GoSeqIt

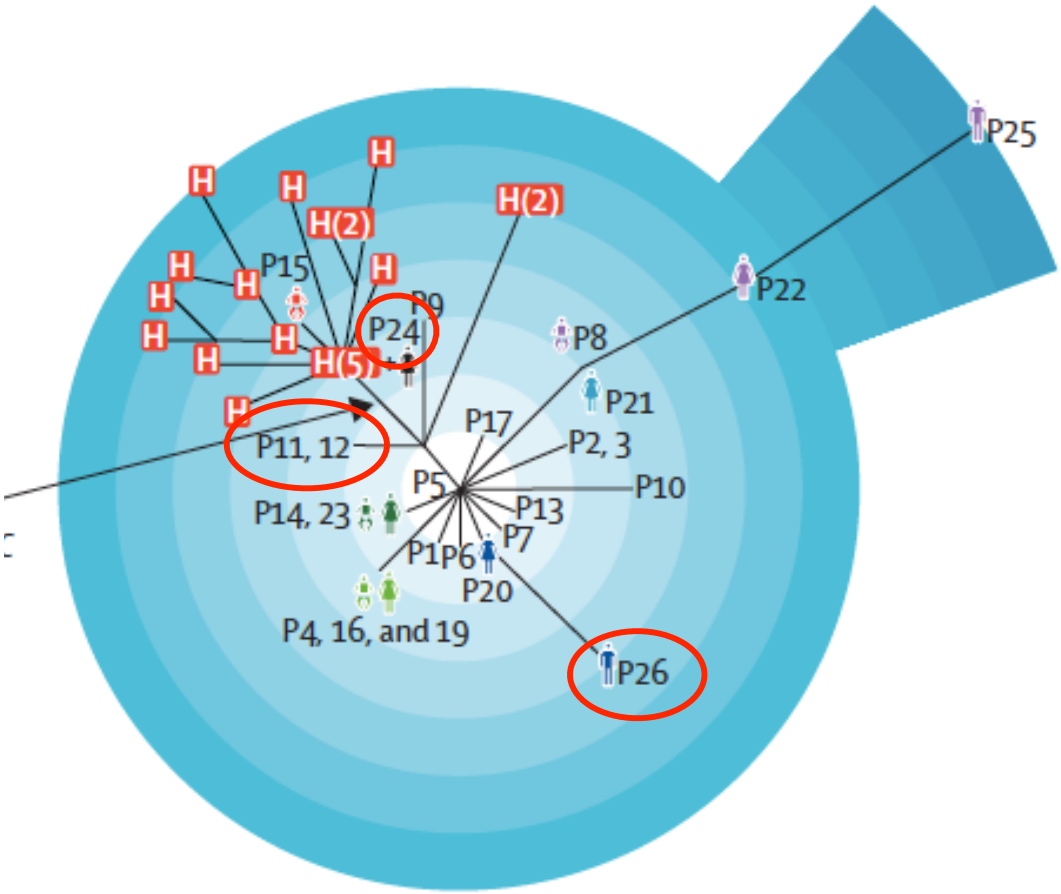
Ex. 7 Recap

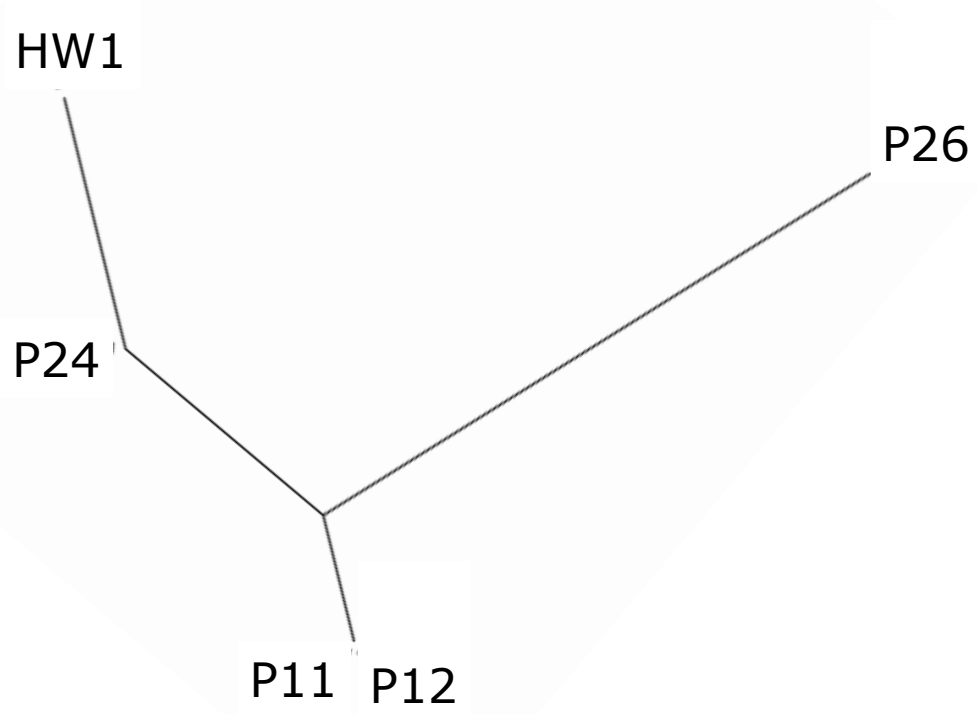
Q1: How many nucleotide differences are there between the HW1 and P26 strains?

Answer: 9

```
5
HW_1.fastq      0.00000000 5.00000000 5.00000000 2.00000000 9.00000000
P11_1.fastq     5.00000000 0.00000000 0.00000000 3.00000000 6.00000000
P12_1.fastq     5.00000000 0.00000000 0.00000000 3.00000000 6.00000000
P24_1.fastq     2.00000000 3.00000000 3.00000000 0.00000000 7.00000000
P26_1.fastq     9.00000000 6.00000000 6.00000000 7.00000000 0.00000000
```

Q2. On your local computer, open the tree.newick file in Figtree (one of the programs you were asked to install from home). Under “Layouts” in the upper left side, select the “radial three layout”. Compare the overall tree topology with the corresponding tree that you can see below from the paper by Harris et al.. Do the trees look similar?





Q3: If you open the output file nctree.log, you can see the length of the reference genome (# Length of chromosome 1) and the number of positions used for phylogeny (# Number of positions used for phylogeny in chromosome 1). Ideally, the number of positions used for phylogeny should be at least 80%. Is it?

```
# Reading inputfile
# Length of chromosome 1: 2832299
# Number of strains: 5
# Number of chromosomes: 1
# Time used:  0  seconds
# Cast input sequences in numpy
# Chromosome: 1
# Number of positions used for phylogeny in chromosome 1: 2746472
# Time used:  12  seconds
# Calculating pairwise distances
# Writing output in neighbor format
# Finishing. Time used:  12  seconds
```

