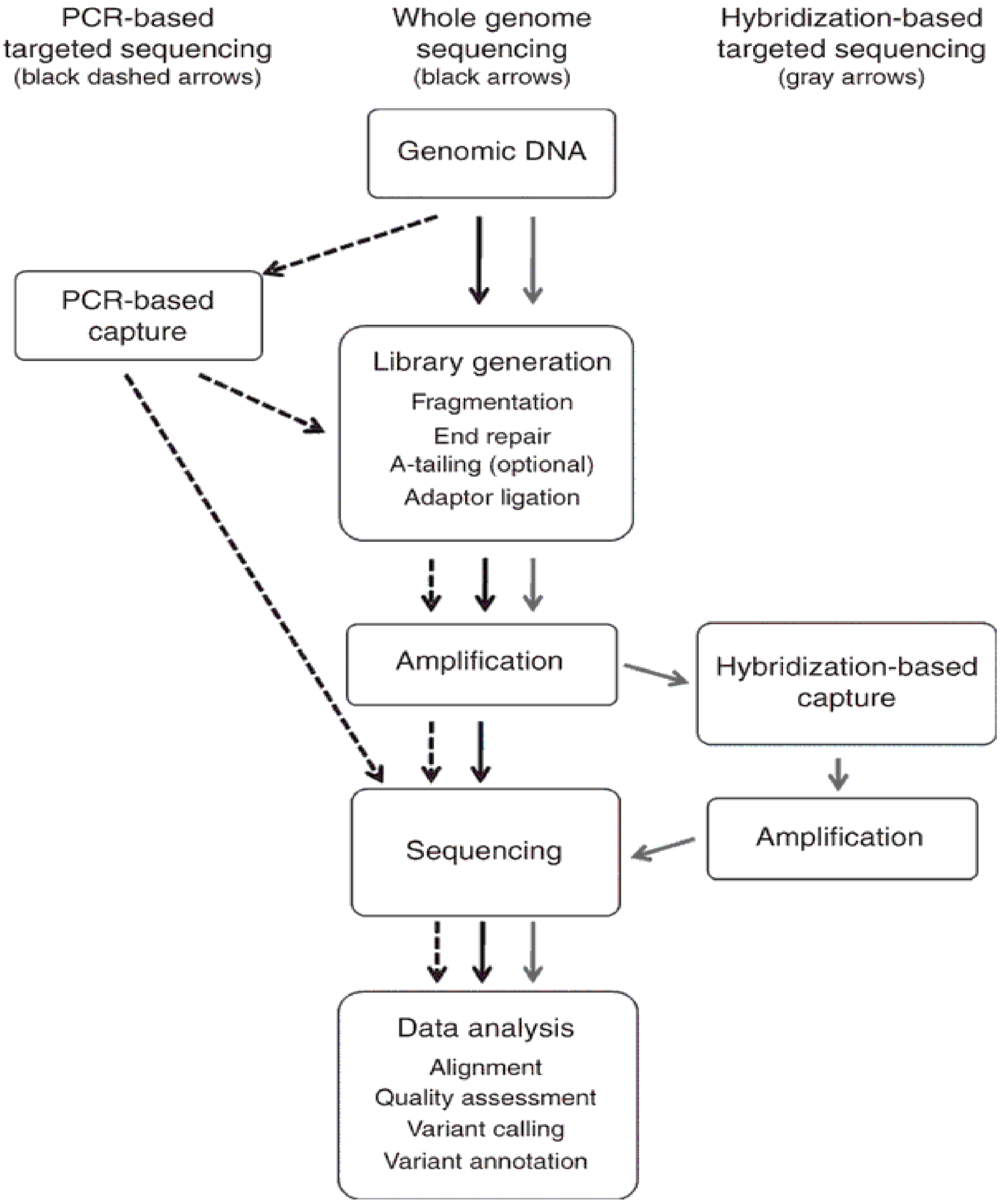


Figure 1: NGS overview (Rehm *et al*., 2013)

3.3 Platform Design*:*

Data used in this study is shown in the table below that is downloaded from TCGA portal website (https://tcga-data.nci.nih.gov/tcga/tcgaPlatformDesign.jsp)

Platform Center	TCGA Platform Code	Platform Name
Broad.mit.edu	ILLUMINA GA_DNASeq	Illumina Genome Analyzer DNA Sequencing

Table 2: Data center

3.4 The Cancer Genome Atlas (TCGA):

The Cancer Genome Atlas stores 34 cancer types high level data that is generated by the various laboratories like Broad Institute, HGSC, WUSTL. These laboratories generate data from many samples of cancer patient in the form of clinical information, CNV(SNP arrays), DNA methylated data, micro RNA-seq data, somatic mutations, expression protein data, RNA seqV2 data, CNV(low Pass DNASeq), protected mutations data, copy number data. The data is present in different types of levels like Level 1, Level 2, Level 3 and each level has its own meaning when it comes to its data type. This data is present for tumor samples and also for normal samples. TCGA has unique ID for the samples.

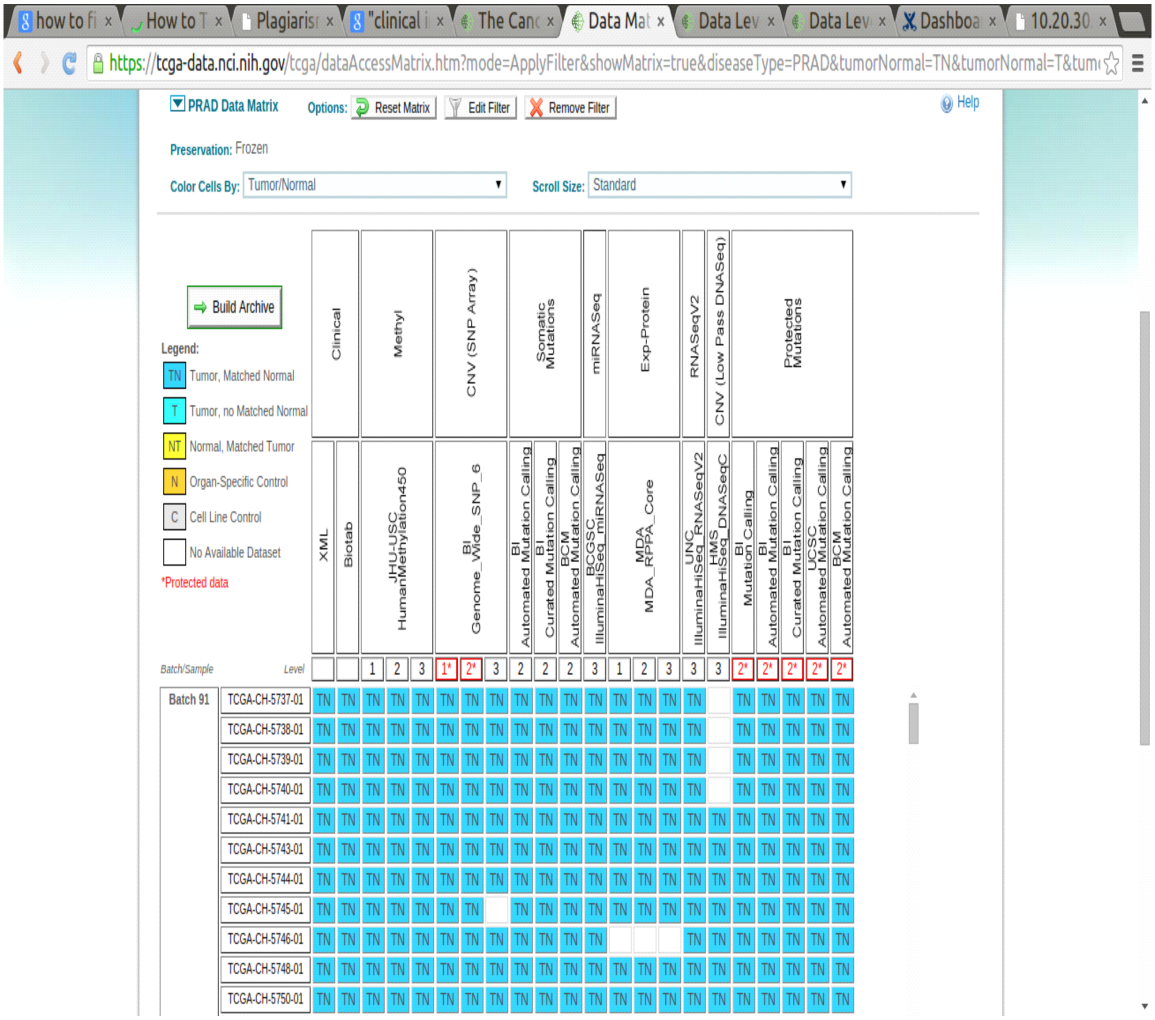
In this study somatic mutation data is used that is downloaded from TCGA. Detail of data is below in the table.

TCGA provides data free available if it is not downloading able then you are not allowed to access that data because this data is protected. To access this data you need special permissions.

In TCGA data we have special bar codes for each sample. The specification of these bar codes is available at https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode. Data can be downloaded for tumor and normal type samples. In this study I download **Level 2** data of prostate adenocarcinoma (PRAD) of **17 tumor matched with normal samples of batch 184**. Level 2 data is preprocessed data that is available on TCGA data portal. Level 3 data is not available as I mentioned it in the above table.

Data Types	*Cancer types applicable*	*Data type Name*	*Level 1*	*Level 2*	*Level 3*	*Important Metadata*	*How to retrieve data*
Mutations	*All*	*Somatic mutation*	*Whole exome sequence*	*Somatic mutation for each participant*	*n/a*	*The mutation data do not have a standard MAGE-TAB archive associated with them yet. The latest mutated file specifications areavailable on the wiki.*	*Data Matrix &Bulk Download: Select 'Somatic Mutations for Data Type* *File Search: Select 'DNA Mutations for Data Category* *Publication MAF Search*

Table 2*:* Data types and data levels at TCGA for Exome of various cancers



PRAD Data Matrix

Preservation: Frozen

Color Cells By: Tumor/Normal

Scroll Size: Standard

Legend:

- TN Tumor, Matched Normal
- T Tumor, no Matched Normal
- NT Normal, Matched Tumor
- N Organ-Specific Control
- C Cell Line Control
- □ No Available Dataset

*Protected data

		Clinical		Methyl			CNV (SNP Array)			Somatic Mutations			miRNASeq	Exp-Protein			RNASeqV2	CNV (Low Pass DNASeq)	Protected Mutations				
		XML	Biotab	JHU-USC HumanMethylation450			BI Genome_Wide_SNP_6			BI Automated Mutation Calling	BI Curated Mutation Calling	BCM Automated Mutation Calling	BCGSC IlluminaHiSeq_miRNASeq	MDA MDA_RPPA_Core			UNC IlluminaHiSeq_RNASeqV2	HMS IlluminaHiSeq_DNASeqC	BI Mutation Calling	BI Automated Mutation Calling	BI Curated Mutation Calling	UCSC Automated Mutation Calling	BCM Automated Mutation Calling
Batch/Sample	Level			1	2	3	1*	2*	3	2	2	2	3	1	2	3	3	3	2*	2*	2*	2*	2*
Batch 91	TCGA-CH-5737-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN	TN
	TCGA-CH-5738-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN	TN
	TCGA-CH-5739-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN	TN
	TCGA-CH-5740-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN	TN
	TCGA-CH-5741-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5743-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5744-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5745-01	TN	TN	TN	TN	TN	TN	TN		TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5746-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN				TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5748-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN
	TCGA-CH-5750-01	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN	TN

Figure 2: Data portal web face

3.5 Analysis of exome level 2 data:

Exome Level 2 data of consists of the variations found in the exome region after preprocessing of the raw data and then match this data with reference genome. Then to find out or prioritizing the genes responsible for a particular disease here in this study DisGeNEt database and ANNOVAR package is used. After that annotation of these genes is also done using ANNOVAR package. Perl language is used to analyze level 2 data using ANNOVAR package and also perl scripts are used to covert maf files into vcf formats. EXCEL tools also used to filter genes from DisGeNET database.

ANNOVAR package is used to annotate functions of the variations up to date. Annovar provides various databases like refgene, cytoband, genomicSuperDups, Cosmic68, 1000g2012apr, esp6500, phastConslelements46way, snp138, ljb26.

Annovar package have perl scripts that uses these databases and helps in filtering those alterations that are responsible for the study of diseases. PhastContrastelements46way database helps in finding those alterations that fall in conserved regions. These variations are most likely responsible for disease. Then genomicSuperDups databases are used to filter out duplicated variations and it filtered our gene list further. Further databases are used to annotate these filtered genes.

3.6 Prostate adenocarcinoma

Prostate cancer is multi-factorial complex disease. Prostate cancer is more prevalent in North America, Australia, Europe and New Zealand (Jemal *et al.*, 2010). About 250,000 patients detected positively for prostate cancer in United States of America alone in 2015. However prostate cancer rate is slow in Asian countries. In India prostate cancer occurrence rate is very low 3.3/100,000(Shen *et al.*, 2010). In Indian cities Delhi, Pune, Kolkata and Thiruvananthapuram prostate cancer is the second leading cancer. Mumbai and Bangaluru are the third leading sites. There is no standard clinical practice for prostate cancer yet however TP53(Balmukhanov *et al.*, 2013), GST family(Ecke *et al.*, 2010), CPY19A1(Kanda *et al.*, 2015), PTEN and AR genes are probable genes responsible for prostate cancer.

3.7 Pipeline used in current study:

(a) Download TCGA data of exome in **maf** format

(b) Convert maf in to **vcf** format

(c) Then map these vcf files with the **DisGeNET** database

(d) Convert vcf files in to **avinput** format

(e) Filter genes by **phastConselements46way** database

(f) Further filtering of genes with **genomicSuperdups** database

(g) Validation of these filtered genes with **cosmics** database

(h) Gene annotation of these filtered genes

(I) In the end get the cytobands for the genes

METHODOLOGY

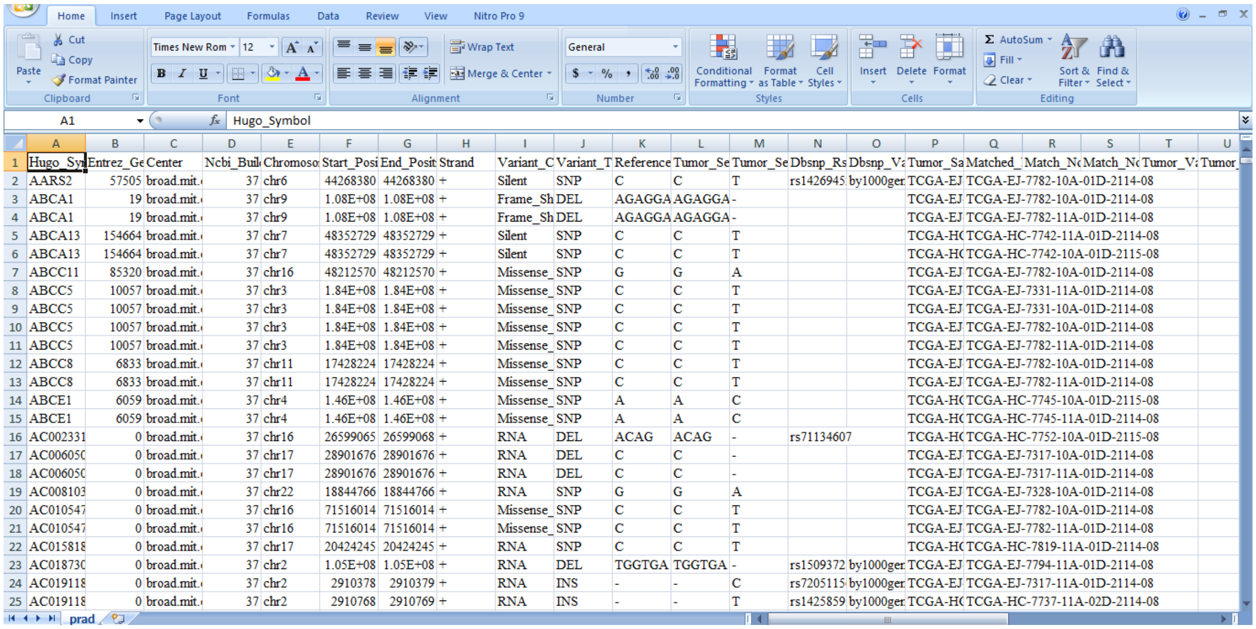
4.1 Data Retrieval

Exome Data downloaded from TCGA data portal in the form of Somatic mutations that are in the form of BI Automated Mutation Calling for 17 same of tumor matched with normal provided with their TCGA IDs for Prostate Adenocarcinoma (PRAD).

