Investigating the phylogenetic relationships between Neanderthals, modern humans, and a "mystery sequence"

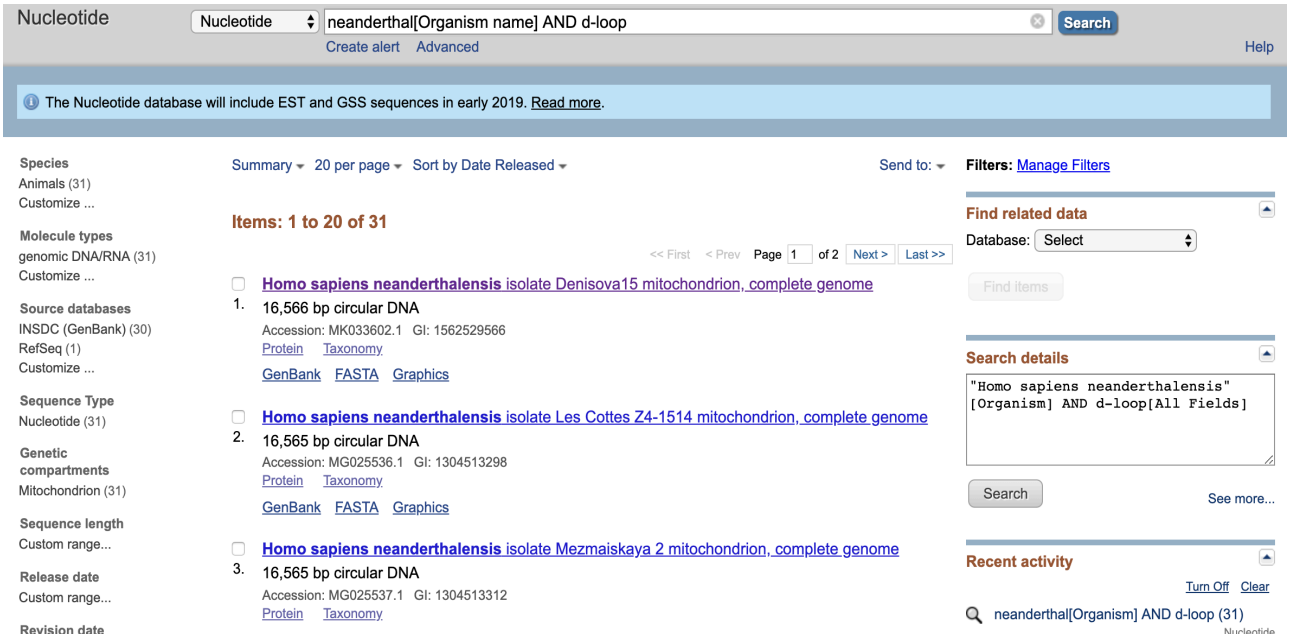
Ensure you answer each of the italicized points
(space is provided)

Its recommended you use the following resources:
Clustalx (to be downloaded and installed – for Linux, Mac OSX, or Windows): To perform multiple sequence alignments and some Neighbor Joining phylogenetic tree construction. http://www.clustal.org/.
Note there are other better alignment methods, like Muscle, that you should also be aware of.
iTOL (on the web): To visualize a tree. http://itol.embl.de/
NCBI nucleotide (on the web): To obtain sequences. http://www.ncbi.nlm.nih.gov/nucleotide

First, we need to choose a marker gene. We will choose the hypervariable regions (acronym: HVR or HPV – not human papilloma virus!) in human and Neanderthal sequences from the D-loop in the mitochondrial genome.
The mitochondrial genome is circular, but its presented in the linear format. Therefore, HVR-I is approximately at positions 16020-16070, and HVR-II is at approximately 1-600.
To fetch the marker genes, go to NCBI nucleotide. Use the following search: neanderthal[Organism name] AND d-loop

How many entries are listed in total? (watch out – they may be listed over multiple pages)

Number of entries: **31 entries**



Click on a few of the sequences and examine the text you see (in GenBank format).
Click on the "D-loop" part in the annotations, see which regions "light up", particularly for the larger sequences. Note that if obtaining any sequences for such analysis, D-loop sequences should be obtained in FASTA format as that is the format used by Clustalx (and many other multiple sequence alignment programs). Obtain the D-loop region from one of the most recent sequences (sort by date released).

