DGS PROJECT

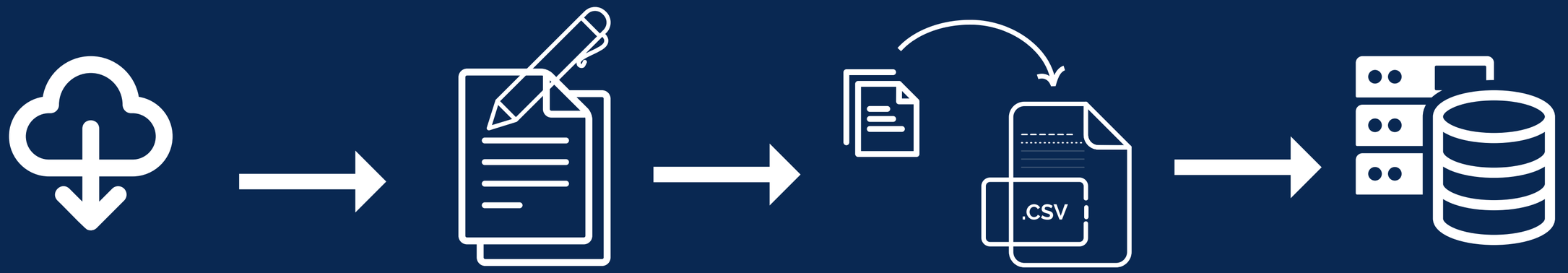
ALBERTO SÁNCHEZ S.

2023-2024

DGS PROJECT

ALBERTO SÁNCHEZ S.

2023-2024



Download Annotation Transform Store

Download

```
file_path = “./files/clinvar.vcf”
url = ”https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar.vcf.gz”


response = requests.get(url)

if response.status_code == 200:
  with open(file_path, ’wb’) as file:
    file.write(response.content)


  with gzip.open(file_path, ’rb’) as f_in:
    with open(file_path[:-3], ’wb’) as f_out:
      f_out.write(f_in.read())


  os.remove(file_path)
```

Download

```
file_path = “./files/clinvar.vcf”
url = ”https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar.vcf.gz”
```

```
response = requests.get(url)

if response.status_code == 200:
  with open(file_path, ’wb’) as file:
    file.write(response.content)
```

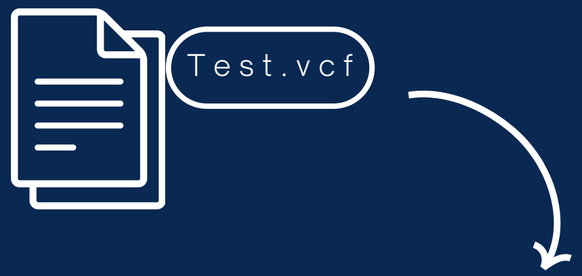
→ Download the file

```
  with gzip.open(file_path, ’rb’) as f_in:
    with open(file_path[:-3], ’wb’) as f_out:
      f_out.write(f_in.read())
```

→ Extract the content

```
  os.remove(file_path)
```

→ Remove the old zip file



Annotation

```
$java -jar ./snpEff/snpEff.jar GRCH37.75 test.vcf > testAnnEff.vcf
```

