Obtaining the sequences

1. Visit the NCBI website (https://www.ncbi.nlm.nih.gov/), and search for the sequences using the keywords ou[au], ciesielski[au] and V3 (the gene region).
2. Download a selection of V3 (gene region) sequences. Sequences starting FLD are from the dentist; FLP from his patients, and FLQ and LC are local controls (V3 sequences from HIV isolated from other HIV-positive individuals in that part of Florida). Get all the dentist sequences, at least 3 sequences from each patient, and at least 10 local controls total. Save the sequences in a FASTA-formatted plain text file.
3. Rename the sequence headers in a sensible way, so you will be able to easily interpret the tree later on (e.g., start each name with a tag to denote if its from a patient, the dentist, or a local control).

Sequence alignment and editing

1. We'll use the alignment program mafft to align the sequences. On bluecrystal, mafft is installed, but need to make it available by loading its module using the `module load` command; e.g. `module load apps/mafft-7.402`. (To see a list of available modules, type `module avail`).
2. Align the sequences using your favourite mode in mafft. If you don't yet have a favourite, the --auto flag can be used to pick a "sensible" default:

```
mafft --auto mySequences > myAlignment
```

If you are working with the people sitting near you, arrange it so that you use different alignment modes, and compare (see https://mafft.cbrc.jp/alignment/software/algorithms/algorithms.html for more on the various modes and their properties. L-ins-i is often a good choice for datasets that are not too large.

3. Copy your alignment to your local machine (with scp), and inspect it using an alignment viewer (if you can, install JalView or Seaview locally. If you can't, or don't want to, you can use an online, web browser-based tool such as MView). Are there unreliable regions? How confident are you of the hypotheses of homology expressed by the alignment?

4. Use the "automated1" mode in the alignment editing program trimal (also available via module load) to delete the poorly aligning parts of your alignment. The command to use will be something like:
trimAl -in myAlignment -phylip -automated1 -out trimmedAlignment

5. How do the trimmed and original alignments compare?

Maximum likelihood tree inference (using IQ-Tree)

1. Infer a phylogeny for both alignments using the best-fitting substitution model in IQ-Tree (also available via module load). IQ-Tree is a maximum likelihood program with a very flexible choice of models. The -m MFP option computes the AIC and BIC scores for a range of models, and then picks the model with the lowest BIC as the best.

```
iqtree -s myAlignment -m MFP -bb 1000
```

This command will select the best-fitting model according to BIC from a set of candidates, then infer a maximum likelihood tree, assessing support using 1000 rapid bootstrap replicates.

2. Which model was selected for each alignment? What do the different components of the model “code” mean?

3. Copy the tree files (*.contree) to your local machine, and visualise them using a tree viewing program. You can use FigTree or, if you don’t have one installed, the web browser-based tree viewer at itol.embl.de.

4. Did the dentist infect his patients? How confident are you?