Note the accession number of the sequence you included, and the D-loop region sequence, here in the text box, in FASTA format:

Sequence, in FASTA format: join (1-574 and 16021-16066)
(this will cover multiple lines of text in the text box)

```
>MK033602.1:1-574,16021-16566 Homo sapiens neanderthalensis isolate
Denisova15 mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTGTGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAGCATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACGATTAAATGTCTGCACAGCCGCTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCCTCCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAA
AAACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATAC
ACTTTTAACAGTCACCCCCTAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTA
ATCTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACCGCTGCTAACCCCATACC
CCGAGCCAACCAAACCCCAAAGACACCCCCCACATTCTTTCATGGGGGAGCAGATTTGGG
TACCACCCAAGTATTGACTCACCCATCAACAACCCTATGTATTTCGTACATTACTGCCAG
CCACCATGAATATTGTACAGTACCATAATTACTTGACTACCTGTAATACATAAAAACCTA
ATCCACATCAAACCCCCTCCCCATGCTTACAAGCAAGCACAGCAATCAACCTTCAACTGC
ATACATCAACCACAACTCCAAAGACACCCTTACACCCACTAGGATATCAACAAACCTACC
CACCCTCGACAGTACATAGCACATAAAGTCATTTACCGTACATAGCACATTATAGTCAAA
TCCCTTCTCGCCCCCATGGATGACCCCCCTCAGATAGGGGTCCCTTGATCACCATCCTCC
GTGAAATCAATATCCCGCACAAGAGTGCTACTCTCCTCGCTCCGGGCCCATAACACTTGG
GGGTAGCTAAAGTGAACTGTATCCGACATCTGGTTCCTACTTCAGGGTCATAAAGCCTAA
ATAGCCCACACGTTCCCCTTAAATAAGACATCACGATG
```

Add this sequence above to the set of sequences below, conveniently provided on the next page to save you time. ☺

```
>nean1 gi|111146900|gb|DQ859014.1| Homo sapiens neanderthalensis from Spain control region, partial
sequence; mitochondrial
AGCAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTACAGTACCATAATTACTTGACTACCTGCAGTACATAAAAACCTAATCCACACCAA
CCCCCCCCCCCCCATGCTTACAAGCAAGCACAGCAATCAACCTTCAACTGTCATACATCAACTACAACTCCAAAGACGCCCTTACACCCACTAGGATATCAACAAA
CCTACCCACCCTTGACAGTACATAGCACATAAAGTCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGCCCCCATGGATGACCCC
>nean2 gi|111035029|gb|DQ836132.1| Homo sapiens neanderthalensis from Italy control region, partial
sequence; mitochondrial
TTCTTTCATGGGGGAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTAC
AGTACCATAATTACTTGACTACCTGTAGTACATAAAAACCTAATCCACATCAACCCCCCCCCCCCATGCTTACAAGCAAGCACAGCAATCAACCTTCAACTGTCA
TACATCAACTACAACTCCAAAGACGCCCTTACACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGCACATAAAGCCATTTACCGTACATAGC
ACATTACAGTCAAATCCCTTCTCATCCCCATAGATGACCCCCCTCAAATAGGGGTCCCTTGAT
```