Annotation

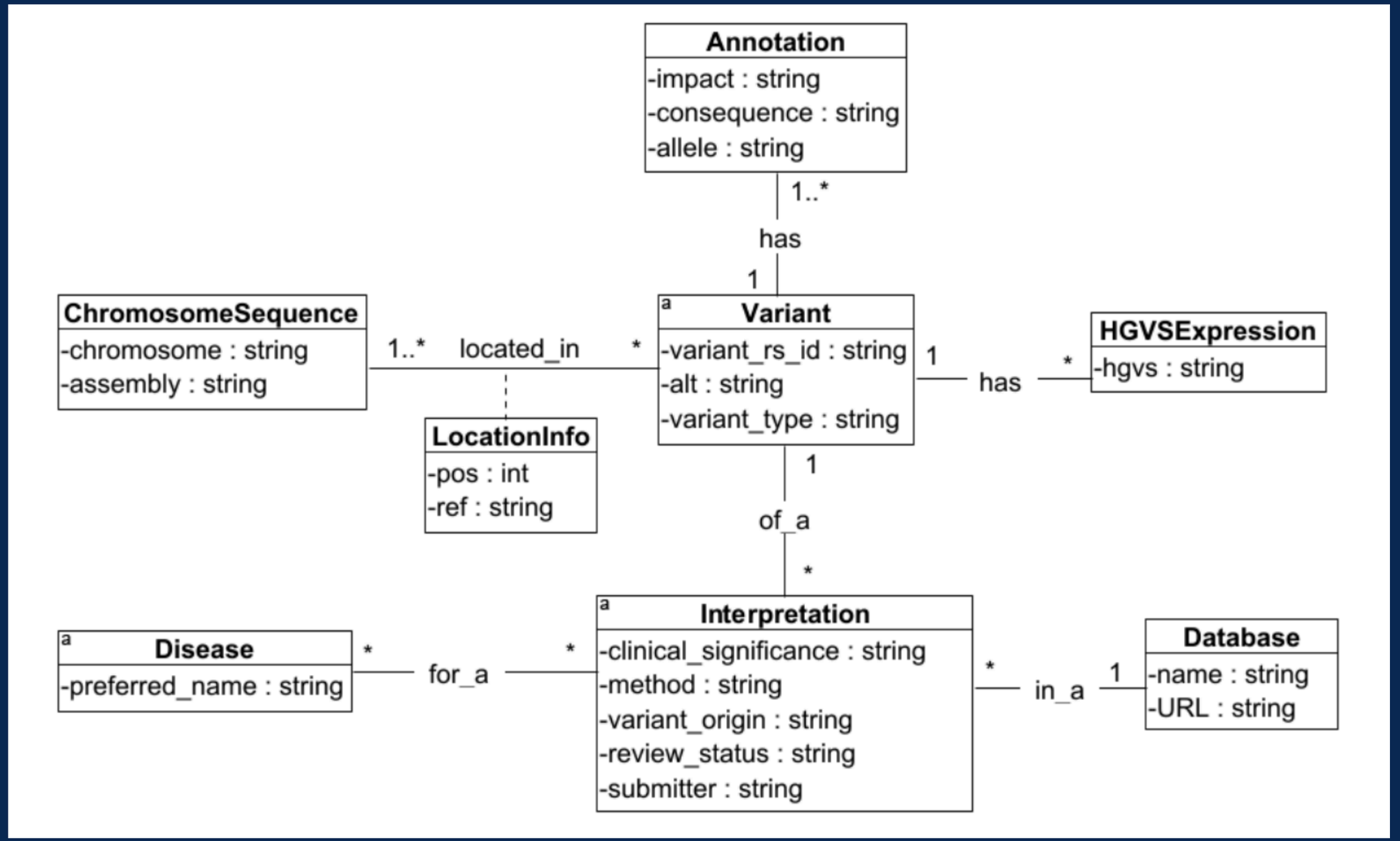
```
$java -jar ./snpEff/snpEff.jar GRCH37.75 test.vcf > testAnnEff.vcf
```

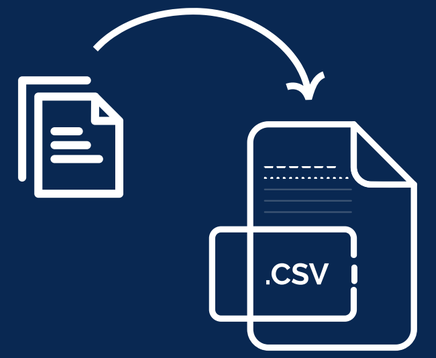

```
$java -jar ./snpEff/SnpSift.jar annotate clinvar.vcf testAnnEff.vcf > testAnnSift.vcf
```

Annotation


```
$java -jar ./snpEff/snpEff.jar GRCH37.75 test.vcf > testAnnEff.vcf
```

