Kanza Batool Haider
Data Science, MS student
kanzahaider@std.sehir.edu.tr
Date: 08/05/2018

Thesis Project Synopsis

Project Title: Automatically annotate input vcf file for cancer variants using public databases.

Project Supervisor: Prof. Mehmet Baysan

Department: Data Science, MS- *Istanbul Sehir University*

Expected time duration for completion of project: ~5 months

Project Summary:

Prerequisites: UNIX/Linux

Python 3 or higher

Samtools, Bcftools, hugo_gene_symbols, hgvs *(and more)*

Background Overview: Given that the Genomic Pipeline for analysing genomic sequence data has been already developed and functional in the Baysan Lab of Bioinformatics which intakes a raw FASTQ file (text file produced by sequencing machines) and ends with a VCF files (detailed files of discovered SNPs/Indels). The next step involves to process the vcf file consisting of variants to indicate for the endangered cancerous mutations.

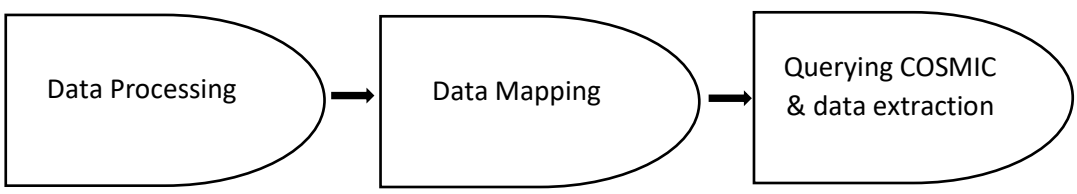
Hence, this project tends to provide the automatic tool which is specific for cancer variant annotations that shall be linked to the genomic pipeline making a complete one-shot solution for doctors and medical researchers to resort for sequence interpretations for any deleterious cancer indicators and a detailed report.

Project Description/Scope:

Genetic differences among individual include single nucleotide polymorphism (SNP), intragenic deletion and insertion (indels) and structural variants such as copy number variants. Some of these gives the rise to tumour as there happens complex interplay among mutations.

The *automatic annotation tool* shall have the ability to identify variants across conserved regions through the sample input VCF file. Global efforts in the research has made it feasible to identify variants and their characteristics which are stored in the publicly accessible data resources like 1000 genomes, dbSNP, RefSeq that are some of the credible data resources. Additionally, the specialised databases such as COSMIC and TCGA provides cancer specific curated data. Hence, the variants present in the VCF input file shall be mapped across the sequenced, referenced and curated specialized TCGA datasets and for more comprehensive detail for each variant, the variant shall be linked to the particular curated gene and mutation causing protein change (HGVSp) found in the COSMIC database.

The project development shall involve following phases:



Kanza Batool Haider
Data Science, MS student
kanzahaider@std.sehir.edu.tr
Date: 08/05/2018

Project Phases

-1 Data Processing : *Database download + converting GRCh37 database into hg19 + input vcf and add Hugo Symbols + HGVSp values (*protein sequence name value*) to variants in the vcf file*

-2 Data Mapping: *map variants of vcf file for the exact variants present in the databases + add corresponding annotations from the database into the vcf file.*

-3 Querying COSMIC and data extraction: *make a query in COSMIC + extract information + link information with each variant + compilation of the report and output*

Types of annotation to be reported:

The tool shall be reporting following information to variants in input vcf

Variant Classification	Variant Type	Mutation type
Non coding	SNP	Somatic/NA
Missense	Indel	
Synonymous	Structural Variant	
Frameshift		
Intron variant		
Insertion		
3 prime UTR insertion		
Stop gained		
Non coding exon variant		
5 Prime utr variant		
3'Flank		
Targeted_Region		
In_Frame_Del		
RNA		

Different Types of Cancers included:

There are 225 types of cancers which resides under the following:

PanCancer studies	Cell lines	Adrenal Gland
Ampulla of Vater	Biliary Tract	Bladder/Urinary Tract
Bone	Breast	CNS/Brain
Cervix	Esophagus/Stomach	Eye
Head and Neck	Kidney	Liver
Lungs	Lymphoid	Myeloid
Ovary/Fallopian tube	Pancreas	Peripheral Nervous System
Pleura	Prostate	Skin
Soft tissue	Testis	Thymus
Thyroid	Uterus	

Kanza Batool Haider
Data Science, MS student
kanzahaider@std.sehir.edu.tr
Date: 08/05/2018

Public Data Base to be used:

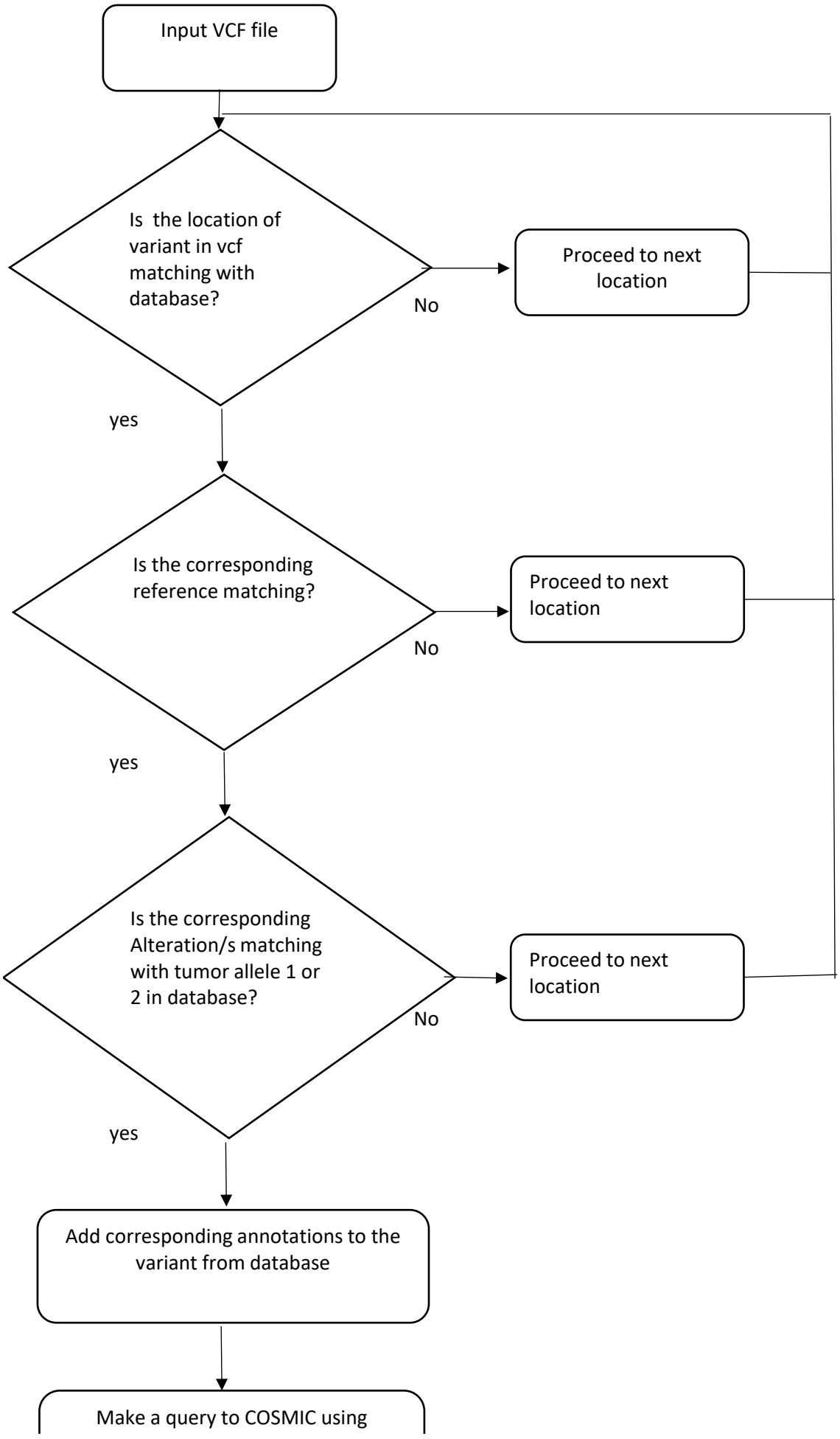
For the project, the help shall be taken from two specialised curated databases. These are:

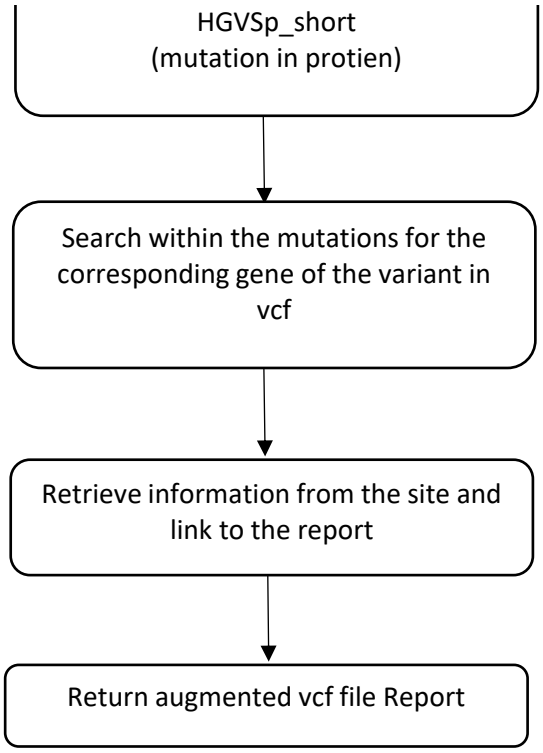
The cBioPortal for Cancer Genomics : is an open-access, open-source resource for interactive exploration of multidimensional cancer genomics data sets. The datasets are directly from TCGA data center partly via Broad Firehose which are updated regularly. All the mutation data (VCF or MAF) are processed through an internal pipeline to annotate the variant effects in a consistent way across studies. http://www.cbioportal.org/data_sets.jsp