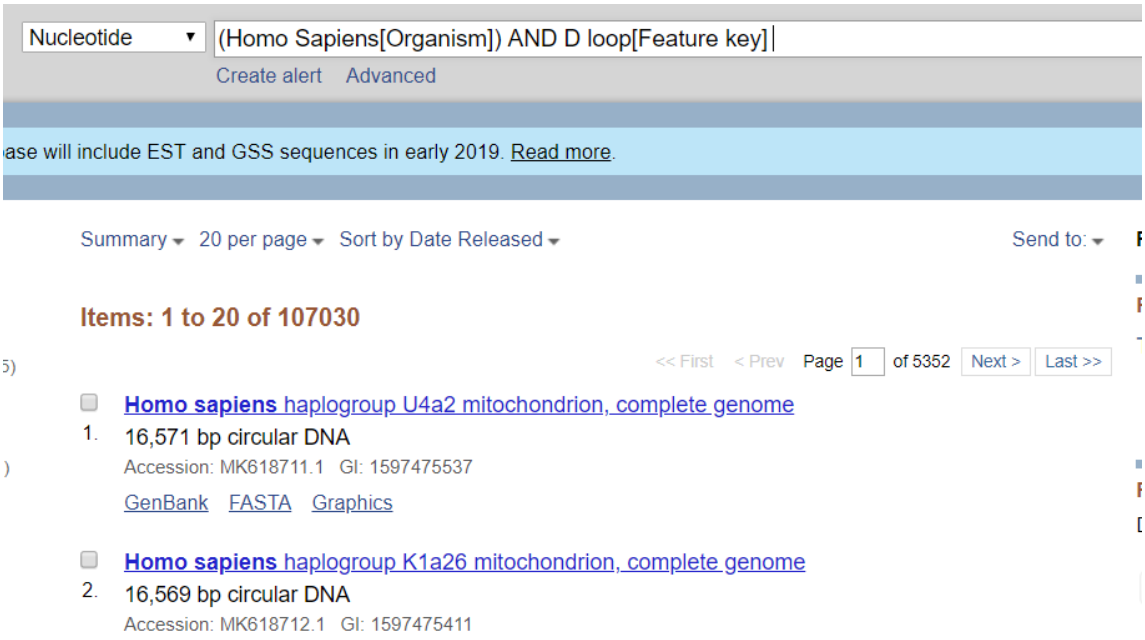
```
>nean3 gi|28557455|gb|AY149291.1| Homo sapiens neanderthalsensis mitochondrial D-loop hypervariable
region I, partial sequence
GTTCTTTCATGGGGGAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAGCAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTA
CAGTACCATAATTACTTGACTACCTGCAGTACATAAAAACCTAATCCACATCAACCCCCCCCCCCCATGCTTACAAGCAAGCACAGCAATCAACCTTCAACTGTC
ATACATCAACTACAACTCCAAAGACACCCTTACACCCACTAGGATATCAACAAACCTACCCACCCTTGACAGTACATAGCACATAAAGTCATTTACCGTACATAG
CACATTACAGTCAAATCCCTTCTCGCCCCCATGGATGACCCC
>nean4 gi|11141612|gb|AF282971.1|AF282971 Homo sapiens neanderthalensis mitochondrial hypervariable
region I sequence
GTTCTTTCATGGGGGAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAGCAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTA
CAGTACCATAATTACTTGACTACCTGCAGTACATAAAAACCTAATCCACATCAACCCCCCCCCCCCATGCTTACAAGCAAGCACAGCAATCAACCTTCAACTGTC
ATACATCAACTACAACTCCAAAGACGCCCTTACACCCACTAGGATATCAACAAACCTACCCACCCTTGACAGTACATAGCACATAAAGTCATTTACCGTACATAG
CACATTACAGTCAAATCCCTTCTCGCCCCCATGGATGACCCC
```

Next, we need some modern human sequences. This time, go the NCBI nucleotide site, but click on "Advanced". Here you can build your query using words and phrases in different fields**. Use "Homo sapiens" (in which field?) and D-loop (in the "Feature Key" field) What was the query you used (including what field for the Homo sapiens search)?**

Query: (Homo Sapiens[Organism]) AND d-loop[Feature key]

How many sequences are retrieved now?