Data Type	**Exome sequencing**
Level	**2**
Center/Platform	**Illumina Genome Analyzer DNA Sequencing**
Batch	**184**
Disease	**PRAD**

Table 3: Data Detail for Prostate Adenocarcinoma (PRAD)

Data downloaded in the maf format as look below



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Hugo_Sy	Entrez_Ge	Center	Ncbi_Buil	Chromoso	Start_Posi	End_Posit	Strand	Variant_C	Variant_T	Reference	Tumor_Se	Tumor_Se	Dbsnp_Rs	Dbsnp_Va	Tumor_Sa	Matched_	Match_N	Match_N	Tumor_Va	Tumor_
2	AARS2	57505	broad.mit.	37	chr6	44268380	44268380	+	Silent	SNP	C	C	T	rs1426945	by1000ger	TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
3	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
4	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
5	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-HC	TCGA-HC-7742-11A-01D-2114-08				
6	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-HC	TCGA-HC-7742-10A-01D-2115-08				
7	ABCC11	85320	broad.mit.	37	chr16	48212570	48212570	+	Missense_	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
8	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-11A-01D-2114-08				
9	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-10A-01D-2114-08				
10	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
11	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
12	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
13	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
14	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-HC	TCGA-HC-7745-10A-01D-2115-08				
15	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-HC	TCGA-HC-7745-11A-01D-2114-08				
16	AC002331	0	broad.mit.	37	chr16	26599065	26599068	+	RNA	DEL	ACAG	ACAG	-	rs71134607		TCGA-HC	TCGA-HC-7752-10A-01D-2115-08				
17	AC006050	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-10A-01D-2114-08				
18	AC006050	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
19	AC008103	0	broad.mit.	37	chr22	18844766	18844766	+	RNA	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7328-10A-01D-2114-08				
20	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
21	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
22	AC015818	0	broad.mit.	37	chr17	20424245	20424245	+	RNA	SNP	C	C	T			TCGA-HC	TCGA-HC-7819-11A-01D-2114-08				
23	AC018730	0	broad.mit.	37	chr2	1.05E+08	1.05E+08	+	RNA	DEL	TGGTGA	TGGTGA	-	rs1509372	by1000ger	TCGA-EJ	TCGA-EJ-7794-11A-01D-2114-08				
24	AC019118	0	broad.mit.	37	chr2	2910378	2910379	+	RNA	INS	-	-	C	rs7205115	by1000ger	TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
25	AC019118	0	broad.mit.	37	chr2	2910768	2910769	+	RNA	INS	-	-	T	rs1425859	by1000ger	TCGA-HC	TCGA-HC-7737-11A-02D-2114-08				

Figure 3: maf file

4.2 Download Human Reference Genome hg19

Download the Human Reference Genome from the link (http://hgdownload.cse.ucsc.edu/downloads.html#human). Reference genome downloaded in the form of **bit** format.

To convert **bit** format into **fasta** format I use **twobittofa** utility tool of UCSC genome browser.

4.3 Converting maf files in to vcf files

Converting maf file into vcf files a perl script **maf2vcf.pl** is developed by ckandoth license to apache2. In this script a slight change is done manually by changing the location of the reference genome by giving the reference genome path that is in the user system. Then run the command.

4.4 Download the ANNOVAR software

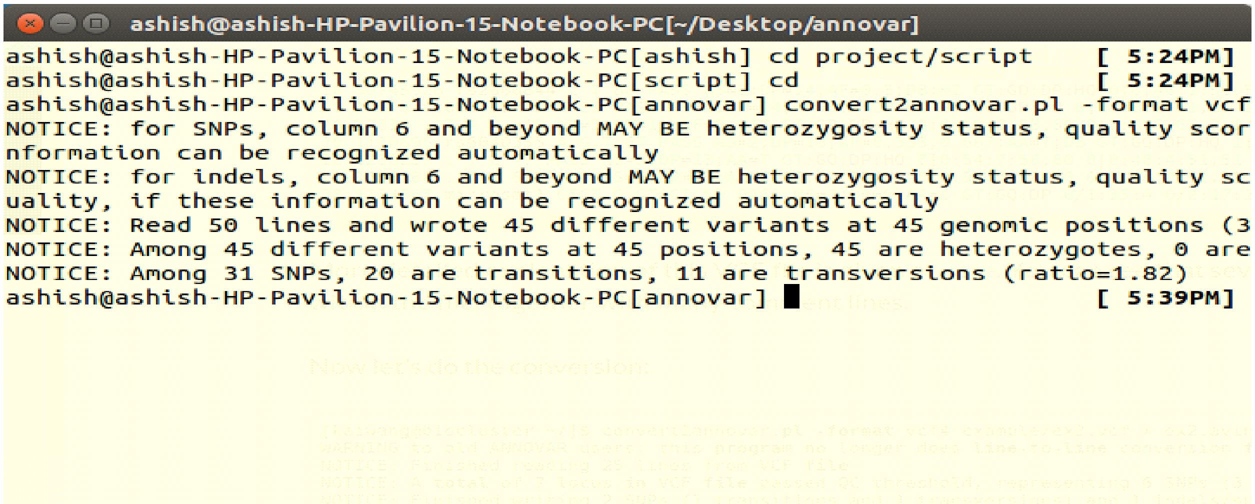
ANNOVAR software is downloaded from the annovar web site but to download this software we have to register first and also needed an academic mail id. After registration a link is send to the mail id from where annovar can be easily downloaded.

4.5 Download DISGENET database

4.6 Downloaded the required databases that are needed in Annovar software

4.7 Convert the vcf files into .avinput format

Annovar uses only .avinput files so convert the vcf files into .avinput format by using convert2annovar.pl

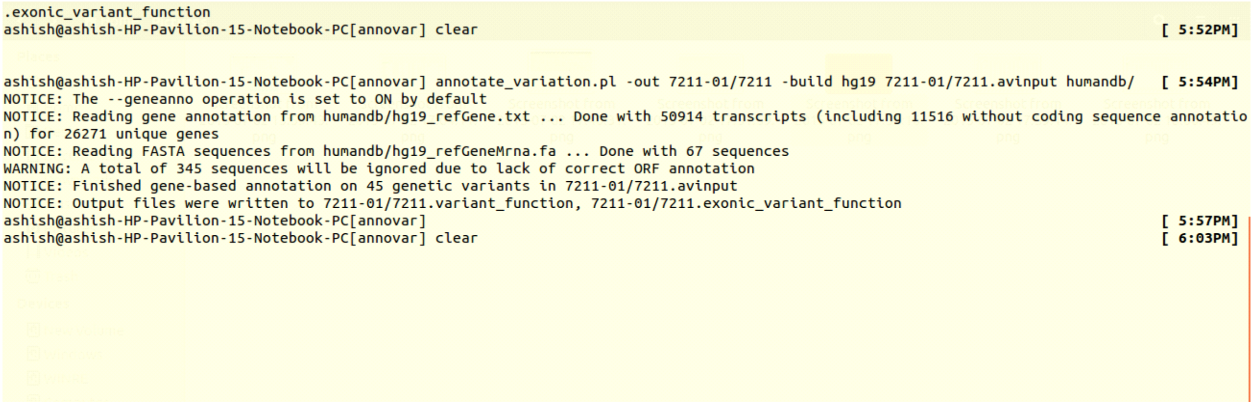


```
ashish@ashish-HP-Pavilion-15-Notebook-PC[ashish] cd project/script    [ 5:24PM]
ashish@ashish-HP-Pavilion-15-Notebook-PC[script] cd                   [ 5:24PM]
ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar] convert2annovar.pl -format vcf
NOTICE: for SNPs, column 6 and beyond MAY BE heterozygosity status, quality scor
nformation can be recognized automatically
NOTICE: for indels, column 6 and beyond MAY BE heterozygosity status, quality sc
uality, if these information can be recognized automatically
NOTICE: Read 50 lines and wrote 45 different variants at 45 genomic positions (3
NOTICE: Among 45 different variants at 45 positions, 45 are heterozygotes, 0 are
NOTICE: Among 31 SNPs, 20 are transitions, 11 are transversions (ratio=1.82)
ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar] █                   [ 5:39PM]
```

Figure 4 : convert2annovar.pl

4.8 Gene based annotation of .avinput for extracting the names of the genes

To get the names of the genes that are filtered after conversion of maf files into vcf files is done by annotate_variation.pl



```
.exonic_variant_function
ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar] clear                                                     [ 5:52PM]

ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar] annotate_variation.pl -out 7211-01/7211 -build hg19 7211-01/7211.avinput humandb/   [ 5:54PM]
NOTICE: The --geneanno operation is set to ON by default
NOTICE: Reading gene annotation from humandb/hg19_refGene.txt ... Done with 50914 transcripts (including 11516 without coding sequence annotatio
n) for 26271 unique genes
NOTICE: Reading FASTA sequences from humandb/hg19_refGeneMrna.fa ... Done with 67 sequences
WARNING: A total of 345 sequences will be ignored due to lack of correct ORF annotation
NOTICE: Finished gene-based annotation on 45 genetic variants in 7211-01/7211.avinput
NOTICE: Output files were written to 7211-01/7211.variant_function, 7211-01/7211.exonic_variant_function
ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar]                                                           [ 5:57PM]
ashish@ashish-HP-Pavilion-15-Notebook-PC[annovar] clear                                                     [ 6:03PM]
```

Figure 6: .avinput script format

Now run the same scripts for all 17 vcf files and then get the all genes name in the excel sheet.
Then mapping vcf files genes with the DisGeNET database.
This script annotate_variation.pl -out 7211-01/7211 -build hg19 7211-01/7211.avinput humandb/ output two files 7211-01/7211.variant_function and 7211-01/7211.exonic_variant_function.
After this step mapping .variant_function files genes name of the entire 17 sample we got filtered genes for PRAD.

4.9 Filtered variations in conserved region

To get the variation in conserved regions we have region based annotation to use the PhastConselements46way database which gives score and we filtered the variation above 400 threshold. This can be done by annotate_variation.pl -regionanno -build hg19 -out 7211-01/7211 -dbtype phastconselements46way 7211-01/7211.avinput humandb/.

4.10 Segmented duplicated variations filtered

This can be done first by downloading the genomicSuperDups database and then run the following perl script annotate_variation.pl -regionanno -build hg19 -out 7211-01/7211-dbype genomicSuperDups 7211-01/7211.avinput huamndb.
After this step I filtered out more genes from the above filtered gene (conserved region) by removing the variations from genomicSuperDups database.

4.11 Further filtration is done by polyphen2 annotation

Download HVAR database and then map the .avinput files with this database by annotate_variation.pl -filter -dbtype ljb23_pp2hvar -buildver hg19 -out 7211-01/7211 7211/7211.avinput humandb.

Probably damaging	= .909 – 1
Possibly damaging	= .447 - .908
Benign	= 0 - .446

4.14 Mapping with cosmic database

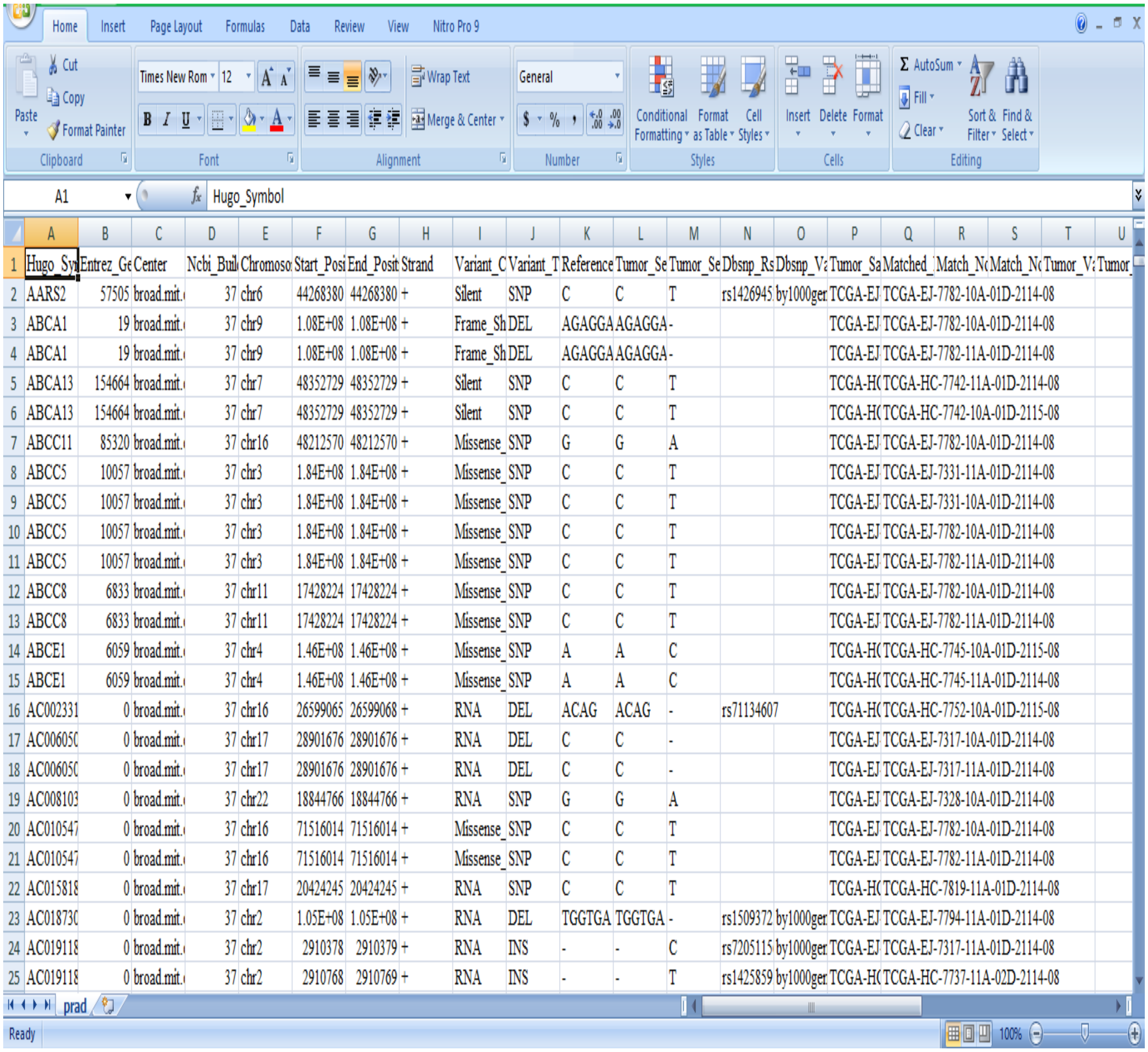
Cosmic database is Catalogue of Somatic mutation in cancers. Annovar uses this database and tell us that these mutations occur in cancer that is present in literature and how many times.

4.15 Getting cytobands

To get the location and the bands of variation download the cytoband database. Then uses the perl script annotate_variation.pl -regionanno -build hg19 -out 7211-01/7211 -dbtype cytoband 7211-01/7211.avinput humandb.

