week 26 redo tailocin

ensure your basic knowledge is solid.. focus on understand biological and also algorithm like MCMC and dating analysis, make sure you know exactly what each of used tool did ....

ref book:

decoding genomes

rereredo:

1. make sure the strand is reversed if its - done (the ref is strand - and my extracted ref HTF and TFA are strand +, checked p25.C2 is the same as ref, so keep 14 contigs and reverse them to strand +, and also when mapped minimap fragments are -, reverse to +)
2. select based on mapping quality (total length and mapping match base number, select based on the matched base proportion like $(($matchedbase/$totallengthofHTForTFA))). done (confimed by p26.E7 and B7) check example TFA_p23.B8 in HB0737
3. lable the TF and HTF inside the tailocin region as well
4. tape measure querying the seq by minimap2 and also by coordinate of the end of first and the start of second contig, and build a MSA/ try blast -n ...
5. 33.ESP why no HTF? check the contig coordinate, and maybe using removal_At_bam to extract the mapping reads id...? minimap didnt capture HTF, and the TFA is TFA_p23.B8   513   0   293 .
6. HB0814 indeed has little info about TFA
7. build the MSA with conserved regions or all of the regions.. try both done
8. is there any sample has no tfa or HTF? yes, for example, here in the right panel, at the bottom 30.ESP has no hyplotypes for both HTF and TFA, 64.GBR has no TFA, HB0814 has no TFA, and at the top 33.ESP has no HTF.
9. read the textbook and paper Hernán sent. read tailocin paper in science

Monday: paper read, then run tape measure

tape measure

step1: extract the tape measure from all samples (tailocin region in /msa) that have them using the coordinate of minigap

step2: for those have Ns in the tape measure region: 1 if its modern, go to the assembly to see if the contig is continuing, if so extract the unmapped seq add to the MSA of tape measure,

step3: if they are historical, using the MSA seq to fish (add the haplotypes to fasta ref genome and map again, then do the same as before for TFA, assembly minimap...)

step4: put them together as a MSA fasta and then use omega to align... then build a tree.

reredo:

new hyplotypes: 7 HTF (hypothetic tailocin fiber protein) and 7 TFA (tailocin fiber assembly protein)

reverse the minimap, map the tailocin reference to long contigs both h and m.

make sure all the strand are consistent when mapping, reverse strand - seq when mapping

make sure the seq given by minimap is the same you extract from fasta with samtools faidx (0-20 and 1-20 give the same length)

modify the plot by making chunk more separate, label OTU5 and nonOTU5

select the longest (highest covered proportion/ nonN base proportion) HFA

nextweek:

1. tailocin redo

need to change: minimap to five hyplotypes, and modern use minimap to capture the assemblies. make sure bam are all mapped raeds,

A. make sure bam are mapped reads: -F4 **done**

B. **spades —merge -1 -2 for 10, -s for HB and PL, add —careful and -k 21, 33** (21,33,55 also ok but PL0001 were not detected among the 23 good quality tailocins)

→ 31/46 historical samples success to assembly (15 samples failed to assembly contigs, and all the 15 are low coverage tailocins detected before...) done

```
if [[ "$samplename" == *.* ]]; then
    subset_fastq1="${trimmed_fastq_dir}/${samplename}_subsetR1.fastq.gz"
    subset_fastq2="${trimmed_fastq_dir}/${samplename}_subsetR2.fastq.gz"
    spades.py --merge "$subset_fastq" -1 $subset_fastq1 -2 $subset_fastq2  --carefu
l -o "$assembly_dir/$samplename" -k 21,33
  else
  # single end has not meta just use multicell/isolate as default
    spades.py -s "$subset_fastq"  -o "$assembly_dir/$samplename"  --careful -k 21,3
3
fi

with 21,33 k and --careful, there are 15 / 46 have no contigs since
'Invalid kmer coverage histogram, make sure that the coverage is indeed uniform
== Error ==  system call for: "['/SAN/ugi/aMetagenomics/jiajucui/tmpbigfile/minicon
da3newdirtostore/envs/phylogeny_snp/bin/spades-core', '/SAN/ugi/plant_genom/jiajucu
i/4_mapping_to_pseudomonas/tailocin_extract/assemblies/PL0027/K21/configs/config.in
fo', '/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract/assem
blies/PL0027/K21/configs/careful_mode.info']" finished abnormally, OS return value:
21
None
'

all 15 are low coverage tailocins detected before...
```

Summary the samples:

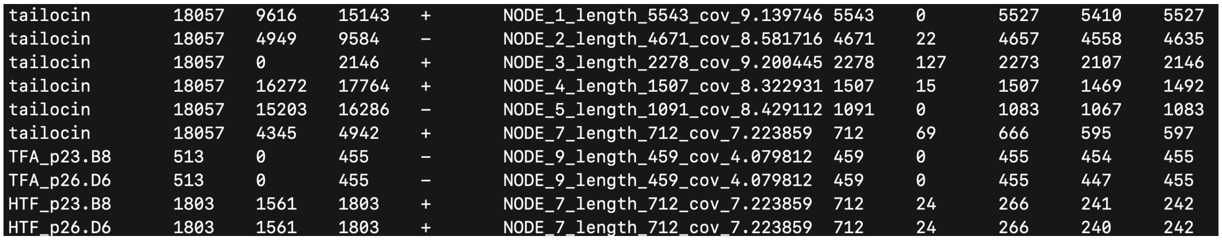
```
46 historical in total:
  30 have contigs, 1 (27.ESP_1975) has 0 contig and 15 failed to assembly since not
uniform coverage.

    40 OTU5 (10 of them failed to assembly, 1 (27.ESP_1975) has no contig) and  6 n
onOTU5 (4 failed):
  below we can know the covered proportion to tailocin region of the six nonOTU5:
  HB0828 fail to assembly
    HB0863 fail to assembly
    PL0066 fail to assembly
    PL0108 0.0485
    PL0203 fail to assembly
    PL0258 0.295

    compare to last week when using artifical doubled r1 and r2 reads (wrong):
    HB0828 0.28
    HB0863 no tailocin
    PL0066 no
    PL0108 0.045
    PL0203 0.023
    PL0258 0.265


    among 30 historical samples that have contigs, 25 passed 65% covered proportion
(careful aboout hyplotypes...)

    after map the raw fasta of 76 availible modern samples to the tailocin region a
nd hyplotypes:
    in total 46+76=122 historical and modern samples, 72 are above 65% covered prop
ortion ( 47 modern (all of them are OTU5) and 25 historical),
    #p12_H7 has 0.63 covered proprotion, it is a nonOTU5 but always inside OTU5 whe
n we tried to find outgroups for rerooting.


    later check the topology on first all 127, and then only the 65% set (72 sample
s)
```

C. minimap the contigs to reference region including both tailocin and hyplotypes, get .paf and mask the fasta gaps with Ns

```
  #  Run minimap2 to get the mapping
  #sergio's idea: map the reference short chunk of hyplotypes to the long contig, in principle works for both historical and modern samples.
  $tools/minimap2/minimap2 -cx asm5 "$reference_genome" "$contig_file" > "$paf_file"
```

example of paf:

ref name / ref length / ref map start / end / strand / contig name /
ref length / start / end /. mapped base/ alignmentlength



```
tailocin      18057   9616    15143   +       NODE_1_length_5543_cov_9.139746 5543    0       5527    5410    5527
tailocin      18057   4949    9584    -       NODE_2_length_4671_cov_8.581716 4671    22      4657    4558    4635
tailocin      18057   0       2146    +       NODE_3_length_2278_cov_9.200445 2278    127     2273    2107    2146
tailocin      18057   16272   17764   +       NODE_4_length_1507_cov_8.322931 1507    15      1507    1469    1492
tailocin      18057   15203   16286   -       NODE_5_length_1091_cov_8.429112 1091    0       1083    1067    1083
tailocin      18057   4345    4942    +       NODE_7_length_712_cov_7.223859  712     69      666     595     597
TFA_p23.B8    513     0       455     -       NODE_9_length_459_cov_4.079812  459     0       455     454     455
TFA_p26.D6    513     0       455     -       NODE_9_length_459_cov_4.079812  459     0       455     447     455
HTF_p23.B8    1803    1561    1803    +       NODE_7_length_712_cov_7.223859  712     24      266     241     242
HTF_p26.D6    1803    1561    1803    +       NODE_7_length_712_cov_7.223859  712     24      266     240     242
```

regarding the strand:

1. reverse the ref haplotypes, make them the same strand as tailocin region (checked with extracted refp25c2 TFA and HTF, the 7 TFAs and 7 HTFs are reversed):

```
while read -r region name; do
  if [ "$name" == "tailocin" ]; then
    samtools faidx "$reference_genome" "$region" | sed "1s/.*/>$name/" >> "$tailocin_fasta"
  else
    samtools faidx "$reference_genome" "$region" | seqtk seq -r | sed "1s/.*/>$name/" >> "$tailocin_fasta"
  fi
done < "$tailocin_region"
```

2. reverse the mapped alignment if they have ‘-’ strand:

```

    if [[ "$strand" == "-" ]]; then
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | seqtk seq -r - | tail -n +2 | tr -d '\n' )
    else
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | tail -n +2 | tr -d '\n')
    fi

    segment_start=$((ref_start + cumulative_starts["$ref_name"]))
    replace_sequence $segment_start "$contig_seq"
```

the ref used to minimap:

```
less ../../tailocin_extract/tailocin_region.fa.fai



tailocin        18057   10      60      61
TFA_p23.B8      513     18380   513     514
TFA_p26.D6      513     18906   513     514
TFA_p21.F9      498     19432   498     499
TFA_p25.C2      513     19943   513     514
TFA_p5.D5       513     20468   513     514
TFA_p25.A12     546     20995   546     547
TFA_p7.G11      546     21554   546     547
HTF_p23.B8      1803    22113   1803    1804
HTF_p26.D6      1803    23929   1803    1804
HTF_p21.F9      1245    25745   1245    1246
HTF_p25.C2      1803    27003   1803    1804
HTF_p5.D5       1803    28818   1803    1804
HTF_p25.A12     1383    30635   1383    1384
HTF_p7.G11      1830    32031   1830    1831
```