COSMIC: the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer. https://cancer.sanger.ac.uk/cosmic

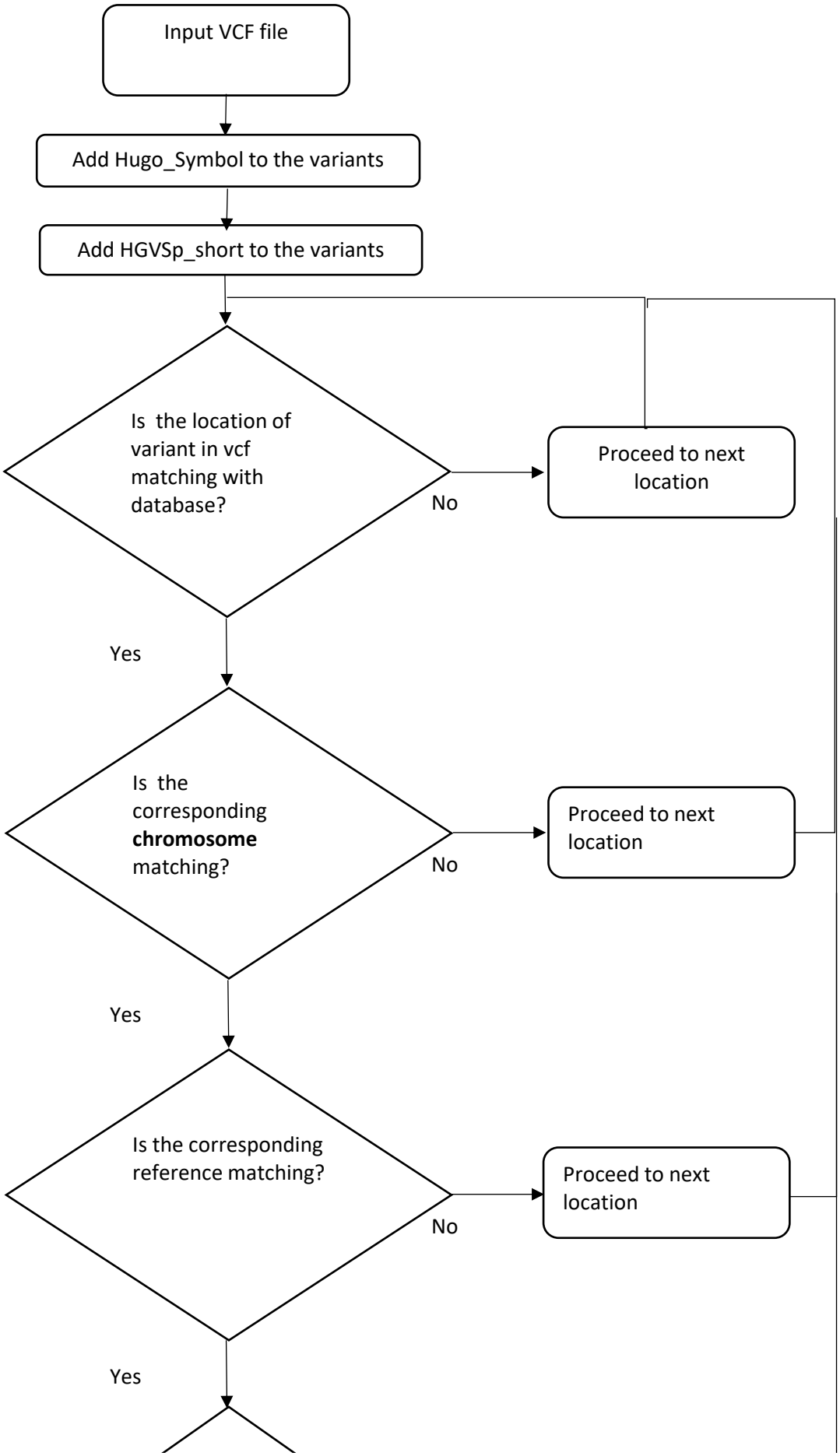
Project Flowchart:

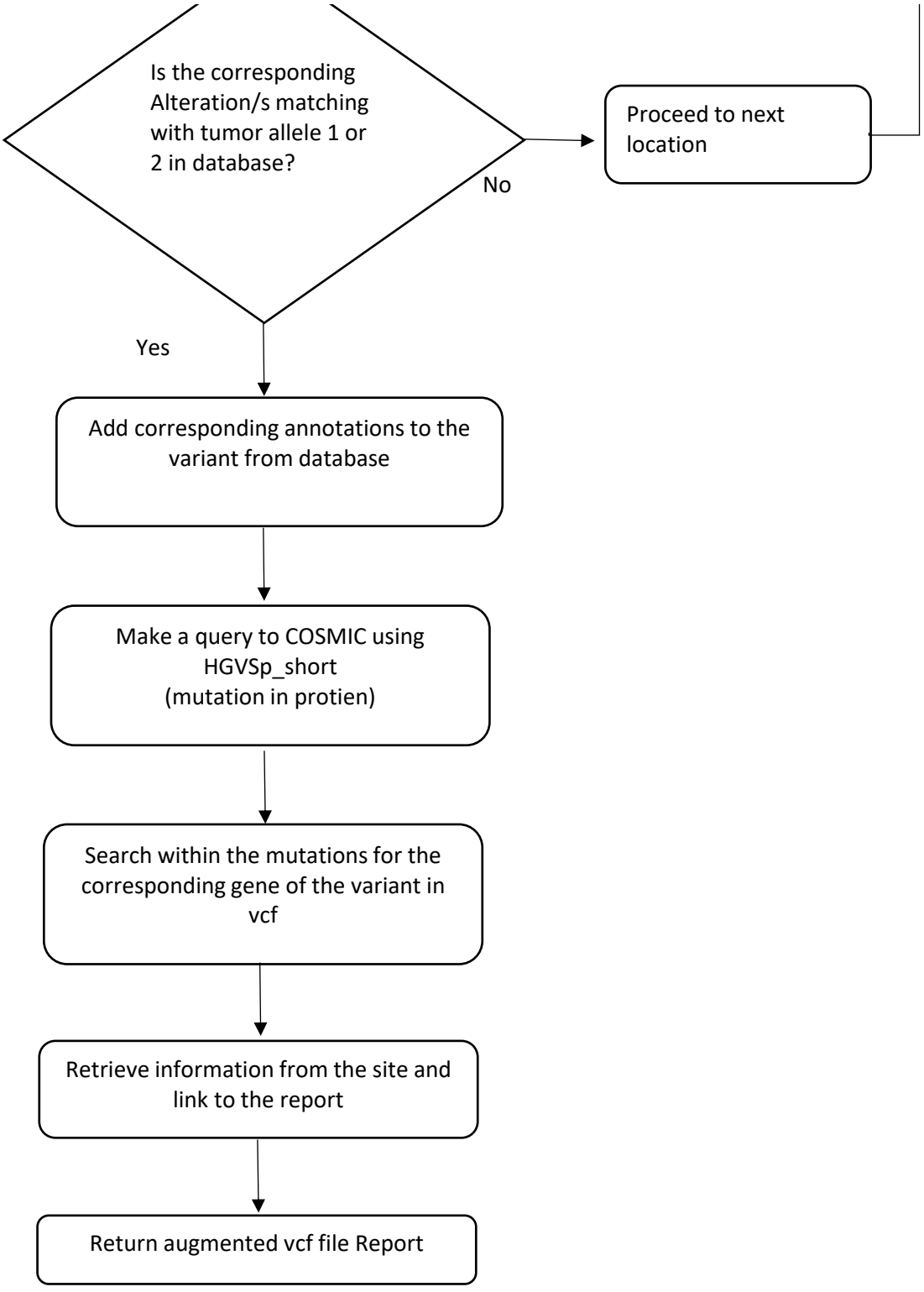
There are two approaches for the project. Please refer to the two different approaches shown in Flowcharts in the next page.


Input VCF file
Is the location of variant in vcf matching with database?
No
Proceed to next location
yes
Is the corresponding reference matching?
No
Proceed to next location
yes
Is the corresponding Alteration/s matching with tumor allele 1 or 2 in database?
No
Proceed to next location
yes
Add corresponding annotations to the variant from database
Make a query to COSMIC using


HGVSp_short
(mutation in protien)
Search within the mutations for the corresponding gene of the variant in vcf
Retrieve information from the site and link to the report
Return augmented vcf file Report

Alternative Approach




Is the corresponding Alteration/s matching with tumor allele 1 or 2 in database?
No
Proceed to next location
Yes
Add corresponding annotations to the variant from database
Make a query to COSMIC using HGVSp_short (mutation in protien)
Search within the mutations for the corresponding gene of the variant in vcf
Retrieve information from the site and link to the report
Return augmented vcf file Report