KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018

WEEK 4- PUBLIC DATA BASES

Topics covered:

- What is a Public Databases
- Different Databases available
- Example of Public Data base- RefSeq (Nucliec Acid Database)
- Two specialized databases
  - TCGA
  - COSMIC
- Input And Output For Annotation Of A Sample Vcf File
  - Case1: The data base file is in the VCF format
  - Case 2: The data base is not in the format of a VCF file

Background: For my project of annotating cancer mutations, I require to gather information from publicly available resources that are annotated for cancer mutations.

Public Databases: are repositories for nucleotide sequence data from all organisms.

There are three different types of genomic databases

Nucleic Acid Databases:

- RefSeq
- HapMap

These are repositories for nucleotide sequence data from all organism

Gene Expression Databases

- These databases collect genome sequences, annotate and analyses them and provide public access.
- Ensemble: provide automatic annotation databases for human, mouse other vertebrates and eukaryote genomes.
- 1000 Genomes Project: The genomes of more than a thousand anonymous participants and made publicly available.

Amino Acid/ Protein Databases

- Swiss Prot
- UniProt

Specialized Databases

- TCGA- The Cancer Genome Atlas
- COSMIC- Catalogue Of Somatic Mutations In Cancer

KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018

We are interested only in the annotated public databases. Annotated public databases means that following are the jobs performed and information added into the database:

1- Identifying portions of the genome that do not code for proteins

2- Identifying elements on the genome, a process called gene prediction, and

3- Attaching biological information to these elements.

Annotations are performed over these databases using annotation tools in silico and manual curation through experts.

RefSeq

- RefSeq database is a non-redundant set of reference standards derived from the INSDC databases that includes chromosomes, complete genomic molecules (organelle genomes, viruses, and plasmids), intermediate assembled genomic contigs, curated genomic regions, mRNAs, RNAs, and proteins.
- RefSeq also includes annotation which is provided by computation and manual curation

RefSeq Record File Type

The RefSeq release consists of data records stored in the form of **.fna** files NOT VCF

RefSeq processing first produces a comprehensive set of ASN.1 files. These initial files (*.bna.gz files) are further processed to export the records by molecule and format type (creating files such as *.genomic.fna.gz, *.protein.faa.gz, etc.).

RefSeq categories

Category	Description
NC	Complete genomic molecules
NG	Incomplete genomic region
NM	mRNA
NR	ncRNA
NP	Protein
XM	predicted mRNA model
XR	predicted ncRNA model
XP	predicted Protein model (eukaryotic sequences)

WP predicted Protein model (prokaryotic sequences)