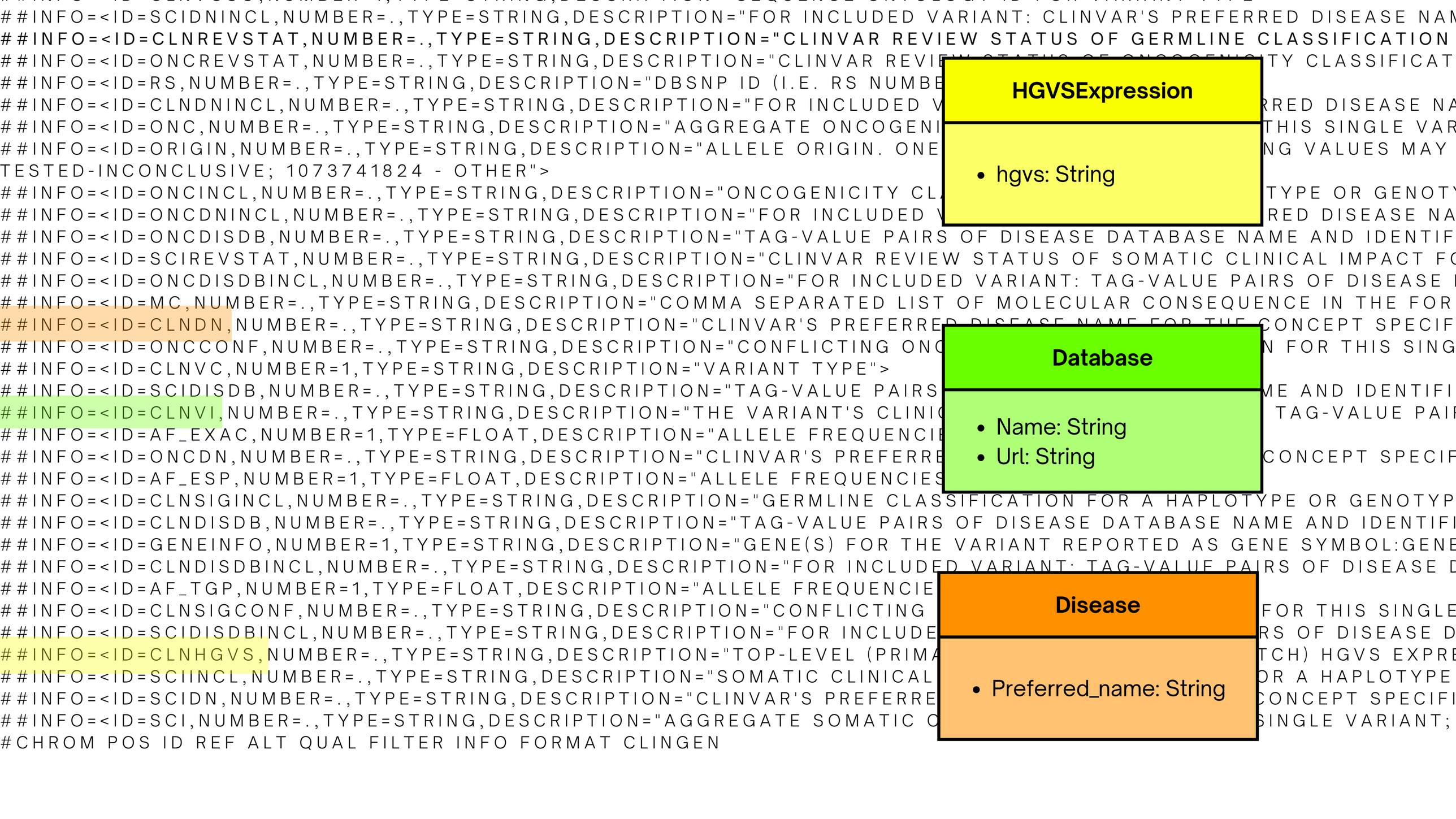

```
$java -jar ./snpEff/SnpSift.jar annotate clinvar.vcf testAnnEff.vcf > testAnnSift.vcf
```