RESULTS

5.1 Converting maf files into vcf files



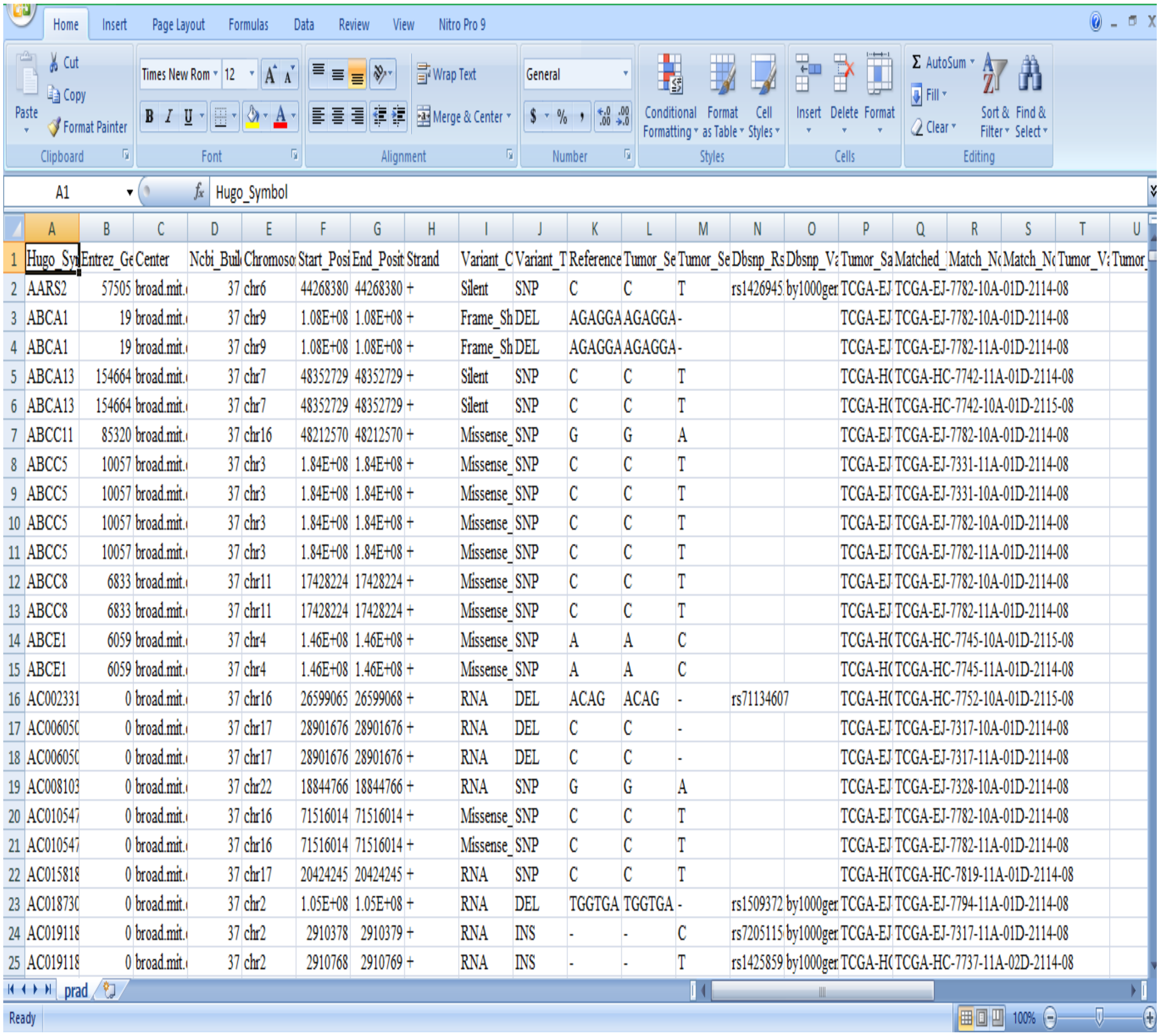
A1 fx Hugo_Symbol

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Hugo_Sy	Entrez_Ge	Center	Ncbi_Buil	Chromoso	Start_Posi	End_Posit	Strand	Variant_C	Variant_T	Reference	Tumor_Se	Tumor_Se	Dbsnp_Rs	Dbsnp_Va	Tumor_Sa	Matched_	Match_No	Match_No	Tumor_Va	Tumor_
2	AARS2	57505	broad.mit.	37	chr6	44268380	44268380	+	Silent	SNP	C	C	T	rs1426945	by1000ger	TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
3	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
4	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
5	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-H(	TCGA-HC-7742-11A-01D-2114-08				
6	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-H(	TCGA-HC-7742-10A-01D-2115-08				
7	ABCC11	85320	broad.mit.	37	chr16	48212570	48212570	+	Missense_	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
8	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-11A-01D-2114-08				
9	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-10A-01D-2114-08				
10	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
11	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
12	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
13	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
14	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-H(	TCGA-HC-7745-10A-01D-2115-08				
15	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-H(	TCGA-HC-7745-11A-01D-2114-08				
16	AC002331	0	broad.mit.	37	chr16	26599065	26599068	+	RNA	DEL	ACAG	ACAG	-	rs71134607		TCGA-H(	TCGA-HC-7752-10A-01D-2115-08				
17	AC00605(	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-10A-01D-2114-08				
18	AC00605(	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
19	AC00810	0	broad.mit.	37	chr22	18844766	18844766	+	RNA	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7328-10A-01D-2114-08				
20	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
21	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
22	AC015818	0	broad.mit.	37	chr17	20424245	20424245	+	RNA	SNP	C	C	T			TCGA-H(	TCGA-HC-7819-11A-01D-2114-08				
23	AC01873(	0	broad.mit.	37	chr2	1.05E+08	1.05E+08	+	RNA	DEL	TGGTGA	TGGTGA	-	rs1509372	by1000ger	TCGA-EJ	TCGA-EJ-7794-11A-01D-2114-08				
24	AC019118	0	broad.mit.	37	chr2	2910378	2910379	+	RNA	INS	-	-	C	rs7205115	by1000ger	TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
25	AC019118	0	broad.mit.	37	chr2	2910768	2910769	+	RNA	INS	-	-	T	rs1425859	by1000ger	TCGA-H(	TCGA-HC-7737-11A-02D-2114-08				

Figure 6: vcf file

5.2 Converting vcf files into .avinput files

Annovar needs .avinput files for processing so we have to convert vcf into .avinput format.

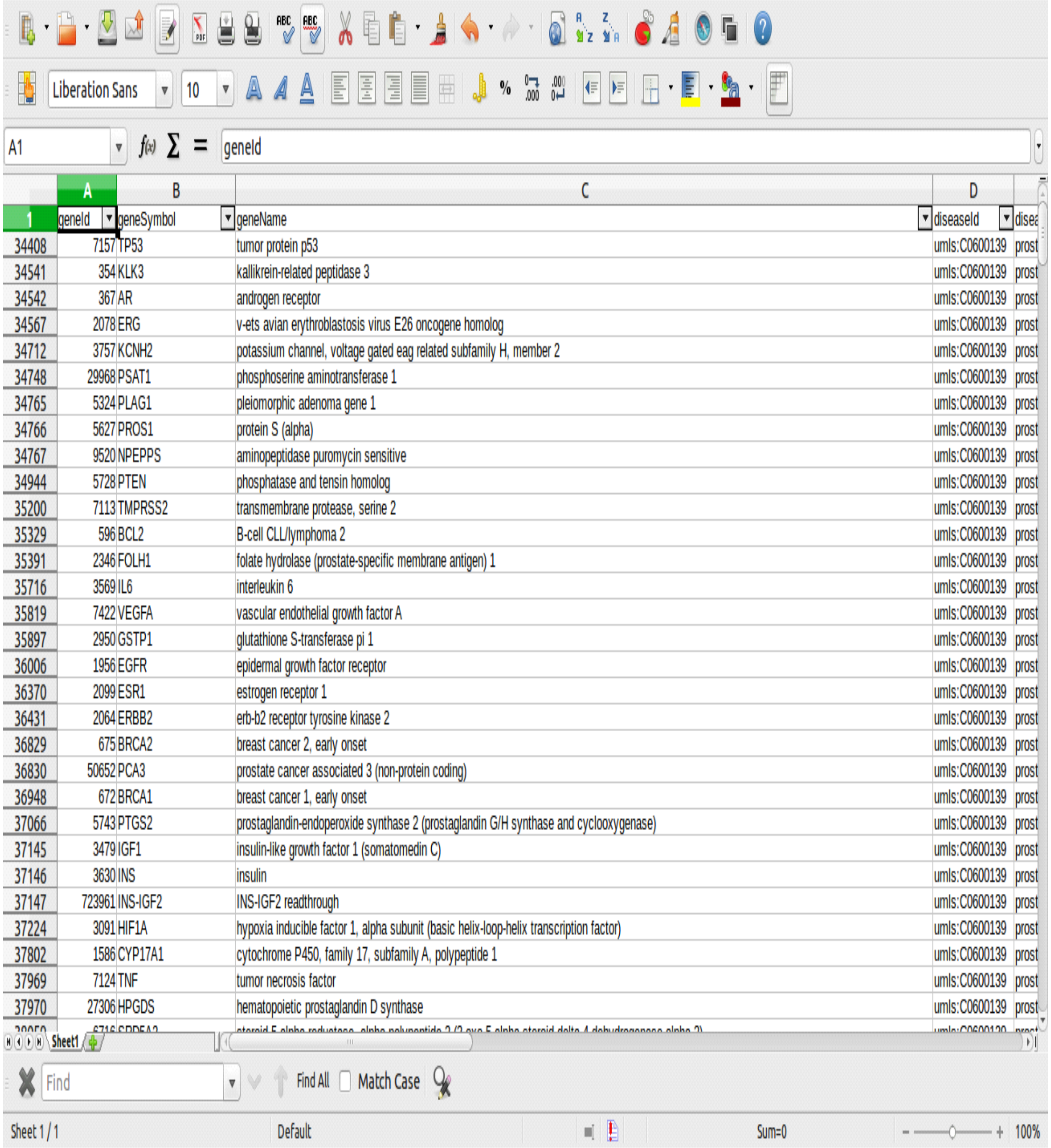


A1 | Hugo_Symbol

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Hugo_Sy	Entrez_Ge	Center	Ncbi_Buil	Chromoso	Start_Posi	End_Posit	Strand	Variant_C	Variant_T	Reference	Tumor_Se	Tumor_Se	Dbsnp_Rs	Dbsnp_V	Tumor_Sa	Matched_	Match_N	Match_N	Tumor_V	Tumor_
2	AARS2	57505	broad.mit.	37	chr6	44268380	44268380	+	Silent	SNP	C	C	T	rs1426945	by1000ger	TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
3	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
4	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
5	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-H(	TCGA-HC-7742-11A-01D-2114-08				
6	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-H(	TCGA-HC-7742-10A-01D-2115-08				
7	ABCC11	85320	broad.mit.	37	chr16	48212570	48212570	+	Missense_	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
8	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-11A-01D-2114-08				
9	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-10A-01D-2114-08				
10	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
11	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
12	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
13	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
14	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-H(	TCGA-HC-7745-10A-01D-2115-08				
15	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-H(	TCGA-HC-7745-11A-01D-2114-08				
16	AC002331	0	broad.mit.	37	chr16	26599065	26599068	+	RNA	DEL	ACAG	ACAG	-	rs71134607		TCGA-H(	TCGA-HC-7752-10A-01D-2115-08				
17	AC00605(	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-10A-01D-2114-08				
18	AC00605(	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
19	AC00810:	0	broad.mit.	37	chr22	18844766	18844766	+	RNA	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7328-10A-01D-2114-08				
20	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
21	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
22	AC015818	0	broad.mit.	37	chr17	20424245	20424245	+	RNA	SNP	C	C	T			TCGA-H(	TCGA-HC-7819-11A-01D-2114-08				
23	AC01873(	0	broad.mit.	37	chr2	1.05E+08	1.05E+08	+	RNA	DEL	TGGTGA	TGGTGA	-	rs1509372	by1000ger	TCGA-EJ	TCGA-EJ-7794-11A-01D-2114-08				
24	AC019118	0	broad.mit.	37	chr2	2910378	2910379	+	RNA	INS	-	-	C	rs7205115	by1000ger	TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
25	AC019118	0	broad.mit.	37	chr2	2910768	2910769	+	RNA	INS	-	-	T	rs1425859	by1000ger	TCGA-H(	TCGA-HC-7737-11A-02D-2114-08				

Figure 7: .avinput file

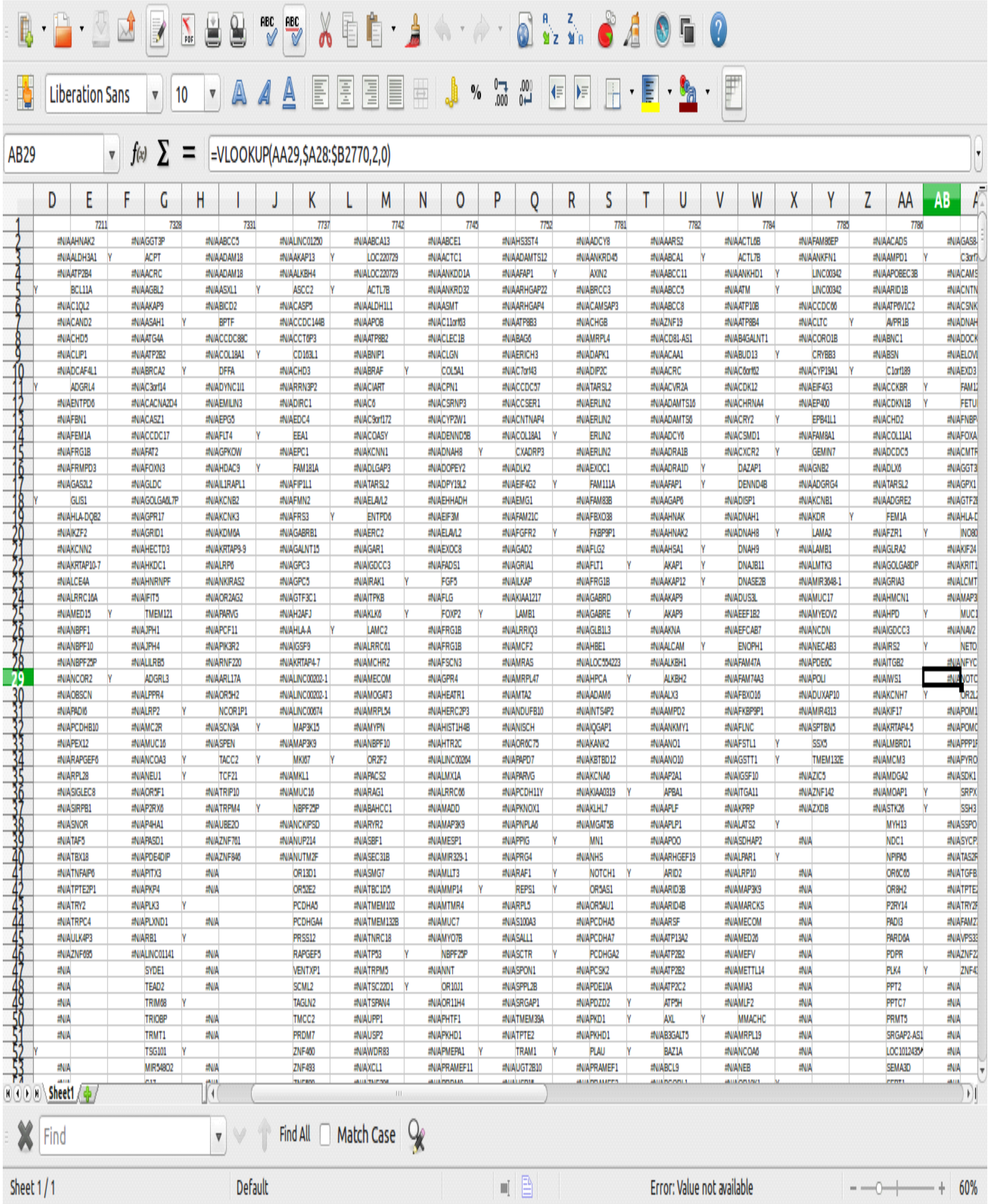
5.3 DisGeNET database for Prostate Cancer



