Commands

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

# DEBUGGING
for file in $TAGS; do $prokka --outdir $all_final_output/prokka/$file --force --prefix $file
$all_initial_input/$file; done
```

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""
```

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
$main_path/standard_hmm_profiles/${i} $database; done
```

Step 4) Creating final fastas contain 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 (database should be indexed)
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
```

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 soft links 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
```

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
```

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
```

Step 8) Blast and filtering blast results by 50% identity

```
INDIR=$out_dir_cdhit
blast_dir=$""$all_final_output/blast""
filtered_blast_dir=$""$all_final_output/blast/blast_filtered""
mkdir $blast_dir

tags=$(ls $INDIR/*.cleaned_fasta)
new_tags=$(ls blast_dir/*)
mkdir $filtered_blast_dir

# blast
for i in $tags; do blastp -query $INDIR/$i -db nr -evalue 1e-5 -num_threads 6 -outfmt 6 -out
$blast_dir/${i%.*}; done

# filter blast
for tag in $new_tags; do awk -F" " 'int($3) > 50' $out_dir_cdhit/$tag >
$filtered_blast_dir/filtered_$tag; done
```

Step 9) esl-sfetch on the blast filtered ids

```
INDIR=$""$all_final_output/blast/blast_filtered""
blast_fetch=$""$all_final_output/blast_fetch""
mkdir blast_fetch

id_dir==$""$blast_fetch/ids""
mkdir id_dir

tables=$(ls $INDIR)
for i in $tables; do grep -v "^#" $INDIR/${i} | awk '{print $2}' >> $id_dir/$i.ids; done

ids=$(ls $id_dir)
for i in $ids; do esl-sfetch -f $database $id_dir/${i} > $blast_fetch/${i%.fasta}; done
```

Step 10) IQTREE