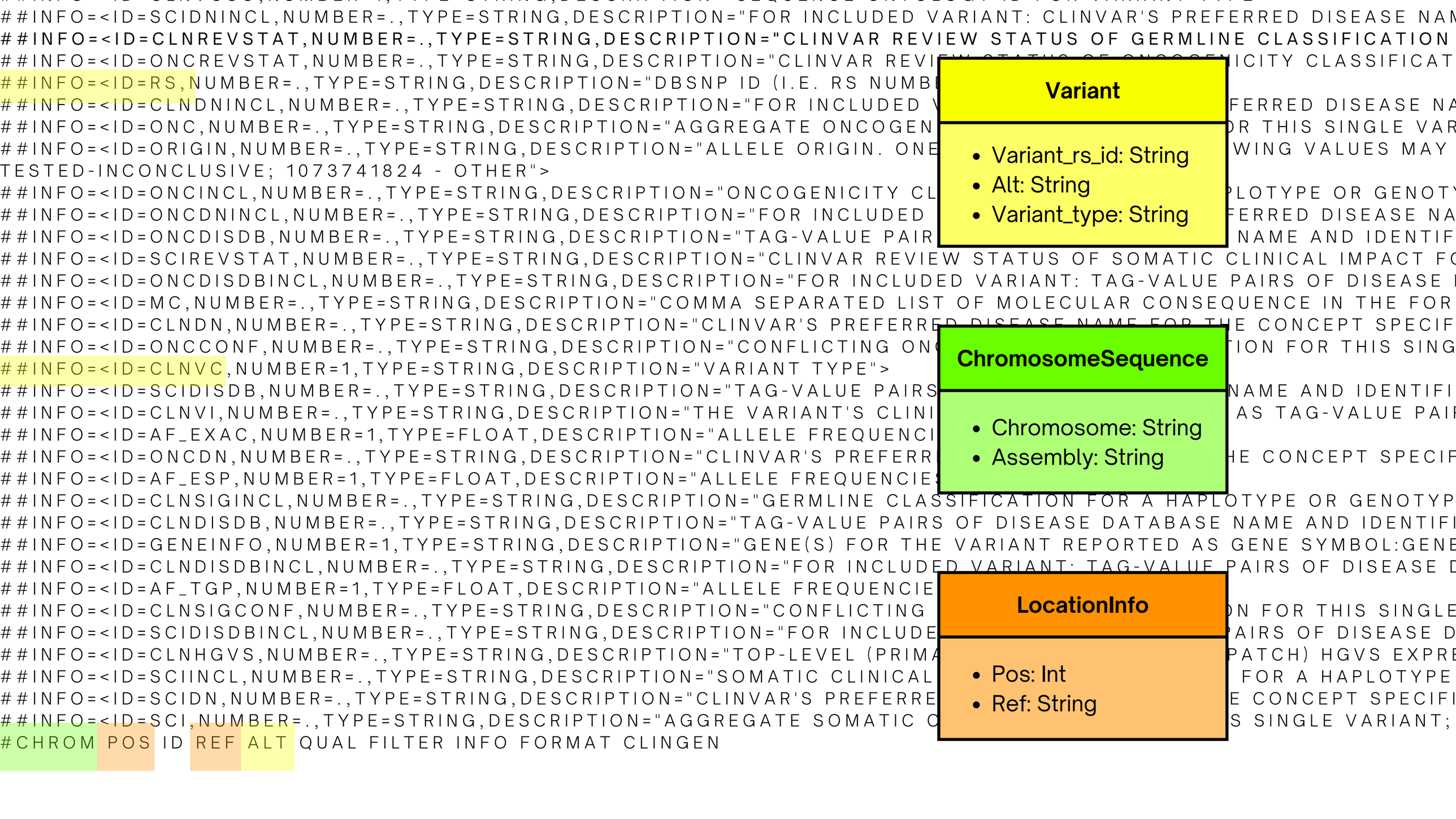

Annotation
-impact : string
-consequence : string
-allele : string
1..*
has
1
ChromosomeSequence
-chromosome : string
-assembly : string
1..*
located_in
*
a Variant
-variant_rs_id : string
-alt : string
-variant_type : string
1
has
*
HGVSExpression
-hgvs : string
LocationInfo
-pos : int
-ref : string
1
of_a
*
a Disease
-preferred_name : string
*
for_a
*
a Interpretation
-clinical_significance : string
-method : string
-variant_origin : string
-review_status : string
-submitter : string
*
in_a
1
Database
-name : string
-URL : string