the script to assembly, minimap and build MSA:

```
#! /bin/bash
#$ -l tmem=100G
#$ -l h_vmem=100G
#$ -l h_rt=24:30:0
#$ -S /bin/bash
#$ -N vcftofastaandclustalotree
#$ -o /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/logs/
#$ -e /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/logs/
mkdir -p /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/logs




# Activate the conda environment
source /home/jiajucui/miniconda3/bin/activate phylogeny_snp

# Define the base directories
bam_dir="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_46"
fastq_dir="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_46_fastq"
output_dir="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract"
mkdir -p $output_dir
tmp_dir="${output_dir}/tmp"
readids_dir="${output_dir}/readids"
```

```
trimmed_fastq_dir="${output_dir}/trimmed_tailocin_fastq/"
assembly_dir="${output_dir}/assemblies"
mapping_dir="${output_dir}/mappings"
msa_dir="${output_dir}/msa"
new_tree_dir="${output_dir}/tree"
nocontigs_file="${msa_dir}/nocontigs.txt"
contigmapping_file="${mapping_dir}/contigmapping.txt"
haplotype_dir="${output_dir}/haplotype_selected"
nonN_file="${haplotype_dir}/nonN_TFAandHTF.txt"

# rm -r msa/ mappings/ assemblies/ contig_stats/ tree/
rm -r $assembly_dir $mapping_dir $msa_dir $new_tree_dir $haplotype_dir
#rm -r $mapping_dir $msa_dir $new_tree_dir $vcf_dir

mkdir -p $assembly_dir  $mapping_dir $msa_dir $new_tree_dir $haplotype_dir

tailocin_fasta="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract/tailocin_region.fa"
reference_genome="/SAN/ugi/plant_genom/jiajucui/1_initial_data/reference_genome_Ps_with_tailocin_haplotypes/Pseudomonas.plate25.C2.pilon.contigs_renamed.with_Tail_Fiber_Haps.fasta"
tailocin_region="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract/regions.txt"

#22:15502-33559 but all the 40 genes
#TFA_p23.B8:1-513
#TFA_p26.D6:1-513
#TFA_p21.F9:1-498
#TFA_p25.C2:1-513
#TFA_p5.D5:1-513
#TFA_p25.A12:1-546
#TFA_p7.G11:1-546
#HTF_p23.B8:1-1803
#HTF_p26.D6:1-1803
#HTF_p21.F9:1-1245
#HTF_p25.C2:1-1803
#HTF_p5.D5:1-1803
#HTF_p25.A12:1-1383
#HTF_p7.G11:1-1830
#in the formated_region, all the TFA and HTF are reversed, make sure all of them are strand + like tailocin first
#I double checked with omega and found the *p25.C2 are the reference indeed but in reverse strand, so do the reverse and dont need to include the extracted ref chunk for TFA and HTF. only 7 and 7
# Read the tailocin region and process it
>"$tailocin_fasta"
while read -r region name; do
  if [ "$name" == "tailocin" ]; then
    samtools faidx "$reference_genome" "$region" | sed "1s/.*/>$name/" >> "$tailoci
```

```
n_fasta"
  else
    samtools faidx "$reference_genome" "$region" | seqtk seq -r | sed "1s/.*/>$nam
e/" >> "$tailocin_fasta"
  fi
done < "$tailocin_region"
# Initialize or clear the nocontigs_file and contigmapping_file
> "$nocontigs_file"
> "$contigmapping_file"
> "$nonN_file"

# Tools path
tools=/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/tools




# Define regions and sequences
declare -A segment_lengths
segment_lengths["tailocin"]=18057
#segment_lengths["TFA_22_fragment_1_21_sp|P03740|TFA_LAMBD_Tail_fiber_assembly_prot
ein"]=$((18230 - 17718 + 1))
segment_lengths["TFA_p23.B8"]=513
segment_lengths["TFA_p26.D6"]=513
segment_lengths["TFA_p21.F9"]=498
segment_lengths["TFA_p25.C2"]=513
segment_lengths["TFA_p5.D5"]=513
segment_lengths["TFA_p25.A12"]=546
segment_lengths["TFA_p7.G11"]=546
#segment_lengths["HTF_22_fragment_1_22_hypothetical_protein"]=$((20043 - 18241 +
1))
segment_lengths["HTF_p23.B8"]=1803
segment_lengths["HTF_p26.D6"]=1803
segment_lengths["HTF_p21.F9"]=1245
segment_lengths["HTF_p25.C2"]=1803
segment_lengths["HTF_p5.D5"]=1803
segment_lengths["HTF_p25.A12"]=1383
segment_lengths["HTF_p7.G11"]=1830
segment_names=(
    "tailocin"
    "TFA_p23.B8"
    "TFA_p26.D6"
    "TFA_p21.F9"
    "TFA_p25.C2"
    "TFA_p5.D5"
    "TFA_p25.A12"
    "TFA_p7.G11"
    "HTF_p23.B8"
    "HTF_p26.D6"
    "HTF_p21.F9"
    "HTF_p25.C2"
    "HTF_p5.D5"
    "HTF_p25.A12"
    "HTF_p7.G11"
)



#SPAdes
for bam in "$bam_dir"/*.bam; do
  samplename=$(basename "$bam" .mapped_to_Pseudomonas.dd.q20.bam)
  subset_fastq="${trimmed_fastq_dir}/${samplename}_subset.fastq.gz"
  # for 143 samples_cat_merged use --12
  # --12 <file_name> File with interlaced forward and reverse paired-end reads.
  # --merged <file_name> File with merged paired reads. If the properties of the library permit, overlapping paired-end reads can be merged using special software.
  # Non-empty files with (remaining) unmerged left/right reads (separate or interlaced) must be provided for the same library for SPAdes to correctly detect the original read length.
  # but no additional files, --12 is ok
  # --isolate - isolate (standard) bacterial data;
  # --meta The --meta mode is designed for metagenomic data, which typically involves a heterogeneous mixture of reads. also good for dealing with a mixture of read qualities
  # --only-error-correction Performs read error correction only.
  # --only-assembler Runs assembly module only. if you have high quality reads, but here we have a mix of qualities, better run error correction
  # checked --meta gave more info in low quality reads like in120, it was nothing but here 30 contigs
  # Determine the correct raw FASTQ file name pattern
  #but the PL0042 process is frozen even using -t 2 and require 50Gb, so remove --meta for it, and it works
  if [[ "$samplename" == *.* ]]; then
    subset_fastq1="${trimmed_fastq_dir}/${samplename}_subsetR1.fastq.gz"
    subset_fastq2="${trimmed_fastq_dir}/${samplename}_subsetR2.fastq.gz"

    spades.py --merge "$subset_fastq" -1 $subset_fastq1 -2 $subset_fastq2  --careful -o "$assembly_dir/$samplename" -k 21,33
  else
  # single end has not meta just use multicell/isolate as default
    spades.py -s "$subset_fastq"  -o "$assembly_dir/$samplename"  --careful -k 21,33
  fi
  # Step 2: Check the number of contigs in each assembly
  contig_file="$assembly_dir/$samplename/contigs.fasta"
#  contig_count=$(grep -c "^>" "$contig_file")
#  echo "$samplename: $contig_count contigs" >> "$contig_stats_dir/contig_counts.txt"
```

```
  # Check if the contig file exists
  if [[ ! -f "$contig_file" ]]; then
    echo "$samplename" >> "$nocontigs_file"
    echo "Contig file $contig_file does not exist. Sample name $samplename added to $nocontigs_file."
    continue
  fi

  reference_genome="${output_dir}/tailocin_region.fa"


  # Define your output files
  paf_file="${mapping_dir}/${samplename}_mapped.paf"
  fasta_out="${msa_dir}/${samplename}_tailocin_region.fasta"
  #  Run minimap2 to get the mapping
  #sergio's idea: map the reference short chunk of hyplotypes to the long contig, in principle works for both historical and modern samples.
  #$tools/minimap2/minimap2 -cx asm5 "$reference_genome" "$contig_file" > "$paf_file"
  mkdir -p ${msa_dir}/fake
  fasta_outfake="${msa_dir}/fake/${samplename}_fake.fasta"
  fasta_out1="${msa_dir}/${samplename}_tailocin_region_allconcatenated.fasta"
  $tools/minimap2/minimap2 -cx asm5 "$contig_file" "$reference_genome" > "$paf_file"


  if [[ $(less "$paf_file" | wc -l) -eq 0 ]]; then
    echo "$samplename nomappedcontig" >> "$nocontigs_file"
    echo "Contig file $contig_file does not exist. Sample name $samplename added to $nocontigs_file."
    continue
  fi
  # Calculate cumulative start positions
  declare -A cumulative_starts
  cumulative_starts["tailocin"]=0
  for i in "${!segment_names[@]}"; do
    if [[ $i -gt 0 ]]; then
      prev_segment="${segment_names[$((i-1))]}"
      cumulative_starts["${segment_names[$i]}"]=$((cumulative_starts["$prev_segment"] + segment_lengths["$prev_segment"]))
    fi
  done

  total_length=0
  for length in "${segment_lengths[@]}"; do
    total_length=$((total_length + length))
  done
  final_sequence=$(printf 'N%.0s' $(seq 1 $total_length))
  fakefinal_sequence=$(printf 'N%.0s' $(seq 1 $total_length))
  replace_sequence() {
    local start=$1
    local seq=$2
    final_sequence="${final_sequence:0:start}${seq}${final_sequence:$(($start + ${#seq}))}"
  }

  tailocin_count=0
  htf_count=0
  taf_count=0
  declare -a htf_list
  declare -a tfa_list
  declare -A nonN_counts

  while read -r line; do
    ref_name=$(echo "$line" | awk '{print $1}')
    ref_start=$(echo "$line" | awk '{print $3}')
    ref_end=$(echo "$line" | awk '{print $4}')
    strand=$(echo "$line" | awk '{print $5}')
    contig_name=$(echo "$line" | awk '{print $6}')
    contig_start=$(echo "$line" | awk '{print $8}')
    contig_end=$(echo "$line" | awk '{print $9}')

    if [[ "$strand" == "-" ]]; then
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | seqtk seq -r - | tail -n +2 | tr -d '\n' )
    else
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | tail -n +2 | tr -d '\n')
    fi

    segment_start=$((ref_start + cumulative_starts["$ref_name"]))
    replace_sequence $segment_start "$contig_seq"

    if [[ "$ref_name" == "tailocin" ]]; then
        tailocin_count=$((tailocin_count + 1))
    elif [[ "$ref_name" == HTF* ]]; then
        htf_count=$((htf_count + 1))
        htfname=${ref_name%%:*}
        htf_list+=("$htfname")
    elif [[ "$ref_name" == TFA* ]]; then
        tfa_count=$((tfa_count + 1))
        tfaname=${ref_name%%:*}
        tfa_list+=("$tfaname")
    fi
  done < "$paf_file"
  # Write the final sequence to fasta_out with individual segment headers
  echo ">${samplename}" > "$fasta_out1"
  echo "$final_sequence" >> "$fasta_out1"

  ## Write the final sequence to fasta_out with individual segment headers
  {
    for segment in "${segment_names[@]}"; do
      start=${cumulative_starts[$segment]}
      length=${segment_lengths[$segment]}
      segment_sequence=${final_sequence:$start:$length}
      echo ">$segment"
      echo "$segment_sequence"
    done
  } > "$fasta_out"

 # if [[ $tailocin_count -eq 0 && $htf_count -eq 0 && $tfa_count -eq 0 ]]; then
 #   rm $fasta_out $fasta_out1
 #   echo "$samplename no mapped contig"
 # fi

  # Append to the combined MSA file before formatting individual segments
  cat "$fasta_out1" >> "$msa_dir/all_historicalfa_samples_tailocin.fasta"

  # Output the mapping summary and lists to the contigmapping file
  {
    echo "$samplename: tailocin mapped contigs = $tailocin_count, TFA mapped contig
s = $tfa_count, HTF mapped contigs = $htf_count"

    echo "HTF List:"
    for htf_segment in "${htf_list[@]}"; do
        echo "$htf_segment"
    done

    echo "TFA List:"
    for tfa_segment in "${tfa_list[@]}"; do
        echo "$tfa_segment"
    done
  } >> "$contigmapping_file"

  # Process segments to calculate non-N counts and proportions
  # Process segments to calculate non-N counts and proportions
  fakefornonN_replace_sequence() {
    local start=$1
    local seq=$2
    fakefinal_sequence="${fakefinal_sequence:0:start}${seq}${fakefinal_sequenc
e:$(($start + ${#seq}))}"
   }

  while read -r line; do
```

```
    ref_name=$(echo "$line" | awk '{print $1}')
    ref_start=$(echo "$line" | awk '{print $3}')
    ref_end=$(echo "$line" | awk '{print $4}')
    strand=$(echo "$line" | awk '{print $5}')
    contig_name=$(echo "$line" | awk '{print $6}')
    contig_start=$(echo "$line" | awk '{print $8}')
    mappedbase=$(echo "$line" | awk '{print $10}')
    contig_end=$((contig_start + mappedbase))
    if [[ "$strand" == "-" ]]; then
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | seqtk seq -r - | tail -n +2 | tr -d '\n' )
    else
      contig_seq=$(samtools faidx "$contig_file" "$contig_name:$contig_start-$contig_end" | tail -n +2 | tr -d '\n')
    fi

    segment_start=$((ref_start + cumulative_starts["$ref_name"]))
    fakefornonN_replace_sequence $segment_start "$contig_seq"

  done < "$paf_file"
 ## Write the final sequence to fasta_out with individual segment headers
  {
    for segment in "${segment_names[@]}"; do
      start=${cumulative_starts[$segment]}
      length=${segment_lengths[$segment]}
      segment_sequence=${fakefinal_sequence:$start:$length}
      echo ">$segment"
      echo "$segment_sequence"
    done
  } > "$fasta_outfake"

  {
    for segment in "${tfa_list[@]}"; do
      sequence=$(grep -A1 ">${segment}" "$fasta_outfake" | tail -n1)
      nonN_count=$(echo "$sequence" | tr -cd 'ATCGatcg' | wc -c)
      proportion=$(echo "scale=5; $nonN_count/${segment_lengths[$segment]}" | bc)
      nonN_counts["$segment"]=$proportion
      echo "$samplename >${segment}:$proportion"
    done
    for segment in "${htf_list[@]}"; do
      sequence=$(grep -A1 ">${segment}" "$fasta_outfake" | tail -n1)
      nonN_count=$(echo "$sequence" | tr -cd 'ATCGatcg' | wc -c)
      proportion=$(echo "scale=5; $nonN_count/${segment_lengths[$segment]}" | bc)
      nonN_counts["$segment"]=$proportion
      echo "$samplename >${segment}:$proportion"
    done
  } >> "$nonN_file"
  longest_tfa=$(printf "%s\n" "${!nonN_counts[@]}" | grep "^TFA" | while read segment; do echo "$segment ${nonN_counts[$segment]}"; done | sort -k2,2nr | head -n1 | awk '{print $1}')
  longest_htf=$(printf "%s\n" "${!nonN_counts[@]}" | grep "^HTF" | while read segment; do echo "$segment ${nonN_counts[$segment]}"; done | sort -k2,2nr | head -n1 | awk '{print $1}')
  # Filter out segments that are still all Ns
  if [[ -n "$longest_tfa" ]]; then
    longest_tfa_seq=$(grep -A1 ">${longest_tfa%%:*}" "$fasta_out" | tail -n1)
    longest_tfa_nonN=$(echo "$longest_tfa_seq" | tr -cd 'ATCGatcg' | wc -c)
    if [[ "$longest_tfa_nonN" -eq 0 ]]; then
      longest_tfa=""
    fi
  fi

  if [[ -n "$longest_htf" ]]; then
    longest_htf_seq=$(grep -A1 ">${longest_htf%%:*}" "$fasta_out" | tail -n1)
    longest_htf_nonN=$(echo "$longest_htf_seq" | tr -cd 'ATCGatcg' | wc -c)
    if [[ "$longest_htf_nonN" -eq 0 ]]; then
      longest_htf=""
    fi
  fi

  # Create the final FASTA file with the selected segments
  final_fasta="${haplotype_dir}/${samplename}.final.fasta"
  {
    echo ">tailocin"
    grep -A1 ">tailocin" "$fasta_out" | tail -n1

    if [[ -n "$longest_tfa" ]]; then
      echo ">${longest_tfa%%:*}"
      grep -A1 ">${longest_tfa%%:*}" "$fasta_out" | tail -n1
    fi

    if [[ -n "$longest_htf" ]]; then
      echo ">${longest_htf%%:*}"
      grep -A1 ">${longest_htf%%:*}" "$fasta_out" | tail -n1
    fi
  } > "$final_fasta"

 #only longest all other Ns
 # Create the final concatenated FASTA file
  final2_fasta="${haplotype_dir}/${samplename}.markexceptlongest.fasta"
  {
  echo ">${samplename}"
  final_sequence=$(printf 'N%.0s' $(seq 1 $total_length))
  # Retain the tailocin sequence
  tailocin_sequence=$(grep -A1 ">tailocin" "$fasta_out" | tail -n1)
  tailocin_start=${cumulative_starts["tailocin"]}
  replace_sequence $tailocin_start "$tailocin_sequence"
  # Retain the longest TFA sequence and mark others as Ns
  for segment in "${tfa_list[@]}"; do
    segment_start=${cumulative_starts["$segment"]}
    if [[ "$segment" == "${longest_tfa%%:*}" ]]; then
      tfa_sequence=$(grep -A1 ">${segment}" "$fasta_out" | tail -n1)
      replace_sequence $segment_start "$tfa_sequence"
    fi
  done

  # Retain the longest HTF sequence and mark others as Ns
  for segment in "${htf_list[@]}"; do
    segment_start=${cumulative_starts["$segment"]}
    if [[ "$segment" == "${longest_htf%%:*}" ]]; then
      htf_sequence=$(grep -A1 ">${segment}" "$fasta_out" | tail -n1)
      replace_sequence $segment_start "$htf_sequence"
    fi
  done

  echo "$final_sequence"
  } > "$final2_fasta"

 # if [[ $tailocin_count -eq 0 && $htf_count -eq 0 && $tfa_count -eq 0 ]]; then
 #   rm $final_fasta
 #   echo "$samplename no mapped contig"
 # fi

  echo "The final sequence has been saved to $final_fasta"
done


# Create output files for HTF and TFA
htf_fasta="${haplotype_dir}/all_HTF_samples.fasta"
tfa_fasta="${haplotype_dir}/all_TFA_samples.fasta"
> "$htf_fasta"
> "$tfa_fasta"

# Iterate over each final fasta file and extract HTF and TFA sequences
for final_fa in "${haplotype_dir}"/*.final.fasta; do
  samplename=$(basename "$final_fa" .final.fasta)

  # Extract HTF sequence
  grep -A1 ">HTF" "$final_fa" | sed "s/^>/>${samplename}|/" >> "$htf_fasta"

  # Extract TFA sequence
  grep -A1 ">TFA" "$final_fa" | sed "s/^>/>${samplename}|/" >> "$tfa_fasta"
done

echo "HTF and TFA multi-sample FASTA files have been generated in $haplotype_dir."
```

```
#filter with 50% proportion covered
#!/bin/bash

# Define input and output directories

# Create filtered output files for HTF and TFA
filtered_htf_fasta="${haplotype_dir}/filtered_HTF_samples.fasta"
filtered_tfa_fasta="${haplotype_dir}/filtered_TFA_samples.fasta"
> "$filtered_htf_fasta"
> "$filtered_tfa_fasta"

# Threshold for non-N proportion
threshold=0.65

# Filter HTF sequences
while read -r line; do
  if [[ $line == ">"* ]]; then
    header=$line
    sequence=""
  else
    sequence=$line
    nonN_count=$(echo "$sequence" | tr -cd 'ATCGatcg' | wc -c)
    total_length=${#sequence}
    proportion=$(echo "scale=5; $nonN_count / $total_length" | bc)
    if (( $(echo "$proportion >= $threshold" | bc -l) )); then
      echo "$header" >> "$filtered_htf_fasta"
      echo "$sequence" >> "$filtered_htf_fasta"
    fi
  fi
done < "${haplotype_dir}/all_HTF_samples.fasta"

# Filter TFA sequences
while read -r line; do
  if [[ $line == ">"* ]]; then
    header=$line
    sequence=""
  else
    sequence=$line
    nonN_count=$(echo "$sequence" | tr -cd 'ATCGatcg' | wc -c)
    total_length=${#sequence}
    proportion=$(echo "scale=5; $nonN_count / $total_length" | bc)
    if (( $(echo "$proportion >= $threshold" | bc -l) )); then
      echo "$header" >> "$filtered_tfa_fasta"
      echo "$sequence" >> "$filtered_tfa_fasta"
    fi
  fi
done < "${haplotype_dir}/all_TFA_samples.fasta"
```

