Hands-On Two
Investigating Inherited Diseases

The purpose of these exercises is to introduce bioinformatics databases and tools. We investigate an important human gene and see how mutations give rise to inherited diseases.

You will learn how to:

- Translate a DNA sequence into a sequence of amino acids
- Look for mutations in this sequence and determine how they cause inherited human diseases.

A) Translating a DNA sequence

Points to remember

- The genetic code is a triplet code, groups of 3 letters (bases) code for amino acids
- The coding sequence (CDS) of a gene generally begins with the start codon ATG. ATG encodes the amino acid methionine (MET, M)
- There are three stop codons (TAA, TGA, TAG) which indicate the end of proteins

Let us try this exercise by hand.

Consider the following DNA sequence:

>CDS of human beta globin

```
atggtgcacctgactcctgaggagaagtctgccgttactgccctgtggggcaaggtgaac
gtggatgaagttggtggtgaggccctgggcaggttgctggtggtctacccttggacccag
aggttctttgagtcctttggggatctgtccactcctgatgctgttatgggcaaccctaag
gtgaaggctcatggcaagaaagtgctcggtgcctttagtgatggcctggctcacctggac
aacctcaagggcacctttgccacactgagtgagctgcactgtgacaagctgcacgtggat
cctgagaacttcaggctcctgggcaacgtgctggtctgtgtgctggcccatcactttggc
aaagaattcaccccaccagtgcaggctgcctatcagaaagtggtggctggtgtggctaat
gccctggcccacaagtatcactaa
```

1) What is the nucleotide in position 20 in this sequence (HBB)? _____
2) Translate the first 27 bases of the sequence:

```
atg gtg cac ctg act cct gag gag aag
___ ___ ___ ___ ___ ___ ___ ___ ___
```

Computers are generally quicker and more accurate.

- Open HandsOn_Sequences.txt
- Copy onto the clipboard the sequence (only): “CDS of human beta globin”
- Go to the ExPasy translate tool at http://web.expasy.org/translate/
- Paste the sequence in the box
- Go to “Output format:” (under the box) and choose “Includes nucleotide sequence”
- Click the “TRANSLATE SEQUENCE” button.

What you should see:

- The DNA sequence with the amino acid sequence written out underneath it.
- Bioinformaticians mostly use the "single letter code", in which each amino acid is represented by a single letter.
- There are six possible translations because there are six possible reading frames depending on where you start reading the sequence.
- The first reading frame with a Methionine at the beginning and a dash (indicating a stop codon) at the end is the correct translation.

Let us save the sequence for later use:

- Click on the link "5'3' Frame 1"
- Now you should see only the amino acid sequence (without the DNA sequence).
- Copy the whole amino acid sequence to the clipboard.
- Replace the stop codon (Stop) by *.
- Name the file Protein_wildtype.txt. We will use it in the next part of the problem when we look at mutations.

B) Looking at the effects of mutations (Part One)

Mutations in the DNA sequence affect the resulting protein in different ways. In this exercise, you will translate the DNA sequence with a mutation found in individuals with the disease called Sickle Cell Anemia and compare the resulting protein with the one you found by translating the human beta globin gene, HBB (the sequence of part A).

Consider the following variant of the human beta globin gene:

>CDS of human beta globin mutant 1

```
atggtgcacctgactcctgtggagaagtctgccgttactgccctgtggggcaaggtgaac
gtggatgaagttggtggtgaggccctgggcaggttgctggtggtctacccttggacccag
aggttctttgagtcctttggggatctgtccactcctgatgctgttatgggcaaccctaag
gtgaaggctcatggcaagaaagtgctcggtgcctttagtgatggcctggctcacctggac
aacctcaagggcacctttgccacactgagtgagctgcactgtgacaagctgcacgtggat
cctgagaacttcaggctcctgggcaacgtgctggtctgtgtgctggcccatcactttggc
aaagaattcaccccaccagtgcaggctgcctatcagaaagtggtggctggtgtggctaat
gccctggcccacaagtatcactaa
```

3) What is the nucleotide in position 20 in this sequence? _____ .

In this case, a single nucleotide has been changed for another: an A changed into a T in position 20. Let us study the effects of this substitution.

4) Translate the first 27 bases of the mutant sequence:

```
atg gtg cac ctg act cct gtg gag aag
___ ___ ___ ___ ___ ___ ___ ___ ___
```

Now let us use the computer to translate the mutant sequence. We could go to the ExPasy translate tool at http://web.expasy.org/translate/ but instead, we will use translate.py that takes as input the CDS of mutant 1, copied from HandsOn_Sequences.txt and saved in CDS_mutant.txt.