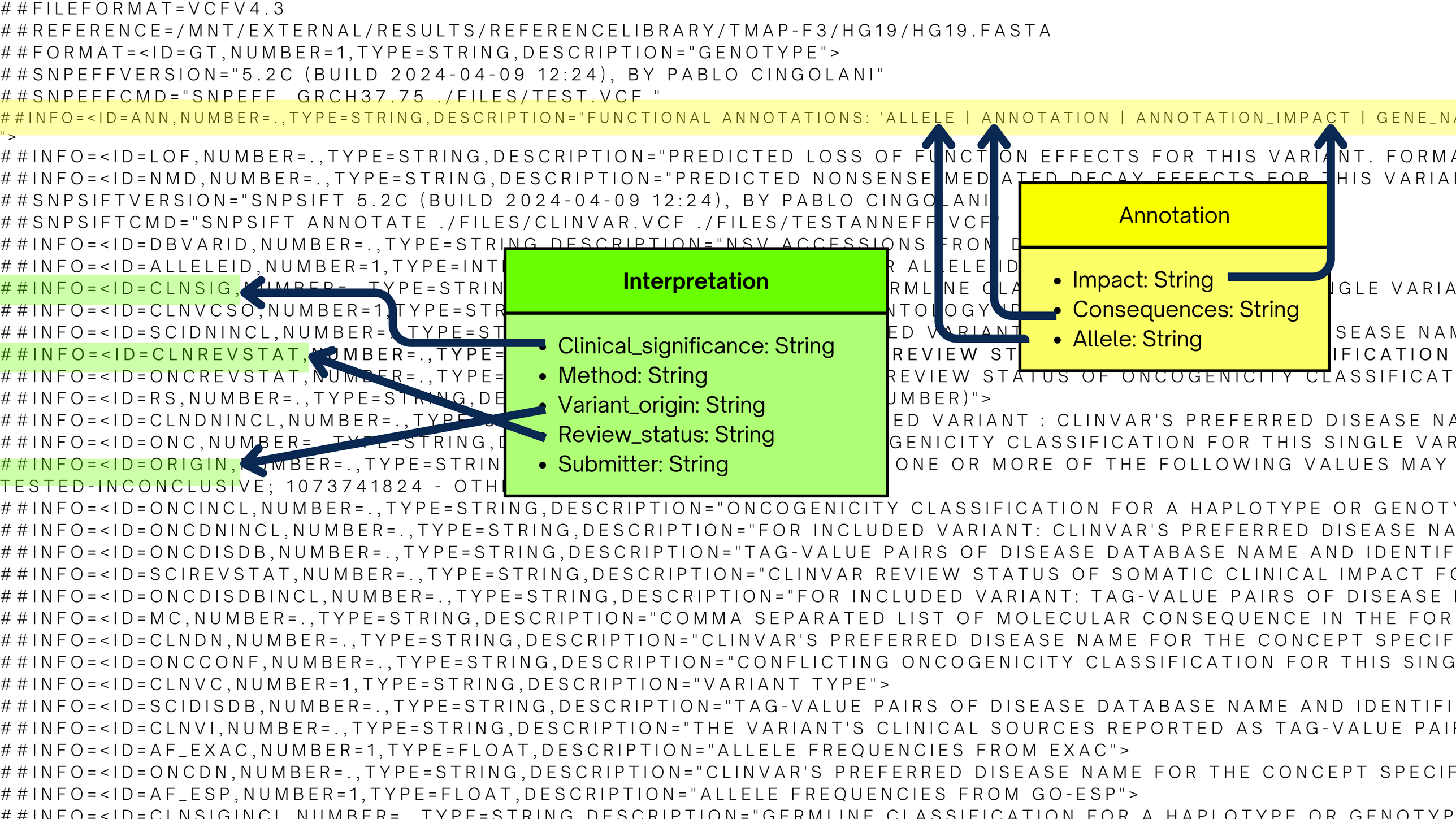

.csv

Transform

##INFO=<ID=SCIDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: CLINVAR'S PREFERRED DISEASE NAM
##INFO=<ID=CLNREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF GERMLINE CLASSIFICATION
##INFO=<ID=ONCREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF ONCOGENICITY CLASSIFICAT
##INFO=<ID=RS,NUMBER=.,TYPE=STRING,DESCRIPTION="DBSNP ID (I.E. RS NUMBER)">
##INFO=<ID=CLNDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT : CLINVAR'S PREFERRED DISEASE NA
##INFO=<ID=ONC,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE ONCOGENICITY CLASSIFICATION FOR THIS SINGLE VAR
##INFO=<ID=ORIGIN,NUMBER=.,TYPE=STRING,DESCRIPTION="ALLELE ORIGIN. ONE OR MORE OF THE FOLLOWING VALUES MAY
TESTED-INCONCLUSIVE; 1073741824 - OTHER">
##INFO=<ID=ONCINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="ONCOGENICITY CLASSIFICATION FOR A HAPLOTYPE OR GENOTY
##INFO=<ID=ONCDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: CLINVAR'S PREFERRED DISEASE NA
##INFO=<ID=ONCDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIF
##INFO=<ID=SCIREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF SOMATIC CLINICAL IMPACT FO
##INFO=<ID=ONCDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE
##INFO=<ID=MC,NUMBER=.,TYPE=STRING,DESCRIPTION="COMMA SEPARATED LIST OF MOLECULAR CONSEQUENCE IN THE FOR
##INFO=<ID=CLNDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRED DISEASE NAME FOR THE CONCEPT SPECIF
##INFO=<ID=ONCCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING ONCOGENICITY CLASSIFICATION FOR THIS SING
##INFO=<ID=CLNVC,NUMBER=1,TYPE=STRING,DESCRIPTION="VARIANT TYPE">
