Huong Phan
BIOL/CS 383: Bioinformatics
Understanding mtDNA variation

1. What do you notice about the lengths (and colors) of the matches (bars) as you look from the top to the bottom? What does this mean?

The length of every bar is 20, meaning the query was able to find hits that have the same length as the length of each primer, which is 20 nucleotides each. The blue color of every bar means that the query was able to find all the hits that matches exactly to both forward primers (F) and reverse primers (R), which is represented by the score of 40 (20 nucleotides in F + 20 nucleotides in R = 40).

2. What is the E value of the most significant hit(s) and what does this mean?

E value of the most significant hits is 0.13, meaning in 100 random searches the sequence would be expected to turn up 13 matches by chance. Since E value is low, there is high probability that the hit is related to the query.

3. What is the source and size of the sequence in which your BLAST hit is located?

Hit: Homo sapiens isolate H3-11 mitochondrion, complete genome
Accession: AY495156.2
Source: mitochondrion Homo sapiens (human)
Size: 16569 bp

4. To which positions do the primers match in the subject sequence?

Hit: Homo sapiens isolate H3-11 mitochondrion, complete genome
Accession: AY495156.2
Positions: 15971, 15990

5. What is the difference between the coordinates?

19

6. Note that the actual length of the amplified fragment includes both ends, so add 1 nucleotide to the result that you obtained in Step 4.c. to obtain the exact length of the PCR product amplified by the two primers.

Exact length of the PCR product amplified by the two primers = 19 + 1 = 20

7. How many features does the amplicon span? What are they called? What is the size of each feature? Be sure to include all of the sequences of each feature in your calculations.

51 features, including 1 D-loop, 22 tRNAs, 2 rRNAs, 13 CDSs, 13 genes

Features - Positions - Length

D-loop - 16028..16569,1..578 - 542 nt + 578 nt

tRNA-Phe - 578..648 - 71 nt

12S ribosomal RNA - 649..1602 - 954 nt

tRNA-Val - 1603..1671 - 69 nt

16S ribosomal RNA - 1672..3229 - 1558

tRNA-Leu - 3230..3304 - 75 nt

Gene ND1 - 3307..4262 - 956 nt

NADH dehydrogenase subunit 1 - 3307..4262 - 956 nt

tRNA-Ile - 4263..4331 - 69 nt

tRNA-Gln - 4329..4400 - 72 nt

tRNA-Met - 4402..4469 - 68 nt

Gene ND2 - 4470..5511 - 1042 nt

NADH dehydrogenase subunit 2 - 4470..5511 - 1042 nt

tRNA-Trp - 5512..5579 - 68 nt

tRNA-Ala - 5587..5655 - 69 nt

tRNA-Asn - 5657..5729 - 73 nt

tRNA-Cys - 5761..5826 - 76 nt

tRNA-Tyr - 5826..5891 - 86 nt

Gene COX1 - 5904..7445 - 1545 nt

cytochrome c oxidase subunit I - 5904..7445 - 1545 nt

tRNA-Ser - 7445..7516 - 72 nt

tRNA-Asp - 7518..7585 - 68 nt

Gene COX2 - 7586..8269 - 683 nt

cytochrome c oxidase subunit II - 7586..8269 - 683 nt

tRNA-Lys - 8295..8364 - 70 nt

Gene ATP8 - 8366..8572 - 207 nt

ATP synthase F0 subunit 8 - 8366..8572 - 207 nt

Gene ATP6 - 8527..9207 - 681 nt

Gene COX3 - 9207..9987 - 781 nt

cytochrome c oxidase subunit III - 9207..9987 - 781 nt

tRNA-Gly - 9991..10058 - 68 nt

Gene ND3 - 10059..10404 - 346 nt

NADH dehydrogenase subunit 3 - 10059..10404 - 346 nt
tRNA-Arg - 10405..10469 - 65 nt
Gene ND4L - 10470..10766 - 297 nt
NADH dehydrogenase subunit 4L - 10470..10766 - 297 nt
Gene ND4 - 10760..12137 - 1378 nt
NADH dehydrogenase subunit 4 - 10760..12137 - 1378 nt
tRNA-His - 12138..12206 - 69 nt
tRNA-Ser - 12207..12265 - 59 nt
tRNA-Leu - 12266..12336 - 71 nt
Gene ND5 - 12337..14148 - 1812 nt
NADH dehydrogenase subunit 5 - 12337..14148 - 1812 nt
Gene ND6 - 14149..14673 - 525 nt
NADH dehydrogenase subunit 6 - 14149..14673 - 525 nt
tRNA-Glu - 14674..14742 - 69 nt
Gene CYTB - 14747..15881 - 1135 nt
cytochrome b - 14747..15881 - 1135 nt
tRNA-Thr - 15888..15953 - 66 nt
tRNA-Pro - 15955..16023 - 69 nt

8. The largest feature that you identified in Step 3.a. has several names. Do some research to learn and understand these names.