- Run translate.py that takes as input file CDS_mutant.txt and outputs the translation of the mutant 1 CDS saved in Protein_mutant.txt.
- Open Protein_wildtype.txt (from part A) and Protein_mutant.txt.

5) Can you find the difference between the two amino acid sequences? _____.

It is hard to see the difference between the two sequences. The difference is a change of one amino acid for another.

- To help you see the difference, run find_substitution.py that takes as input Protein_wildtype.txt and Protein_mutant.py and outputs the position(s) where the sequences differ and the substitution(s).

This replacement of a hydrophilic glutamate residue (E) in position 7, with a hydrophobic valine residue (V) creates a hydrophobic spot on the outside of the protein. This causes the hemoglobin molecules to stick together with very serious consequences.

C) Looking at the effects of mutations (Part Two)

In this exercise, you will translate the DNA sequence with a mutation found in individuals with the disease called Beta Thalassemia and compare the resulting protein with the one you found by translating the human beta globin gene, HBB (the sequence of part A).

Consider the following variant of the human beta globin gene:

>CDS of human beta globin mutant 2

```
atggtgcacctgactcctgggagaagtctgccgttactgccctgtggggcaaggtgaac
gtggatgaagttggtggtgaggccctgggcaggttgctggtggtctacccttggacccag
aggttctttgagtcctttggggatctgtccactcctgatgctgttatgggcaaccctaag
gtgaaggctcatggcaagaaagtgctcggtgcctttagtgatggcctggctcacctggac
aacctcaagggcacctttgccacactgagtgagctgcactgtgacaagctgcacgtggat
cctgagaacttcaggctcctgggcaacgtgctggtctgtgtgctggcccatcactttggc
aaagaattcaccccaccagtgcaggctgcctatcagaaagtggtggctggtgtggctaat
gccctggcccacaagtatcactaa
```

A single nucleotide has been deleted from the wild-type (the CDS of human beta globin of Part A) to produce the above mutant. To find out what effect this mutation has on the beta globin protein, we will need to translate it into an amino acid sequence as we did before.

Let us first translate by hand the first 57 bases:

```
atg gtg cac ctg act cct ggg aga agt ctg ccg tta ctg ccc tgt ggg
___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___ ___

gca agg tga
___ ___ ___
```

Now let us use the computer to translate the mutant sequence.

- Open HandsOn_Sequences.txt.
- Using the mouse, copy the mutant sequence (only): the human beta globin CDS mutant 2.
- Go to the ExPasy translate tool at http://web.expasy.org/translate/
- Paste the sequence in the box
- Go to "Output format" (under the box) and choose "Includes nucleotide sequence"
- Click the "TRANSLATE SEQUENCE" button.
- Click on the link "5'3' Frame 1". Now you should see only the amino acid sequence (without the DNA sequence).
- Save the amino acid sequence in the same file as part A).

Carefully examine the second (the mutant 2) amino acid sequence and compare it to the one you obtained in part A.

6) Can you find the difference between the two amino acid sequences? _____.

What you should see:

- the first 6 amino acids are the same
- these are followed by 12 amino acids, namely: GRSLPLLPCGAR which were not in the normal protein
- followed by a stop codon
- after the first stop codon, we have amino acids, which are different from the normal beta globin sequence, however once a stop codon has been reached the rest of the sequence is not important. This type of mutation is called a frame shift mutation and results in truncated or shortened proteins because of the stop codon.

Mutations like this are common in people with beta thalassemia. As humans have two copies of each gene, people with this mutation are likely to have low levels of beta globin produced from one normal copy of the gene.

D) Using the UCSC Genome Browser

We now use the genome database available through the University of California at Santa Cruz (UCSC) to investigate the beta globin gene with and without the mutations.

- Go to the UCSC Genome Browser: http://genome.ucsc.edu
- Click on the "Genome Browser" link under "Our tools".

This will bring you to the "UCSC Genome Browser Gateway". This tool allows you to search the human genome and retrieve gene information.

- Keep the default version "Dec. 2013 (GRCh38/hg38)" under "Human Assembly". Note that GRC stands for Genome Reference Consortium and h for human.

- Type "HBB" (this is the standard notation for the Hemoglobin beta chain) in the "Position/Search Term" window.
- Click on the "GO" blue button.

The search results are returned on a new page. We are interested in the "RefSeq Genes". This is the link to the "reference sequence": the sequence that is considered the authoritative version.

- Click on the HBB reference sequence link: HBB at chr11:5225466-5227071 to view the HBB gene on chromosome 11.

You will be taken to a new page that displays a graphical window in the center. Near the top of the page is an illustration of the entire human chromosome 11.

7) What is the red line on the illustration of human chromosome 11 indicating? _____ .

One of the top lines in the main panel (the graphical window) represents the "GENCODE V24 Comprehensive Transcript Set".

