Phylogenetics HW2 Report

Methods

Multiple alignment was performed using online MAFFT with default settings. To create an input file for MAFFT, the sequences were merged into a single file using code:

```python
import os

def combine_fasta_files(folders, output_file):
    with open(output_file, 'w') as outfile:
        for folder in folders:
            if not os.path.isdir(folder):
                print(f"Folder not found: {folder}")
                continue

            for filename in os.listdir(folder):
                if filename.endswith('.fasta') or filename.endswith('.fa'):
                    file_path = os.path.join(folder, filename)
                    with open(file_path, 'r') as infile:
                        outfile.write(infile.read())

folders_1 = ['/content/Human']
output_file_1 = '/content/human_combined_sequences.fasta'
folders_2 = ['/content/Human', '/content/Neanderthal', '/content/Denisova']
output_file_2 = '/content/all_combined_sequences.fasta'

combine_fasta_files(folders_1, output_file_1)
combine_fasta_files(folders_2, output_file_2)
```

Phylogeny was reconsted using MEGA software: Neighbor-joining method, with bootstrap to test phylogeny (No. of replications equals 500), Jukes-Cantor substitution model, other settings were default. Manually resized original trees are presented, other settings are default. Pairwise distances were calculated using MEGA, with the same settings as for phylogenetic trees. The output was converted into column-type .xls files. Necessary sample pairs were filtered using Excel.

Results

1. The phylogenetic tree was constructed using human mtDNA data:
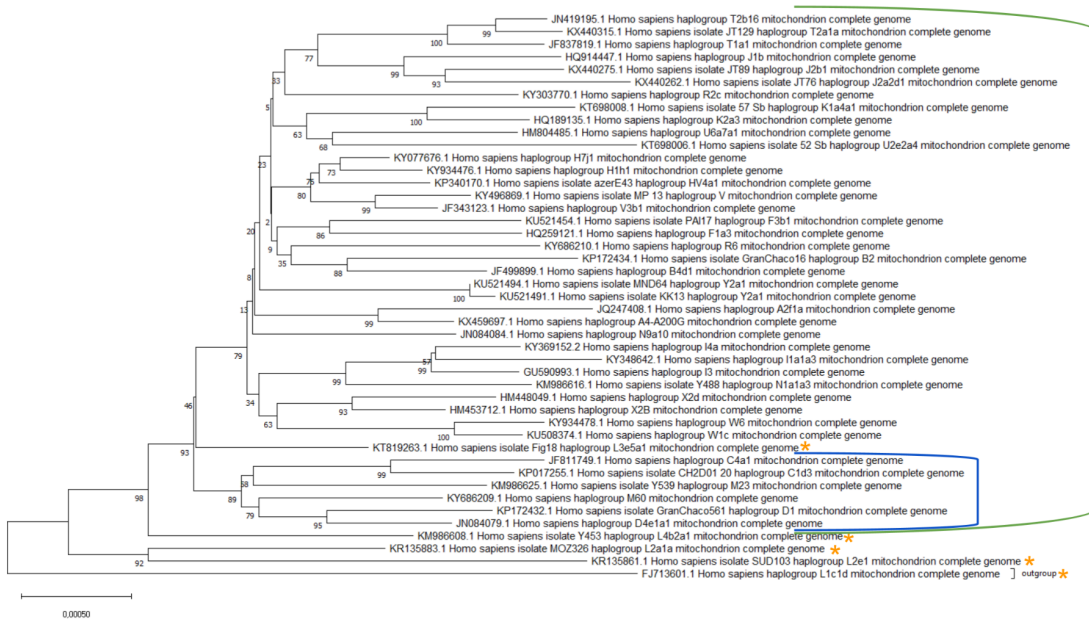
Asterixes (*) mark the mtDNA of African (L-) haplotypes. The subgroup whose recent common ancestor is probably the most recent ancestor of all non-Africans is highlighted in green.

2. Based on a strict molecular clock model, the time since divergence (T) between two species can be calculated as:

$$T = \frac{number\ of\ molecular\ differences\ (mutations)\ between\ species}{rate\ of\ molecular\ change\ (mutations\ per\ unit\ time)}$$

To evaluate the age of the mitochondrial Eve, we can calculate T between two less related human mtDNA samples. The number of *mutations per site (ms)* was calculated as a distance between the FJ713601.1 (haplogroup L1c1d, Central African) and KX440262.1 (haplogroup J2a2d1, Tunisian), which was the highest pairwise distance. The MtDNA evolutionary rate of 1.91×10^{-8} *mutations per site per year (msy)* was chosen (*Cabrera V. M., Scientific Reports, 2021*). Based on our samples, the mtEve age is approximately equal to:

$$T_{mtEve} = \frac{0.005818}{1.91 \times 10^{-8}} \approx 304,618\ years$$

This is aligned with recent data on the oldest Homo sapiens Remains from Morocco dated to 315,000 years ago (*Hublin JJ. et al., Nature, 2017*). If we choose the mutation rate calculated for mitogenomes of the most recent *period* (4.33×10^{-8} msy), the mtEve age is approximately 135,307 years.