RefSeq record style

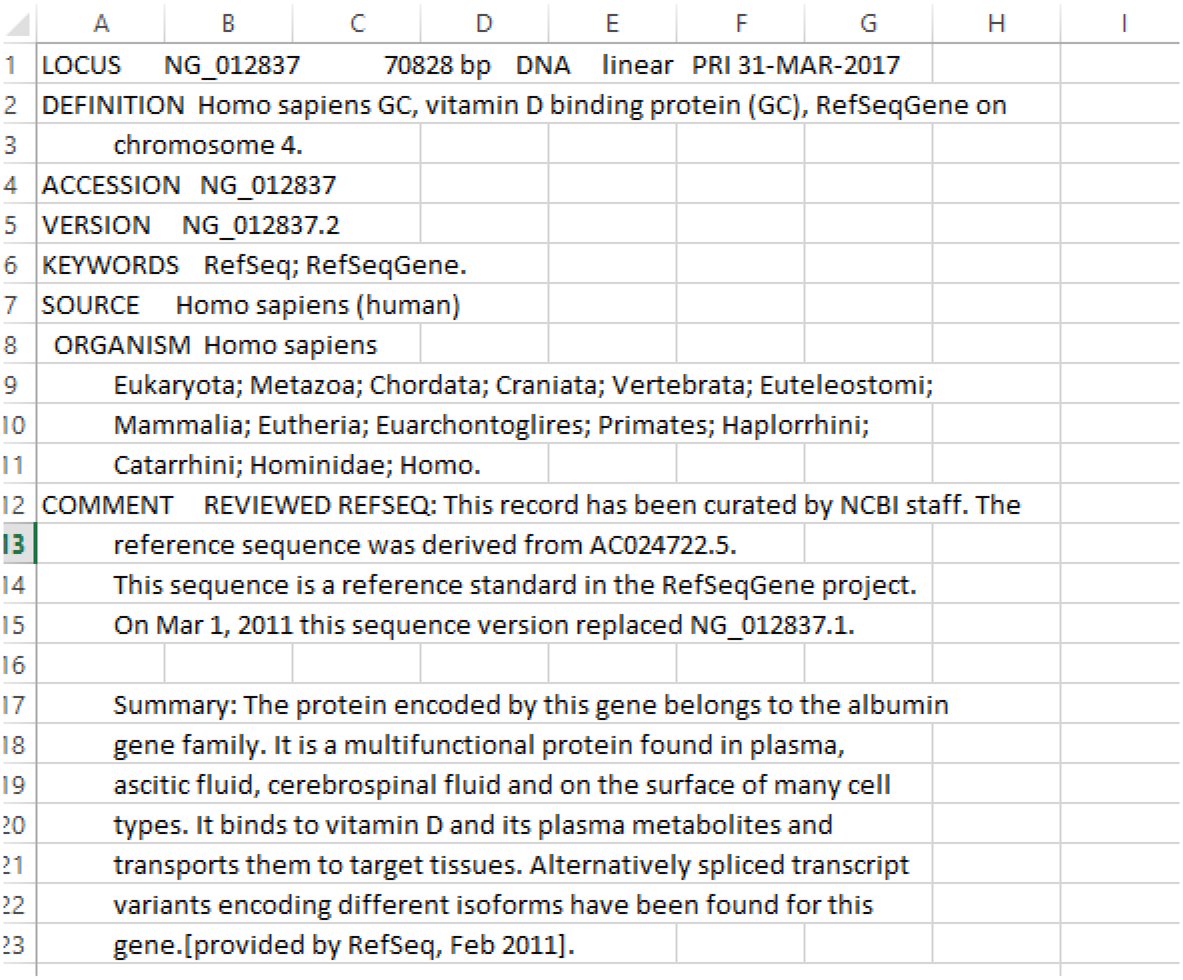
Accession format: distinct accession number format that begins with two characters followed by an underscore (e.g., NP_).

Comment: identifies the record status, the source accession(s) used to derive the RefSeq sequence, and the collaborating group, if any.

Nomenclature: use official nomenclature for the gene feature, when available.

db_xrefs: inclusion of db_xrefs on the gene or other features provides links to other sources of information, such as OMIM, Gene, UniProt, CCDS, CDD, and model-organism databases.