##INFO=<ID=SCIDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIFI
##INFO=<ID=CLNVI,NUMBER=.,TYPE=STRING,DESCRIPTION="THE VARIANT'S CLINICAL SOURCES REPORTED AS TAG-VALUE PAI
##INFO=<ID=AF_EXAC,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIES FROM EXAC">
##INFO=<ID=ONCDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRED DISEASE NAME FOR THE CONCEPT SPECIF
##INFO=<ID=AF_ESP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIES FROM GO-ESP">
##INFO=<ID=CLNSIGINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="GERMLINE CLASSIFICATION FOR A HAPLOTYPE OR GENOTYP
##INFO=<ID=CLNDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIFI
##INFO=<ID=GENEINFO,NUMBER=1,TYPE=STRING,DESCRIPTION="GENE(S) FOR THE VARIANT REPORTED AS GENE SYMBOL:GENE
##INFO=<ID=CLNDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE D
##INFO=<ID=AF_TGP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIES FROM TGP">
##INFO=<ID=CLNSIGCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING GERMLINE CLASSIFICATION FOR THIS SINGLE
##INFO=<ID=SCIDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE D
##INFO=<ID=CLNHGVS,NUMBER=.,TYPE=STRING,DESCRIPTION="TOP-LEVEL (PRIMARY ASSEMBLY, ALT, OR PATCH) HGVS EXPRE
##INFO=<ID=SCIINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="SOMATIC CLINICAL IMPACT CLASSIFICATION FOR A HAPLOTYPE
##INFO=<ID=SCIDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRED DISEASE NAME FOR THE CONCEPT SPECIFI
##INFO=<ID=SCI,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE SOMATIC CLINICAL IMPACT FOR THIS SINGLE VARIANT;
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CLINGEN