Number of sequences: 107030



Clearly, you need to narrow down your search….

To save you the time, I'll provide you with some human sequences!
(listed over two pages)

>gi|38885248|gb|AY390635.1| Homo sapiens clone thetaA4 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCCCCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGTCC
>gi|38885246|gb|AY390634.1| Homo sapiens clone alphaD3 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCCCCTCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGTCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGTCCCT
>gi|38885245|gb|AY390633.1| Homo sapiens clone alphaC3 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCCCCTCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGTCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGTCC
>gi|38885244|gb|AY390632.1| Homo sapiens clone ZC7 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCCCCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGTCC
>gi|38885243|gb|AY390631.1| Homo sapiens clone XB3 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCCCCTCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGTCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCC
>gi|38885242|gb|AY390630.1| Homo sapiens clone WR3 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCCCCTCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGTCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGG
>gi|38885240|gb|AY390628.1| Homo sapiens clone WJ4 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCTCCTCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCA
CCCCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGTCC
>gi|38885241|gb|AY390629.1| Homo sapiens clone WQ1 control region, partial sequence; mitochondrial
CCCCCCCCCCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGCCACCCCTCACCCACTAGGATACCAACAAACCT
ACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGCCCCCATGGATGACCCCCCTCAGATAGGGG
>gi|38885239|gb|AY390627.1| Homo sapiens clone WH3 control region, partial sequence; mitochondrial
TAGTACATAAAAACCCAATCCACATCAAAACCCCCTCCCCATGCTTACAAGCAAGTACAGCAATCAACCCTCAACTATCACACATCAACTGCAACTCCAAAGTCA
CCCCTCACCCATTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCACATTACAGTCAAATCCCTTCTCGTC
CCCATGGATGACCCCCCTCAGATAGGGGT

You now need an outgroup to root the tree. *What is a suitable species that you should choose?*

Species: We can choose any species of primates which diverged from humans and Neanderthals in early time *Pan troglodytes*

How would you phrase your NCBI nucleotide query to retrieve the sequence?

Query: (Pan troglodytes[Organism]) AND d-loop[Feature key]

Sequence retrieved:
>Pan trogolydytes (OutgroupRoot) schweinfurthii isolate TSWR_1TSWR D-loop, partial sequence; mitochondrial
TACATTACTGCCAGCCACCATGAATATTGTACAGTACTATAATCACTCAACTACCTATAATACATCAAAC
CCACCCCACATTACAACCTCCACCCTATGCTTACAAGCACGCACAACAATTAACCCTCAACTGTCACACA
TAAAACACAACTCCAAAGACATTCCTCCCCCACCCCGATACCAACAGACCTATACTCTCTTAACAGTACA
TAGTACATACAACCGTACACCATACATAGCACATTACAGTCAAATCCATCCTCGCCCCCACGGATGCCCC
CCCTCAGATAGGGG

Now include also in your FASTA file of sequences the following *mystery sequence* from Papua New Guinea.

>mystery sequence
TTCTTTCATGGGGAAGCAGATTTGGGTACCACCCAAGTATTGACTCACCCATCAACAACCGCTATGTATTTCGTACATTACTGCCAGCCACCATGAATATTGTAC
GGTACCATAAATACTTGACCACCTGTAGTACATAAAAACCCAATCCACATCAACCCCCCCCCCCCCATGCTTACAAGCAAGTACAGCAACCAACCCTCAACTATCA

CACATCAACTGCAGCTCCAAAGCCACCTCTCACCCACTAGGATACCAACAAACCTACCCACCCTTAACAGTACATAGTACATAAAGCCATTTACCGTACATAGCA
CATTACAGTCAAATCCCTTCTCGTCCCCATGGATGACCCCCCTCAGATAGGGGT

AND

Also include the D-loop from the following sequence:
"Homo heidelbergensis mitochondrion, complete genome - NCBI Reference Sequence: NC_023100.1"