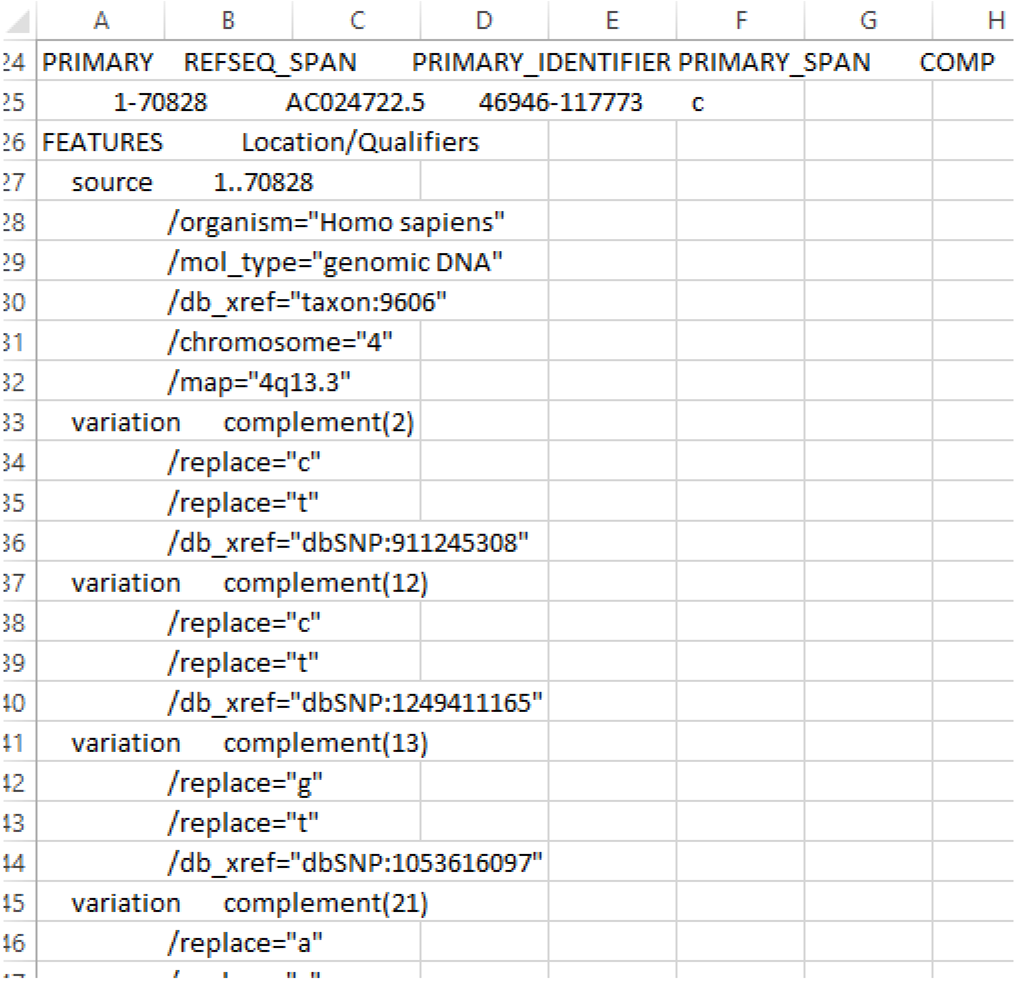
DBSOURCE: protein records indicate REFSEQ as the DBSOURCE



```
LOCUS       NG_012837          70828 bp    DNA     linear   PRI 31-MAR-2017
DEFINITION  Homo sapiens GC, vitamin D binding protein (GC), RefSeqGene on
            chromosome 4.
ACCESSION   NG_012837
VERSION     NG_012837.2
KEYWORDS    RefSeq; RefSeqGene.
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
COMMENT     REVIEWED REFSEQ: This record has been curated by NCBI staff. The
            reference sequence was derived from AC024722.5.
            This sequence is a reference standard in the RefSeqGene project.
            On Mar 1, 2011 this sequence version replaced NG_012837.1.

            Summary: The protein encoded by this gene belongs to the albumin
            gene family. It is a multifunctional protein found in plasma,
            ascitic fluid, cerebrospinal fluid and on the surface of many cell
            types. It binds to vitamin D and its plasma metabolites and
            transports them to target tissues. Alternatively spliced transcript
            variants encoding different isoforms have been found for this
            gene.[provided by RefSeq, Feb 2011].
```

KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018



	A	B	C	D	E	F	G	H
24	PRIMARY REFSEQ_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP							
25	1-70828 AC024722.5 46946-117773 c							
26	FEATURES Location/Qualifiers							
27	source 1..70828							
28	/organism="Homo sapiens"							
29	/mol_type="genomic DNA"							
30	/db_xref="taxon:9606"							
31	/chromosome="4"							
32	/map="4q13.3"							
33	variation complement(2)							
34	/replace="c"							
35	/replace="t"							
36	/db_xref="dbSNP:911245308"							
37	variation complement(12)							
38	/replace="c"							
39	/replace="t"							
40	/db_xref="dbSNP:1249411165"							
41	variation complement(13)							
42	/replace="g"							
43	/replace="t"							
44	/db_xref="dbSNP:1053616097"							
45	variation complement(21)							
46	/replace="a"							

SPECIALIZED DATABASES

TCGA- The Cancer Genome Atlas

Is the catalogue of genetic mutations responsible for cancer, using genome sequencing and bioinformatics. TCGA applies high-throughput genome analysis techniques to improve our ability to diagnose, treat, and prevent cancer through a better understanding of the genetic basis of this disease.