##INFO=<ID=SCIDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: CLINVAR'S PREFERRED DISEASE NAM
##INFO=<ID=CLNREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF GERMLINE CLASSIFICATION
##INFO=<ID=ONCREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF ONCOGENICITY CLASSIFICAT
##INFO=<ID=RS,NUMBER=.,TYPE=STRING,DESCRIPTION="DBSNP ID (I.E. RS NUMBE
##INFO=<ID=CLNDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED V
##INFO=<ID=ONC,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE ONCOGENI
##INFO=<ID=ORIGIN,NUMBER=.,TYPE=STRING,DESCRIPTION="ALLELE ORIGIN. ONE
TESTED-INCONCLUSIVE; 1073741824 - OTHER">
##INFO=<ID=ONCINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="ONCOGENICITY CL
##INFO=<ID=ONCDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED V
##INFO=<ID=ONCDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIF
##INFO=<ID=SCIREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF SOMATIC CLINICAL IMPACT FO
##INFO=<ID=ONCDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE
##INFO=<ID=MC,NUMBER=.,TYPE=STRING,DESCRIPTION="COMMA SEPARATED LIST OF MOLECULAR CONSEQUENCE IN THE FOR
##INFO=<ID=CLNDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRED DISEASE NAME FOR THE CONCEPT SPECIF
##INFO=<ID=ONCCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING ONC
##INFO=<ID=CLNVC,NUMBER=1,TYPE=STRING,DESCRIPTION="VARIANT TYPE">
##INFO=<ID=SCIDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS
##INFO=<ID=CLNVI,NUMBER=.,TYPE=STRING,DESCRIPTION="THE VARIANT'S CLINIC
##INFO=<ID=AF_EXAC,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIE
##INFO=<ID=ONCDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRE
##INFO=<ID=AF_ESP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIES
##INFO=<ID=CLNSIGINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="GERMLINE CLASSIFICATION FOR A HAPLOTYPE OR GENOTYP
##INFO=<ID=CLNDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIFI
##INFO=<ID=GENEINFO,NUMBER=1,TYPE=STRING,DESCRIPTION="GENE(S) FOR THE VARIANT REPORTED AS GENE SYMBOL:GENE
##INFO=<ID=CLNDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE D
##INFO=<ID=AF_TGP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIE
##INFO=<ID=CLNSIGCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING
##INFO=<ID=SCIDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDE
##INFO=<ID=CLNHGVS,NUMBER=.,TYPE=STRING,DESCRIPTION="TOP-LEVEL (PRIMA
##INFO=<ID=SCIINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="SOMATIC CLINICAL
##INFO=<ID=SCIDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRE
##INFO=<ID=SCI,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE SOMATIC C
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CLINGEN
HGVSExpression
hgvs: String
Database
Name: String
Url: String
Disease
Preferred_name: String


##INFO=<ID=SCIDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: CLINVAR'S PREFERRED DISEASE NAM
##INFO=<ID=CLNREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF GERMLINE CLASSIFICATION
##INFO=<ID=ONCREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF ONCOGENICITY CLASSIFICAT
##INFO=<ID=RS,NUMBER=.,TYPE=STRING,DESCRIPTION="DBSNP ID (I.E. RS NUMBE
##INFO=<ID=CLNDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED V
##INFO=<ID=ONC,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE ONCOGEN
##INFO=<ID=ORIGIN,NUMBER=.,TYPE=STRING,DESCRIPTION="ALLELE ORIGIN. ONE
TESTED-INCONCLUSIVE; 1073741824 - OTHER">
##INFO=<ID=ONCINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="ONCOGENICITY CL
##INFO=<ID=ONCDNINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED
##INFO=<ID=ONCDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIR
##INFO=<ID=SCIREVSTAT,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR REVIEW STATUS OF SOMATIC CLINICAL IMPACT FO
##INFO=<ID=ONCDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE
##INFO=<ID=MC,NUMBER=.,TYPE=STRING,DESCRIPTION="COMMA SEPARATED LIST OF MOLECULAR CONSEQUENCE IN THE FOR
##INFO=<ID=CLNDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRED DISEASE NAME FOR THE CONCEPT SPECIF
##INFO=<ID=ONCCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING ON
##INFO=<ID=CLNVC,NUMBER=1,TYPE=STRING,DESCRIPTION="VARIANT TYPE">
##INFO=<ID=SCIDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS
##INFO=<ID=CLNVI,NUMBER=.,TYPE=STRING,DESCRIPTION="THE VARIANT'S CLINI
##INFO=<ID=AF_EXAC,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCI
##INFO=<ID=ONCDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERR
##INFO=<ID=AF_ESP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIES
##INFO=<ID=CLNSIGINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="GERMLINE CLASSIFICATION FOR A HAPLOTYPE OR GENOTYP
##INFO=<ID=CLNDISDB,NUMBER=.,TYPE=STRING,DESCRIPTION="TAG-VALUE PAIRS OF DISEASE DATABASE NAME AND IDENTIFI
##INFO=<ID=GENEINFO,NUMBER=1,TYPE=STRING,DESCRIPTION="GENE(S) FOR THE VARIANT REPORTED AS GENE SYMBOL:GENE
##INFO=<ID=CLNDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDED VARIANT: TAG-VALUE PAIRS OF DISEASE D
##INFO=<ID=AF_TGP,NUMBER=1,TYPE=FLOAT,DESCRIPTION="ALLELE FREQUENCIE
##INFO=<ID=CLNSIGCONF,NUMBER=.,TYPE=STRING,DESCRIPTION="CONFLICTING
##INFO=<ID=SCIDISDBINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="FOR INCLUDE
##INFO=<ID=CLNHGVS,NUMBER=.,TYPE=STRING,DESCRIPTION="TOP-LEVEL (PRIMA
##INFO=<ID=SCIINCL,NUMBER=.,TYPE=STRING,DESCRIPTION="SOMATIC CLINICAL
##INFO=<ID=SCIDN,NUMBER=.,TYPE=STRING,DESCRIPTION="CLINVAR'S PREFERRE
##INFO=<ID=SCI,NUMBER=.,TYPE=STRING,DESCRIPTION="AGGREGATE SOMATIC C
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT CLINGEN
Variant
Variant_rs_id: String
Alt: String
Variant_type: String
ChromosomeSequence
Chromosome: String
Assembly: String
LocationInfo
Pos: Int
Ref: String