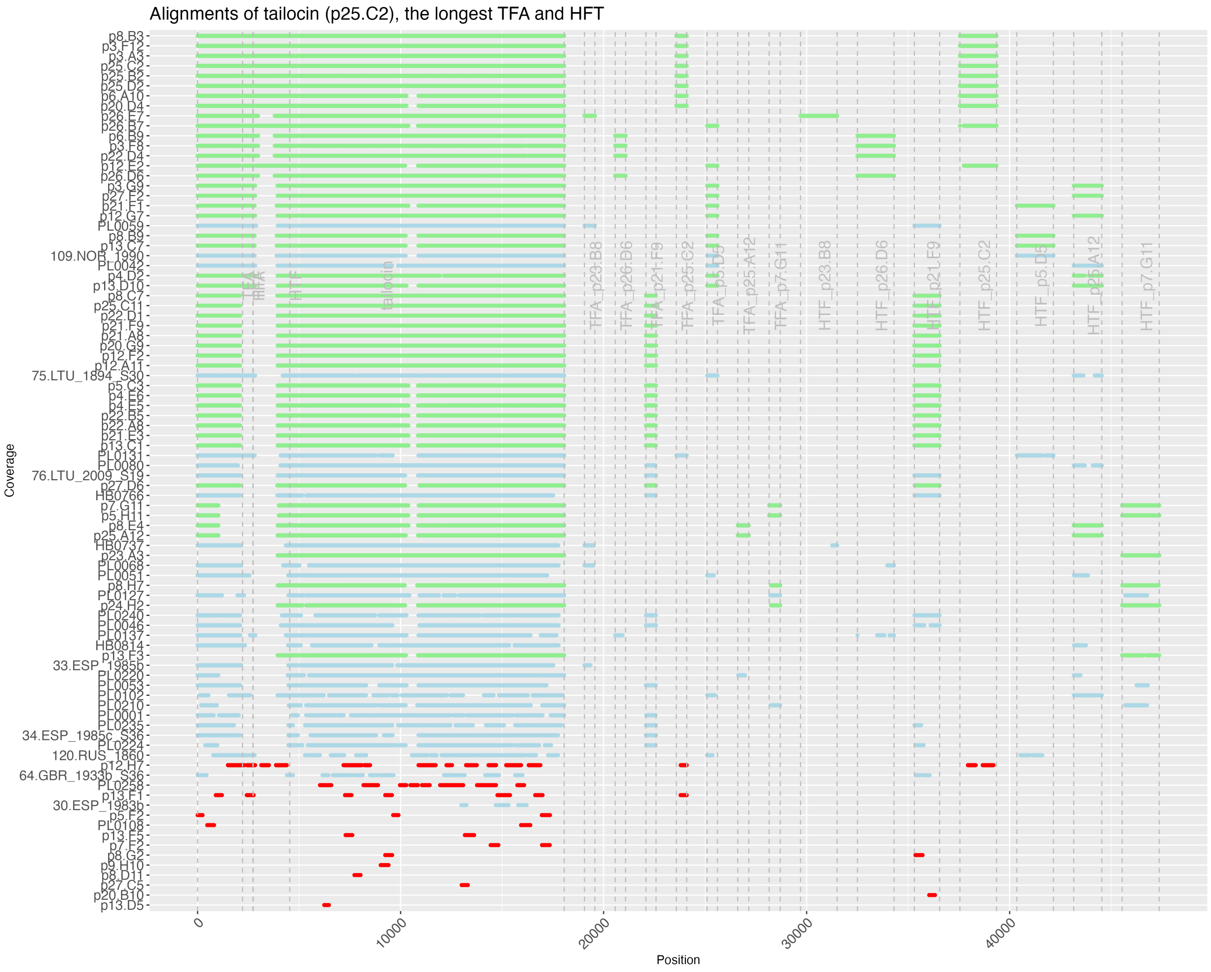
```
echo "Filtered HTF and TFA multi-sample FASTA files have been generated in $haploty
pe_dir."
```

D. modern samples with assemblies to use minimap to map to the ref region and hyplotypes...

use assemblies, now we have 76 available

```
the script is similar to using minimap to map the fasta to tailocin and five hyplot
ypes, then align the seq, fill the gaps with Ns...
```

then the results of mapping:



summary the TFA and HTF presence/absence:

88 samples in total (30h and 58 modern) (13nonOTU5, 28 hOTU5, 47 mOTU5)

```
sample	TFA	HTF	group
p8.G2	NULL	HTF_p21.F9	noTFA
p23.A3	NULL	HTF_p7.G11	noTFA
```

```
p20.B10 NULL    HTF_p21.F9      noTFA
p13.F3  NULL    HTF_p7.G11      noTFA
HB0814  NULL    HTF_p25.A12     noTFA
64.GBR_1933b_S36        NULL    HTF_p21.F9      noTFA
PL0001  TFA_p21.F9      NULL    noHTF
p13.F1  TFA_p25.C2      NULL    noHTF
34.ESP_1985c_S36        TFA_p21.F9      NULL    noHTF
33.ESP_1985b    TFA_p23.B8      NULL    noHTF
PL0258  NULL    NULL    noboth
PL0108  NULL    NULL    noboth
p9.H10  NULL    NULL    noboth
p8.D11  NULL    NULL    noboth
p7.F2   NULL    NULL    noboth
p5.F2   NULL    NULL    noboth
p27.C5  NULL    NULL    noboth
p13.F5  NULL    NULL    noboth
p13.D5  NULL    NULL    noboth
30.ESP_1983b    NULL    NULL    noboth
PL0240  TFA_p21.F9      HTF_p21.F9
PL0235  TFA_p21.F9      HTF_p21.F9
PL0224  TFA_p21.F9      HTF_p21.F9
PL0220  TFA_p25.A12     HTF_p25.A12
PL0210  TFA_p7.G11      HTF_p7.G11
PL0137  TFA_p26.D6      HTF_p26.D6
PL0131  TFA_p25.C2      HTF_p5.D5
PL0127  TFA_p7.G11      HTF_p7.G11
PL0102  TFA_p5.D5       HTF_p25.A12
PL0080  TFA_p21.F9      HTF_p25.A12
PL0068  TFA_p23.B8      HTF_p26.D6
PL0059  TFA_p23.B8      HTF_p21.F9
PL0053  TFA_p21.F9      HTF_p7.G11
PL0051  TFA_p5.D5       HTF_p25.A12
PL0046  TFA_p21.F9      HTF_p21.F9
PL0042  TFA_p5.D5       HTF_p25.A12
p8.H7   TFA_p7.G11      HTF_p7.G11
p8.E4   TFA_p25.A12     HTF_p25.A12
p8.C7   TFA_p21.F9      HTF_p21.F9
p8.B9   TFA_p5.D5       HTF_p5.D5
p8.B3   TFA_p25.C2      HTF_p25.C2
p7.G11  TFA_p7.G11      HTF_p7.G11
p6.B9   TFA_p26.D6      HTF_p26.D6
p6.A10  TFA_p25.C2      HTF_p25.C2
p5.H11  TFA_p7.G11      HTF_p7.G11
p5.C3   TFA_p21.F9      HTF_p21.F9
p4.E6   TFA_p21.F9      HTF_p21.F9
p4.E5   TFA_p21.F9      HTF_p21.F9
p4.D2   TFA_p5.D5       HTF_p25.A12
p3.G9   TFA_p5.D5       HTF_p25.A12
p3.F8   TFA_p26.D6      HTF_p26.D6
```