The GDC data portal is linked with TCGA that consists of the database which is cancer specific. The annotations has been performed using the different techniques such as:

Mutect2 annotation **11,396 files**

Varscan2 annotation **11,395**

KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018

MuSE annotation **11,387**

SomaticSniper annotation **11,398**

FM simple somatic annotation **18,004**

****The output data is available in the format of VCF file**

COSMIC

COSMIC is the Catalogue of Somatic Mutations in Cancer. Some of the useful databases found are the following:

COSMIC Complete Mutation Data (Targeted Screens): A tab separated table of the complete curated COSMIC dataset (targeted screens) from the current release. It includes all coding point mutations, and the negative data set.

COSMIC Mutation Data (Genome Screens): A tab separated table of coding point mutations from genome wide screens (including whole exome sequencing).

Structural Genomic Rearrangements: All structural variants from the current release in a tab separated table.

All Mutations in Census Genes: All coding mutations in genes listed in the Cancer Gene Census (http://cancer.sanger.ac.uk/census) in a tab separated table.

Non coding variants: A tab separated table of all non-coding mutations from the current release.

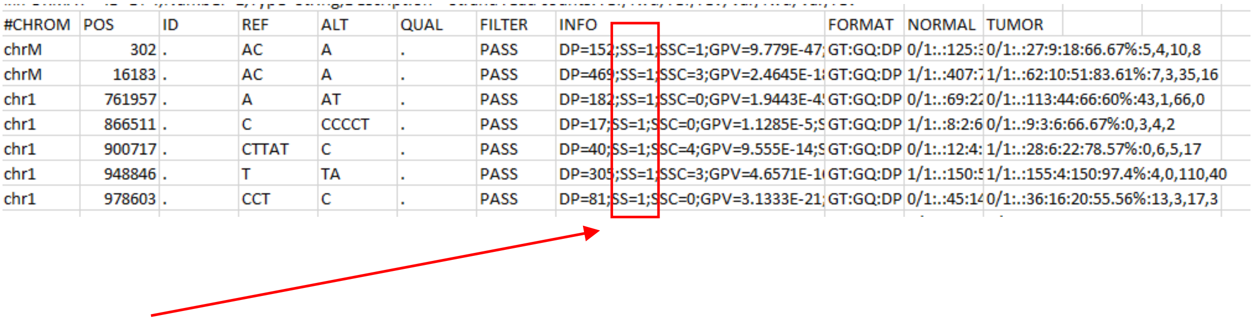
Cancer Gene Census: A list of all cancer census genes from the current release in a comma separated table. The census table is exported from http://cancer.sanger.ac.uk/census and the format is the same.

INPUT AND OUTPUT FOR ANNOTATION of a sample VCF file

CASE 1: The data base file is in the VCF format

Since the TCGA databases are in the format of VCF, the number of steps involved to output a file consisting of annotations is assumed to be less.

For example: Following is the database file from varscan2 mutated TCGA database



#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	NORMAL	TUMOR
chrM	302	.	AC	A	.	PASS	DP=152;SS=1;SSC=1;GPV=9.779E-47;	GT:GQ:DP	0/1:.:125:3	0/1:.:27:9:18:66.67%:5,4,10,8
chrM	16183	.	AC	A	.	PASS	DP=469;SS=1;SSC=3;GPV=2.4645E-1	GT:GQ:DP	1/1:.:407:7	1/1:.:62:10:51:83.61%:7,3,35,16
chr1	761957	.	A	AT	.	PASS	DP=182;SS=1;SSC=0;GPV=1.9443E-4	GT:GQ:DP	0/1:.:69:22	0/1:.:113:44:66:60%:43,1,66,0
chr1	866511	.	C	CCCCT	.	PASS	DP=17;SS=1;SSC=0;GPV=1.1285E-5;S	GT:GQ:DP	1/1:.:8:2:6	0/1:.:9:3:6:66.67%:0,3,4,2
chr1	900717	.	CTTAT	C	.	PASS	DP=40;SS=1;SSC=4;GPV=9.555E-14;S	GT:GQ:DP	0/1:.:12:4:	1/1:.:28:6:22:78.57%:0,6,5,17
chr1	948846	.	T	TA	.	PASS	DP=305;SS=1;SSC=3;GPV=4.6571E-1	GT:GQ:DP	1/1:.:150:5	1/1:.:155:4:150:97.4%:4,0,110,40
chr1	978603	.	CCT	C	.	PASS	DP=81;SS=1;SSC=0;GPV=3.1333E-21;	GT:GQ:DP	0/1:.:45:14	0/1:.:36:16:20:55.56%:13,3,17,3

KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018

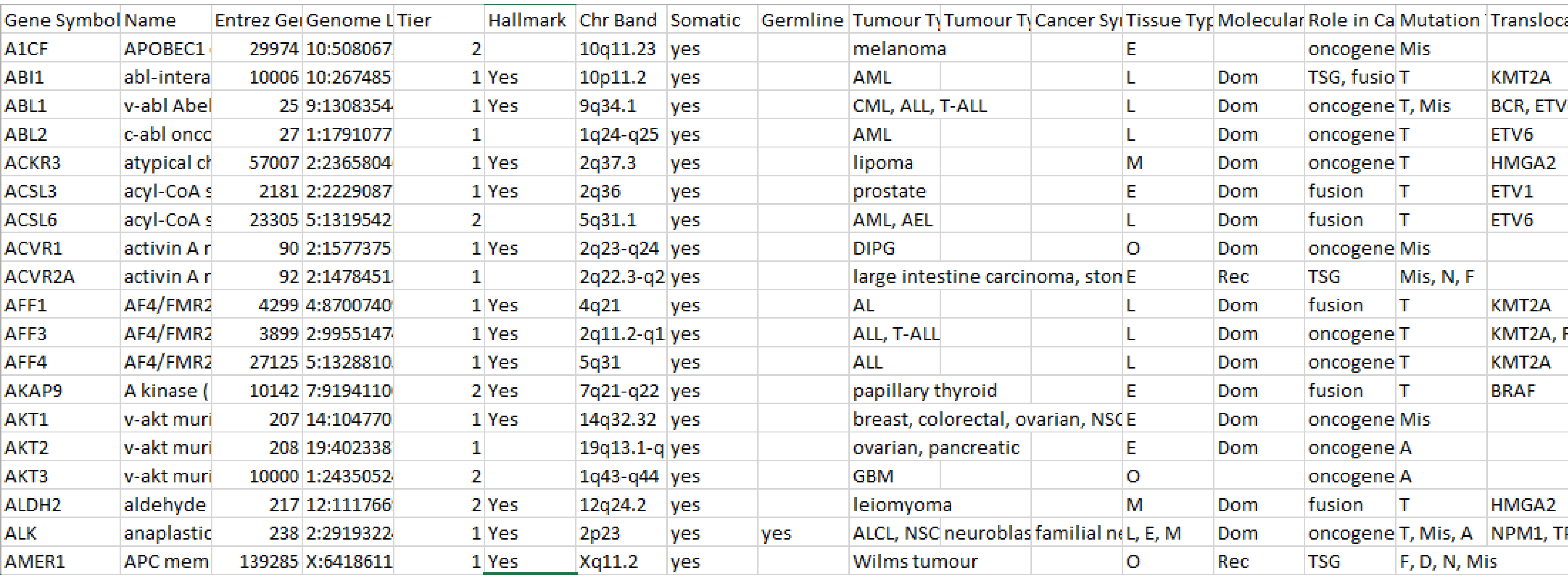
Here Somatic status of variant are represented as 0=Reference,1=Germline,2=Somatic,3=LOH, or 5=Unknown)

In order to automatically annotate a sample file:

- Traversing of the sample VCF file has to be performed that looks out for the matching first 5 columns from the database file. i.e. CHROM, POS, ID, REF and ALT
- If the first five columns of the sample VCF file matches with the first five columns of the database file then the corresponding INFO column information shall be retained.
- A new file shall be created with consisting of the sample VCF file information and new columns consisting of annotation information