>NC_023100.1:1-576 Homo heidelbergensis mitochondrion, complete genome
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTTTGGGGG
GTGTGCACGCGATAGCATTGCGAAACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTC
CTGCCCCATCTCATTATTTATCGCACCTACGTTCAATATTACAGGCGAGCATANNTACTAAAGTGTGTTA
ATTAATTAATGCTTGTAGGACATAATAATAACNATTNAATGNCTGCACAGCCGCTTTCCACACAGACATC
ACAACNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNGCTTCTGGCCACAGCACTTAAACACATCTCTGCCA
AACCCCAAAAACAAAGAACCCTAACACCAGCCTAGCCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAATCTCATTAATA
CAACCCCCGCCCATNCTACCCANNACACACACACCGCTGCTAACTCCATACCCCGAACCAACCAAACCCC
AAAGACACCCCCCANA

What bioinformatics search tool could you use to find the portion of this sequence that would be best to align, if the sequence was much longer than the other sequences*?*

Bioinformatics tool: We can use Blastn and run it against the GenBank nr/nt database if the sequence is not very long. An alternative to Blastn could be Magic-Blast which uses innovative techniques that include the optimization of a spliced alignment score and selective masking during seed selection. Also we could any other pipeline method where the core of the searching algorithm is RapSearch2. Improved algorithms and preset pipelines are the way to go for longer reads.

Why could a really long sequence be a problem?

Why: The time for returning a query would increase exponentially overtime if the sequence read is very long. It might take days to get our results if we are comparing whole sequence read to older submitted genomes or other references. A lot of K-mers will be used to find similar sequences and it would take a very long time and there is no guarantee of getting a good result.

Take the human, Neanderthal and outgroup sequences, plus the mystery sequence and heidelbergensis sequence D-loop. Group them in a single multi-FASTA file. Use that as input for Clustalx.

Generate a multiple sequence alignment in Clustalx. Examine the alignment. Remember that for tree-drawing methods like Neighbor Joining, provided in Clustalx, the only part of the multiple sequence alignment used is the columns of the aligned nucleotides where there are no gaps in any of the sequences.
Now generate a NJ tree.
Upload tree files to iTOL. Root the tree using the outgroup species. Consider adding some more primate HVR-1 sequences to your data. Generate a new multiple alignment and unrooted tree. Visualize it using iTOL.

Provide your phylogenetic trees below (you may use up to three pages) and add a paragraph addressing the following questions (you may answer each question in turn if you wish below each question):

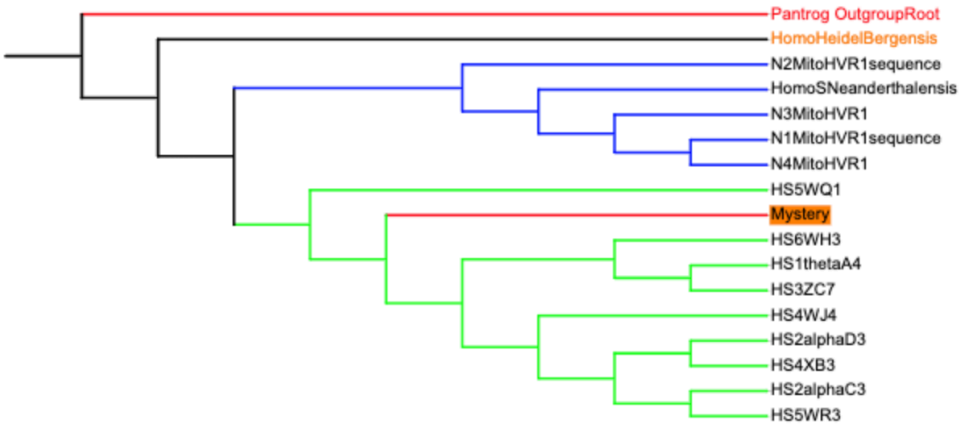


Figure 1: Denotes a rooted evolutionary phylogenetic tree using Mitochondrion HVR-1 sequences of Homo Sapiens (Green Clade), Neanderthals (Blue Clade) and an outgroup Pan Troglodytes (red node). Orange node denotes Homo Heidal Bergensis(orange). Mystery sequence is highlighted at orange and Node color is red. Rooted at Pan Troglodytes(outgroup).

