# Instruction for cdot podot generation

work directory: /isilon/R_and_D/user_folders/sjung/project/oncomine

input file: Oncomine_Myeloid_v2.20230321.GX5.hotspots.bed

manually create test.bed file: test.bed

Run work.ipynb

```python
import pandas as pd
df=pd.read_csv('test.bed',sep='\t')
df.fillna('-',inplace=True)
df['start']=df['start'].astype(int)+1
df['ID3']=df['ref'].astype(str)+"/"+df['alt'].astype(str)
df['ID4']=1
df[['chr','start','stop','ID3','ID4']].to_csv('vep_input.txt',index=False,sep='\t',header=False)
```

prepare an input file for VEP: vep_input.txt

[Variant Effect Predictor - Homo_sapiens - GRCh37 Archive browser 112]

Run VEP twice with different options

1. Ensembl/GENCODE transcripts and HGVS only
    a. output: ensembl.txt ensembl.txt
2. RefSeq transcripts and HGVS only
    a. output: refseq.txt refseq.txt

prepare neo transcript file (received from Sage): neo_transcript.txt

intersect the VEP outputs and the neo transcript file

```python
# intersect between vep files and our transcripts
import pandas as pd
ref=pd.read_csv('neo_transcript.txt',sep='\t')
df=pd.read_csv('refseq.txt',sep='\t') #,usecols=['Location','SYMBOL','Feature','HGVSc','HGVSp','Feature_original',])
df['Feature2']=df['Feature'].str.split('.').str[0]
df_merged=pd.merge(ref,df,how='inner',on=['SYMBOL','Feature2'])
df2=pd.read_csv('ensembl.txt',sep='\t')#,usecols=['Location','SYMBOL','Feature','HGVSc','HGVSp'])
df2['Feature2']=df2['Feature'].str.split('.').str[0]
df2_merged=pd.merge(ref,df2,how='inner',on=['SYMBOL','Feature2'])
# merge refseq and ensembl
df_concat=pd.concat([df_merged,df2_merged])
df_concat.to_csv('refseq_ensembl_combined.txt',sep='\t',index=False)
```

output: `refseq_ensembl_combined.txt`

```python
#manually copy over 4th column from check.txt to test.bed for now
ref=pd.read_csv('test.bed',sep='\t')
x =pd.read_csv('refseq_ensembl_combined.txt',sep='\t',usecols=['Location','SYMBOL','Feature','UPLOADED_ALLELE','HGVSc','HGVSp'])
x['chr']=x['Location'].str.split(':').str[0]
x['temp']=x['Location'].str.split(':').str[1]
x['start']=x['temp'].str.split('-').str[0].astype(int) - 1
x['start']=x['start'].astype(str)
x['stop']=x['temp'].str.split('-').str[1]
```

```
x.to_csv('temp2.txt',sep='\t',index=False)
x =pd.read_csv('temp2.txt',sep='\t')
x['chr']=x['chr'].astype(str)

xx=pd.merge(ref,x,how='left',on=['chr','stop','UPLOADED_ALLELE'])
xx.to_csv('inspect2.txt',sep='\t')
xx.drop_duplicates(inplace=True)
xx.to_csv('final.txt',sep='\t')
```