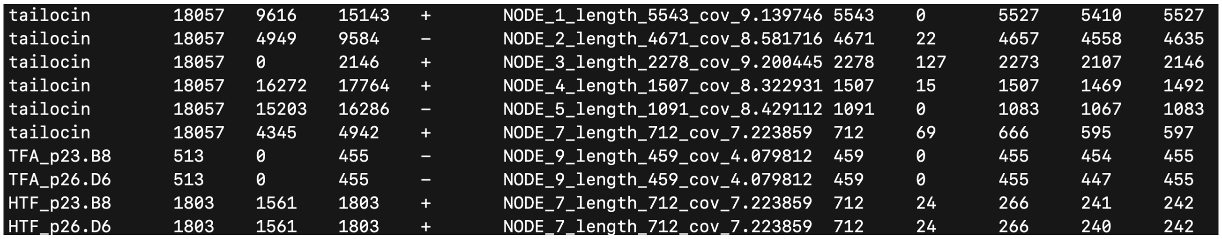
```
p3.F12  TFA_p25.C2      HTF_p25.C2
p3.A3   TFA_p25.C2      HTF_p25.C2
p27.F2  TFA_p5.D5       HTF_p25.A12
p27.D6  TFA_p21.F9      HTF_p21.F9
p26.E7  TFA_p23.B8      HTF_p23.B8
p26.D6  TFA_p26.D6      HTF_p26.D6
p26.B7  TFA_p5.D5       HTF_p25.C2
p25.D2  TFA_p25.C2      HTF_p25.C2
p25.C2  TFA_p25.C2      HTF_p25.C2
p25.C11 TFA_p21.F9      HTF_p21.F9
p25.B2  TFA_p25.C2      HTF_p25.C2
p25.A12 TFA_p25.A12     HTF_p25.A12
p24.H2  TFA_p7.G11      HTF_p7.G11
p22.D4  TFA_p26.D6      HTF_p26.D6
p22.D1  TFA_p21.F9      HTF_p21.F9
p22.B5  TFA_p21.F9      HTF_p21.F9
p22.A8  TFA_p21.F9      HTF_p21.F9
p21.F9  TFA_p21.F9      HTF_p21.F9
p21.F1  TFA_p5.D5       HTF_p5.D5
p21.E3  TFA_p21.F9      HTF_p21.F9
p21.A8  TFA_p21.F9      HTF_p21.F9
p20.G9  TFA_p21.F9      HTF_p21.F9
p20.D4  TFA_p25.C2      HTF_p25.C2
p13.D10 TFA_p5.D5       HTF_p25.A12
p13.C7  TFA_p5.D5       HTF_p5.D5
p13.C1  TFA_p21.F9      HTF_p21.F9
p12.H7  TFA_p25.C2      HTF_p25.C2
p12.G7  TFA_p5.D5       HTF_p25.A12
p12.F2  TFA_p21.F9      HTF_p21.F9
p12.E2  TFA_p5.D5       HTF_p25.C2
p12.A11 TFA_p21.F9      HTF_p21.F9
HB0766  TFA_p21.F9      HTF_p21.F9
HB0737  TFA_p23.B8      HTF_p23.B8
76.LTU_2009_S19 TFA_p21.F9      HTF_p21.F9
75.LTU_1894_S30 TFA_p5.D5       HTF_p25.A12
120.RUS_1860    TFA_p5.D5       HTF_p5.D5
109.NOR_1990    TFA_p5.D5       HTF_p5.D5
```

example to check if the pipeline did a right job: TFA_p23.B8 in HB0737

```
in two things:
1. make sure the strand is reversed if it is '-' done (the ref is strand - and my e
xtracted ref HTF and TFA are strand +, checked p25.C2 is the same as ref, so keep 1
4 contigs and reverse them to strand +, and also when mapped minimap fragments are
-, reverse to +)
2. select based on mapping quality (total length and mapping match base number, sel
ect based on the matched base proportion like $(($matchedbase/$totallengthofHTForTF
A))). done (confimed by p26.E7 and B7) check example TFA_p23.B8 in HB0737
```

first the mapping paf:

ref name / ref length / ref map start / end / strand / contig name /
ref length / start / end /. mapped base/ alignmentlength



```
tailocin     18057   9616    15143   +       NODE_1_length_5543_cov_9.139746 5543    0       5527    5410    5527
tailocin     18057   4949    9584    -       NODE_2_length_4671_cov_8.581716 4671    22      4657    4558    4635
tailocin     18057   0       2146    +       NODE_3_length_2278_cov_9.200445 2278    127     2273    2107    2146
tailocin     18057   16272   17764   +       NODE_4_length_1507_cov_8.322931 1507    15      1507    1469    1492
tailocin     18057   15203   16286   -       NODE_5_length_1091_cov_8.429112 1091    0       1083    1067    1083
tailocin     18057   4345    4942    +       NODE_7_length_712_cov_7.223859  712     69      666     595     597
TFA_p23.B8   513     0       455     -       NODE_9_length_459_cov_4.079812  459     0       455     454     455
TFA_p26.D6   513     0       455     -       NODE_9_length_459_cov_4.079812  459     0       455     447     455
HTF_p23.B8   1803    1561    1803    +       NODE_7_length_712_cov_7.223859  712     24      266     241     242
HTF_p26.D6   1803    1561    1803    +       NODE_7_length_712_cov_7.223859  712     24      266     240     242
```

```
 TFA_p23.B8 mapped fragment is from 0 to 455 and the strand is -, check if pipe rev
erse it:

less HB0737/contigs.fasta | grep 'NODE_9_length_459_cov_4.079812' -A8 (the ref leng
th is 513, the contig length is 459 and mapped length is 0:455)
>NODE_9_length_459_cov_4.079812
TTGAGATGTGGGTGTTATTTGAAGAAATGAAAGATATCTATGGTGAGGTGCCTTTTGCTG
CGTCTCCCAAAGATTCCGAGCCTCACGGCGTCGACCTGCTTAACCGTGCTGTCGCTGGTG
AGTTTGGCGAGGTACTGGAGCCCACCGAGCAAACGGTATTAACGCTGGTTACGCTCCAGC
GGGAAGCCTTTTCAGCGACAGCCACTGCCAGAATCAACGAGTTGGTTGCTGAACTGGATA
TGCTGCAAGACGCTACGGCGTTGAAAATGGAGACTGAAGCGCAAGTGAACTCCTTGCCAG
CGATACAGGCCGAGCTCAATGCGTTCCGTCTTTATCGCGTGCAACTTTCCCAGCTTGAAA
CGTTGGAAGGTTATCCGGCGAATGTCGATTGGCCTGTGGCTCCGGCAAAGCCGTTTGTGT
ATGTGCAGCCGGTCGAAGAAGCCGTGTCTGCTTAA  AACA

less HB0737/contigs.fasta | grep 'NODE_9_length_459_cov_4.079812' -A8 | seqtk seq -
r
>NODE_9_length_459_cov_4.079812
TGTT  TTAAGCAGACACGGCTTCTTCGACCGGCTGCACATACACAAACGGCTTTGCCGGAGCCACAGGCCAATCGACATTCG
CCGGATAACCTTCCAACGTTTCAAGCTGGGAAAGTTGCACGCGATAAAGACGGAACGCATTGAGCTCGGCCTGTATCGCTGGC
AAGGAGTTCACTTGCGCTTCAGTCTCCATTTTCAACGCCGTAGCGTCTTGCAGCATATCCAGTTCAGCAACCAACTCGTTGAT
TCTGGCAGTGGCTGTCGCTGAAAAGGCTTCCCGCTGGAGCGTAACCAGCGTTAATACCGTTTGCTCGGTGGGCTCCAGTACCT
CGCCAAACTCACCAGCGACAGCACGGTTAAGCAGGTCGACGCCGTGAGGCTCGGAATCTTTGGGAGACGCAGCAAAAGGCACC
TCACCATAGATATCTTTCATTTCTTCAAATAACACCCACATCTCAA

and check the final fasta:
less ../msa/HB0737_tailocin_region.fasta | grep 'TFA_p23.B8' -A1
>TFA_p23.B8
TTAAGCAGACACGGCTTCTTCGACCGGCTGCACATACACAAACGGCTTTGCCGGAGCCACAGGCCAATCGACATTCGCCGGAT
AACCTTCCAACGTTTCAAGCTGGGAAAGTTGCACGCGATAAAGACGGAACGCATTGAGCTCGGCCTGTATCGCTGGCAAGGAG
TTCACTTGCGCTTCAGTCTCCATTTTCAACGCCGTAGCGTCTTGCAGCATATCCAGTTCAGCAACCAACTCGTTGATTCTGGC
AGTGGCTGTCGCTGAAAAGGCTTCCCGCTGGAGCGTAACCAGCGTTAATACCGTTTGCTCGGTGGGCTCCAGTACCTCGCCAA
```

```
ACTCACCAGCGACAGCACGGTTAAGCAGGTCGACGCCGTGAGGCTCGGAATCTTTGGGAGACGCAGCAAAAGGCACCTCACCATAGATATCTTTCATTTCTTCAAATAACACCCACATCTCAANNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN


the same. Correct.
And check if the selection of the best matched TFA/HTF is based on mapping quality.
In principle, TFA_p24.B8 has the best quality since the mapped base is 454/513 and
TFA_p26.D6 is 447/513.
Check the final fasta:


 less ../haplotype_selected/HB0737.final.fasta | grep '>'
>tailocin
>TFA_p23.B8
>HTF_p23.B8

correct!
```

then with these selected TFA and HTF, start to build a tree:

E. tree

```
#!/bin/bash
#$ -l tmem=8G
#$ -l h_vmem=8G
#$ -l h_rt=4:30:0
#$ -S /bin/bash
#$ -N m87otree
#$ -o /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/logs/
#$ -e /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/logs/

# Remember first PL0042 cant use --meta, check readme
# and PL0066 and PL0222 have zero contig, need to be excluded in MSA

# Activate the conda environment
source /home/jiajucui/miniconda3/bin/activate phylogeny_snp

# Define the base directories
houtput_dir="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract"
moutput_dir="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_modern85"
tools=/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/tools/
#vim ../../tailocin_modern85/haplotype_selected/all_
#all_HTF_samples.fasta                   all_modern76fa_markexceptlongest.fasta  all_TFA_samples.fasta
```

```
#(base) [jiajucui@pchuckle step3_combinemodern_tree]$ vim ../../tailocin_modern85/haplotype_selected/all_
#filtered_HTF_samples.fasta
new_tree_dir="${houtput_dir}/tree"

#bash /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/modern85_extractfastafromvcf.sh
#cat and remove the Ns, since we believe clustalo omega could do a good job to align them to the right place, may be Ns could mislead since the ref hyplotypes have differen length.
cat $houtput_dir/haplotype_selected/all_TFA_samples.fasta    $moutput_dir/haplotype_selected/all_TFA_samples.fasta  | sed 's/N//g'> $houtput_dir/tree/handm_all_TFA_samples.fasta
TFAfinal_fasta=$houtput_dir/tree/handm_all_TFA_samples_clustalo.fasta
$tools/clustalo -i $houtput_dir/tree/handm_all_TFA_samples.fasta -o $TFAfinal_fasta --force
#--force: Overwrites the output file if it already exists.
cat $houtput_dir/haplotype_selected/all_HTF_samples.fasta    $moutput_dir/haplotype_selected/all_HTF_samples.fasta | sed 's/N//g'> $houtput_dir/tree/handm_all_HTF_samples.fasta
HTFfinal_fasta=$houtput_dir/tree/handm_all_HTF_samples_clustalo.fasta
$tools/clustalo -i $houtput_dir/tree/handm_all_HTF_samples.fasta --force -o $HTFfinal_fasta

cat $houtput_dir/haplotype_selected/filtered_TFA_samples.fasta    $moutput_dir/haplotype_selected/filtered_TFA_samples.fasta | sed 's/N//g'> $houtput_dir/tree/handm_filtered_TFA_samples.fasta
TFAfilter_fasta=$houtput_dir/tree/handm_filtered_TFA_samples_clustalo.fasta
$tools/clustalo -i $houtput_dir/tree/handm_filtered_TFA_samples.fasta --force -o $TFAfilter_fasta

cat $houtput_dir/haplotype_selected/filtered_HTF_samples.fasta    $moutput_dir/haplotype_selected/filtered_HTF_samples.fasta | sed 's/N//g'> $houtput_dir/tree/handm_filtered_HTF_samples.fasta
HTFfilter_fasta=$houtput_dir/tree/handm_filtered_HTF_samples_clustalo.fasta
$tools/clustalo -i $houtput_dir/tree/handm_filtered_HTF_samples.fasta --force -o $HTFfilter_fasta

cd $new_tree_dir
iqtree -s $HTFfinal_fasta -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/tailocin_HTF"
iqtree -s $TFAfinal_fasta -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/tailocin_TFA"
iqtree -s $HTFfilter_fasta -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/filter_HTF"
iqtree -s $TFAfilter_fasta -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/filter_TFA"

mkdir -p $new_tree_dir/fullinfo/
```

```
cd $new_tree_dir/fullinfo/

TFAfinal_fasta2=$houtput_dir/tree/fullinfo/handm_all_TFA_samples_clustalo.fasta
HTFfinal_fasta2=$houtput_dir/tree/fullinfo/handm_all_HTF_samples_clustalo.fasta
TFAfilter_fasta2=$houtput_dir/tree/fullinfo/handm_filtered_TFA_samples_clustalo.fasta
HTFfilter_fasta2=$houtput_dir/tree/fullinfo/handm_filtered_HTF_samples_clustalo.fasta
# Step 5: Build a phylogenetic tree
bash /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/step3_combinemodern_tree/fullinfoMSA.sh $HTFfinal_fasta $HTFfinal_fasta2
bash /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/step3_combinemodern_tree/fullinfoMSA.sh $HTFfilter_fasta $HTFfilter_fasta2
bash /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/step3_combinemodern_tree/fullinfoMSA.sh $TFAfinal_fasta $TFAfinal_fasta2
bash /SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/shfortailocin/step3_combinemodern_tree/fullinfoMSA.sh $TFAfilter_fasta $TFAfilter_fasta2




iqtree -s $HTFfinal_fasta2 -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/fullinfo/ftailocin_HTF"
iqtree -s $TFAfinal_fasta2 -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/fullinfo/ftailocin_TFA"
iqtree -s $HTFfilter_fasta2 -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/fullinfo/ffilter_HTF"
iqtree -s $TFAfilter_fasta2 -nt AUTO -bb 1000 -alrt 1000 -pre "$new_tree_dir/fullinfo/ffilter_TFA"


echo "Analysis complete. Check the directories for results."
```

results:

TFA: 25 historical and 47 modern samples has TFA:

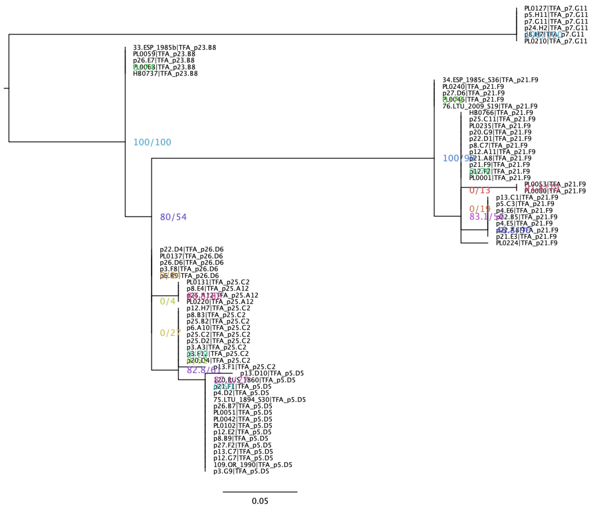
```
(base) [jiajucui@pchuckle fullinfo]$ less handm_all_TFA_samples_clustalo.fasta | grep '>' -c
72
(base) [jiajucui@pchuckle fullinfo]$ less handm_all_TFA_samples_clustalo.fasta | grep '>p' -c
47
```

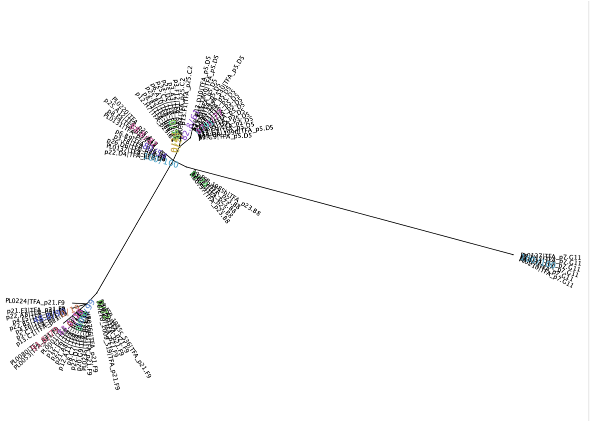
fullinfo MSA:

```
>109.OR_1990|TFA_p5.D5
ATCCAGCTCAGCAACCAGCTCGTTGATTCTGGCAGTGGCTGTTGCTGAAAAGGCTT
>120.RUS_1860|TFA_p5.D5
```

```
ATCCAGCTCAGCAACCAGCTCGTTGATTCTGGCAGTGGCTGTTGCTGAAAAGGCTT
>33.ESP_1985b|TFA_p23.B8
ATCCAGTTCAGCAACCAACTCGTTGATTCTGGCAGTGGCTGTCGCTGAAAAGGCTT
...
```

tree:





TFA: 22 historical and 45 modern samples has TFA (including p25.C2) (the presence was determined by ≥ 65% covered proportion to the specific haplotype):

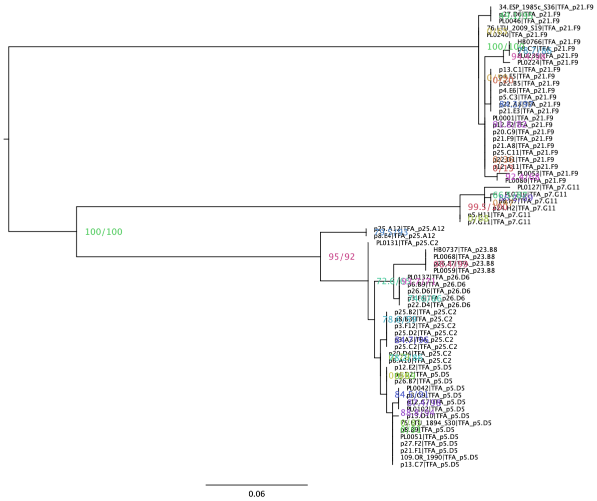
```
(base) [jiajucui@pchuckle fullinfo]$ less handm_filtered_TFA_samples_clustalo.fasta |
grep '>' -c
67
(base) [jiajucui@pchuckle fullinfo]$ less handm_filtered_TFA_samples_clustalo.fasta |
grep '>p' -c
45
```

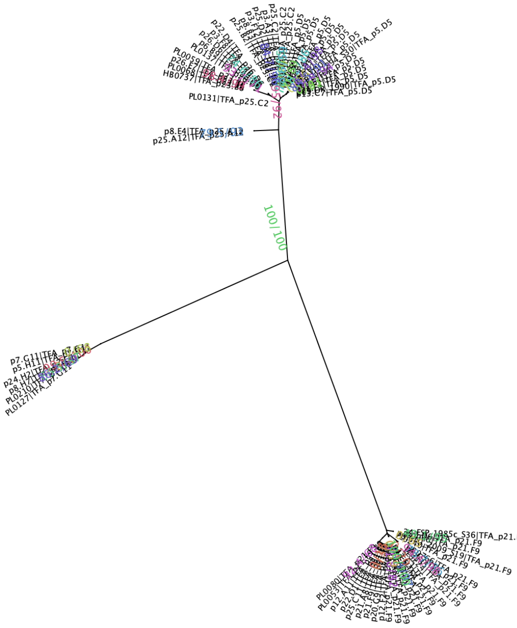
full info MSA：

```
>109.OR_1990|TFA_p5.D5
CGGAGCCACAGGCCAATCGACCTTCGCTGGATAACCTTCCAACGTTTCAAGCTGGGCAAGTTGCACGCGATAAAGACGGAACGCAT
TGAGCTCGGCCTGTATCGCAGGCAAGGAGTTCACTTGCGATTCAGTCTCCATCTTCAACGCCGTGGCGTCTTGCAGCGTATCCAGC
TCAGCAACCAGCTCGTTGATTCTGGCAGTGGCTGTTGCTGAAAAGGCTTCCCGCTGGAGCGTAACCAGCGTTAATACCGTTTGCTC
GGTGGGCTCCAGTACCTCGCCAAA
>34.ESP_1985c_S36|TFA_p21.F9
AGGTGCTACTGGCCACTCGAACTTGGTCGGGTAACCTGCCAGCGTATCGATCTGGGAAAGCCGCACGCGATAAACGCGGAACGCAT
AGAGTTCAGCATTGACTGCCGGTACGGAGTTTATTTGCTGTTGGGTCGCGAGATTCATCGAAATGGCGTCTTGCAACATATCCAGC
TTGGCCACCAACTCGTTGATGCGGGCAGTGGCTGTCGCGGACAAAGCATCCCGCTGGTTTGTTACCTGCGCCAGGATCGTTTGCTC
GGTAGGCTCAAGAACCGGCCCGAA
>75.LTU_1894_S30|TFA_p5.D5
```

```
CGGAGCCACAGGCCAATCGACCTTCGCTGGATAACCTTCCAACGTTTCAAGCTGGGCAAGTTGCACGCGATAAAGACGGAACGCAT
TGAGCTCGGCCTGTATCGCAGGCAAGGAGTTCACTTGCGATTCAGTCTCCATCTTCAACGCCGTGGCGTCTTGCAGCGTATCCAGC
TCAGCAACCAGCTCGTTGATTCTGGCAGTGGCTGTTGCTGAAAAGGCTTCCCGCTGGAGCGTAACCAGCGTTAATACCGTTTGCTC
GGTGGGCTCCAGTACCTCGCCAAA
...
```

tree:





HTF: 24 h and 50 m but full info has no base

```
(base) [jiajucui@pchuckle fullinfo]$ less handm_all_HTF_samples_clustalo.fasta | grep
'>' -c
74
(base) [jiajucui@pchuckle fullinfo]$ less handm_all_HTF_samples_clustalo.fasta | grep
'>p' -c
50
```

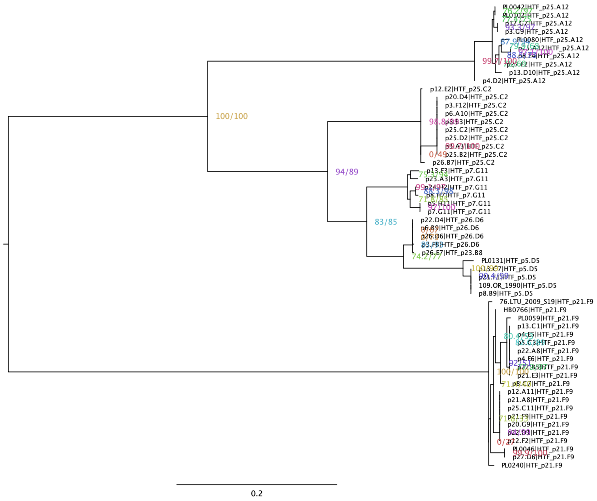
set a threshold of 65% covered proportion: 10 h and 47 m

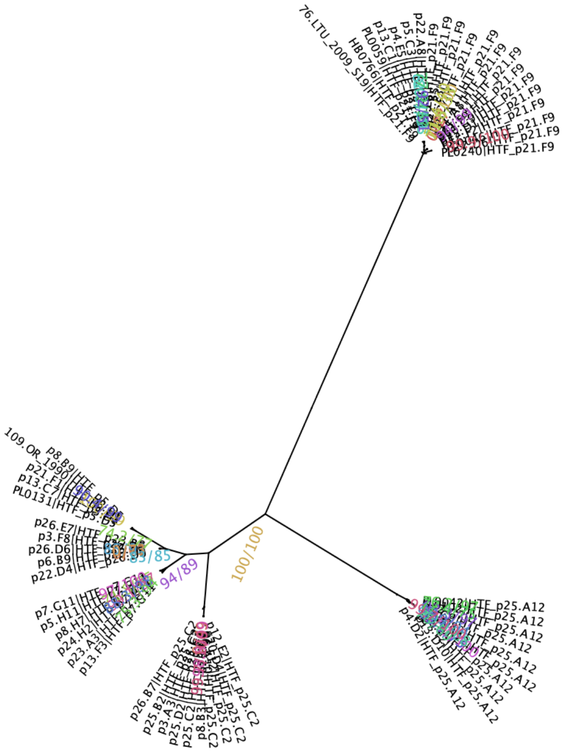
```
(base) [jiajucui@pchuckle fullinfo]$ less handm_filtered_HTF_samples_clustalo.fasta |
grep '>' -c
57
(base) [jiajucui@pchuckle fullinfo]$ less handm_filtered_HTF_samples_clustalo.fasta |
grep '>p' -c
47
```