ND5, the longest sequence has other names such as mitochondrially encoded NADH dehydrogenase 5, MTND5, NADH dehydrogenase subunit 5, NADH-ubiquinone oxidoreductase chain 5, NADH-ubiquinone oxidoreductase, subunit ND5, NADH5, NU5M_HUMAN. According to NIH, “the *MT-ND5* gene provides instructions for making a protein called NADH dehydrogenase 5. This protein is part of a large enzyme complex known as complex I, which is active in mitochondria.”[1], which explains why this gene is called MT-ND5 and NADH dehydrogenase 5.

9. List the genes that are transcribed from the forward strand and those that are transcribed from the reverse strand. On which strand are the most genes found?

Forward strand - tRNA-Phe, 12S ribosomal RNA, tRNA-Val, 16S ribosomal RNA, tRNA-Leu, gene ND1, NADH dehydrogenase subunit 1, tRNA-Ile, tRNA-Met, gene ND2, NADH dehydrogenase subunit 2, tRNA-Trp, gene COX1, cytochrome c oxidase subunit I, tRNA-Asp,

[1] https://ghr.nlm.nih.gov/gene/MT-ND5#

gene COX2, cytochrome c oxidase subunit II, tRNA-Lys, gene ATP8, ATP synthase F0 subunit 8, gene ATP6, ATP synthase F0 subunit 6, gene COX3, cytochrome c oxidase subunit III, tRNA-Gly, gene ND3, NADH dehydrogenase subunit 3, tRNA-Arg, gene ND4L, NADH dehydrogenase subunit 4L, gene ND4, NADH dehydrogenase subunit 4, tRNA-His, tRNA-Ser, tRNA-Leu, gene ND5, NADH dehydrogenase subunit 5, gene CYTB, cytochrome b, tRNA-Thr.

Backward strand - D-loop, tRNA-Gln, tRNA-Ala, tRNA-Asn, tRNA-Cys, tRNA-Tyr, tRNA-Ser, gene ND6, NADH dehydrogenase subunit 6, tRNA-Glu, tRNA-Pro.

10. What kind of product do most of the genes produce? For what biological function are these products used?

Transport proteins that are part of respiratory chains, acting like enzymes that bind to specific molecules for catalysis of chemical reactions and electron transport for respiration and ATP production.

11. What is the spacing like between two adjacent genes? Is there much intergenic DNA between two adjacent genes?

Some overlaps, some picks up right where the previous gene stops. There is not much intergenic DNA between two adjacent genes.

12. What do you notice? What does this tell you about the structure and origin of mitochondrial genes?

The nucleotide coordinates for the gene and the coding sequence are the same. That means all DNA parts are genes, that encodes for a protein without any non-coding part, meaning that mitochondrial genes are a compact structure and do not contain non-coding DNA. It originates from a single copy of mitochondrial gene of our mother.

13. How would you interpret a lane in which you observe primer dimer but no bands, as described in Step 3?

The lane contains the marker.

14. Propose a biological reason for the high mutation rate of mtDNA.

The process of mtDNA replication is random and not as accurate as the process of nucleus DNA replication because of an imbalance in dNTPs, making mtDNA replication much less accurate.[2]

[2] https://www.nature.com/scitable/topicpage/mtdna-and-mitochondrial-diseases-903

15. The high mutability of the mitochondrial genome means that it evolves more quickly than the nuclear genome. This makes the mitochondrial control region a laboratory for the study of the DNA evolution. However, can you think of any drawbacks to this high mutation rate when studying evolution?

Mutation is central to evolution, in which it creates genetic variations that either increases or decreases one's ability to adapt to the changing environment. High mutation rate means that there are a lot of different mutations created at the same time, so it would be difficult to examine all the mutations and categorize them to either increase or decrease one's fitness.

15. Would you expect any difference in the mutations of the control region sequence in the mitochondrial genome versus the chromosome 11 insertion? What implication does this have for the study of human evolution?

Mutation rate of the control region sequence in the mitochondrial genome is likely to be faster than that in the chromosome 11 insertion because it remains in the mitochondria and when it's replicated, the process is likely to be random and disorganized so a lot of copying errors will happen, leading to a higher mutation rate. When control region sequence is added to chromosome containing nucleus DNA, the replication process in DNA will result in fewer copying errors due to the presence of DNA polymerase. We can compare mutation rate of the control region sequence in two different situations to study human evolution.

16. What does the amplitude of each peak represent? What do you notice about the amplitude of the peaks at the beginning of the sequence?

The amplitude of each peak represent the strength of the flourescent signal for each of A, T, C or G nucleotides. The amplitude of the peaks at the beginning of the sequence were not distinct enough to identify the correct nucleotides.

17. In what parts(s) of your sequence read are the most Ns found? What do you notice about the trace at N positions?

The beginning of the sequence. At N positions, there is either no distinct peak or there are too many different weak florescent signals from different source DNA so it's impossible to identify an appropriate nucleotide.

