Bioinformatics, BSC 4930/6932, Midterm 1 Name: William Holt  
Take-home practical

**Write your name at the top right of this page.**

This midterm is out of 100 points. 65 of those points were from the in-class exam sheet that you already completed. The remaining 35 are from this take-home, open-book practical exam distributed on Canvas that is due on **Tuesday, 20 February, at 5 pm**.

You are welcome to work together on this practical. However, **the work that you submit must be your own**. Submitting identical work will be considered plagiarism and you will receive zero points for this part of the exam.

This practical is largely based on coding in R, so will be similar to the homeworks that you have completed. As Bioinformaticians, we often re-use code, and you are welcome to do so here. I recognize that much of the code will be identical to both that of your classmates and to the code you’ve already written. That is fine. For every line (or couple of lines) of code, you must write a comment line describing what your code is doing. This is what must be your own work. See the answer key to homework 5 as an example. The “#” at the start of a line of code makes it a comment line.

For any questions in this practical, put your answers in a comment in your R script following the relevant code (so, write your answers in an R script, not here). Save this R script to your Bioinformatics GitHub page. Submit a link to your GitHub page in the Midterm 1 assignment on Canvas. Make sure that the GitHub repository is public (otherwise I can’t access it).

To make a GitHub repository public, go to your repository on the website, click ‘settings’, scroll to the bottom under ‘Danger Zone’ and click ‘Change visibility’, then ‘change to public’. There will be a few more steps to verify that you want to make this change.

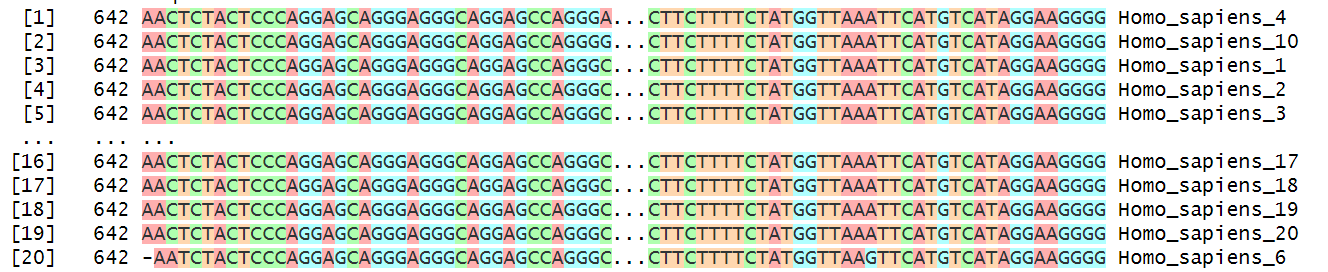
Push all your scripts, data files, and results plots to GitHub.

The questions in this practical are on the following page.

Take Home practical questions:

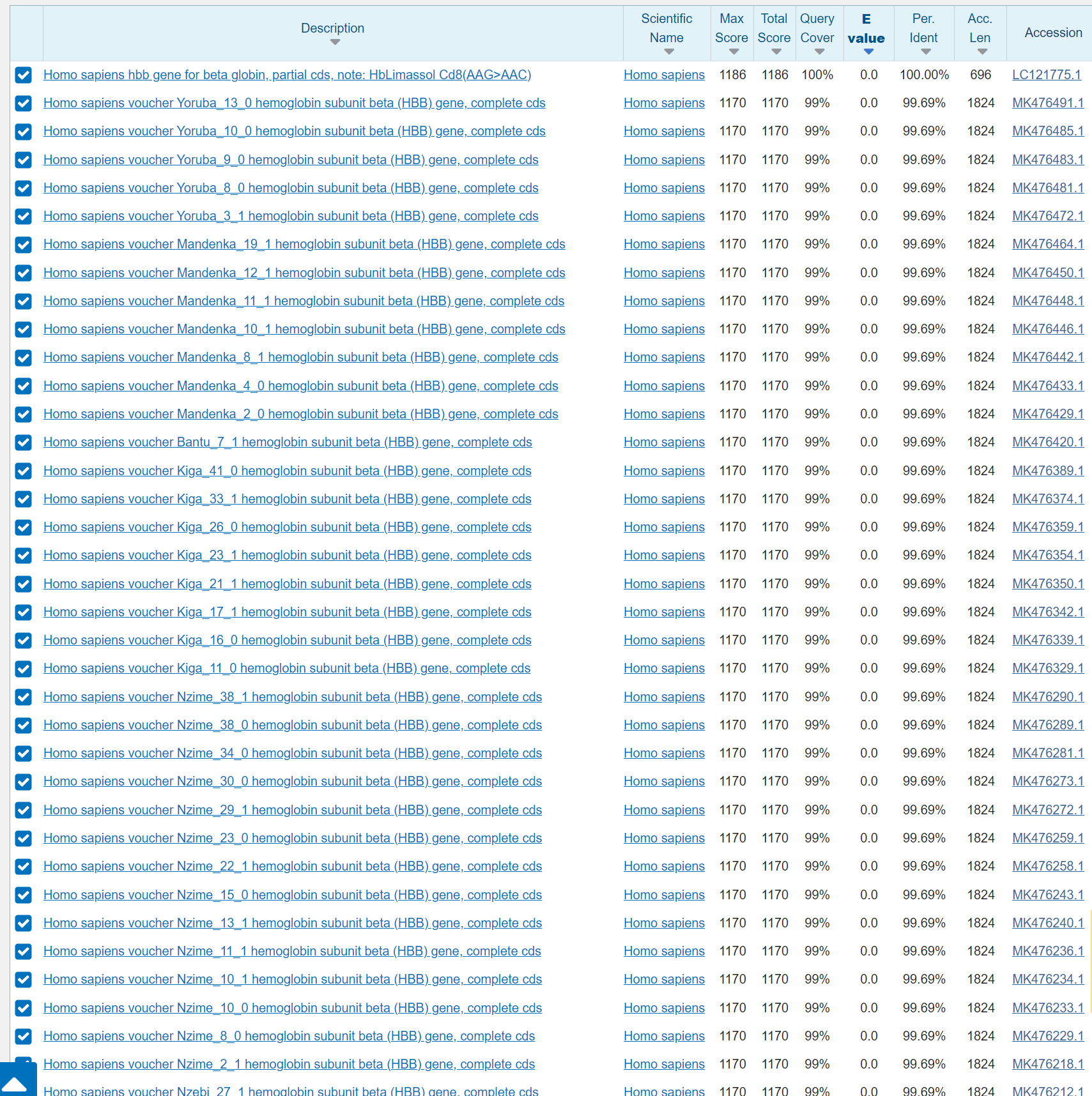
Your research lab is conducting a project that involves sequencing a gene in a population of people. You’ve sequenced this gene in 20 people so far, and you want to find out if there is any variation in this gene in your population. There is a file on Canvas containing all the DNA data from your sequencing efforts. Go ahead and download it to your Bioinformatics folder.

1. Import and align your DNA sequences
2. Check to see how different your samples are from one another. Are any of them different from the rest? If so, what kinds of mutations do you observe in this individual (or individuals)?



Sequences 4,10, and 6 have at least 1 mutated base. The bases are different than most of the other aligned sequences.

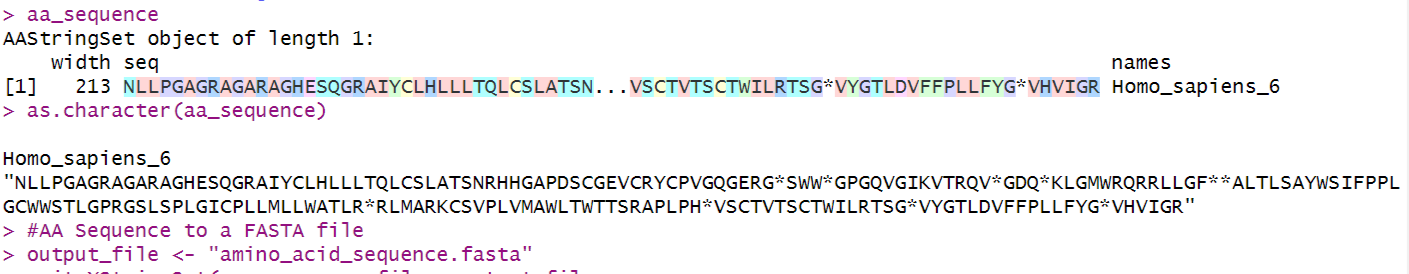
1. You suspect that an individual (or individuals) in this population might have some mutations in this gene, but you don’t know what this gene might be. Compare your sequences to a database to figure out what the gene is. Export your data, paste it into the relevant database search engine, and add your results to a comment line in R. What is the gene? What is the accession number of the best match to your search?