full info

```
>109.OR_1990|HTF_p5.D5
GAAGTTATTGCCGGTGGCTTTAAACCAGCCCAATTTGATCGGGTGATTTTGCTGACCTACCCCGCCACCCTGCTGAAAGTCAGCCG
CTCGCCGCATATAAGGTTGGTTAAGATCCCCACTAGCCAGTCCTACGTAAGTGATGCTGTCGGCTAATGCTCTGCTCCCGACTCGT
GTATCAACTTCGCTCCGTAAACGTCAGAAAAATTGGCTTTGGGATCGAAGTTAGCGGAGTACCACAGGGTGCCCAAATCGGAACCG
ACTGTGGCTTTAAGGTTAGAGCCAGACCAGCCGATTTTGATTGAATTACGCGTCCTGCGCACGGGTCCAGGCGTCCGCTGCTGCCA
AATAGATCCAGTTCTGCGCCGGGTTGTCCTGGTTCTTGACCAGTACCCGGTCGCCTGCCACCAGCGTGACGTCGTCGATGGTCTGC
AAGCCGCTCAAGCCGATCGCCATGGTCGTGGCGCAGCGCACGGACTTTTTGTAGTCCGAGGCGGCGAGGCCTAGGATGGCCCTGTG
CAGTTGCGTGACATCCGCTTCGCTGGGCACCAGACCGGCCCCCAGAATCACGTTCAAAATTTCCTGCGTCACCGAGTTGCCCCACT
GCGCCGGAATCAGCGAGCCGGGGGTGCCGGTCGCCGGGTTTTCATCTACAAACTTGCCGCTGACCA
>76.LTU_2009_S19|HTF_p21.F9
CCAATAACGGAACCGATCTTCAAACTGGCCGTTATTAACGAGCTTTGCCCCCTGCCGGTTTCTGTCAATAATCGACAAATCAGGAG
CAAAAGCAATGGAGACGGCCGCATTATCAAATGAGATGGGCCAATTGAACGGGATCACACTCGTTCATCCATTGCATGAGTGTTCC
GTTGGGCAGCCTCTCCGGAGACACCCGCAAGAGCGGTGCAATACTTCAGTGCGATGCTGCCCCCCAGCAAGCGCCATTCACCAGCC
ACCCGAGATAAAAGCGCCGAATCACCCGGCCCAATACAATCGGAATTGTGCGTCCTGCGCTCGGGTCCAGGCGTCTGCTGCTGCCA
AATAGATCCAGTTCTGCGCCGGGTTGTCCTGGTTCTTGACCAGTACCCGGTCGCCTGCCACCAGCGTGACGTCGTCGATGGTCTGC
AAGCCGCTCAAGCCGATCGCCATGGTCGTGGCGCAGCGCACGGACTTTTTGTAGTCCGAGGCGGCGAGGCCTAGGATGGCCCTGTG
CAGTTGCGTGACATCCGCTTCGCTGGGCACCAGACCGGCCCCCAGAATCACGTTCAAAATTTCCTGCGTCACCGAGTTGCCCCACT
GCGCCGGAATCAGCGAGCCGGGGGTGCCGGTCGCCGGGTTTTCATCTACAAACTTGCCGCTGACCA
>HB0766|HTF_p21.F9
CCAATAACGGAACCGATCTTCAAACTGGCCGTTATTAACGAGCTTTGCCCCCTGCCGGTTTCTGTCAATAATCGACAAATCAGGAG
CAAACGCAATGGAGACGGCCGCATTATCAAATGAGATGGGCCAATTGAACGGGATCACACTCGTTCATCCATTGCATGAGTGTTCC
GTTGGGCAGCCTCTCCGGAAACACCCGCAAGAGCGGTGCAATACTTCAGTGCGATGCTGCCCCCCAGCAAGCGCCATTCACCAGCC
ACCCGAGATAAAAGCGCCGAATCACCCGGCCCAATACAATCGGAATTGTGCGTCCTGCGCTCGGGTCCAGGCGTCTGCTGCTGCCA
AATAGATCCAGTTCTGCGCCGGGTTGTCCTGGTTCTTGACCAGTACCCGGTCGCCTGCCACCAGCGTGACGTCGTCGATGGTCTGC
AAGCCGCTCAAGCCGATCGCCATCGTCGTAGCGCAGCGCACGGACTTTTTGTAGTCCGAGGCGGCGAGGCCGAGGATGGCCCTGTG
CAGTTGCGTGACATCCGCTTCGCTGGGCACCAGACCGGCCCCCAGAATCACGTTCAGAATCTCCTGCGTCACCGAGTTTCCCCACT
GCGCCGGAATCAGCGAGCCGGGGGTGCCGGTCGCCGGGTTTTCATCTACAAACTTGCCGCTGACCA
...
```

the tree:





the small gap should around tailocin:10249-10856 (check p4.E6 p5.C3)

15502+10249=25751. 15502+10856=26358 the coordinate in ref: 22:25751-26358

#within 22 25175 27703 22_fragment_1_29_REFSEQ_hypothetical_protein
#samtools faidx /SAN/ugi/plant_genom/jiajucui/1_initial_data/reference_genome_Ps/Pseudomonas.OTU5_ref.fasta "22:25751-26358" > ./minigapseq_p25.C2.fasta

```
22  14507   15106   Anthranilate_synthase_component_2
22  15502   33558   Full_Tailocin
22  15502   16011   22_fragment_1_18_sp|P00726|SPAN1_LAMBD_Spanin_inner_membrane_subun
it
```

```
22  16008   16553   22_fragment_1_19_K03791_putative_chitinase
22  16577   17581   22_fragment_1_20_hypothetical_protein

22  17718   18230   22_fragment_1_21_sp|P03740|TFA_LAMBD_Tail_fiber_assembly_protein
22  18241   20043   22_fragment_1_22_hypothetical_protein


22  20054   20653   22_fragment_1_23_Uncharacterised_protein_conserved_in_bacteria_DUF2313
22  20641   21681   22_fragment_1_24_Baseplate_J-like_protein
22  21671   22069   22_fragment_1_25_Phage_protein_GP46
22  22069   22578   22_fragment_1_26_Bacteriophage_Mu_Gp45_protein
22  22575   23681   22_fragment_1_27_sp|P10312|BPD_BPP2_Probable_baseplate_hub_protein
22  23685   25178   22_fragment_1_28_sp|P71389|VPN_HAEIN_Mu-like_prophage_FluMu_DNA_circularization_protein
22  25175   27703   22_fragment_1_29_REFSEQ_hypothetical_protein
22  27834   28130   22_fragment_1_30_Phage_tail_assembly_chaperone_proteins_E_or_41_or_14
22  28127   28474   22_fragment_1_31_Phage_tail_tube_protein
22  28542   30038   22_fragment_1_32_sp|P44233|VPL_HAEIN_Mu-like_prophage_FluMu_tail_sheath_protein
22  30057   30242   22_fragment_1_33_Protein_of_unknown_function_(DUF2635)
22  30239   30829   22_fragment_1_34_REFSEQ_hypothetical_protein
22  30916   31254   22_fragment_1_35_hypothetical_protein
22  31235   31624   22_fragment_1_36_hypothetical_protein
22  31929   32366   22_fragment_1_37_hypothetical_protein
22  32542   33153   22_fragment_1_38_lexA_repressor_LexA_[EC:3.4.21.88]
22  33302   33559   22_fragment_1_39_REFSEQ_hypothetical_protein
22  33648   33830   hypothetical_protein
001.p11.B8.fa   1   498 22_fragment_1_21_Haplotype1
001.p12.G4.fa   1   513 22_fragment_1_21_Haplotype2
001.p12.D2.fa   1   513 22_fragment_1_21_Haplotype3
001.p21.B5.fa   1   513 22_fragment_1_21_Haplotype4
001.p12.C5.fa   1   513 22_fragment_1_21_Haplotype5
```

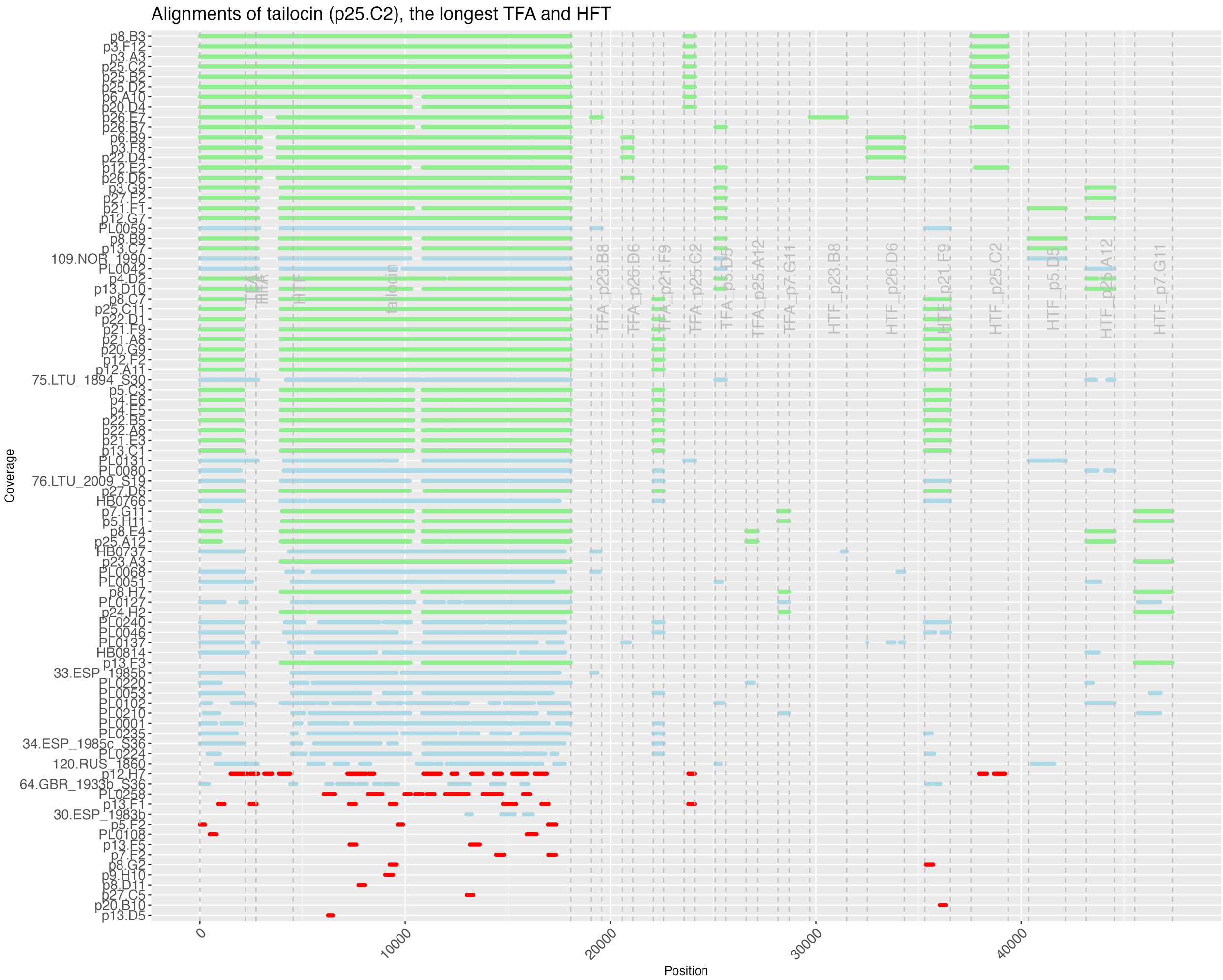
tape measure

step1: extract the tape measure from all samples (tailocin region in /msa) that have them using the coordinate of minigap

step2: for those have Ns in the tape measure region: 1 if its modern, go to the assembly to see if the contig is continuing, if so extract the unmapped seq add to the MSA of tape measure,

step3: if they are historical, using the MSA seq to fish (add the haplotypes to fasta ref genome and map again, then do the same as before for TFA, assembly minimap...)

step4: put them together as a MSA fasta and then use omega to align... then build a tree.



step1: extract the tape measure from all samples that have them in the tailocin region, using the coordinate of minigap

p13.C7: 10249:10880

109: 10249:10880

p12.E2: 10249:10856

76.: 10249:10856

p8.P9: 10249:10880

....