Click on the black "HBB" in the left margin of "GENCODE v24 Comprehensive Transcript Set".
This takes you to the "Human Gene HBB (ENST00000335295.4) Description and Page Index" page which contains a lot of information on HBB and annotations compiled from many different sources and websites.

8) According to "RefSeq Summary (NM_000518)", what is the order of the genes in the beta-globin cluster? _____.

9) Is this gene known under other names? _____.

10) Under what heading does this page have graphical representations of the secondary structure of the 5'UTR and 3'UTR? _____.

11) a) What is the size of the 5'UTR? _____.
b) What is the size of the 3'UTR? _____.

- Go back to the Genome Browser main panel that showed the HBB gene on the map of chromosome 11.

The dark blue line in the main panel represents the HBB reference sequence "NCBI RefSeq Genes". The thicker parts of the line are exons and the thin lines are introns. The direction of the arrow shows the direction of transcription.

Above the chromosome illustration are a series of "move", "zoom in" and "zoom out" navigational tools.
Let us zoom out 30x (i.e., 30 times).
- Click on "3x" to the right of "zoom out"
- Click on "10x" to the right of "zoom out"
- Move a little to the right with one click of the right arrow ">".

This brings up a cluster of interesting HB genes (HBD, HBG1, HBG2) in the main panel.

12) What are the HBD and HBG1, HBG2 genes on human chromosome 11?
_____.

13) Hypothesize why these HB genes are found in a cluster.
_____ .

- Go back to the Genome Browser main panel that showed the HBB gene on the map of chromosome 11.

- Click on the blue "HBB" in the left margin of "NCBI RefSeq genes" (in the main window).
  This brings you to the "RefSeq Gene HBB" page where most of the data is from NCBI.
- Scroll down to the link "Genomic Sequence from assembly" under "Links to sequence:".
- Click on the "Genomic Sequence from assembly" link. This brings you to a formatting page that allows you to structure the way you want the DNA sequence to be displayed. Most of the default settings are fine. Read (and do not change) the selected ones.
- Make sure the radio button for "Exons in upper case, everything else in lower case" under "Sequence Formatting Options:" is selected.
- Hit the "get DNA" button at the bottom of the page. Now you finally have the HBB gene sequence in FASTA format.

14) a) How many exons does the gene have? _____ .
b) How many introns does the gene have? _____ .

15) What are the first 2 bases of each intron? _____.
16) What are the last 2 bases of each intron? _____.

17) Are there more DNA bases in the introns or the exons of the HBB gene? _____ .

Let us now identify the DNA sequence for the sickle cell mutant hemoglobin allele.

- Go back to the Genome Browser main panel that showed the HBB gene on the map of chromosome 11.

To get the sequence of the mutant allele, we need to add information about genetic variations to the map (before we zoomed out).

- Scroll down to the section entitled "Variation". This is one of the tracks of the UCSC Genome Browser. You might have to click on "+" to the left of "Variation" to expand this track.
- Click on "Common SNPs(147)". SNPs are single nucleotide polymorphisms, in other words, single base mutations.
- In the new page: "Common SNPs(147) Track Settings", expand "Coloring Options" by clicking on the "+" next to it.

18) What is the default color of "Coding – Synonymous" SNPs? _____.

19) What is the default color of "Coding – Non-Synonymous" SNPs? _____.

- Go back to the Genome Browser main panel that showed the HBB gene on the map of chromosome 11.
- Scroll down to the section entitled "Variation".
- Set the "Common SNPs(147)" pulldown to "pack".
- Hit the "refresh" button to the right of "Variation".

Now the main panel has more information in it. At the bottom of the main panel there is a section entitled "Simple Nucleotide Polymorphisms". Each of these is a known mutation of the human HBB gene. We are interested in the "rs334" polymorphism. This is the sickle cell mutation. The horizontal bar to the right of "rs334" indicates the location of the SNP in HBB. You might have to be patient while looking for "rs334" among all the SNPs.

20) Do we have a "synonymous" or "non-synonymous" mutation? _____.
Why? _____.

21) Is the sickle cell mutation in an intron or an exon? _____ .

22) Is the mutation at the beginning or end of the gene? _____ .

- Click on the "rs334" polymorphism. We are taken to a page entitled "Simple Nucleotide Polymorphisms (dbSNP 144) Found in >= 1% of Samples" summarizing the information about this SNP.

It contains the following three lines:

Strand: -
Observed: A/C/G/T
Reference allele: A

23) What is meant by "Strand: - "? _____ . Was it expected? _____ .
Why? _____ .

24) What is meant by "Observed: A/C/G/T"? _____ . Was it expected? _____ .
Why? _____ .

_____