Figure 2: Denotes an unrooted evolutionary phylogenetic tree using Mitochondrion HVR-1 sequences of Homo Sapiens (Green Clade), Neanderthals (Blue Clade) and an outgroup Pan Troglodytes (red node). Orange node denotes Homo Heidal Bergensis(orange). Mystery sequence is highlighted at orange and Node color is red.

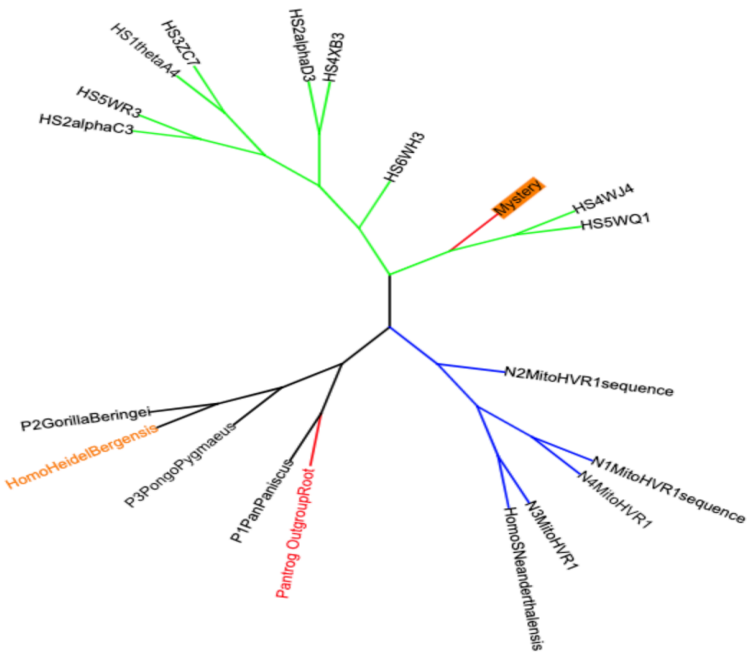


Figure 3: Denotes an unrooted evolutionary phylogenetic tree using Mitochondrion HVR-1 sequences of Homo Sapiens (Green Clade), Neanderthals (Blue Clade) and an outgroup Pan Troglodytes (red node). Orange node denotes Homo Heidal Bergensis(orange). Mystery sequence is highlighted at orange and Node color is red. Primates are added such as Gorilla berinei , Pongo Pygmaeus and Pan Paniscus.