CASE 2: The data base is not in the format of a VCF file

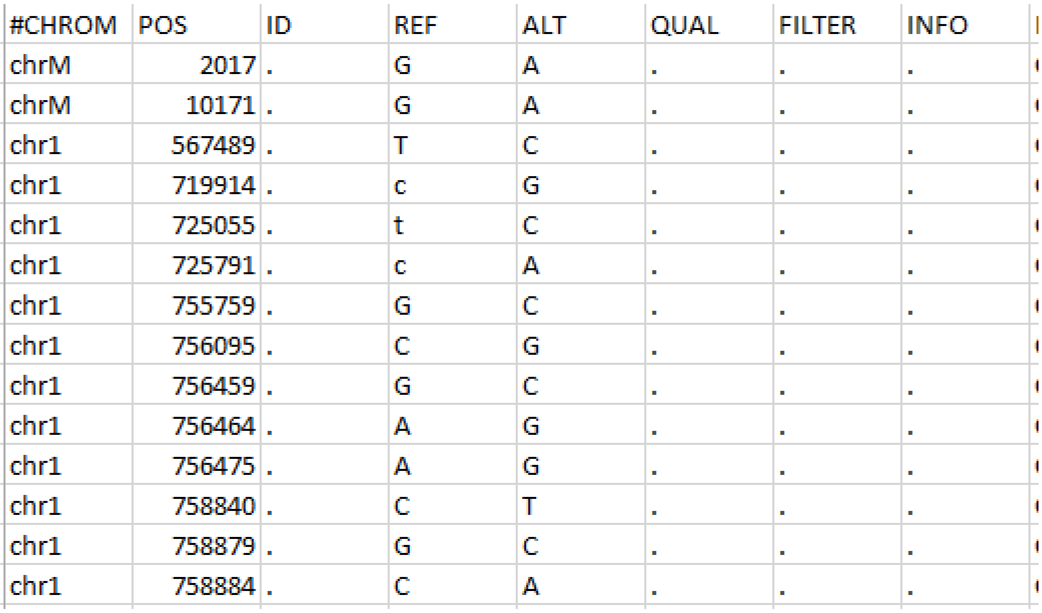
Such as in the case of the COSMIC data base, where if we look at the data Cancer Gene Census, the data is the csv format that looks like this:



Gene Symbol	Name	Entrez Ge	Genome L	Tier	Hallmark	Chr Band	Somatic	Germline	Tumour Ty	Tumour Ty	Cancer Sy	Tissue Typ	Molecular	Role in Ca	Mutation	Transloca
A1CF	APOBEC1	29974	10:508067	2		10q11.23	yes		melanoma			E		oncogene	Mis	
ABI1	abl-intera	10006	10:267485	1	Yes	10p11.2	yes		AML			L	Dom	TSG, fusio	T	KMT2A
ABL1	v-abl Abel	25	9:1308354	1	Yes	9q34.1	yes		CML, ALL, T-ALL			L	Dom	oncogene	T, Mis	BCR, ETV
ABL2	c-abl oncc	27	1:1791077	1		1q24-q25	yes		AML			L	Dom	oncogene	T	ETV6
ACKR3	atypical ch	57007	2:2365804	1	Yes	2q37.3	yes		lipoma			M	Dom	oncogene	T	HMGA2
ACSL3	acyl-CoA s	2181	2:2229087	1	Yes	2q36	yes		prostate			E	Dom	fusion	T	ETV1
ACSL6	acyl-CoA s	23305	5:1319542	2		5q31.1	yes		AML, AEL			L	Dom	fusion	T	ETV6
ACVR1	activin A r	90	2:1577375	1	Yes	2q23-q24	yes		DIPG			O	Dom	oncogene	Mis	
ACVR2A	activin A r	92	2:1478451	1		2q22.3-q2	yes		large intestine carcinoma, ston			E	Rec	TSG	Mis, N, F	
AFF1	AF4/FMR2	4299	4:8700740	1	Yes	4q21	yes		AL			L	Dom	fusion	T	KMT2A
AFF3	AF4/FMR2	3899	2:9955147	1	Yes	2q11.2-q1	yes		ALL, T-ALL			L	Dom	oncogene	T	KMT2A, F
AFF4	AF4/FMR2	27125	5:1328810	1	Yes	5q31	yes		ALL			L	Dom	oncogene	T	KMT2A
AKAP9	A kinase (	10142	7:9194110	2	Yes	7q21-q22	yes		papillary thyroid			E	Dom	fusion	T	BRAF
AKT1	v-akt muri	207	14:104770	1	Yes	14q32.32	yes		breast, colorectal, ovarian, NSC			E	Dom	oncogene	Mis	
AKT2	v-akt muri	208	19:402338	1		19q13.1-q	yes		ovarian, pancreatic			E	Dom	oncogene	A	
AKT3	v-akt muri	10000	1:2435052	2		1q43-q44	yes		GBM			O		oncogene	A	
ALDH2	aldehyde	217	12:111766	2	Yes	12q24.2	yes		leiomyoma			M	Dom	fusion	T	HMGA2
ALK	anaplastic	238	2:2919322	1	Yes	2p23	yes	yes	ALCL, NSC	neuroblas	familial ne	L, E, M	Dom	oncogene	T, Mis, A	NPM1, TF
AMER1	APC mem	139285	X:6418611	1	Yes	Xq11.2	yes		Wilms tumour			O	Rec	TSG	F, D, N, Mis	