	A	B	C	D	
1	geneId	geneSymbol	geneName	diseaseId	disea
34408	7157	TP53	tumor protein p53	umls:C0600139	prost
34541	354	KLK3	kallikrein-related peptidase 3	umls:C0600139	prost
34542	367	AR	androgen receptor	umls:C0600139	prost
34567	2078	ERG	v-ets avian erythroblastosis virus E26 oncogene homolog	umls:C0600139	prost
34712	3757	KCNH2	potassium channel, voltage gated eag related subfamily H, member 2	umls:C0600139	prost
34748	29968	PSAT1	phosphoserine aminotransferase 1	umls:C0600139	prost
34765	5324	PLAG1	pleiomorphic adenoma gene 1	umls:C0600139	prost
34766	5627	PROS1	protein S (alpha)	umls:C0600139	prost
34767	9520	NPEPPS	aminopeptidase puromycin sensitive	umls:C0600139	prost
34944	5728	PTEN	phosphatase and tensin homolog	umls:C0600139	prost
35200	7113	TMPRSS2	transmembrane protease, serine 2	umls:C0600139	prost
35329	596	BCL2	B-cell CLL/lymphoma 2	umls:C0600139	prost
35391	2346	FOLH1	folate hydrolase (prostate-specific membrane antigen) 1	umls:C0600139	prost
35716	3569	IL6	interleukin 6	umls:C0600139	prost
35819	7422	VEGFA	vascular endothelial growth factor A	umls:C0600139	prost
35897	2950	GSTP1	glutathione S-transferase pi 1	umls:C0600139	prost
36006	1956	EGFR	epidermal growth factor receptor	umls:C0600139	prost
36370	2099	ESR1	estrogen receptor 1	umls:C0600139	prost
36431	2064	ERBB2	erb-b2 receptor tyrosine kinase 2	umls:C0600139	prost
36829	675	BRCA2	breast cancer 2, early onset	umls:C0600139	prost
36830	50652	PCA3	prostate cancer associated 3 (non-protein coding)	umls:C0600139	prost
36948	672	BRCA1	breast cancer 1, early onset	umls:C0600139	prost
37066	5743	PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	umls:C0600139	prost
37145	3479	IGF1	insulin-like growth factor 1 (somatomedin C)	umls:C0600139	prost
37146	3630	INS	insulin	umls:C0600139	prost
37147	723961	INS-IGF2	INS-IGF2 readthrough	umls:C0600139	prost
37224	3091	HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	umls:C0600139	prost
37802	1586	CYP17A1	cytochrome P450, family 17, subfamily A, polypeptide 1	umls:C0600139	prost
37969	7124	TNF	tumor necrosis factor	umls:C0600139	prost
37970	27306	HPGDS	hematopoietic prostaglandin D synthase	umls:C0600139	prost
38050	6716	SRD5A2	steroid-5-alpha-reductase, alpha polypeptide 2 (3-oxo-5 alpha-steroid delta 4-dehydrogenase alpha 2)	umls:C0600139	prost

Figure 8: DisGeNET database

5.4 Filtering genes after mapping with DisGeNET database



AB29 =VLOOKUP(AA29,$A28:$B2770,2,0)

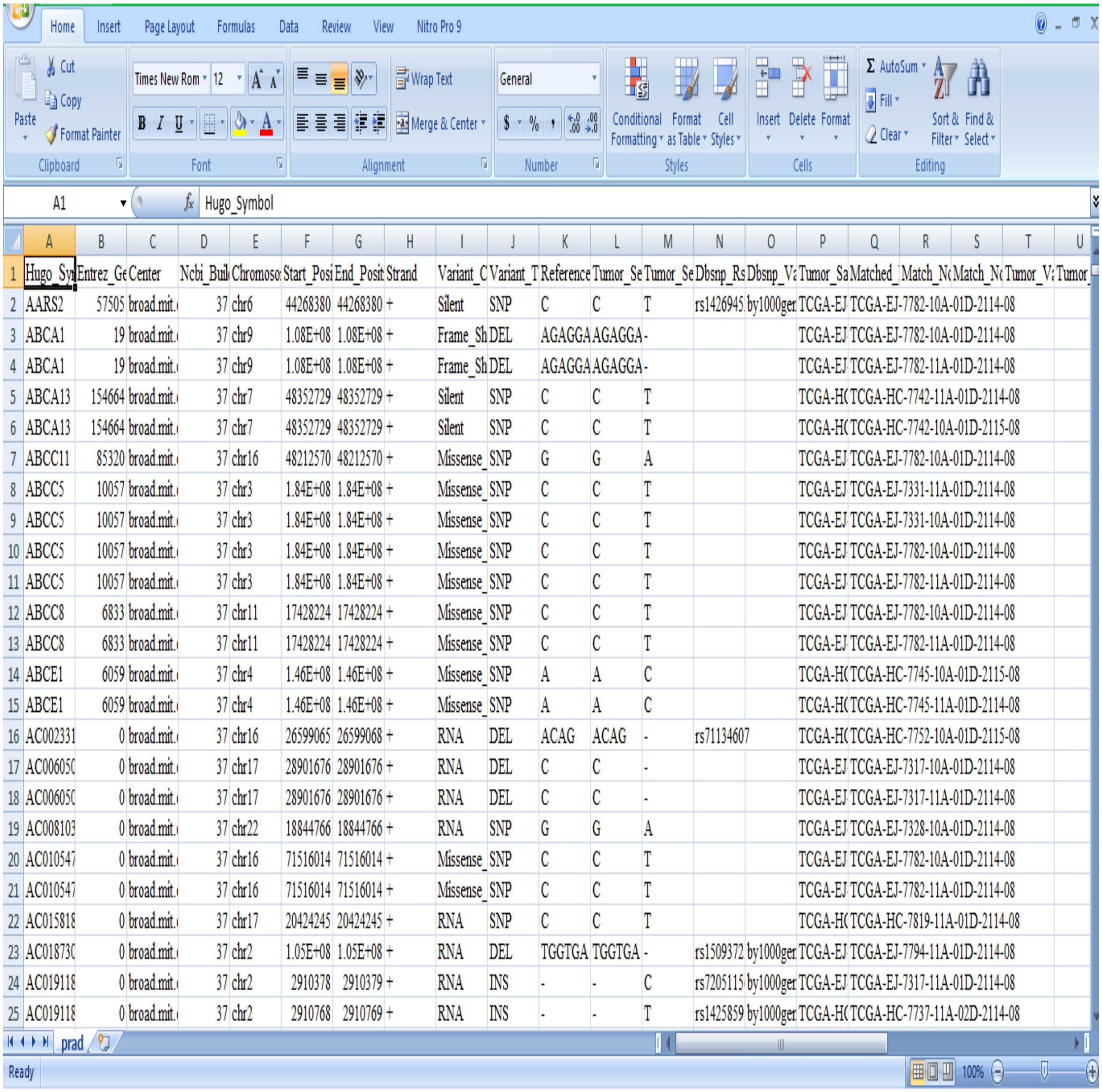
	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1		7211		7328		7331		7737		7742		7745		7752		7781		7782		7784		7785		7786		
2	#N/A	AHNAK2	#N/A	GGT3P	#N/A	ABCC5	#N/A	LINC01250	#N/A	ABCA13	#N/A	ABCE1	#N/A	HS3ST4	#N/A	ADCY8	#N/A	AARS2	#N/A	ACTL6B	#N/A	FAM86EP	#N/A	ACADS	#N/A	GAS8-
3	#N/A	ALDH3A1	Y	ACPT	#N/A	ADAM18	#N/A	AKAP13	Y	LOC220729	#N/A	ACTC1	#N/A	ADAMTS12	#N/A	ANKRD45	#N/A	ABCA1	Y	ACTL7B	#N/A	ANKFN1	#N/A	AMPD1	Y	C3orf7
4	#N/A	ATP2B4	#N/A	ACRC	#N/A	ADAM18	#N/A	ALKBH4	#N/A	LOC220729	#N/A	ANKDD1A	#N/A	AFAP1	Y	AXIN2	#N/A	ABCC11	#N/A	ANKHD1	Y	LINC00342	#N/A	APOBEC3B	#N/A	CAMS
5	Y	BCL11A	#N/A	AGBL2	#N/A	ASXL1	Y	ASCC2	Y	ACTL7B	#N/A	ANKRD32	#N/A	ARHGAP22	#N/A	BRCC3	#N/A	ABCC5	#N/A	ATM	Y	LINC00342	#N/A	ARID1B	#N/A	CNTN
6	#N/A	C1QL2	#N/A	AKAP9	#N/A	BICD2	#N/A	CASP5	#N/A	ALDH1L1	#N/A	ASMT	#N/A	ARHGAP4	#N/A	CAMSAP3	#N/A	ABCC8	#N/A	ATP10B	#N/A	CCDC66	#N/A	ATP6V1C2	#N/A	CSNK
7	#N/A	CAND2	#N/A	ASAH1	Y	BPTF	#N/A	CCDC144B	#N/A	APOB	#N/A	C11orf63	#N/A	ATP8B3	#N/A	CHGB	#N/A	ZNF19	#N/A	ATP8B4	#N/A	CLTC	Y	AVPR1B	#N/A	DNAH
8	#N/A	CHD5	#N/A	ATG4A	#N/A	CCDC88C	#N/A	CCT6P3	#N/A	ATP8B2	#N/A	CLEC1B	#N/A	BAG6	#N/A	MRPL4	#N/A	CD81-AS1	#N/A	B4GALNT1	#N/A	CORO1B	#N/A	BNC1	#N/A	DOCK
9	#N/A	CLIP1	#N/A	ATP2B2	#N/A	COL18A1	Y	CD163L1	#N/A	BNIP1	#N/A	CLGN	#N/A	ERICH3	#N/A	DAPK1	#N/A	ACAA1	#N/A	BUD13	Y	CRYBB3	#N/A	BSN	#N/A	ELOV
10	#N/A	DCAF4L1	#N/A	BRCA2	Y	DFFA	#N/A	CHD3	#N/A	BRAF	Y	COL5A1	#N/A	C7orf43	#N/A	DIP2C	#N/A	ACRC	#N/A	C6orf62	#N/A	CYP19A1	Y	C1orf189	#N/A	EXD3
11	Y	ADGRL4	#N/A	C3orf14	#N/A	DYNC1I1	#N/A	RRN3P2	#N/A	CIART	#N/A	CPN1	#N/A	CCDC57	#N/A	TARSL2	#N/A	ACVR2A	#N/A	CDK12	#N/A	EIF4G3	#N/A	CCKBR	Y	FAM1
12	#N/A	ENTPD6	#N/A	CACNA2D4	#N/A	EMILIN3	#N/A	DIRC1	#N/A	C6	#N/A	CSRNP3	#N/A	CCSER1	#N/A	ERLIN2	#N/A	ADAMTS16	#N/A	CHRNA4	#N/A	EP400	#N/A	CDKN1B	Y	FETU
13	#N/A	FBN1	#N/A	CASZ1	#N/A	EPG5	#N/A	EDC4	#N/A	C9orf172	#N/A	CYP2W1	#N/A	CNTNAP4	#N/A	ERLIN2	#N/A	ADAMTS6	#N/A	CRY2	Y	EPB41L1	#N/A	CHD2	#N/A	FNBP
14	#N/A	FEM1A	#N/A	CCDC17	#N/A	FLT4	Y	EEA1	#N/A	COASY	#N/A	DENND5B	#N/A	COL18A1	Y	ERLIN2	#N/A	ADCY6	#N/A	CSMD1	#N/A	FAM8A1	#N/A	COL11A1	#N/A	FOXA
15	#N/A	FRG1B	#N/A	FAT2	#N/A	GPKOW	#N/A	EPC1	#N/A	KCNN1	#N/A	DNAH8	Y	CXADRP3	#N/A	ERLIN2	#N/A	ADRA1B	#N/A	CXCR2	Y	GEMIN7	#N/A	DCDC5	#N/A	CMTR
16	#N/A	FRMPD3	#N/A	FOXN3	#N/A	HDAC9	Y	FAM181A	#N/A	DLGAP3	#N/A	DOPEY2	#N/A	DLK2	#N/A	EXOC1	#N/A	ADRA1D	Y	DAZAP1	#N/A	GNB2	#N/A	DLX6	#N/A	GGT3
17	#N/A	GAS2L2	#N/A	GLDC	#N/A	IL1RAPL1	#N/A	FIP1L1	#N/A	TARSL2	#N/A	DPY19L2	#N/A	EIF4G2	Y	FAM111A	#N/A	AFAP1	Y	DENND4B	#N/A	ADGRG4	#N/A	TARSL2	#N/A	GPX1
18	Y	GLIS1	#N/A	GOLGA6L7P	#N/A	KCNB2	#N/A	FMN2	#N/A	ELAVL2	#N/A	EHHADH	#N/A	EMG1	#N/A	FAM83B	#N/A	AGAP6	#N/A	DISP1	#N/A	KCNB1	#N/A	ADGRE2	#N/A	GTF2
19	#N/A	HLA-DQB2	#N/A	GPR17	#N/A	KCNK3	#N/A	FRS3	Y	ENTPD6	#N/A	EIF3M	#N/A	FAM21C	#N/A	FBXO38	#N/A	AHNAK	#N/A	DNAH1	#N/A	KDR	Y	FEM1A	#N/A	HLA-D
20	#N/A	IKZF2	#N/A	GRID1	#N/A	KDM6A	#N/A	GABRB1	#N/A	ERC2	#N/A	ELAVL2	#N/A	FGFR2	Y	FKBP9P1	#N/A	AHNAK2	#N/A	DNAH8	Y	LAMA2	#N/A	FZR1	Y	INO80
21	#N/A	KCNN2	#N/A	HECTD3	#N/A	KRTAP9-9	#N/A	GALNT15	#N/A	GAR1	#N/A	EXOC8	#N/A	GAD2	#N/A	FLG2	#N/A	AHSA1	Y	DNAH9	#N/A	LAMB1	#N/A	GLRA2	#N/A	KIF24
22	#N/A	KRTAP10-7	#N/A	HKDC1	#N/A	LRP6	#N/A	GPC3	#N/A	IGDCC3	#N/A	FADS1	#N/A	GRIA1	#N/A	FLT1	Y	AKAP1	Y	DNAJB11	#N/A	LMTK3	#N/A	GOLGA8DP	#N/A	KRIT1
23	#N/A	LCE4A	#N/A	HNRNPF	#N/A	NKIRAS2	#N/A	GPC5	#N/A	IRAK1	Y	FGF5	#N/A	ILKAP	#N/A	FRG1B	#N/A	AKAP12	Y	DNASE2B	#N/A	MIR3648-1	#N/A	GRIA3	#N/A	LCMT
24	#N/A	LRRC16A	#N/A	IFIT5	#N/A	OR2AG2	#N/A	GTF3C1	#N/A	ITPKB	#N/A	FLG	#N/A	KIAA1217	#N/A	GABRD	#N/A	AKAP9	#N/A	DUS3L	#N/A	MUC17	#N/A	HMCN1	#N/A	MAP3
25	#N/A	MED15	Y	TMEM121	#N/A	PARVG	#N/A	H2AFJ	#N/A	KLK6	Y	FOXP2	Y	LAMB1	#N/A	GABRE	Y	AKAP9	#N/A	EEF1B2	#N/A	MYEOV2	#N/A	HPD	Y	MUC1
26	#N/A	NBPF1	#N/A	JPH1	#N/A	PCF11	#N/A	HLA-A	Y	LAMC2	#N/A	FRG1B	#N/A	LRRIQ3	#N/A	GLB1L3	#N/A	AKNA	#N/A	EFCAB7	#N/A	NCDN	#N/A	IGDCC3	#N/A	NAV2
27	#N/A	NBPF10	#N/A	JPH4	#N/A	PIK3R2	#N/A	IGSF9	#N/A	LRRC61	#N/A	FRG1B	#N/A	MCF2	#N/A	HBE1	#N/A	ALCAM	Y	ENOPH1	#N/A	NECAB3	#N/A	IRS2	Y	NETO
28	#N/A	NBPF25P	#N/A	LILRB5	#N/A	RNF220	#N/A	KRTAP4-7	#N/A	MCHR2	#N/A	FSCN3	#N/A	MRAS	#N/A	LOC554223	#N/A	ALKBH1	#N/A	FAM47A	#N/A	PDE6C	#N/A	ITGB2	#N/A	NFYC
29	#N/A	NCOR2	Y	ADGRL3	#N/A	ARL17A	#N/A	LINC00202-1	#N/A	MECOM	#N/A	GPR4	#N/A	MRPL47	#N/A	HPCA	Y	ALKBH2	#N/A	FAM74A3	#N/A	POLI	#N/A	IWS1	#N/A	NOTC
30	#N/A	OBSCN	#N/A	LPPR4	#N/A	OR5H2	#N/A	LINC00202-1	#N/A	MOGAT3	#N/A	HEATR1	#N/A	MTA2	#N/A	ADAM6	#N/A	ALX3	#N/A	FBXO16	#N/A	DUXAP10	#N/A	KCNH7	Y	OR2L
31	#N/A	PADI6	#N/A	LRP2	Y	NCOR1P1	#N/A	LINC00674	#N/A	MRPL54	#N/A	HERC2P3	#N/A	NDUFB10	#N/A	INTS4P2	#N/A	AMPD2	#N/A	FKBP9P1	#N/A	MIR4313	#N/A	KIF17	#N/A	POM1
32	#N/A	PCDHB10	#N/A	MC2R	#N/A	SCN9A	Y	MAP3K15	#N/A	MYPN	#N/A	HIST1H4B	#N/A	NISCH	#N/A	IQGAP1	#N/A	ANKMY1	#N/A	FLNC	#N/A	SPTBN5	#N/A	KRTAP4-5	#N/A	POMC
33	#N/A	PEX12	#N/A	MUC16	#N/A	SPEN	#N/A	MAP3K9	#N/A	NBPF10	#N/A	HTR2C	#N/A	OR6C75	#N/A	KANK2	#N/A	ANO1	#N/A	FSTL1	Y	SSX5	#N/A	LMBRD1	#N/A	PPP1R
34	#N/A	RAPGEF6	#N/A	NCOA3	Y	TACC2	Y	MKI67	Y	OR2F2	#N/A	LINC00264	#N/A	PAPD7	#N/A	KBTBD12	#N/A	ANO10	#N/A	GSTT1	Y	TMEM132E	#N/A	MCM3	#N/A	PYRO
35	#N/A	RPL28	#N/A	NEU1	Y	TCF21	#N/A	MKL1	#N/A	PACS2	#N/A	LMX1A	#N/A	PARVG	#N/A	KCNA6	#N/A	AP2A1	#N/A	IGSF10	#N/A	ZIC5	#N/A	MDGA2	#N/A	SDK1
36	#N/A	SIGLEC8	#N/A	OR5F1	#N/A	TRIP10	#N/A	MUC16	#N/A	RAG1	#N/A	LRRC66	#N/A	PCDH11Y	#N/A	KIAA0319	Y	APBA1	#N/A	ITGA11	#N/A	ZNF142	#N/A	MOAP1	Y	SRPX
37	#N/A	SIRPB1	#N/A	P2RX6	#N/A	TRPM4	Y	NBPF25P	#N/A	BAHCC1	#N/A	MADD	#N/A	PKNOX1	#N/A	KLHL7	#N/A	APLF	#N/A	KPRP	#N/A	ZXDB	#N/A	STK26	Y	SSH3
38	#N/A	SNOR	#N/A	P4HA1	#N/A	UBE2O	#N/A	NCKIPSD	#N/A	RYR2	#N/A	MAP3K9	#N/A	PNPLA6	#N/A	MGAT5B	#N/A	APLP1	#N/A	LATS2	Y			MYH13	#N/A	SSPO
39	#N/A	TAF5	#N/A	PASD1	#N/A	ZNF761	#N/A	NUP214	#N/A	SBF1	#N/A	MESP1	#N/A	PPIG	Y	MN1	#N/A	APOO	#N/A	SDHAP2	#N/A			NDC1	#N/A	SYCP
40	#N/A	TBX18	#N/A	PDE4DIP	#N/A	ZNF846	#N/A	NUTM2F	#N/A	SEC31B	#N/A	MIR329-1	#N/A	PRG4	#N/A	NHS	#N/A	ARHGEF19	#N/A	LPAR1	Y			NPIPA5	#N/A	TAS2R
41	#N/A	TNFAIP6	#N/A	PITX3	#N/A			OR13D1	#N/A	SMG7	#N/A	MLLT3	#N/A	RAF1	Y	NOTCH1	Y	ARID2	#N/A	LRP10	#N/A			OR6C65	#N/A	TGFB
42	#N/A	TPTE2P1	#N/A	PKP4	#N/A			OR52E2	#N/A	TBC1D5	#N/A	MMP14	Y	REPS1	Y	OR5AS1	#N/A	ARID3B	#N/A	MAP3K9	#N/A			OR8H2	#N/A	TPTE
43	#N/A	TRY2	#N/A	PLK3	Y			PCDHA5	#N/A	TMEM102	#N/A	MTMR4	#N/A	RPL5	#N/A	OR5AU1	#N/A	ARID4B	#N/A	MARCKS	#N/A			P2RY14	#N/A	TRY2
44	#N/A	TRPC4	#N/A	PLXND1	#N/A			PCDHGA4	#N/A	TMEM132B	#N/A	MUC7	#N/A	S100A3	#N/A	PCDHA5	#N/A	ARSF	#N/A	MECOM	#N/A			PADI3	#N/A	FAM2
45	#N/A	ULK4P3	#N/A	RB1	Y			PRSS12	#N/A	TNRC18	#N/A	MYO7B	#N/A	SALL1	#N/A	PCDHA7	#N/A	ATP13A2	#N/A	MED26	#N/A			PARD6A	#N/A	VPS3
46	#N/A	ZNF695	#N/A	LINC01141	#N/A			RAPGEF5	#N/A	TP53	Y	NBPF25P	#N/A	SCTR	Y	PCDHGA2	#N/A	ATP2B2	#N/A	MEFV	#N/A			PDPR	#N/A	ZNF2
47	#N/A			SYDE1	#N/A			VENTXP1	#N/A	TRPM5	#N/A	NNT	#N/A	SPON1	#N/A	PCSK2	#N/A	ATP2B2	#N/A	METTL14	#N/A			PLK4	Y	ZNF4
48	#N/A			TEAD2	#N/A			SCML2	#N/A	TSC22D1	Y	OR10J1	#N/A	SPPL2B	#N/A	PDE10A	#N/A	ATP2C2	#N/A	MIA3	#N/A			PPT2	#N/A	
49	#N/A			TRIM68	Y			TAGLN2	#N/A	TSPAN4	#N/A	OR11H4	#N/A	SRGAP1	#N/A	PDZD2	Y	ATP5H	#N/A	MLF2	#N/A			PPTC7	#N/A	
50	#N/A			TRIOBP	#N/A			TMCC2	#N/A	UPP1	#N/A	PHTF1	#N/A	TMEM39A	#N/A	PKD1	Y	AXL	Y	MMACHC	#N/A			PRMT5	#N/A	
51	#N/A			TRMT1	#N/A			PRDM7	#N/A	USP2	#N/A	PKHD1	#N/A	TPTE2	#N/A	PKHD1	#N/A	B3GALT5	#N/A	MRPL19	#N/A			SRGAP2-AS1	#N/A	
52	Y			TSG101	Y			ZNF460	#N/A	WDR83	#N/A	PMEPA1	Y	TRAM1	Y	PLAU	Y	BAZ1A	#N/A	NCOA6	#N/A			LOC1012435	#N/A	
53	#N/A			MIR548O2	#N/A			ZNF493	#N/A	XCL1	#N/A	PRAMEF11	#N/A	UGT2B10	#N/A	PRAMEF1	#N/A	BCL9	#N/A	NEB	#N/A			SEMA3D	#N/A	