18. Do the nucleotide numbers (positions) correspond to the same nucleotide sequence in both reads? What trick did you use to align them?

I compared my sequence (1) with one of my classmates (7) and the nucleotide numbers. First, I deleted the N at the beginning of both our sequences, then I started to align both of our sequences. First, I try to copy 100 of the first nucleotides of my sequence and search that in the my classmates' sequence, and our 100 first nucleotides were identical. However, when I do the same for the next 100 nucleotides, the search gives me no result, and I began to look at each nucleotide and found that our sequence differ at the 115th, 148th, 158th, 191th (N for my classmate), 209th, 268th, 278th, 284th, 290th (N for my classmate) nucleotide (counted after deletion of the Ns at the beginning). My classmate sequence does not have any N at the end, but I have 2.

19. What sequence feature coincides with the abrupt transition from a clean to a poor-quality sequence?

In a poor-quality sequence, the background DNA signal was nearly as strong as the main DNA signal so it was impossible to determine which is the most prominent fluorescent signal, while in a clean sequence, background DNA gives a very weak signal and thus does not interfere with the main DNA signal.

21. Use these data to determine the average number of mutations, as well as the average percent difference, between pairs. What do these numbers tell us about human variation?

Calculated from row D3 to D23 in Excel sheet, average number of mutations: 5.67

Calculated from row D3 to D23 in Excel sheet, average number of total counts: 303

Calculated from row F3 to F23 in Excel sheet, average number of percent difference, between pairs: 1.89%

There is only an average difference of about 2 nucleotides per 100 nucleotides among modern human genome.

22. If this is so, what is the approximate mutation rate for the mitochondrial control region? Why is your calculation only approximate?

(1.89/100)/200000 = 9.45e-8 base pairs per year

Not enough data was collected, data is not representative of the whole world population since it is only within a class.

23. Pool the class data to determine the average number of differences and average percent difference between Otzi and modern humans. What does this tell you about Otzi and the DNA mutational clock?

Between Otzi and modern humans: average 3.75 differences (only from first 4 rows of data), average 355.5 total counts and average 1.12% difference (only from Khoa's and my data). Otzi and modern humans are quite close in our ancestry and our mutational clock is similar to each other.

24. Pool the class data to determine the average number of differences and average percent difference between Neandertals and modern humans. What does this tell you about the relationship between Neandertals and modern humans?
Between Neandertals and modern humans: average 22.4 differences (from first 10 rows of data), average 359.75 total counts and average 6.88% difference (only from Khoa's and my data).

25. If modern humans arose 200,000 years ago, approximately how long ago did humans and Neandertals share a common mitochondrial ancestor?
6.88/1.89 = 3.64
200,000*3.64 = 728,000 years
=> Approximately around 730,000 years ago humans and Neandertals share a common mitochondrial ancestor.

26. What does the level of genetic variation in Neandertals suggest about Neandertal population changes during the course of their existence on Earth?
They evolve a lot through both acts of survival and through interactions with other groups of species.

27. Denisovans: Pool the class data to determine the average number of differences average percent differences. Does your result suggest that the samples are modern human?
Between Denisovans and modern humans: From first 5 rows of data, average 33.8 differences, average 376 total counts and average 9.01% difference.

28. If modern humans arose 200,000 years ago, approximately how long ago did modern humans and the people from Denisova Cave share a common ancestor?
9.01/1.89 = 4.77
200,000*4.77 = 954,000 years
=> Approximately around 950,000 years ago humans and Neandertals share a common mitochondrial ancestor.

29. Denisovans vs. Neandertals: Pool the class data to determine the average number of differences and average percent differences. What does this suggest? Discuss.

Between Denisovans and Neandertals: average 30.3 differences (from first 6 rows of data), average 294.3 total counts (from last 3 rows of data), and average 10.45% difference (from last 3 rows of data).

30. Does the mitochondrial DNA tell the whole story? Explain.
Mitochondrial DNA does not tell the whole story, as it can only tell how long ago different groups of species share a common mitochondrial ancestor, but do not tell when our genomes diverge from each other or when different groups interbred with each other to create a fitter species.

31. Do a literature search and summarize what we now know about the relationship of these two extinct hominids and modern human populations.
Neandertals and modern humans coexisted in Europe, in which both groups have to compete for food and survival. It was speculated that interbreeding between two groups resulted from acts of violence rather than romance.[3] Neandertals were also found to have 2% of their genomes in modern humans.
Denisovans living in Eurasia interbred with humans from Papua New Guinea, which contributed between 3 to 5 percent of their genetic material to the genomes of Melanesians, a group of Pacific Islanders living in Papua New Guinea.[4]

[3] https://phys.org/news/2017-06-scientists-reconstructing-relationship-modern-humans.html

[4] https://genographic.nationalgeographic.com/denisovan/