Lab Exercise 4

1)

Find in RefSeq the transcript of the **hemoglobin subunit beta** in human. Download it as a FASTA file.

Translate the sequence into protein and write it as a FASTA file (hint: look for the CDS feature in RefSeq).

2)

A)

Download the entire human transcriptome from NCBI's FTP site (hint: you are looking for the compressed FASTA files *human.#.rna.fna.gz*).

B)

What is the longest human transcript?

C)

What is the GC content of that transcript?

D)

Find this transcript in RefSeq and download it in the genbank (.gb) format.

How many exons does this transcript have? Find the one with the highest GC content.

E)

Calculate the amino-acid frequencies in the protein translated from this transcript.

3)

A)

Download the sequence of the entire human chromosome 11 from UCSC. **Use version hg19.**

B)

The human **hemoglobin subunit beta** gene is located at chromosome 11. The GRCh37 coordinates of its exons on that chromosome are: *5248160..5248301, 5247807..5248029, 5246694..5246956* (**on the negative strand**).

Use the sequence of chromosome 11 and these coordinates in order to recover the exon sequences of this transcript. Can you recover the same sequence you downloaded in (1)?

C)

[**Optional**, **Challenging**]

Can you repeat (B) without loading the entire sequence of chromosome 11 into memory?

D)

Extract the intron sequences of the **hemoglobin subunit beta** transcript.

Compare the GC content of the exons with that of the introns.