Using BLAST, the sequences best matched for the HBB gene in *Homo sapiens*. This makes sense because our participants are human. The best matched accession number is “LC121775.1”.

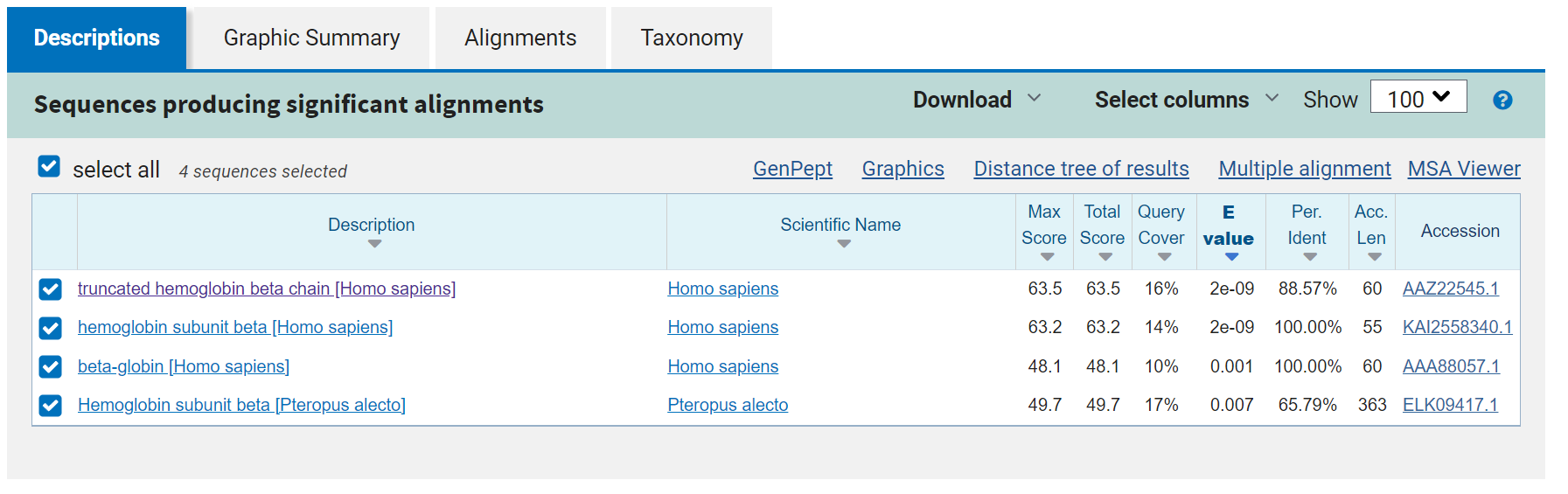
1. Find the individual that is the most different from the rest of the individuals in your dataset. Translate that sequence to protein. Write it to a fasta file.

Based on the previous MSA alignment, “Homo\_sapiens\_6” would probably be the most different from the rest. This is because of its 2 mutated bases that differ from the others and the gap at the beginning of the alignment.