Sheet 1 / 1 Default Error: Value not available 60%

Figure 9: Matching vcf variants with DisGeNET

5.5 Mapping with PhastConelements46way database

Filtering genes after matching with phastconselements46way we get the variations in conserved regions



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	Hugo_Sy	Entrez_Ge	Center	Ncbi_Buil	Chromoso	Start_Posi	End_Posit	Strand	Variant_C	Variant_T	Reference	Tumor_Se	Tumor_Se	Dbsnp_Rs	Dbsnp_Va	Tumor_Sa	Matched_	Match_N	Match_N	Tumor_Va	Tumor_
2	AARS2	57505	broad.mit.	37	chr6	44268380	44268380	+	Silent	SNP	C	C	T	rs1426945	by1000ger	TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
3	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
4	ABCA1	19	broad.mit.	37	chr9	1.08E+08	1.08E+08	+	Frame_Sh	DEL	AGAGGA	AGAGGA	-			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
5	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-HC	TCGA-HC-7742-11A-01D-2114-08				
6	ABCA13	154664	broad.mit.	37	chr7	48352729	48352729	+	Silent	SNP	C	C	T			TCGA-HC	TCGA-HC-7742-10A-01D-2115-08				
7	ABCC11	85320	broad.mit.	37	chr16	48212570	48212570	+	Missense_	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
8	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-11A-01D-2114-08				
9	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7331-10A-01D-2114-08				
10	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
11	ABCC5	10057	broad.mit.	37	chr3	1.84E+08	1.84E+08	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
12	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
13	ABCC8	6833	broad.mit.	37	chr11	17428224	17428224	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
14	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-HC	TCGA-HC-7745-10A-01D-2115-08				
15	ABCE1	6059	broad.mit.	37	chr4	1.46E+08	1.46E+08	+	Missense_	SNP	A	A	C			TCGA-HC	TCGA-HC-7745-11A-01D-2114-08				
16	AC002331	0	broad.mit.	37	chr16	26599065	26599068	+	RNA	DEL	ACAG	ACAG	-	rs71134607		TCGA-HC	TCGA-HC-7752-10A-01D-2115-08				
17	AC00605	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-10A-01D-2114-08				
18	AC00605	0	broad.mit.	37	chr17	28901676	28901676	+	RNA	DEL	C	C	-			TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
19	AC008103	0	broad.mit.	37	chr22	18844766	18844766	+	RNA	SNP	G	G	A			TCGA-EJ	TCGA-EJ-7328-10A-01D-2114-08				
20	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-10A-01D-2114-08				
21	AC010547	0	broad.mit.	37	chr16	71516014	71516014	+	Missense_	SNP	C	C	T			TCGA-EJ	TCGA-EJ-7782-11A-01D-2114-08				
22	AC015818	0	broad.mit.	37	chr17	20424245	20424245	+	RNA	SNP	C	C	T			TCGA-HC	TCGA-HC-7819-11A-01D-2114-08				
23	AC01873	0	broad.mit.	37	chr2	1.05E+08	1.05E+08	+	RNA	DEL	TGGTGA	TGGTGA	-	rs1509372	by1000ger	TCGA-EJ	TCGA-EJ-7794-11A-01D-2114-08				
24	AC019118	0	broad.mit.	37	chr2	2910378	2910379	+	RNA	INS	-	-	C	rs7205115	by1000ger	TCGA-EJ	TCGA-EJ-7317-11A-01D-2114-08				
25	AC019118	0	broad.mit.	37	chr2	2910768	2910769	+	RNA	INS	-	-	T	rs1425859	by1000ger	TCGA-HC	TCGA-HC-7737-11A-02D-2114-08				

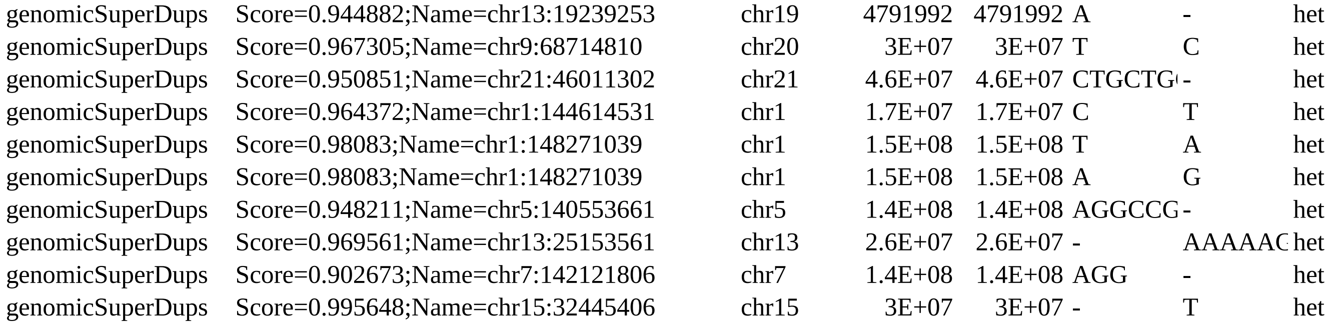
Figure 10: Conserved region variations

5.6 list of genes after mapping with Phastconselements Database

Variations in the genes are filtered on the basis of the threshold score above 400. Variations above this score are likely to fall in conserved region.

ANKHD1, APC, ASAH1, ASCC2, ASXL1, ATM, BRAF, BUD13, CLTC, CXCR2, **CYP19A1**, DNAH8, EIF4G2, FGFR2, FLT4, **FOXA1**, FOXP2, FRS3, FSTL1, FZR1, GPX1, **GSTT1**, HDAC9, HPCA, KDR, KRIT1, LPAR1, MED15, MMP14, MOAP1, NCOA3, NCOR2, NOTCH1, PLK3, PLK4, PMEPA1, PPIG, RAF1, RB1, RCHY1, REPS1, SCN9A, SEMA3A, SIRT1, SLC39A6, STK26, TCN1, TGFBI, THBS1, **TP53**, TRAM1, TSHZ3, TTN.

5.7 Region based annotation with genomicSuperDups



genomicSuperDups	Score=0.944882;Name=chr13:19239253	chr19	4791992	4791992	A	-	het
genomicSuperDups	Score=0.967305;Name=chr9:68714810	chr20	3E+07	3E+07	T	C	het
genomicSuperDups	Score=0.950851;Name=chr21:46011302	chr21	4.6E+07	4.6E+07	CTGCTG	-	het
genomicSuperDups	Score=0.964372;Name=chr1:144614531	chr1	1.7E+07	1.7E+07	C	T	het
genomicSuperDups	Score=0.98083;Name=chr1:148271039	chr1	1.5E+08	1.5E+08	T	A	het
genomicSuperDups	Score=0.98083;Name=chr1:148271039	chr1	1.5E+08	1.5E+08	A	G	het
genomicSuperDups	Score=0.948211;Name=chr5:140553661	chr5	1.4E+08	1.4E+08	AGGCCG	-	het
genomicSuperDups	Score=0.969561;Name=chr13:25153561	chr13	2.6E+07	2.6E+07	-	AAAAAC	het
genomicSuperDups	Score=0.902673;Name=chr7:142121806	chr7	1.4E+08	1.4E+08	AGG	-	het
genomicSuperDups	Score=0.995648;Name=chr15:32445406	chr15	3E+07	3E+07	-	T	het

Figure 11: GenomicSuperDups filtration

After filtering with genomicSuperDups segmented duplicated genetic variants can be removed. These types of variants can be show two non polymorphic types in genome.