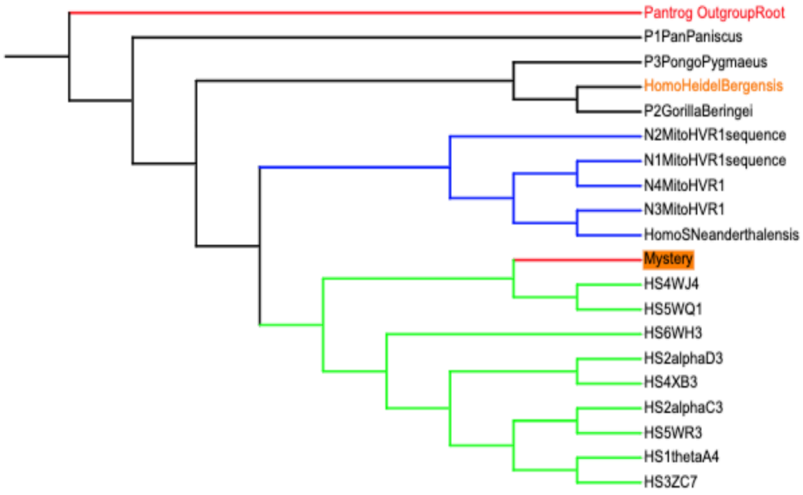


Figure 4: Denotes an rooted evolutionary phylogenetic tree using Mitochondrion HVR-1 sequences of Homo Sapiens (Green Clade), Neanderthals (Blue Clade) and an outgroup Pan Troglodytes (red node). Orange node denotes Homo Heidal Bergensis(orange). Mystery sequence is highlighted at orange and Node color is red. HVR -1 mitochondrion sequences od Primates are added such as Gorilla berinei , Pongo Pygmaeus and Pan Paniscus.

Based on these sequences, are modern humans and Neanderthals the same species or different species? Is there a clear cut answer? Why or why not? What else could you do to investigate this?

Based on the sequences above, and phylogenetic trees generated humans and Neanderthals are different species who have common ancestor. Both the clades are distinctly visible and we can see that they did diverge from each but are not the same species. Most of the studies have focused on the mitochondrial D-loop region, the most variable part of mtDNA due to increased substitution rate than in the rest of the mtDNA genome which serves as a better genetic marker to assess the diversity (3). Since, the control region of the mitochondrial DNA (mtDNA) due to its elevated mutation rate, lack of recombination and maternal inheritance serve as a biomarker in phylogenetic studies would always differentiate between species of same genus (3). For further investigations, we could use some not as variable regions as the mitochondrion to see how similar our results get and we can also check for Single Nucleotide variable region in their genomes to verify our results. Then we would be more sure about the evolutionary history between human and Neanderthals.

For the heidelbergensis sequence, what possible reason(s) are there (ideally with evidence supporting your reason(s) that this sequence is located in that position in your tree?

As our tree suggests that they are quite next to the out group Pan Troglodytes, this means that Homo Heidelbergensis must have diverged before the ancestor of Neanderthals and Homo sapiens diverged from each other(fig. 1 andn fig. 2) and also when we add more primates it stays in that region and more closely related to Gorilla Beringie (black clade). Therefore, there HVR1 regions are more closely related to other still alive primates like gorilla, pongo pygmaeus and pan panicus which also diverged from human ancestor long time ago. It must be a pre – Neanderthal or a pre – 'Homo' specie which diverged into the ancestor of Neanderthals and Modern Humans (2). Its fossil record also suggest that they might have lived around 700,000 – 200,000 years ago (2). Previous researchers do suggest they had a common ancestor which diverged into two species of humans.

What species is the mystery sequence likely from? Why is it located at that position in the tree?

The mystery sequence mostly likely from Homo sapiens as it falls in the green clade (figure 4 and 3 above). It might be a partial sequence of Hyper Variable regions 1 of the human mitochondrion with similar functions as the partial sequences derived from Homo sapiens clone WJ4 control region and the WQ1 specifically that is why it does show a very close relationship to them than rest of the sequences of Homo Sapiens HVR-I control region when we add multiple primates to verify its results. D-loop HVR is a very reliable for analysis of interspecies phylogenetic relationship.

References:

1. Grzegorz M Boratyn, Jean Thierry-Mieg*, Danielle Thierry-Mieg, Ben Busby and Thomas L Madden* *Magic-BLAST, an accurate DNA and RNA-seq aligner for long and short reads*. National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, MD, 20894, USA

2. Perner, Josef, and Frank Esken. 2015. "Evolution of Human Cooperation in Homo Heidelbergensis: Teleology versus Mentalism." *Developmental Review*, Theories of development, 38 (December): 69–88. https://doi.org/10.1016/j.dr.2015.07.005.[98]

3. Gupta A, Bhardwaj A, Supriya, Sharma P, Pal Y, et al. (2015) Mitochondrial DNA- a Tool for Phylogenetic and Biodiversity Search in Equines. J Biodivers Endanger Species S1:006. doi:10.4172/2332-2543.S1-006