p27.D6: 10249:10934

coordinate is around 10249:10880

extract from fasta:

```
#!/bin/bash

# Define the input FASTA file and the output file for the extracted region
input_fasta="../../tailocin_extract/haplotype_selected/all_merged_hmfa_samples_tailocin_markexceptlongest.fasta"
output='/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract/tapemeasure'
mkdir -p $output
step1=$output/step1
mkdir -p $step1
output_fasta="$step1/extracted_region_all.fasta"

# Clear the output file if it already exists
> "$output_fasta"

# Loop through each sequence name in the .fai index file
while read -r line; do
    # Extract the sequence name (first column in .fai file)
    seq_name=$(echo "$line" | cut -f1)

    # Extract the desired region using samtools faidx and append to the output file
    samtools faidx "$input_fasta" "${seq_name}:10249-10880" >> "$output_fasta"
done < "${input_fasta}.fai"

echo "Extraction complete. Extracted regions are saved in $output_fasta."


#!/bin/bash

# Define the input FASTA file and the output files for the extracted regions and those containing 'N's
input_fasta="../../tailocin_extract/haplotype_selected/all_merged_hmfa_samples_tailocin_markexceptlongest.fasta"
output_fasta="$step1/noNs_extracted_region.fasta"
output_modern_ns_fasta="$step1/modernNs.fasta"
output_historical_ns_fasta="$step1/historicalNs.fasta"


# Clear the output files if they already exist
> "$output_fasta"
> "$output_modern_ns_fasta"
> "$output_historical_ns_fasta"

# Loop through each sequence name in the .fai index file
while read -r line; do
    # Extract the sequence name (first column in .fai file)
    seq_name=$(echo "$line" | cut -f1)

    # Extract the desired region using samtools faidx
    extracted_region=$(samtools faidx "$input_fasta" "${seq_name}:10249-10880")

    # Modify the header to "${seq_name}_tape"
    modified_region=$(echo "$extracted_region" | sed "1s/.*/>${seq_name}_tape/")

    # Check if the extracted region contains 'N'
    if echo "$modified_region" | grep -q 'N'; then
        # Append to the appropriate file based on the sequence name prefix
        if [[ $seq_name == p* ]]; then
            echo "$modified_region" >> "$output_modern_ns_fasta"
        else
            echo "$modified_region" >> "$output_historical_ns_fasta"
        fi
    else
        # Append to the output file for extracted regions
        echo "$modified_region" >> "$output_fasta"
    fi
done < "${input_fasta}.fai"

echo "Extraction complete. Extracted regions are saved in $output_fasta. Modern sequences containing 'N' are saved in $output_modern_ns_fasta. Historical sequences containing 'N' are saved in $output_historical_ns_fasta."
```

33 samples have full-length tape measure 10249:10880 (23 m and 10 h)

55 have Ns inside (20 h and 35 m)

```
33 fulllength tape measure:
>p12.A11_tape
TCCCCGTGAAGAAACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGACA
CGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCCC
AGAGACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGCG
ACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACGC
CCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATCG
GCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGT
CGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGA
TCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTGA
TGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGGC
CAATGGCACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCATGACGGGTTCCC
GGATCGCCGCCCATACCGCCTGGAAATACGGT
>p12.F2_tape
TCCCCGTGAAGAAACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGACA
CGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCCC
AGAGACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGCG
ACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACGC
CCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATCG
GCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGT
```

```
CGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGA
TCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTGA
TGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGGC
CAATGGCACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCATGACGGGTTCCC
GGATCGCCGCCCATACCGCCTGGAAATACGGT
>p12.G7_tape
TCCCCGTGAAGAAACCGCTCACGGGGTGCCAGGCAGCCGAAATCAGTTCCATAGGCGACA
CGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCCC
AGAGACCAGAGAAGAAGCTGCTAACGGGTTGCCAGGCAGACGAAATCGTGTCCATCGGCG
ACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACGC
CCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATCG
GCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGT
CGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGA
TCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTGA
TGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGGC
CAATGACACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCATGACGGGTTCCC
GGATCGCCGCCCATACCGCCTGGAAATACGGT
...

10h
>33.ESP_1985b_tape
>HB0737_tape
>HB0766_tape
>HB0814_tape
>PL0042_tape
>PL0051_tape
>PL0059_tape
>PL0068_tape
>PL0080_tape
>PL0220_tape
```

step2: for those have Ns in the tape measure region: 1 if its modern, go to the assembly to see if the contig is continuing, if so extract the unmapped seq add to the MSA of tape measure,

35 modern:

```
less modernNs.fasta | grep '>'
>p12.E2_tape
>p12.H7_tape
>p13.C1_tape
>p13.C7_tape
>p13.D10_tape
>p13.D5_tape
>p13.F1_tape
>p13.F3_tape
>p13.F5_tape
>p20.B10_tape
```

```
>p20.D4_tape
>p21.E3_tape
>p21.F1_tape
>p22.A8_tape
>p22.B5_tape
>p24.H2_tape
>p25.A12_tape
>p25.D2_tape
>p26.B7_tape
>p27.C5_tape
>p27.D6_tape
>p4.E5_tape
>p4.E6_tape
>p5.C3_tape
>p5.F2_tape
>p5.H11_tape
>p6.A10_tape
>p7.F2_tape
>p7.G11_tape
>p8.B9_tape
>p8.D11_tape
>p8.E4_tape
>p8.G2_tape
>p8.H7_tape
>p9.H10_tape
```

extract the gap from assemblies:

1. extract the end of the contig before the tape and the start of the contig after the tape, check if the two contigs is continuing and interrupted in the tape only, then extract the fasta seq with the coordinate of the gap (if the strand is - then reverse)

```
#!/bin/bash

# Define the input FASTA file and the output file for the extracted region
output='/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_extract/tapemeasure'
mkdir -p $output
step2=$output/step2
mkdir -p $step2

input_fasta="$output/step1/modernNs.fasta"

# Function to extract and process each sample
process_sample() {
    local sample=$1
    local paf_file=$2
    local raw_fasta=$3
    tape_real=-1
    tape_start=-1
    tape_end=0
    tape_chr=0
    strand=0
    while read -r line; do
        ref_name=$(echo "$line" | awk '{print $1}')
        ref_start=$(echo "$line" | awk '{print $3}')
        ref_end=$(echo "$line" | awk '{print $4}')
        strand=$(echo "$line" | awk '{print $5}')
        contig_name=$(echo "$line" | awk '{print $6}')
        contig_start=$(echo "$line" | awk '{print $8}')
        contig_end=$(echo "$line" | awk '{print $9}')
       if [[ $strand == "+" ]]; then
        if [[ $ref_name == "tailocin" ]]; then
            if (( (ref_end - 10249) <= 300 && (ref_end - 10249) >= -300 )); then
                #tape_start="$contig_end"
                tape_start=$((10249 - ref_end + contig_end))
                #tape_end=$((tape_real + 631)) wrong eg in p6.A10 the gap length is no
t the same in contig
                tape_chr="$contig_name"
                strand='+'
            fi
            if (( (ref_start - 10880) <= 300 && (ref_start - 10880) >= -300 )); then

                tape_end=$((10880 - ref_start + contig_start))
                tape_chr2="$contig_name"
            fi
        fi
       else
        if [[ $ref_name == "tailocin" ]]; then
            if (( (ref_end - 10249) <= 300 && (ref_end - 10249) >= -300 )); then
                #tape_end="$contig_start"
                tape_end=$((ref_end-10249 + contig_start))
                tape_real=$((tape_end - 631))
                tape_chr="$contig_name"
                strand='-'
            fi
            if (( (ref_start - 10880) <= 300 && (ref_start - 10880) >= -300 )); then
                tape_start=$((ref_start-10880 + contig_end))
                #the start is from 10880 to 10249, check p13.C1, and then I reverse it
to 10249 to 10880.
                #tape_real=$((ref_start - 10880 + contig_end))
                #tape_end=$((tape_real + 631))
                tape_chr2="$contig_name"
            fi
        fi
       fi
    done < "$paf_file"
    if [[ $tape_chr == "$tape_chr2" ]]; then
        tape_coor="${tape_chr}:${tape_start}-${tape_end}"
        if [[ "$strand" == "+" ]]; then
          echo ">${sample}_tape" >> "$step2/modernNs30_tape_MSA.fasta"
          samtools faidx "$raw_fasta" "$tape_coor" |seqtk seq >> "$step2/modernNs30_tape_MSA.fasta"
        else
          echo '-' $sample
          #samtools faidx "$raw_fasta" "$tape_coor" > "$step2/${sample}_tape.fasta"
          echo ">${sample}_tape" >> "$step2/modernNs30_tape_MSA.fasta"
          samtools faidx "$raw_fasta" "$tape_coor" | seqtk seq -r - >> "$step2/modernNs30_tape_MSA.fasta"
        fi
       # Save individual sample tape to a separate file
    fi
}

# Clear the modernNs30_tape_MSA.fasta if it already exists
> "$step2/modernNs30_tape_MSA.fasta"

# Extract sample names from modernNs.fasta and process each sample
grep ">" "$input_fasta" | sed 's/>//' | while read -r sample_tape; do
    sample=$(echo "$sample_tape" | sed 's/_tape//')
    echo $sample
    # Define paths for the PAF file and raw FASTA (you need to adjust these paths as necessary)
    raw_fasta="/SAN/ugi/plant_genom/jiajucui/phylogeny/phylogeny_read2tree/read2treeinput/pankmerwithpan85modernraw/rawfasta30nonOTU5_55OTU5/${sample}.fasta.bgz"
    paf_file="/SAN/ugi/plant_genom/jiajucui/4_mapping_to_pseudomonas/tailocin_modern85/mappings/${sample}_mapped.paf"


    # Process each sample
    process_sample "$sample" "$paf_file" "$raw_fasta"
done

echo "Processing complete. Extracted regions are saved in $step2."
```

xpasy protein ....

blast protein to see the start codon, and the end condon could be missed since it would be where the whole protein ends, and found 3-5 frame in which the 42 until ATC the start codon should be moved to extract the tape measure seq....

results: (the pipeline was double checked by p6.A10 (+) and p7.G11 (-), and also by omega, showing a high quality but partially covered mapping pattern.) ref: https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240708-204823-0736-70542697-p1m&js=pass

22 modern have tape measure

```
>p12.E2_tape
>215:25605-25990
>p13.C1_tape
>44:298607-299115
>p13.C7_tape
>5:25660-26045
>p13.D10_tape
>1:25238-25623
>p13.F3_tape
>131:24628-25013
>p20.D4_tape
>17:25834-26219
>p21.E3_tape
>46:298602-299110
>p21.F1_tape
>63:592331-592839
>p22.A8_tape
>34:298607-299115
>p22.B5_tape
>35:38925-39433
>p24.H2_tape
>79:24549-24934
>p26.B7_tape
>96:61991-62499
>p27.D6_tape
>44:26118-26503
>p4.E5_tape
>161:74977-75485
>p4.E6_tape
>23:298611-299119
>p5.C3_tape
>17:298607-299115
>p5.H11_tape
>37:214082-214590
>p6.A10_tape
>41:25834-26219
>p7.G11_tape
>59:209552-210060
>p8.B9_tape
>5:25672-26057
>p8.E4_tape
>71:17907-18415
>p8.H7_tape
>41:24550-24935

double checked by p6.A10 (+) and p7.G11 (-),
and also by omega, show a high quality but partially covered mapping pattern.
```

```
the rest of (35-22=13): check (11 nonOTU5 and 2 has no continous gap)
      1 >p12.H7_tape nonOTU5
      1 >p13.D5_tape nonOTU5
      1 >p13.F1_tape nonOTU5
      1 >p13.F5_tape nonOTU5
      1 >p20.B10_tape nonOTU5
      1 >p25.A12_tape uncontinous
      1 >p25.D2_tape uncontinous
      1 >p27.C5_tape nonOTU5
      1 >p5.F2_tape nonOTU5
      1 >p7.F2_tape nonOTU5
      1 >p8.D11_tape nonOTU5
      1 >p8.G2_tape nonOTU5
      1 >p9.H10_tape nonOTU5
```

p25.A12

```
tailocin        18057   12144   18057   +       136     18194   1557    7471    5786
tailocin        18057   3943    10249   -       7       24085   115     6427    6137
tailocin        18057   10856   12060   +       136     18194   269     1473    1198
```

p25.D2: the gap is between contigs

```
tailocin        18057   0       10483   -       284     11623   0       10483   10483
tailocin        18057   10517   18057   +       662     67585   57      7597    7539
```

#then we have 33 samples having full-length tape measure 10249:10880 (23 m and 10 h), and 22 modern samples having unmapped and incomplete tape measure.

with these 55 seqs, do mapping against these seqs again and try to find out the msa in historical samples (all 46 together, for now 23 should be the same as before and the rest 23 should have contigs mapping to different tape measure)

step3: if they are historical, using the MSA seq to fish (add the haplotypes to fasta ref genome and map again, then do the same as before for TFA, assembly minimap...)

the pipe is similar to before for TFA and HTF, but the ref here is the 55 tape measure seq. (try kners)

```
5 more historical samples:

less all_HTF_samples.fasta | grep '>'
>109.NOR_1990|p21.F1_tape
>76.LTU_2009_S19|p24.H2_tape
>HB0737|p8.C7_tape x
>HB0766|HB0766_tape x
>PL0046|p27.D6_tape
>PL0051|p4.D2_tape x
>PL0080|p24.H2_tape x
>PL0131|p24.H2_tape
>PL0137|p21.F1_tape
>PL0220|PL0220_tape x


the 10 having tape in step1:

>33.ESP_1985b_tape
>HB0737_tape
>HB0766_tape
>HB0814_tape
>PL0042_tape
>PL0051_tape
>PL0059_tape
>PL0068_tape
>PL0080_tape
>PL0220_tape
```

step4: put them together as a MSA fasta and then use omega to align… then build a tree.

