```
# Sedreh Nassirnia, ITMO University, Crispr Cas Pipeline, 2020


main_path=$'/home/cas_pipeline'
all_initial_input=$'/home/cas_pipeline/all_initial_input'
all_final_output=$'/home/cas_pipeline/all_final_output'

################################################################################

echo """Step 1) annotate bacterial genome (make faa files) and make protein
database"""

prokka setupdb

TAGS=$(ls $all_initial_input/*.fna | xargs -n 1 basename)
mkdir $all_final_output/prokka

# DEBUGGG
for file in $TAGS; do prokka --outdir $all_final_output/prokka/$file --force --
prefix $file $all_init
ial_input/$file; done


echo """Step 2) Creating cas database"""

prokka_dir=$''$all_final_output''
tags=$(find $prokka_dir -name '*.faa')

# concatenating all tags to a database
mkdir $all_final_output/database
touch $all_final_output/database/seqdb
cat $tags > $all_final_output/database/seqdb

# path to the database created from hmmprofiles
database=$""$all_final_output/database/seqdb""

################################################################################

echo """Step 3) Hmm search: Extracting significant hits using HMMSEARCH"""

INDIR=$all_initial_inputs
hmm_search_output=$""$all_final_output/hmmsearch_results""
mkdir $hmm_search_output

# path to standard hmm profiles
std_hmms=$(ls $main_path/standard_hmm_profiles/* | xargs -n 1 basename)

for i in $std_hmms; do hmmsearch --tblout $hmm_search_output/${i}.tbl
$main_path/standard_hmm_profile
s/${i} $database; done

echo """Step 4) Creating final fastas containg sequences from final hits"""

INDIR=$hmm_search_output
lists=$""$hmm_search_output/lists""
cd $INDIR
mkdir $lists
```

```
tables=$(cd $INDIR && ls *.hmm.tbl)
for i in $tables; do grep -v "^#" ${i} | awk '{print $1}' >>
$lists/$i.cleaned_fasta; done

INDIR=$lists
clean_fasta=$""$hmm_search_output/final""
mkdir $clean_fasta

# indexing step
esl-sfetch --index $database

tables=$(cd $INDIR && ls *.hmm.tbl.cleaned_fasta)
for i in $tables; do esl-sfetch -f $database $INDIR/${i} > $clean_fasta/${i
%.fasta}; done


################################################################################

echo """Step 5) cluster the sequences"""

INDIR=$clean_fasta
clustering=$''$all_final_output/clustering''
mkdir $clustering

links=$(ls $INDIR/*.cleaned_fasta)

# this is an important path, it must needs be modified for containerisation
path_to_cdhit=$'/home/cas_pipeline/cdhit-master/psi-cd-hit'

# making softlinks to cd_hit_input folder for all cleaned fastas
cd_hit_input=$path_to_cdhit
ln -s $links $cd_hit_input

TAGS=$(ls $cd_hit_input/*.cleaned_fasta | xargs -n 1 basename)


for i in $TAGS; do cd $path_to_cdhit/; ./psi-cd-hit.pl -i ${i} -o ${i
%.hmm.tbl.cleaned_fasta} -c 0.95
; done

out_dir_cdhit=$""$all_final_output/cdhit""
mkdir $out_dir_cdhit

cp -r $path_to_cdhit/* $out_dir_cdhit

################################################################################

echo """Step 6) Muscle alignment """

INDIR=$out_dir_cdhit
muscle_dir=$""$all_final_output/muscle""
mkdir $muscle_dir

tags=$(ls $INDIR/*.cleaned_fasta | xargs -n 1 basename | sed
's/.hmm.tbl.cleaned_fasta//')
for i in $tags; do muscle -in $i -out $muscle_dir/$i.fasta; done

################################################################################
```

```
echo """Step 7) IQTREE"""

INDIR=$muscle_dir
iqtree_dir=$""$all_final_output/iqtree""
mkdir $iqtree_dir

tags=$(ls $INDIR/*.fasta)
cd $iqtree_dir

for f in $tags; do iqtree -s $f -bb 1000 -alrt 1000 -nt 6; done


# end of script
```