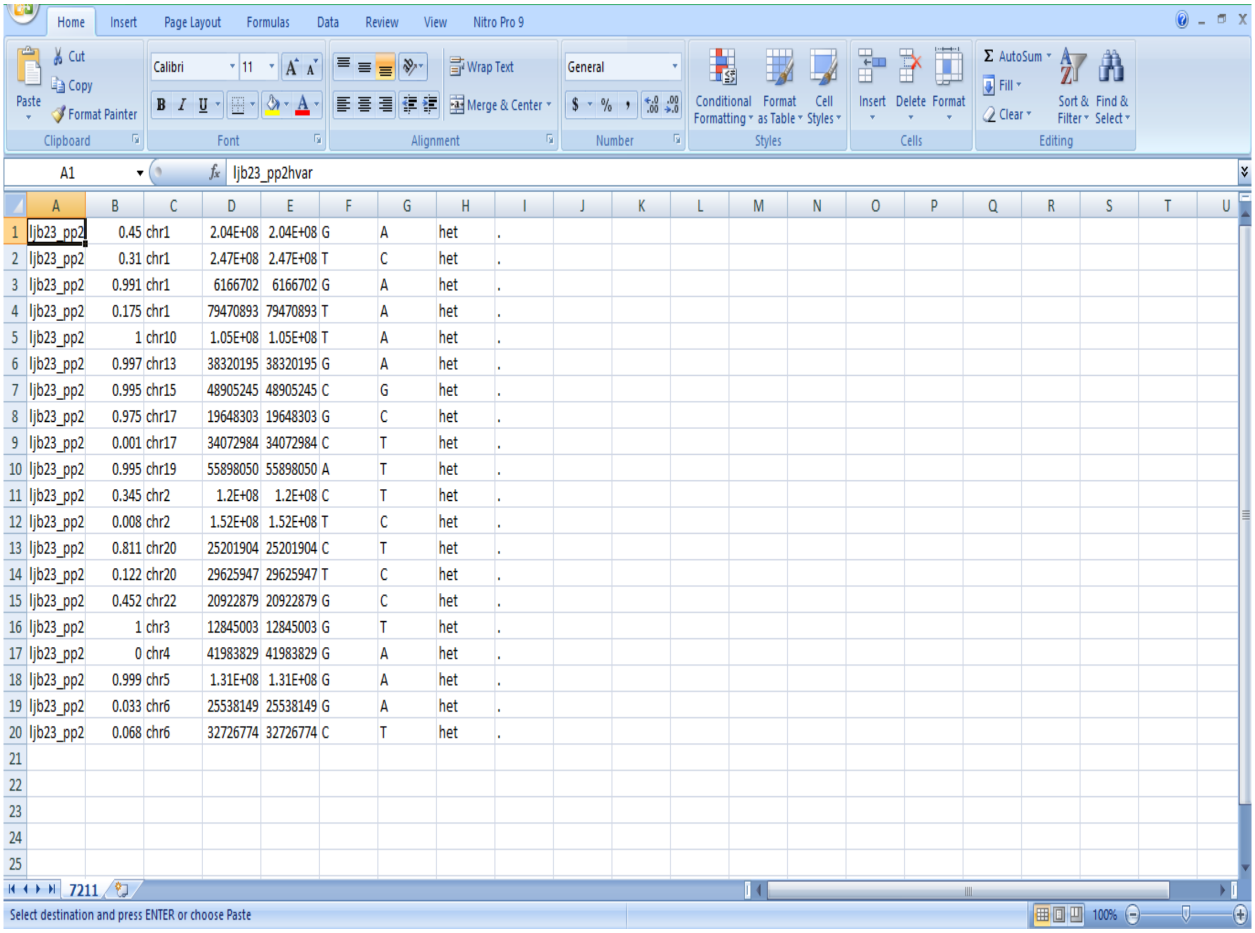
5.8 Filter based annotation with polyphen2 database to know about the damaging effect of disease

This annotation shows the damaging effect of the variants on the basis of the scores.

Probably damaging = .909 – 1

Possibly damaging = .447 - .908

Benign = 0 - .446



A1 ljb23_pp2hvar

	A	B	C	D	E	F	G	H	I
1	ljb23_pp2	0.45	chr1	2.04E+08	2.04E+08	G	A	het	.
2	ljb23_pp2	0.31	chr1	2.47E+08	2.47E+08	T	C	het	.
3	ljb23_pp2	0.991	chr1	6166702	6166702	G	A	het	.
4	ljb23_pp2	0.175	chr1	79470893	79470893	T	A	het	.
5	ljb23_pp2	1	chr10	1.05E+08	1.05E+08	T	A	het	.
6	ljb23_pp2	0.997	chr13	38320195	38320195	G	A	het	.
7	ljb23_pp2	0.995	chr15	48905245	48905245	C	G	het	.
8	ljb23_pp2	0.975	chr17	19648303	19648303	G	C	het	.
9	ljb23_pp2	0.001	chr17	34072984	34072984	C	T	het	.
10	ljb23_pp2	0.995	chr19	55898050	55898050	A	T	het	.
11	ljb23_pp2	0.345	chr2	1.2E+08	1.2E+08	C	T	het	.
12	ljb23_pp2	0.008	chr2	1.52E+08	1.52E+08	T	C	het	.
13	ljb23_pp2	0.811	chr20	25201904	25201904	C	T	het	.
14	ljb23_pp2	0.122	chr20	29625947	29625947	T	C	het	.
15	ljb23_pp2	0.452	chr22	20922879	20922879	G	C	het	.
16	ljb23_pp2	1	chr3	12845003	12845003	G	T	het	.
17	ljb23_pp2	0	chr4	41983829	41983829	G	A	het	.
18	ljb23_pp2	0.999	chr5	1.31E+08	1.31E+08	G	A	het	.
19	ljb23_pp2	0.033	chr6	25538149	25538149	G	A	het	.
20	ljb23_pp2	0.068	chr6	32726774	32726774	C	T	het	.

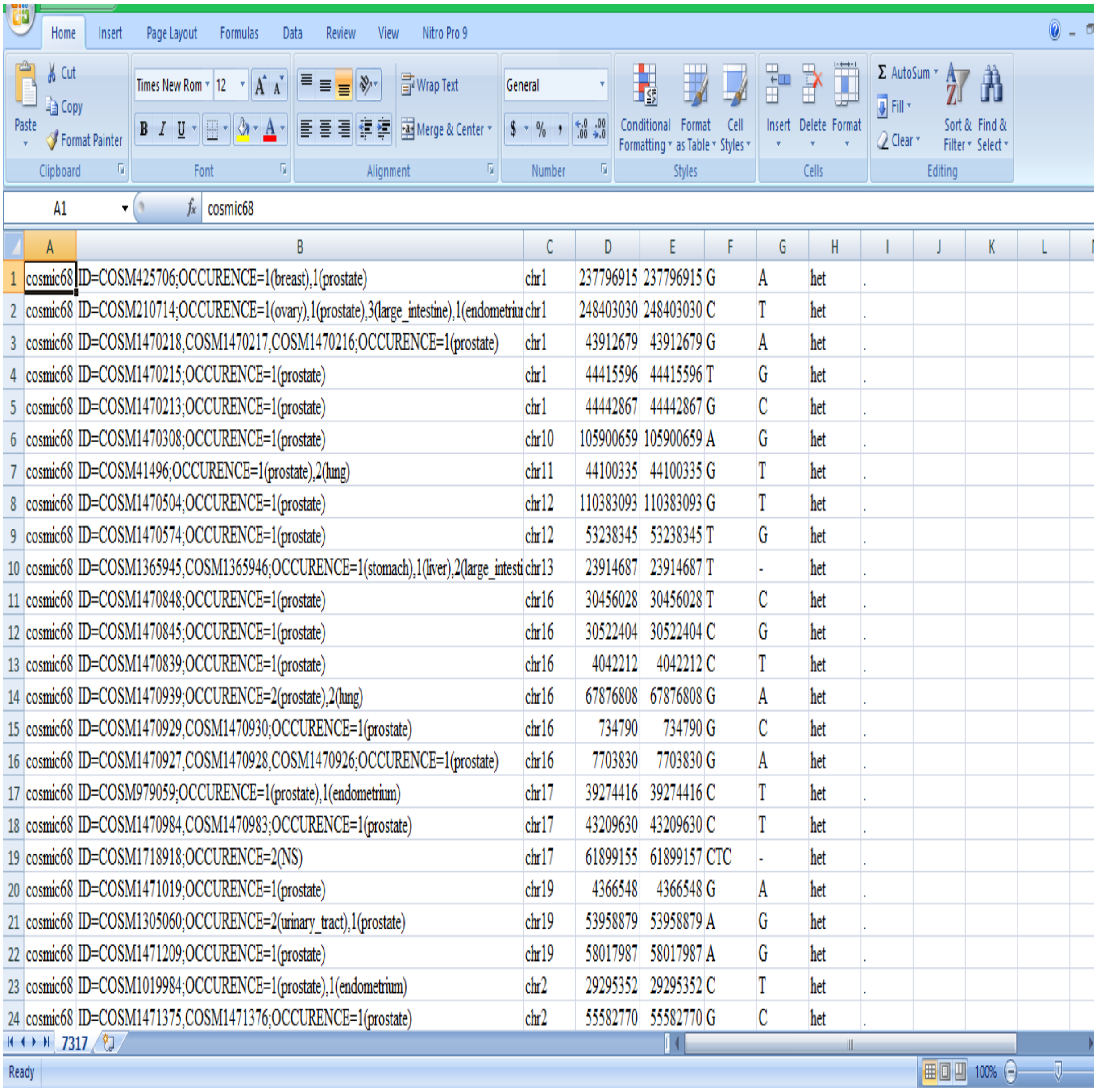
7211

Select destination and press ENTER or choose Paste

Figure 12: Polyphen2 scores

5.9 Mapping input file with cosmic database

Validation of the filtered genes can be done by this database. Cosmic database contains all the somatic mutations that present in literature of the all type of the cancers.

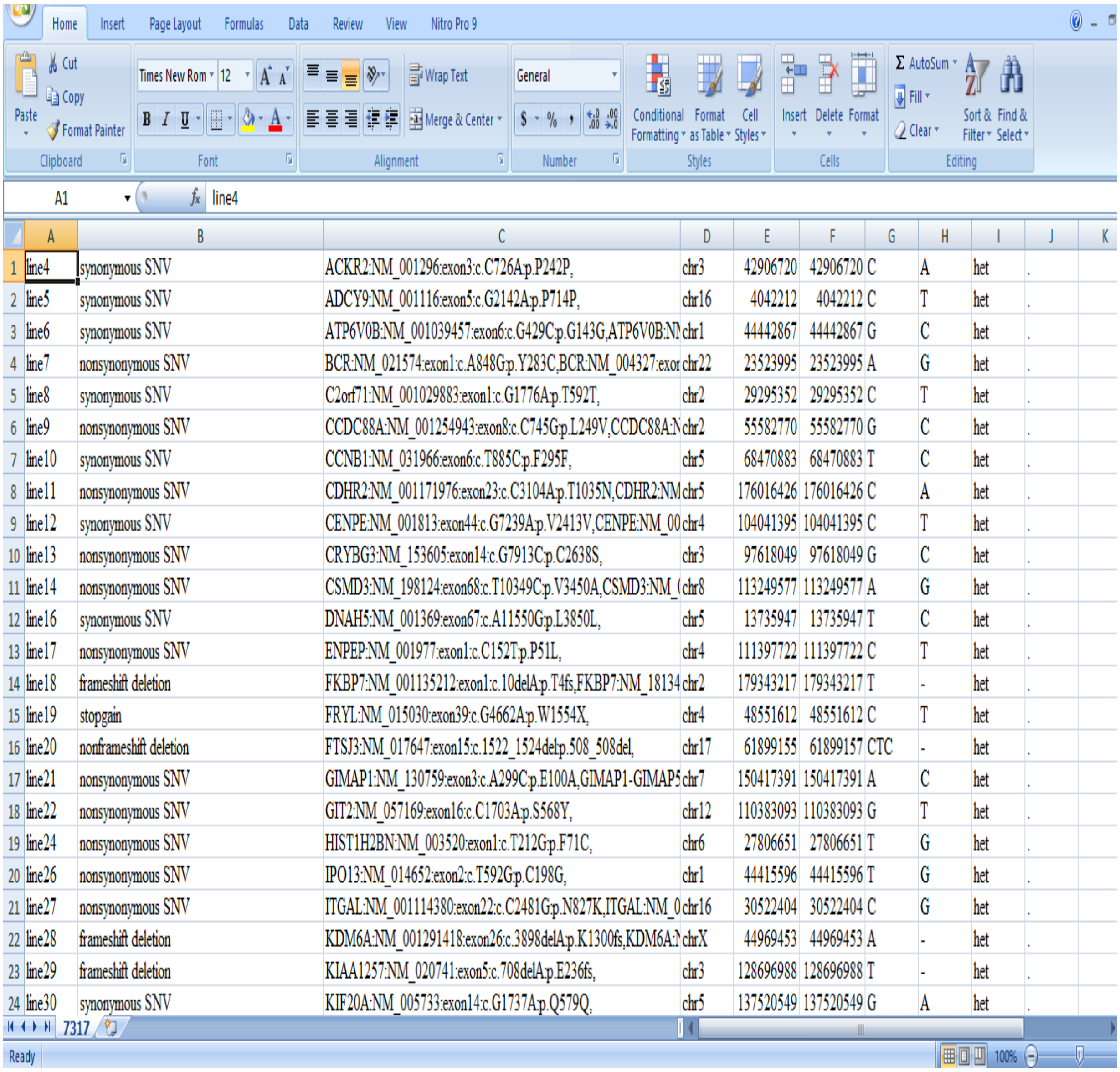


A1 cosmic68

	A	B	C	D	E	F	G	H	I
1	cosmic68	ID=COSM425706;OCCURENCE=1(breast),1(prostate)	chr1	237796915	237796915	G	A	het	.
2	cosmic68	ID=COSM210714;OCCURENCE=1(ovary),1(prostate),3(large_intestine),1(endometriu	chr1	248403030	248403030	C	T	het	.
3	cosmic68	ID=COSM1470218,COSM1470217,COSM1470216;OCCURENCE=1(prostate)	chr1	43912679	43912679	G	A	het	.
4	cosmic68	ID=COSM1470215;OCCURENCE=1(prostate)	chr1	44415596	44415596	T	G	het	.
5	cosmic68	ID=COSM1470213;OCCURENCE=1(prostate)	chr1	44442867	44442867	G	C	het	.
6	cosmic68	ID=COSM1470308;OCCURENCE=1(prostate)	chr10	105900659	105900659	A	G	het	.
7	cosmic68	ID=COSM41496;OCCURENCE=1(prostate),2(lung)	chr11	44100335	44100335	G	T	het	.
8	cosmic68	ID=COSM1470504;OCCURENCE=1(prostate)	chr12	110383093	110383093	G	T	het	.
9	cosmic68	ID=COSM1470574;OCCURENCE=1(prostate)	chr12	53238345	53238345	T	G	het	.
10	cosmic68	ID=COSM1365945,COSM1365946;OCCURENCE=1(stomach),1(liver),2(large_intesti	chr13	23914687	23914687	T	-	het	.
11	cosmic68	ID=COSM1470848;OCCURENCE=1(prostate)	chr16	30456028	30456028	T	C	het	.
12	cosmic68	ID=COSM1470845;OCCURENCE=1(prostate)	chr16	30522404	30522404	C	G	het	.
13	cosmic68	ID=COSM1470839;OCCURENCE=1(prostate)	chr16	4042212	4042212	C	T	het	.
14	cosmic68	ID=COSM1470939;OCCURENCE=2(prostate),2(lung)	chr16	67876808	67876808	G	A	het	.
15	cosmic68	ID=COSM1470929,COSM1470930;OCCURENCE=1(prostate)	chr16	734790	734790	G	C	het	.
16	cosmic68	ID=COSM1470927,COSM1470928,COSM1470926;OCCURENCE=1(prostate)	chr16	7703830	7703830	G	A	het	.
17	cosmic68	ID=COSM979059;OCCURENCE=1(prostate),1(endometrium)	chr17	39274416	39274416	C	T	het	.
18	cosmic68	ID=COSM1470984,COSM1470983;OCCURENCE=1(prostate)	chr17	43209630	43209630	C	T	het	.
19	cosmic68	ID=COSM1718918;OCCURENCE=2(NS)	chr17	61899155	61899157	CTC	-	het	.
20	cosmic68	ID=COSM1471019;OCCURENCE=1(prostate)	chr19	4366548	4366548	G	A	het	.
21	cosmic68	ID=COSM1305060;OCCURENCE=2(urinary_tract),1(prostate)	chr19	53958879	53958879	A	G	het	.
22	cosmic68	ID=COSM1471209;OCCURENCE=1(prostate)	chr19	58017987	58017987	A	G	het	.
23	cosmic68	ID=COSM1019984;OCCURENCE=1(prostate),1(endometrium)	chr2	29295352	29295352	C	T	het	.
24	cosmic68	ID=COSM1471375,COSM1471376;OCCURENCE=1(prostate)	chr2	55582770	55582770	G	C	het	.