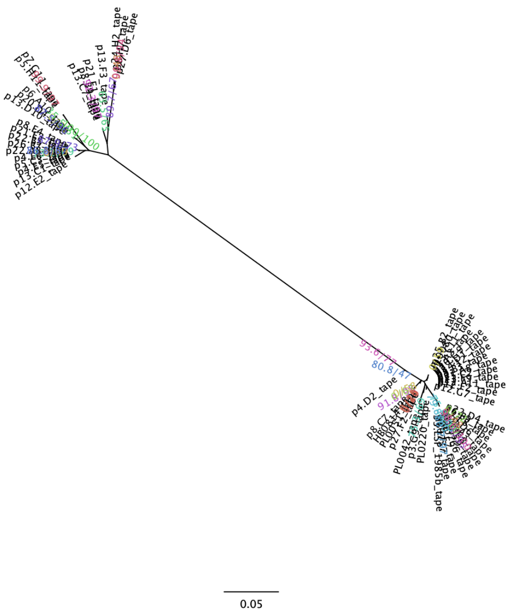
the tree of all 55 (33 having tapemeasure mapped to ref and 22 extracted from contig by myself):

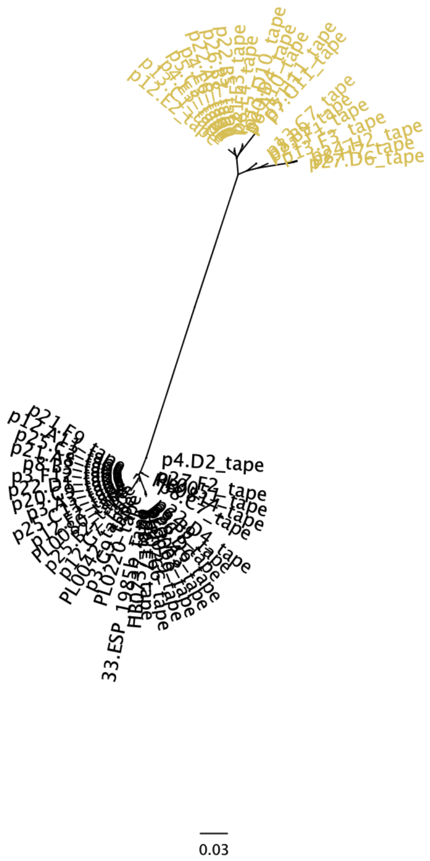
build by omega and then fullinfo

```
>p12.A11_tape
GACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGCGACCAGTCGAACATCGATTGAAAATAACCA
ATCACCGGCGAAGCCAAGGCCTGGATAACGCCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCG
ATCGGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGTCGTCCCATAAGCCGACAAAAAA
TGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAG
CCAGCACCTTGATGCCATCCCACAAGCCAACAAAAAATGCG
>p12.F2_tape
GACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGCGACCAGTCGAACATCGATTGAAAATAACCA
ATCACCGGCGAAGCCAAGGCCTGGATAACGCCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCG
```

```
ATCGGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGTCGTCCCATAAGCCGACAAAAAA
TGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAG
CCAGCACCTTGATGCCATCCCACAAGCCAACAAAAAATGCG
>p12.G7_tape
GACCAGAGAAGAAGCTGCTAACGGGTTGCCAGGCAGACGAAATCGTGTCCATCGGCGACCAGTCGAACATCGATTGAAAATAACCA
ATCACCGGCGAAGCCAAGGCCTGGATAACGCCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCG
ATCGGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGTCGTCCCATAAGCCGACAAAAAA
TGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAG
CCAGCACCTTGATGCCATCCCACAAGCCAACAAAAAATGCG
...
```

the support is ok and the yellow ones are the 22 extracted from contig gaps, which is separated with the 33 that mapped to ref (black ones):





then the tree of all samples:

t/blastx blastx to translate seq to protein (select codon usage bacteria)

read the paper of tape measure.... repeat.... check the MSA,

check week 25 strategy with 70% threshold, pankmer, aMeta, editdistance rmdup, beast2...

1. update aMeta results and the new bams of 4 ref after break
2. check new HPA dataset? stat of host and HPA (PCA?)
3. pankmer hahah
4. all sample nodup and depth recalculation after rmdup
5. paper reading
6. PCA comparison between At Ps HPA and tailocin, use SNP and also mash -s 100,000

also have a look at Hierarchical clustering

about tape measure:

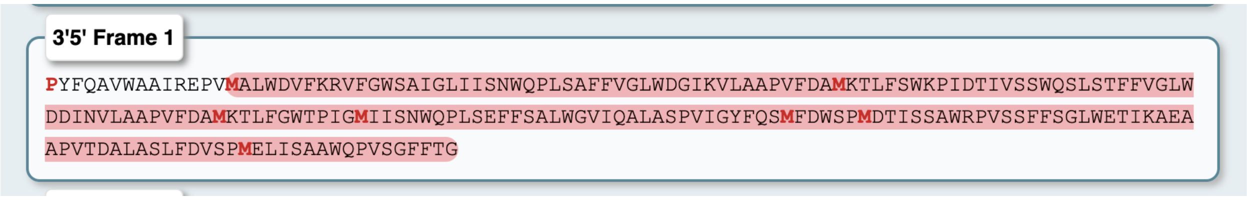
1. xpasy protein ....

samtools faidx haplotype_selected/all_merged_hmfa_samples_tailocin_markexceptlongest.fasta "p25.C2:10249-10880"

> ```
> p25.C2:10249-10880
> ATCCCCGTGAAGAAACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGAC
> ACGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCC
> CAGAGACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGC
> GACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACG
> ```

> CCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATC
> GGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATG
> TCGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCG
> ATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTG
> ATGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGG
> CCAATGGCACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGC
> **CAT**GACGGGTTCC
> CGGATCGCCGCCCATACCGCCTGGAAATACGG

then with xpasy translate we found :



3'5' Frame 1

PYFQAVWAAIREPVMALWDVFKRVFGWSAIGLIISNWQPLSAFFVGLWDGIKVLAAPVFDAMKTLFSWKPIDTIVSSWQSLSTFFVGLW
DDINVLAAPVFDAMKTLFGWTPIGMIISNWQPLSEFFSALWGVIQALASPVIGYFQSMFDWSPMDTISSAWRPVSSFFSGLWETIKAEA
APVTDALASLFDVSPMELISAAWQPVSGFFTG

the strand is from 3 to 5, reversed, and the M (AUG / in DNA TAC) should be the start codon, so we remove P**YFQAVWAAIREPV from the end,**

which is GACGGGTTCCCGGATCGCCGCCCATACCGCCTGGAAATACGG

blast protein to see the start codon, and the end condon could be missed since it would be where the whole protein ends, and found 3-5 frame in which the 42 until TAC the start codon should be moved to extract the tape measure seq....

then the coordinate should be 10249:10838

```
 samtools faidx haplotype_selected/all_merged_hmfa_samples_tailocin_markexceptlongest.
fasta "p25.C2:10249-10838"
>p25.C2:10249-10838
ATCCCCGTGAAGAAACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGAC
ACGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCC
CAGAGACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGC
GACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACG
CCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATC
GGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATG
TCGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCG
ATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTG
ATGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGG
CCAATGGCACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCAT

the length is 590
```

then do step1 and step2 with the new coordinate:

step1:

allgood

34 this time, one more is :

```
>75.LTU_1894_S30_tape
AACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGACACGTCGAACAACG
AGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCCCAGAGACCAGAGA
AGAAGCTGCTAACGGGTTGCCAGGCAGACGAAATCGTGTCCATCGGCGACCAGTCGAGCA
TCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACGCCCCAAAGCGCGC
TGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATCGGCGTCCAGCCAA
ACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGTCGTCCCATAAGC
CGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGATCGGCTTCCAGC
TGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTGATGCCATCCCACA
AGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGGCCAATGGCACTCC
AGCCGAACACCCTCTTGAACACATCCCACAGCGCCATGACGGGTTCCCGG
```

some thing about shift match when querying with coordinated. maybe we could use the p25.C2 fragment to minimap to other samples… think about it later

```
https://www.ebi.ac.uk/jdispatcher/msa/clustalo/summary?jobId=clustalo-I20240710-173944
-0221-34851026-p1m&js=pass
>p12.G7_tape
TCCCCGTGAAGAAACCGCTCACGGGGTGCCAGGCAGCCGAAATCAGTTCCATAGGCGACA
CGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCCC
AGAGACCAGAGAAGAAGCTGCTAACGGGTTGCCAGGCAGACGAAATCGTGTCCATCGGCG
ACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACGC
CCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATCG
GCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATGT
CGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCGA
TCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTGA
TGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGGC
CAATGACACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCATG
>p20.G9_tape
ATCCCCGTGAAGAAACCGCTCACGGGTTGCCAGGCAGCCGAAATCAGTTCCATAGGCGAC
ACGTCGAACAACGAGGCCAACGCATCGGTGACAGGTGCCGCCTCAGCCTTGATGGTTTCC
CAGAGACCAGAGAAGAAGCTGCTAACGGGTCGCCAGGCAGACGAAATCGTGTCCATCGGC
GACCAGTCGAACATCGATTGAAAATAACCAATCACCGGCGAAGCCAAGGCCTGGATAACG
CCCCAAAGCGCGCTGAAAAACTCGGACAACGGTTGCCAGTTCGAGATGATCATGCCGATC
GGCGTCCAGCCGAACAGCGTTTTCATCGCGTCGAAGACCGGCGCGGCCAGAACATTGATG
TCGTCCCATAAGCCGACAAAAAATGTGGACAGCGATTGCCAGCTTGAGACGATCGTGTCG
ATCGGCTTCCAGCTGAACAACGTTTTCATTGCGTCGAACACCGGCGCAGCCAGCACCTTG
ATGCCATCCCACAAGCCAACAAAAAATGCGGACAACGGCTGCCAATTCGAAATGATGAGG
CCAATGGCACTCCAGCCGAACACCCTCTTGAACACATCCCACAGCGCCAT



two different pattern like shift 1 base in omega...
```

step2:

```
less modernNs30_tape_MSA.fasta | grep '>'


>p12.E2_tape
>215:25605-25990
>p13.C1_tape
>44:298607-299115
>p13.C7_tape
>5:25660-26045
>p13.D10_tape
>1:25238-25623
>p13.F3_tape
>131:24628-25013
>p20.D4_tape
>17:25834-26219
>p21.E3_tape
>46:298602-299110
>p21.F1_tape
>63:592331-592839
>p22.A8_tape
>34:298607-299115
>p22.B5_tape
>35:38925-39433
>p24.H2_tape
>79:24549-24934
>p26.B7_tape
>96:61991-62499
>p27.D6_tape
>44:26118-26503
>p4.E5_tape
>161:74977-75485
>p4.E6_tape
>23:298611-299119
>p5.C3_tape
>17:298607-299115
>p5.H11_tape
>37:214082-214590
>p6.A10_tape
>41:25834-26219
>p7.G11_tape
>59:209552-210060
>p8.B9_tape
>5:25672-26057
>p8.E4_tape
```

```
>71:17907-18415
>p8.H7_tape
>41:24550-24935

double checked by p6.A10 (+) and p7.G11 (-),
and also by omega, show a high quality but partially covered mapping pattern.



the rest of (35-22=13): check (11 nonOTU5 and 2 has no continous gap)
      1 >p12.H7_tape nonOTU5
      1 >p13.D5_tape nonOTU5
      1 >p13.F1_tape nonOTU5
      1 >p13.F5_tape nonOTU5
      1 >p20.B10_tape nonOTU5
      1 >p25.A12_tape uncontinous
      1 >p25.D2_tape uncontinous
      1 >p27.C5_tape nonOTU5
      1 >p5.F2_tape nonOTU5
      1 >p7.F2_tape nonOTU5
      1 >p8.D11_tape nonOTU5
      1 >p8.G2_tape nonOTU5
      1 >p9.H10_tape nonOTU5
```

the similar as before.

in summary

we have 55 modern and 30 historical assemblies:

in step 1, we have 34 (23 modern and 11 historical samples) tape measure whose length is 590, mapped to reference p25.C2 tape measure region and extracted from sample assembly.

In step 2, we have 22 out of 35 modern extracted using the coordinate tailocin:10249-10838 (from the start codon M (AUG / in DNA TAC)). 11 in the 13 having no tape measure are nonOTU5 and 2 rest of them has uncontinous contigs in the tape measure region.

step3 last time only 5 more historical out of 19 have tape measure partially mapped to these haplotypes in step1 and 2.

the unique length of the 56 tape measure:

```
less ../../tailocin_extract/tapemeasure/step3/alltapemeasureforfishlength.txt  | cut -d ' ' -f2 | sort | uniq
344
467
590
```

later: optimse the fishing step3 of tailocin

summary all so far in a manuscript

2. read the book, read the paper of tape measure and tailocin
3. generate figures as in paper

alphafold: https://alphafold.ebi.ac.uk/entry/A0A0P9YZE4

expasy: https://web.expasy.org/translate/

blast: https://blast.ncbi.nlm.nih.gov/Blast.cgi