To calculate the age of the most recent ancestor of all non-Africans, we can calculate T between two less related non-African samples (non-L haplogroups).

The number of mutations per site is calculated as the maximum distance between Si and Sj, where Si is any sample within a blue subgroup (see the tree above) and Sj is any sample within a green subgroup except a blue subgroup and a sample L3e5a1. The distance between JF811749.1 (haplogroup C4a1, Turkish) and KT698006.1 (haplogroup U2e2a4, Serbian) was used. Since the migration from Africa is a much earlier event, a mutation rate of 4.33×10^{-8} msy was used:

$$T_{non-Africans} = \frac{0.003751}{4.33 \times 10^{-8}} \approx 86,628 \ years$$

3. The phylogenetic tree with Neanderthal and Denisovan samples:
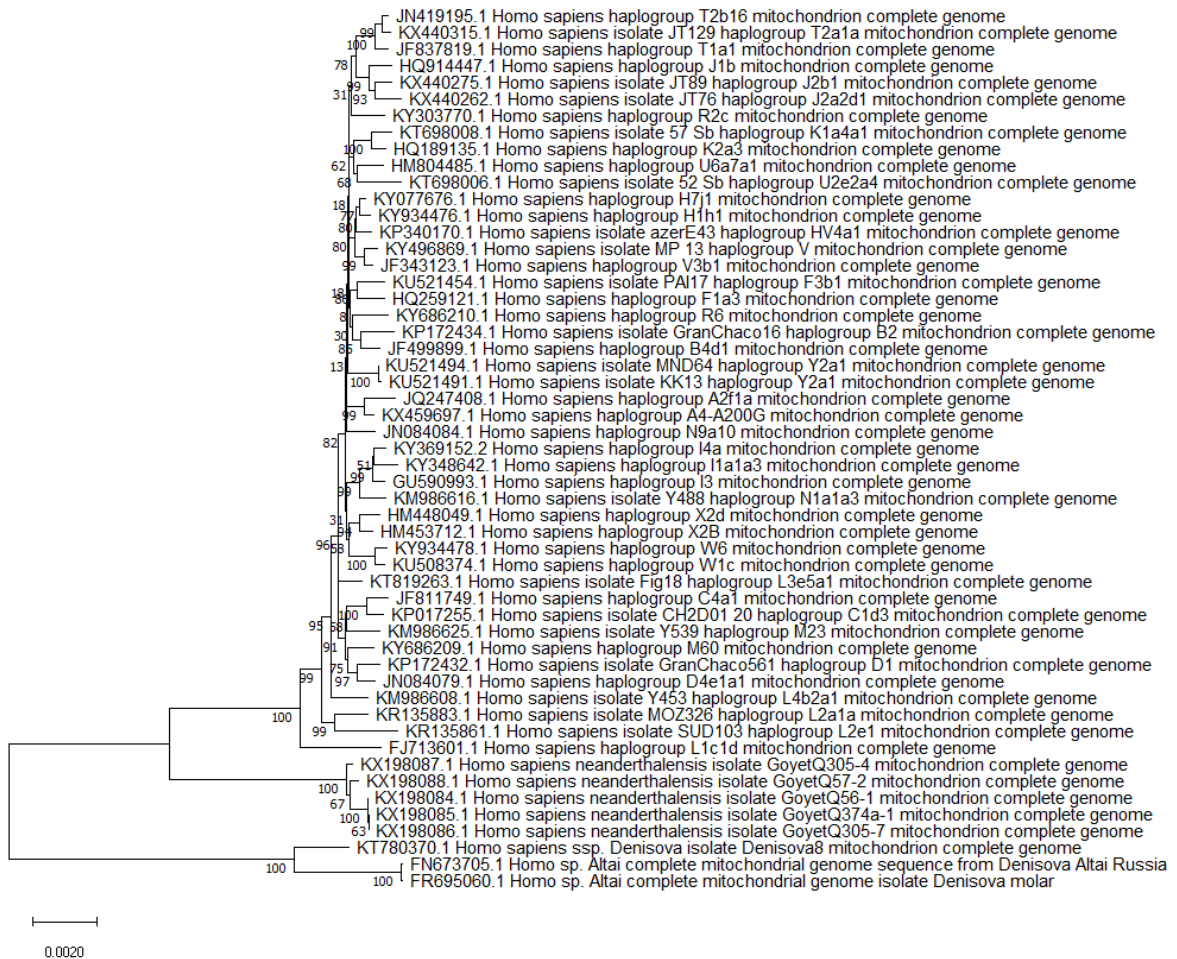
The age of the most recent Neanderthal-modern human ancestor is calculated using the maximum distance between any Neanderthal sample and any human sample (this is a distance for KX440262.1 (haplogroup J2a2d1, Tunisian) and KX198086.1 (neanderthalensis GoyetQ305-7):

$$T_{Neanderthal-Human} = \frac{0.0134057272}{1.91 \times 10^{-8}} \approx 701,870 \ years$$