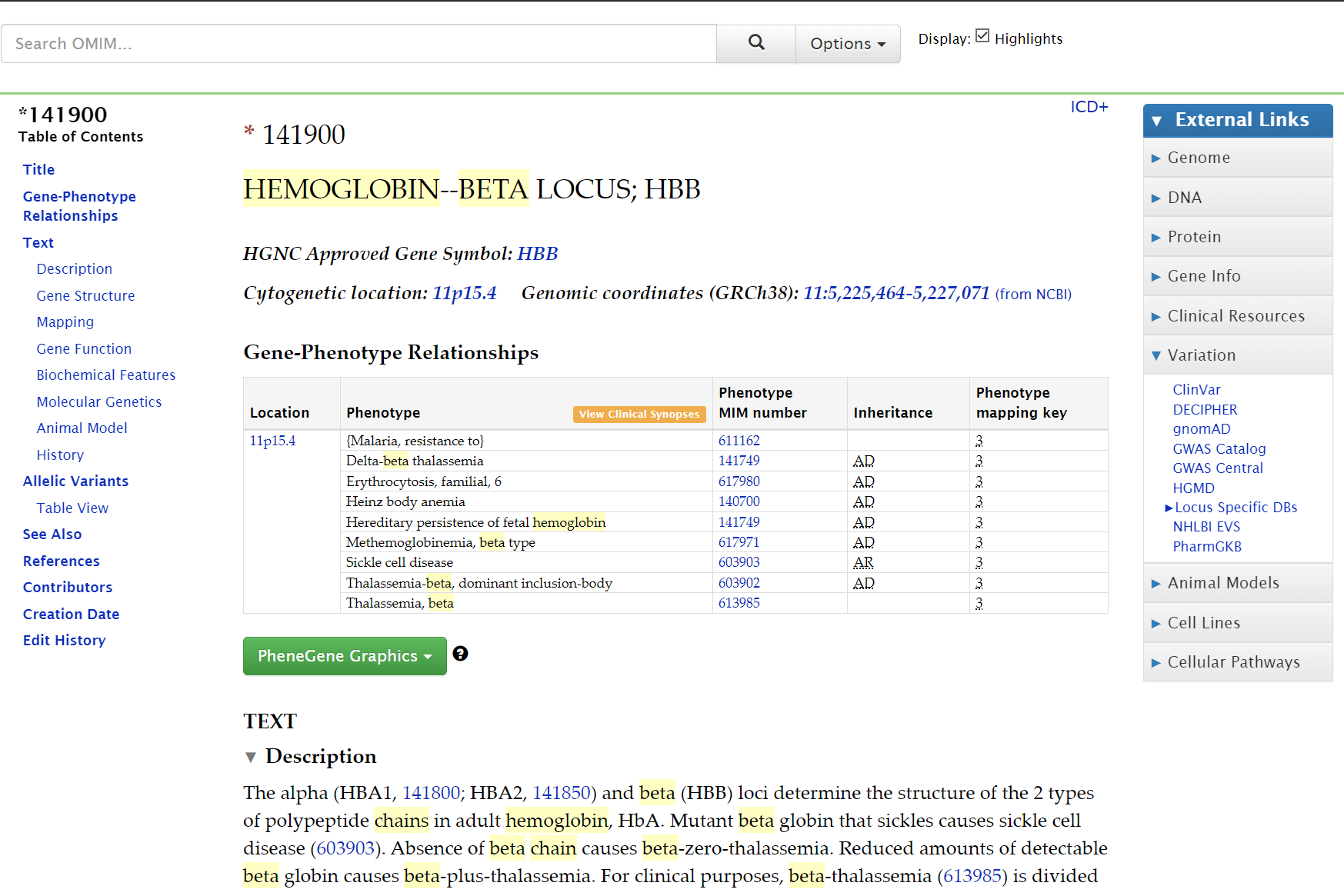
Fasta file is labeled as “amino\_acid\_sequence.fasta”

1. Use a database to figure out what your protein matches to. Click on the record for the best match. What is the accession number of this entry?



I performed a protein BLAST with the “Homo\_sapiens\_6” AA sequence. The entry with the highest total score is labeled as “Truncated hemoglobin beta chain”. Its accession number is “AAZ22545.1”.

1. Either using R or by searching in the database, what disease(s) is this gene associated with? Does this person have the disease?



Putting the accession numbers to try to find a related disease leads to errors in R. However, a search through the OMIM database relating to HBB gives results on multiple diseases associated with HBB mutation.

The person that best matches our AA sequence has a gene mutation causing a “truncated hemoglobin beta chain”, which is based on the BLAST from earlier. In this match, the person has a “-G” mutation, which causes a premature stop codon that leads to the truncation of the protein. The title associated with the BLAST entry was “Beta Zero: A Novel Beta-globin Gene Mutation Found in an Iranian Family”.

Looking for a premature stop codon in the OMIM entry leads to a disease called “Beta-Zero-Thalassemia”, which matches the title and the type of mutation from the BLAST entry. Therefore, based on the match and the database entries, it is likely that the “Homo\_sapiens\_6” person might have beta-zero-thalassemia.

1. What is the 3-dimensional structure of this protein? You can include a screenshot or download of a photo of this structure in your GitHub repository.

I used the AlphaFold website to find the 3-dimensional structure of the non-mutated HBB protein.

A blue ribbon on a white background

Description automatically generated