Annotation
Impact: String
Consequences: String
Allele: String
Interpretation
Clinical_significance: String
Method: String
Variant_origin: String
Review_status: String
Submitter: String

Transform

Read the file

```
path ="./files/testAnnSift.vcf"
with open(path, 'r') as f:
    lines = [l for l in f if not l.startswith('##')]
```

Store the fields

Search the required information

Return de dataframe to then create the csv file

```
def vcf_annotation():
    data = []
    for line in lines:
        if line.startswith('#'):  # Skip header lines
            continue
        fields = line.strip().split('\t')
        chr, pos, _, ref, _, _, info, *_ = fields

        impact = ''
        consequences = ''
        allele = ''
        rs_id = ''

        for field in info.split(';'):
            if field.startswith('ANN='):
                ann_info = field.split('=')[1]
                ann_fields = ann_info.split('|')
                impact = ann_fields[2]
                consequences = ann_fields[1]
                allele = ann_fields[0]
            if field.startswith('RS='):
                rs_id = field.split('=')[1]

        data.append([impact, consequences, allele, rs_id])

    return pd.DataFrame(data, columns=['IMPACT', 'CONSEQUENCES', 'ALLELE', 'RS_ID'])
```

Store

```
# Database credentials
db_name = 'bembogzwheeytajy6wqi'
user = 'ueegmnxqcoesneek'
password = 'M90MAudryp09adcA7Yt0'
host = 'bembogzwheeytajy6wqi-mysql.services.clever-cloud.com'

# Establish the connection
connection = mysql.connector.connect(
    user=user,
    password=password,
    host=host,
    database=db_name
)

# Create a cursor to execute queries
cursor = connection.cursor()

cursor.execute(query)

# Confirm the changes and close the connection
connection.commit()
connection.close()
```

→ Define credentials

→ Connect to the database

→ Make the query

→ Commit and disconnect

Store

Query to create the table

```
annotateTable = '''
CREATE TABLE annotation (
    id INT AUTO_INCREMENT PRIMARY KEY,
    impact TEXT,
    consequences TEXT,
    allele TEXT,
    variant_rs_id VARCHAR(16),
    FOREIGN KEY (variant_rs_id) REFERENCES variant(variant_rs_id)
)
'''
```

Query to make an insertion of the data

```
annotateTable = '''
INSERT INTO annotation (impact, consequences, allele, variant_rs_id)
VALUES ('HIGH', 'frameshift_variant', 'TA', '587781299'),
    ('HIGH', 'frameshift_variant', 'C', '180177102'),
    ('MODERATE', 'missense_variant', 'G', '80356993'),
    ('HIGH', 'splice_acceptor_variant&intron_variant', 'A', '80358158');
'''
```

Thank you for your attention

Alberto Sánchez S.