Answer: Yes, 2746472 / 2832299 = 97.0%

Q4: What is the minimum fraction covered and by which input file?

```
I00001.assimpler.out # frac covered:  0.975306
I00002.assimpler.out # frac covered:  0.975421
I00003.assimpler.out # frac covered:  0.975347
I00004.assimpler.out # frac covered:  0.975180
I00005.assimpler.out # frac covered:  0.975302
```

According to “names”,
I000004 = P24

Ex. 7-extra

Look at run of NDtree that besides the same four test strains also contained the non-outbreak strain, ERR070035.

Modify the extract_fracs.sh script so that it will also extract information from this strain and rerun the script.

extract_fracs_copy.sh

```
#!/bin/bash


awk '/frac covered/ {print FILENAME,$0}' I00001.assimpler.out >> covered_fractions
awk '/frac covered/ {print FILENAME,$0}' I00002.assimpler.out >> covered_fractions
awk '/frac covered/ {print FILENAME,$0}' I00003.assimpler.out >> covered_fractions
awk '/frac covered/ {print FILENAME,$0}' I00004.assimpler.out >> covered_fractions
awk '/frac covered/ {print FILENAME,$0}' I00005.assimpler.out >> covered_fractions
awk '/frac covered/ {print FILENAME,$0}' I00006.assimpler.out >> covered_fractions
```

Q5: What is the fraction of the reference genome that is covered by reads from the ERR070035 strain?

Answer: 90.8%

Questions ?

Evaluation!

Please remember to pick up your course certificates!