Figure 13: Cosmic validation

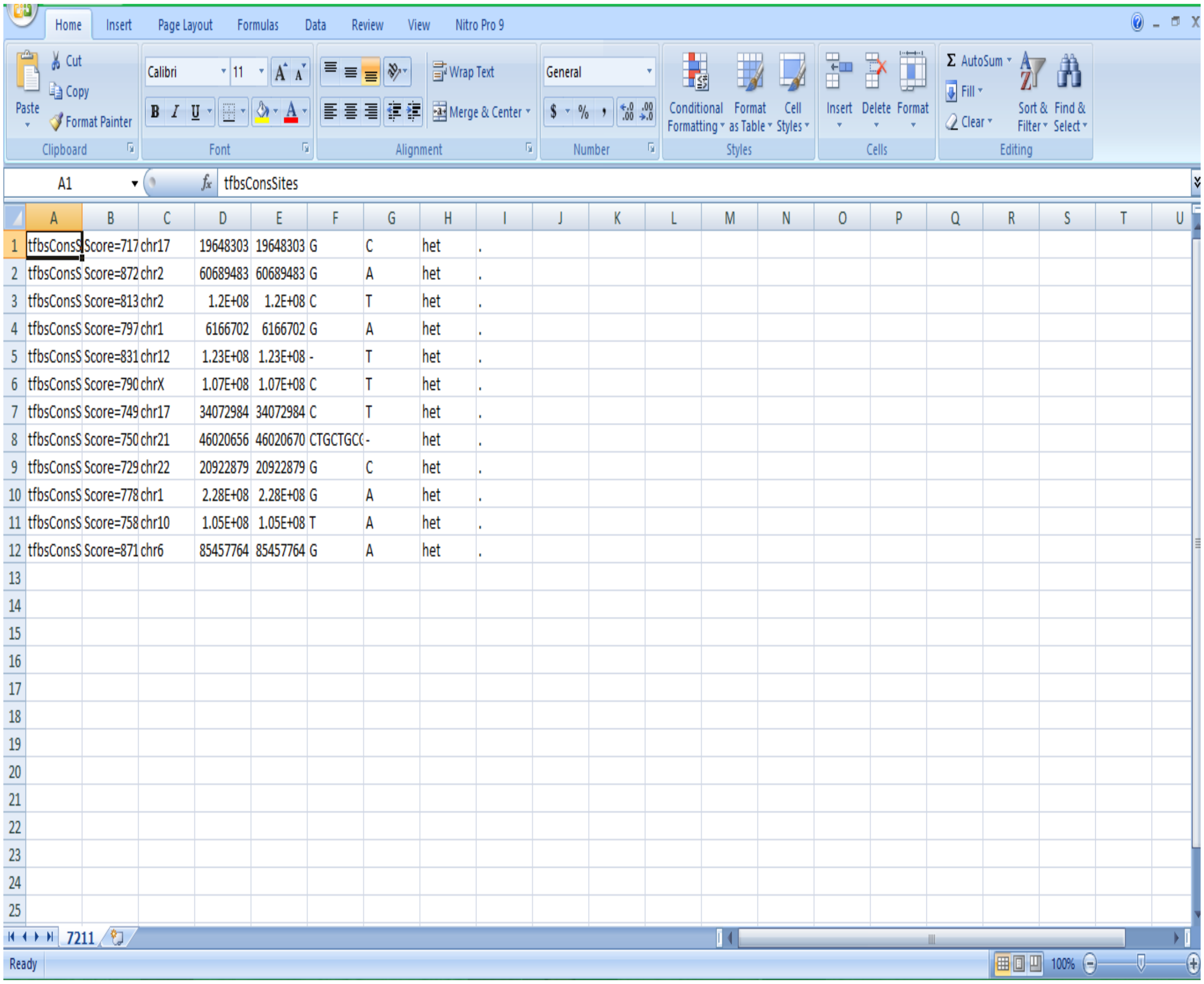
5.10 Gene annotation to know about the nature of the variant and its effect on amino acid change



	A	B	C	D	E	F	G	H	I	J	K
1	line4	synonymous SNV	ACKR2:NM_001296:exon3:c.C726Ap.P242P,	chr3	42906720	42906720	C	A	het	.	
2	line5	synonymous SNV	ADCY9:NM_001116:exon5:c.G2142Ap.P714P,	chr16	4042212	4042212	C	T	het	.	
3	line6	synonymous SNV	ATP6V0B:NM_001039457:exon6:c.G429Cp.G143G,ATP6V0B:NI	chr1	44442867	44442867	G	C	het	.	
4	line7	nonsynonymous SNV	BCR:NM_021574:exon1:c.A848Gp.Y283C,BCR:NM_004327:exor	chr22	23523995	23523995	A	G	het	.	
5	line8	synonymous SNV	C2orf71:NM_001029883:exon1:c.G1776Ap.T592T,	chr2	29295352	29295352	C	T	het	.	
6	line9	nonsynonymous SNV	CCDC88A:NM_001254943:exon8:c.C745Gp.L249V,CCDC88A:N	chr2	55582770	55582770	G	C	het	.	
7	line10	synonymous SNV	CCNB1:NM_031966:exon6:c.T885Cp.F295F,	chr5	68470883	68470883	T	C	het	.	
8	line11	nonsynonymous SNV	CDHR2:NM_001171976:exon23:c.C3104Ap.T1035N,CDHR2:NM	chr5	176016426	176016426	C	A	het	.	
9	line12	synonymous SNV	CENPE:NM_001813:exon44:c.G7239Ap.V2413V,CENPE:NM_00	chr4	104041395	104041395	C	T	het	.	
10	line13	nonsynonymous SNV	CRYBG3:NM_153605:exon14:c.G7913Cp.C2638S,	chr3	97618049	97618049	G	C	het	.	
11	line14	nonsynonymous SNV	CSMD3:NM_198124:exon68:c.T10349Cp.V3450A,CSMD3:NM_(	chr8	113249577	113249577	A	G	het	.	
12	line16	synonymous SNV	DNAH5:NM_001369:exon67:c.A11550Gp.L3850L,	chr5	13735947	13735947	T	C	het	.	
13	line17	nonsynonymous SNV	ENPEP:NM_001977:exon1:c.C152Tp.P51L,	chr4	111397722	111397722	C	T	het	.	
14	line18	frameshift deletion	FKBP7:NM_001135212:exon1:c.10delAp.T4fs,FKBP7:NM_18134	chr2	179343217	179343217	T	-	het	.	
15	line19	stopgain	FRYL:NM_015030:exon39:c.G4662Ap.W1554X,	chr4	48551612	48551612	C	T	het	.	
16	line20	nonframeshift deletion	FTSJ3:NM_017647:exon15:c.1522_1524delp.508_508del,	chr17	61899155	61899157	CTC	-	het	.	
17	line21	nonsynonymous SNV	GIMAP1:NM_130759:exon3:c.A299Cp.E100A,GIMAP1-GIMAP5	chr7	150417391	150417391	A	C	het	.	
18	line22	nonsynonymous SNV	GIT2:NM_057169:exon16:c.C1703Ap.S568Y,	chr12	110383093	110383093	G	T	het	.	
19	line24	nonsynonymous SNV	HIST1H2BN:NM_003520:exon1:c.T212Gp.F71C,	chr6	27806651	27806651	T	G	het	.	
20	line26	nonsynonymous SNV	IPO13:NM_014652:exon2:c.T592Gp.C198G,	chr1	44415596	44415596	T	G	het	.	
21	line27	nonsynonymous SNV	ITGAL:NM_001114380:exon22:c.C2481Gp.N827K,ITGAL:NM_0	chr16	30522404	30522404	C	G	het	.	
22	line28	frameshift deletion	KDM6A:NM_001291418:exon26:c.3898delAp.K1300fs,KDM6A:N	chrX	44969453	44969453	A	-	het	.	
23	line29	frameshift deletion	KIAA1257:NM_020741:exon5:c.708delAp.E236fs,	chr3	128696988	128696988	T	-	het	.	
24	line30	synonymous SNV	KIF20A:NM_005733:exon14:c.G1737Ap.Q579Q,	chr5	137520549	137520549	G	A	het	.	

Figure 14: Exonic variants

5.11 Predict the variations lie in the Transcription factor binding sites



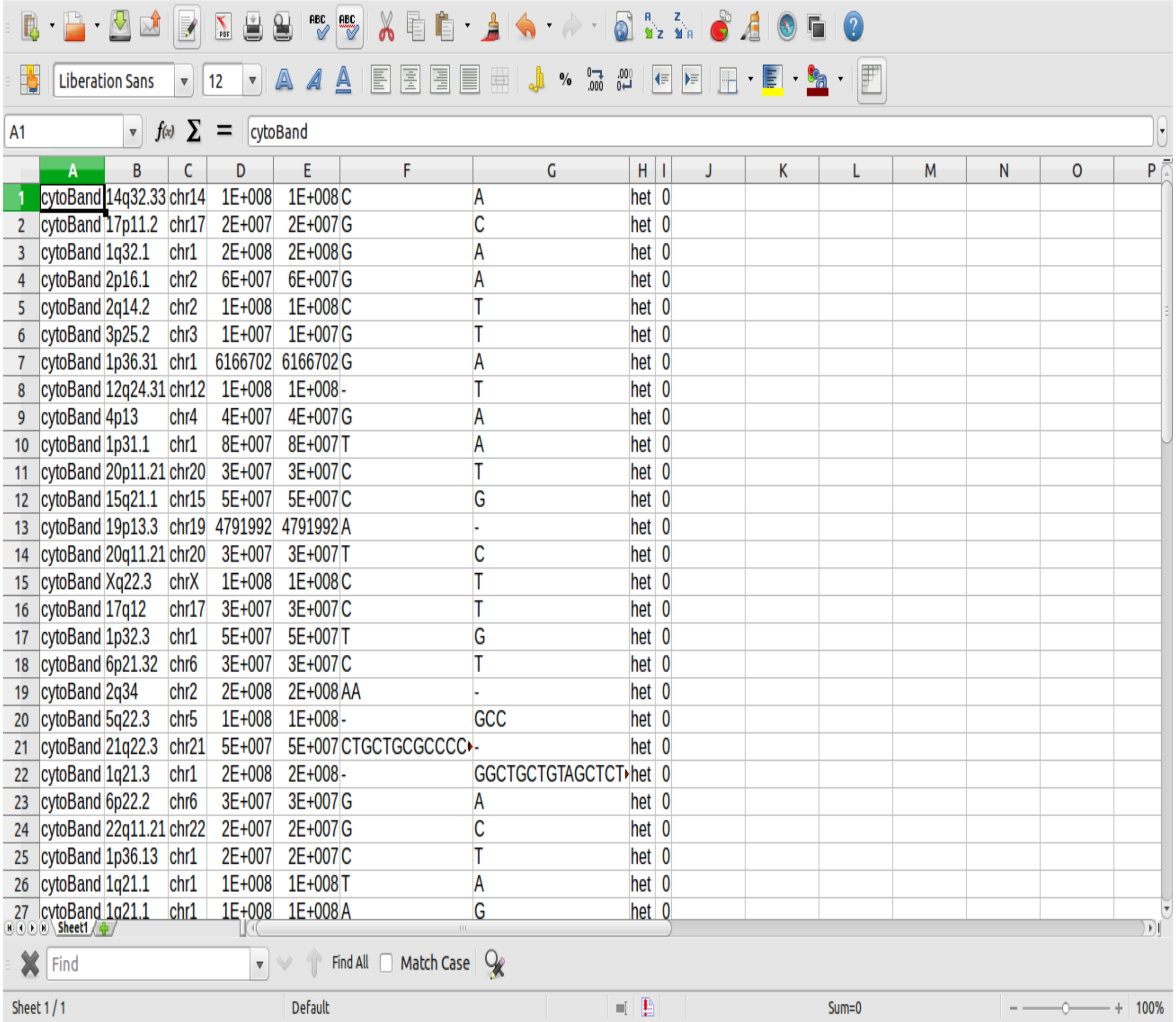
A1 tfbsConsSites

	A	B	C	D	E	F	G	H	I
1	tfbsConsS	Score=717	chr17	19648303	19648303	G	C	het	.
2	tfbsConsS	Score=872	chr2	60689483	60689483	G	A	het	.
3	tfbsConsS	Score=813	chr2	1.2E+08	1.2E+08	C	T	het	.
4	tfbsConsS	Score=797	chr1	6166702	6166702	G	A	het	.
5	tfbsConsS	Score=831	chr12	1.23E+08	1.23E+08	-	T	het	.
6	tfbsConsS	Score=790	chrX	1.07E+08	1.07E+08	C	T	het	.
7	tfbsConsS	Score=749	chr17	34072984	34072984	C	T	het	.
8	tfbsConsS	Score=750	chr21	46020656	46020670	CTGCTGC(	-	het	.
9	tfbsConsS	Score=729	chr22	20922879	20922879	G	C	het	.
10	tfbsConsS	Score=778	chr1	2.28E+08	2.28E+08	G	A	het	.
11	tfbsConsS	Score=758	chr10	1.05E+08	1.05E+08	T	A	het	.
12	tfbsConsS	Score=871	chr6	85457764	85457764	G	A	het	.

Figure 15: Transcription factor binding site

5.12 Getting cytobands

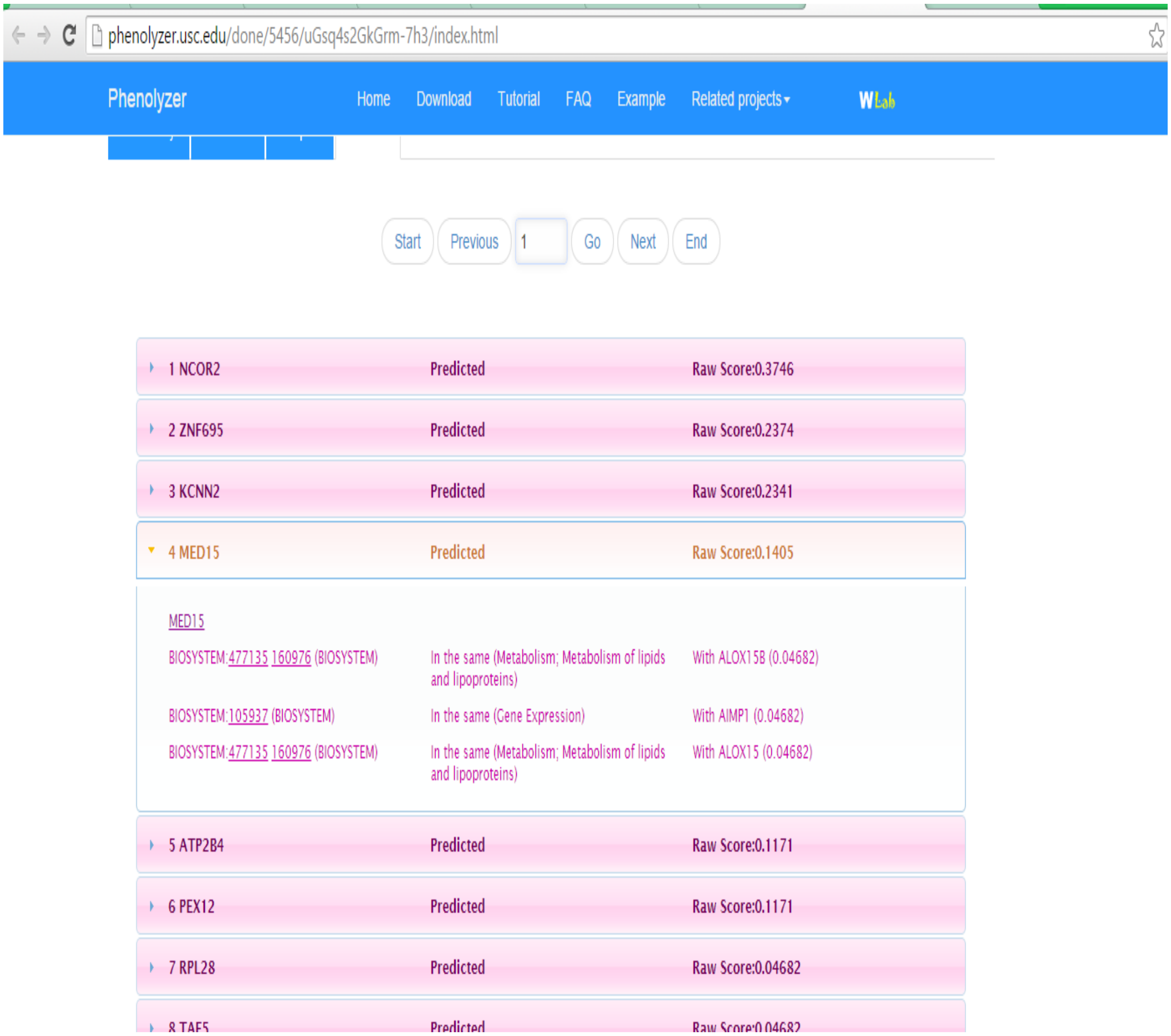
Chromosomes bands that are Geimsa- stained can be achieved by mapping with the cytobands database