However, our sample VCF file looks like this, for example:

KANZA BATOOL HAIDER
WEEK 4 INDEPENDENT STUDY
Supervisor: Prof. Mehmet Baysan
Date: 22 July, 2018



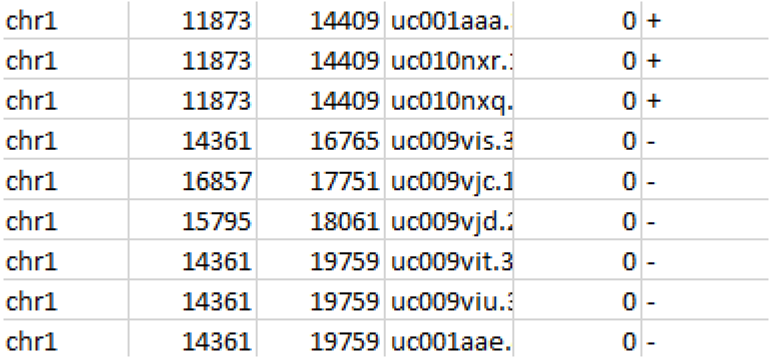
#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
chrM	2017	.	G	A	.	.	.
chrM	10171	.	G	A	.	.	.
chr1	567489	.	T	C	.	.	.
chr1	719914	.	c	G	.	.	.
chr1	725055	.	t	C	.	.	.
chr1	725791	.	c	A	.	.	.
chr1	755759	.	G	C	.	.	.
chr1	756095	.	C	G	.	.	.
chr1	756459	.	G	C	.	.	.
chr1	756464	.	A	G	.	.	.
chr1	756475	.	A	G	.	.	.
chr1	758840	.	C	T	.	.	.
chr1	758879	.	G	C	.	.	.
chr1	758884	.	C	A	.	.	.

Now, inorder to add annotations from the database file we shall have to :

First add the names of the genes into the sample file. Inorder to do that, there are some steps involved:

First download the BED format file with gene names of the GRCh37/hg19 whole genome.

The BED file looks like this:



chr1	11873	14409	uc001aaa.	0	+
chr1	11873	14409	uc010nxr.1	0	+
chr1	11873	14409	uc010nxq.	0	+
chr1	14361	16765	uc009vis.3	0	-
chr1	16857	17751	uc009vjc.1	0	-
chr1	15795	18061	uc009vjd.2	0	-
chr1	14361	19759	uc009vit.3	0	-
chr1	14361	19759	uc009viu.3	0	-
chr1	14361	19759	uc001aae.	0	-

The first three fields are required which are:

chrom - The name of the chromosome (e.g. chr3, chrY, chr2_random) or scaffold (e.g. scaffold10671).

chromStart - The starting position of the feature in the chromosome or scaffold. The first base in a chromosome is numbered 0.

chromEnd - The ending position of the feature in the chromosome or scaffold.

Then, there are some tools/coding involved to add the gene names from the BED file to the VCF file. The possible ways are listed below:

- GATK
- VarriantAnnotator
- Vcftools annotate
- Bcftools
- VCFBed(JAVA)

Once the VCF file is with the gene names, we can now trace with the gene location and alterations matching with the database file. *(need to discuss and do more research)*

Create a new VCF file and add annotated columns indicating mutations as somatic/germline etc taking corresponding information from the database file.