	A	B	C	D	E	F	G	H	I
1	cytoBand	14q32.33	chr14	1E+008	1E+008	C	A	het	0
2	cytoBand	17p11.2	chr17	2E+007	2E+007	G	C	het	0
3	cytoBand	1q32.1	chr1	2E+008	2E+008	G	A	het	0
4	cytoBand	2p16.1	chr2	6E+007	6E+007	G	A	het	0
5	cytoBand	2q14.2	chr2	1E+008	1E+008	C	T	het	0
6	cytoBand	3p25.2	chr3	1E+007	1E+007	G	T	het	0
7	cytoBand	1p36.31	chr1	6166702	6166702	G	A	het	0
8	cytoBand	12q24.31	chr12	1E+008	1E+008	-	T	het	0
9	cytoBand	4p13	chr4	4E+007	4E+007	G	A	het	0
10	cytoBand	1p31.1	chr1	8E+007	8E+007	T	A	het	0
11	cytoBand	20p11.21	chr20	3E+007	3E+007	C	T	het	0
12	cytoBand	15q21.1	chr15	5E+007	5E+007	C	G	het	0
13	cytoBand	19p13.3	chr19	4791992	4791992	A	-	het	0
14	cytoBand	20q11.21	chr20	3E+007	3E+007	T	C	het	0
15	cytoBand	Xq22.3	chrX	1E+008	1E+008	C	T	het	0
16	cytoBand	17q12	chr17	3E+007	3E+007	C	T	het	0
17	cytoBand	1p32.3	chr1	5E+007	5E+007	T	G	het	0
18	cytoBand	6p21.32	chr6	3E+007	3E+007	C	T	het	0
19	cytoBand	2q34	chr2	2E+008	2E+008	AA	-	het	0
20	cytoBand	5q22.3	chr5	1E+008	1E+008	-	GCC	het	0
21	cytoBand	21q22.3	chr21	5E+007	5E+007	CTGCTGCGCCCC	-	het	0
22	cytoBand	1q21.3	chr1	2E+008	2E+008	-	GGCTGCTGTAGCTCT	het	0
23	cytoBand	6p22.2	chr6	3E+007	3E+007	G	A	het	0
24	cytoBand	22q11.21	chr22	2E+007	2E+007	G	C	het	0
25	cytoBand	1p36.13	chr1	2E+007	2E+007	C	T	het	0
26	cytoBand	1q21.1	chr1	1E+008	1E+008	T	A	het	0
27	cytoBand	1q21.1	chr1	1E+008	1E+008	A	G	het	0

Figure 16: CytoBand

5.13 Phenolyzer result: Phenolyzer analysis results show the network and the pathways of the genes in which they involve.



1 NCOR2	Predicted	Raw Score:0.3746
2 ZNF695	Predicted	Raw Score:0.2374
3 KCNN2	Predicted	Raw Score:0.2341
4 MED15	Predicted	Raw Score:0.1405
MED15		
BIOSYSTEM:477135 160976 (BIOSYSTEM)	In the same (Metabolism; Metabolism of lipids and lipoproteins)	With ALOX15B (0.04682)
BIOSYSTEM:105937 (BIOSYSTEM)	In the same (Gene Expression)	With AIMP1 (0.04682)
BIOSYSTEM:477135 160976 (BIOSYSTEM)	In the same (Metabolism; Metabolism of lipids and lipoproteins)	With ALOX15 (0.04682)
5 ATP2B4	Predicted	Raw Score:0.1171
6 PEX12	Predicted	Raw Score:0.1171
7 RPL28	Predicted	Raw Score:0.04682
8 TAF5	Predicted	Raw Score:0.04682

Figure 17: Phenolyzer Details

DISCUSSION

With the advancement in NGS, there is sudden explosion of data in the public centers (Rakyan *et al.*, 2011). Genetic alterations accumulating in the exome region of the genome are the main reason for the occurrence of various forms of cancers (Meyerson *et al*., 2010). Since variations in the exome region directly affect the phenotype of the individuals (Pauline *et al*., 2008), many human diseases involve genetic alterations in the exome region such as cancer, Mendelian disorders and neurological diseases.

Exome sequence analysis of the genetic variations in PRAD patients were carried out in this study. A number of genes were filtered after analysis that carries those variations that might be the cause of the PRAD. From these filtered genes few genes like GSST1, TP53 and CYP19A1 are already known to play the role in PRAD pathway. Similarly, we could conclude that out of the 52 filtered genes some could possibly act as biomarkers.

Segmented duplicated variations that show two non-polymorphic phenotypes were removed from the filtered genes and this was done by matching with genomicSuperDups database. To know about the damaging effect of the variation, polyphen2 was used (Adzhubei *et al.*, 2010). The variations with scores in the range 0.446-0 are benign, score ranges 0.447-0.908 are possibly damaging and score 0.909-1.0 are probably damaging. 52 Filtered genes variations had amino acid changes which were annotated using Refgene database. Type of mutations like frame shift, synonymous and non synonymous were also found out using the same database. Database tfbsConsSites showed that which variants lie in the conserved regions of the transcription factor binding sites on the basis of scores. Cytogenetic bands of each gene were obtained by matching with cytoband database and viewed using Integrative Genomics Viewer. Network pathways of each filtered gene were also annotated. This analysis tells about the Phenolyzer effects of the genes like genes involved in lipid metabolism etc. Validation of the filtered genes was done by COSMIC database. This database includes somatic mutations reported in the literature. This study can be used on cancer and other type of diseases to find out the probable biomarkers.

To validate the results, wet lab experiments are very important for supporting the analysis of genetic variations in the exome regions. After validation of the results, variations can be used as biomarkers for the prostate cancer which could help in prognosis and diagnosis of the cancer. Different phenotypic effects are observed in the same gene that shows different mutations (Boyadjiev *et al.*, 2000). Environmental factors also affect some phenotypes due to the presence of some variations (Hunter, 2005). So along with genetic factors we should also consider the environmental factors for the causation of the cancer.

CONCLUSION

Somatic mutations in Prostate Adenocarcinoma are revealed by processing of the next-generation DNA sequencing data of the exome region. NGS helps in studying the alternative splicing complexity (Martin *et al*., 2011), landscape of mutations in cancer. Exome sequencing used for diagnosis and identification of novel genetic variants and novels genes associated with Mendelian diseases.

In present study 93 genes filtered by mapping vcf files genetic variations with the DisGeNEt database. Further filtration was done by matching with the phastconselelement46way and genomicSuperDups database. After this step we got 52 most probable variations that might alter the pathways that caused prostate cancer.

Annotation of these variations was done by ANNOVAR software package. Annotation includes to find the nature of the variants; like: are variants falling in transcription binding factor sites, class of mutation in which variants lie, to know about the damaging effects of the variants and to get the cytobands of the variations. These annotations use the tnbps, refgene, polyphen2 and cytoband databases respectively. Validation of the filtered genes can be done by matching the variations with the Cosmic database, this database consists of the variations that present in the literature. In this study annotations include gene based, region based and filter base analysis.

Advancement in integrative approaches like somatic mutations analysis with DNA methylations and gene expression studies, we could expect a well developed personalized medicine in future soon (Rabbani *et al.,* 2014).

FUTURE PROSPECTIVE

In this study a computational pipeline is developed for exome sequence analysis to prioritize and annotate the genetic variations found in exome region. Hg19 Human Reference Genome was used for the initial analysis work in this study. This study marks as a proof of concept for integrating the studies of different platforms such as DNA methylation, somatic mutations, gene expressions, and clinical data along with exome sequencing. TCGA has enabled the researchers to carry out a multi-platform analysis as it provides multi-platform data for the samples of same patient. TCGA is maintaining data very well and the architecture is well documented.

In the current study analysis and annotation of somatic mutations obtained by exome sequencing was done by ANNOVAR software. This is another progress for the further development of the pipelines for the analysis of the exome data. Integrative studies are also done on somatic variations with the DNA methylated and gene expression studies. Advancement in NGS and collaboration with statisticians, mathematicians, computer engineers and bioinformaticians could help in better development of the new pipelines that can integrate somatic variations, clinical data, DNA expression, DNA methylated and SNP data to find out the functional biomarkers. These biomarkers can be used for the diagnosis, prognosis and treatment of the complex cancers.

Whole exome sequencing can be used for analysis of other diseases like monogenic disorders, hearing loss, intellectual disabilities, movement disorders, cardiovascular diseases, obesity and diabetes, cancers and hypertensions.

More accurate Bioinformatics tools will be required for studying genetic diseases so that we can counsel the patient, find the more precise diagnosis and treatment for the disease in a personalized manner. This study will also further contribute to the development of personalized medicine. In future, as the consequence of advancement in integrative approaches like somatic mutations analysis with DNA methylations and gene expression studies, we could expect a well developed personalized medicine in future soon.

There are ethical issues also concerned in medical genetic studies. While studying the individual genome, it can reveal the relative DNA information or even exposed the properties of the population. Confidentiality, privacy and return of the results are some ethical issues that should be solved by professionals.

REFERENCES

Martin, JA; Wang, Z (2011). Next-generation transcriptome assembly. Nat Rev Genet. 12 (10):671–682

Vogelstein, B; Papadopoulos, N; Velculescu, VE; Zhou, S; Diaz, LA Jr; Kinzler, KW (2013). Cancer genome landscapes. Science. 339 (6127):15461558

Conrad, DF; Keebler, JE; DePristo, MA; Lindsay, SJ; Zhang, Y; Casals, F; Idaghdour, Y; Hartl, CL; Torroja, C; Garimella, KV; Zilversmit, M; Cartwright, R; Rouleau, GA; Daly, M; Stone, EA; Hurles, ME; Awadalla, P; 1000 Genomes Project (2011). Variation in genome-wide mutation rates within and between human families. Nat Genet. 43(7):712-4

Wang, K; Li, M; Hakonarson, H (2010). ANNOVAR; functional annotation of genetic variants from high throughput sequencing data. Nucleic Acid Res. 38(16).

Perkel, MJ (2010). BIG TROUBLE IN LITTLE SCIENCE Nature Methods. 7: 589–593

Jemal, A: Bray, F; Center, MM; Ferlay, J; War, E; Forman, D (2011). Global cancer statistics. CA Cancer J Clin. 61(2):69-90

Shen, MM; Abate, C (2010). Molecular genetics of prostate cancer: new prospects for old challenges.Genes Dev. GD 24(18):1967-2000

Balmukhanov, T; Khanseitova, A; Nigmatova, V; Ashirbekov, E; Talaeva, S; Aitkhozhina, N. (2013) Polymorphisms at *GSTM* 1, *GSTP* 1, *GSTT* 1 Detoxification Genes Loci and Risk of Breast Cancer in Kazakhstan Population. Advances in Breast Cancer Research. 2(4):114-118

Mamanova, L; Coffey, AJ; Scott, CE; Kozarewa, I; Turner, EH; Kumar, A; Howard, E; Shendure, J; Turner, DJ (2010). Target-enrichment strategies for next-generation sequencing. Nat Methods. 7(2):111-118

Damodaran, S; Berger, MF; Roychowdhury, S (2015). Clinical tumor sequencing: opportunities and challenges for precision cancer medicine. Europe PubMed Central. 35: e175-182

M, Meyerson; S, Gabriel; G, Getz(2010). Advances in understanding cancer genomes through second-generation sequencing. Nature Reviews Genetics. 11:685-696

Pauline,C. Ng ; Samuel, Levy; Jiaqi, Huang; Timothy, B; Stockwell; Brian, P. Walenz; Kelvin, Li; Nelson, Axelrod; Dana, A; Busam, Robert, L; Strausberg, J (2008). Genetic Variation in an Individual Human Exome. PLoS Genet. 4(8): e1000160

Adzhubei, IA; Schmidt, S; Peshkin, L; Ramensky, VE; Gerasimova, A; Bork, P; Kondrashov, AS; Sunyaev, SR (2010). Annotation of functional variation in personal genomes using RegulomeDB. Nat Methods. 7(4):248-249

Boyadjiev, SA; Jabs, EW(2000). Online Mendelian Inheritance in Man (OMIM) as a knowledge base for human developmental disorders. Clin Genet. 57(4):253-66

Hunter, DJ(2005). Gene-environment interactions in human diseases. Nat Rev Genet. 6(4):287-98.
Rakyan VK, Down TA, Balding DJ, *et al.* (2011). Epigenome-wide association studies for common human diseases. Nat Rev Genet; 12:529–41.

Spinelli, A; R, Piazza; Pirola, A; Valletta, S (2014). Whole-Exome Sequencing Data–Identifying Somatic Mutations. Springer Link. 419-427

Rabbani, B; Mahdieh, N; Hosomichi, K; Nakaoka, H; Inoue, I(2012). Next-generation sequencing: impact of exome sequencing in characterizing Mendelian disorders. J. Hum. Genet. **57**, 621–632

B, Rabbani; M, Tekin; N, Mahdieh (2014). The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics. 59: 5–15

Biesecker, L; Mullikin, J. C; Facio, F. M; Turner, C; Cherukuri, P. F; Blakesley, R. W(2009). The ClinSeq Project: piloting large-scale genome sequencing for research in genomic medicine. Genome Res. 19: 1665–1674

Ecke, TH1; Schlechte, HH; Schiemenz, K; Sachs, MD, Lenk, SV; Rudolph, BD; Loening, SA(2010). TP53 gene mutations in prostate cancer progression. Anticancer Res. 30(5):1579-86

Kanda, S; Tsuchiya, N; Narita, S; Inoue, T; Huang, M; Chiba, S; Akihama, S; Saito, M; Numakura, K; Tsuruta, H; Satoh, S; Saito, S; Ohyama, C; Arai, Y; Habuchi, T(2015). Effects of functional genetic polymorphisms in the CYP19A1 gene on prostate cancer risk and survival. Int J Cancer. 136(1):74-82

APPENDIX

```
1. annotate_variation.pl -buildver hg19 -downdb -webfrom annovar refGene humandb/
2. annotate_variation.pl -buildver hg19 -downdb cytoBand humandb/
3. annotate_variation.pl -buildver hg19 -downdb genomicSuperDups humandb/
4. annotate_variation.pl -buildver hg19 -downdb -webfrom annovar esp6500siv2_all humandb/
5. annotate_variation.pl -buildver hg19 -downdb -webfrom annovar 1000g2014oct humandb/
6. annotate_variation.pl -buildver hg19 -downdb -webfrom annovar snp138 humandb/
7. annotate_variation.pl -buildver hg19 -downdb -webfrom annovar ljb26_all humandb/
8. convert2annovar.pl -format vcf4 example/ex2.vcf > ex2.avinput
9. annotate_variation.pl -out ex1 -build hg19 example/ex1.avinput humandb/
10. annotate_variation.pl -build hg19 -downdb phastConsElements46way humandb/
11. annotate_variation.pl -regionanno -build hg19 -out ex1 -dbtype phastConsElements46way example/ex1.avinput humandb/
12. annotate_variation.pl -build hg19 -downdb tfbsConsSites humandb/
13. annotate_variation.pl -regionanno -build hg19 -out ex1 -dbtype tfbsConsSites example/ex1.avinput humandb/
14. annotate_variation.pl -build hg19 -downdb cytoBand humandb/
15. annotate_variation.pl -regionanno -build hg19 -out ex1 -dbtype cytoBand example/ex1.avinput humandb/
16. annotate_variation]$ annotate_variation.pl -build hg19 -downdb genomicSuperDups humandb/
17. annotate_variation.pl -regionanno -build hg19 -out ex1 -dbtype genomicSuperDups example/ex1.avinput humandb/
18. annotate_variation.pl -filter -dbtype ljb23_pp2hvar -buildver hg19 -out ex1 example/ex1.avinput humandb/
```