Project 1: Retrieving the Protein Sequences from the Human Genome

Objectives:

1. Be familiar with the human genome and major genome annotation databases
2. Be familiar with the central dogma
3. Be familiar with alternative splicing and the codon table
4. Be familiar with the FASTA format for storing biological sequences

Task:

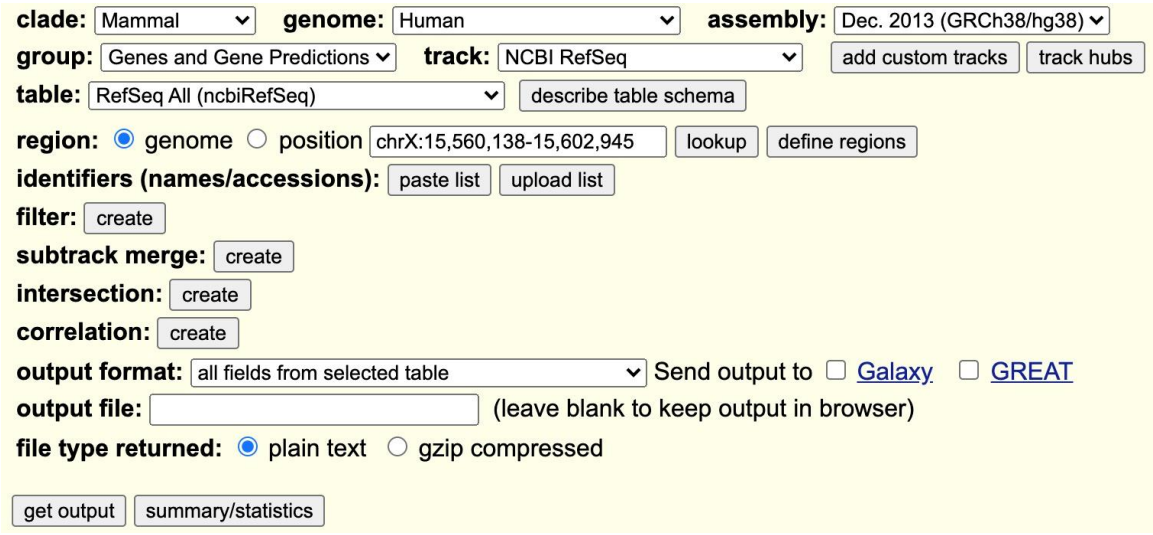
Retrieve the sequences of all proteins encoded in the human genome.

Hits:

(1): Explore the UCSC (U. California Santa Cruz) Genome Browser website (genome.ucsc.edu). Try to find where to download the human genome. (If you can't, here is the link: http://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.fa.gz)

(2): Use the Table browser of the website to obtain human genome annotation. (From the top bar, under "Tools", select "Table Browser").

(3): Make the following selection:



clade: Mammal
genome: Human
assembly: Dec. 2013 (GRCH38/hg38)
group: Genes and Gene Preditions
track: NCBI RefSeq

table: RefSeq All (ncbiRefSeq) (**I strongly recommend you to click "describe table schema" to understand the meaning of the table. This is where I will direct you to if you ask me what does each field of the table mean.**)

region: genome

output file: [make you own selection]

and then click "get output".

(4): Obtain the human codon table from https://www.genscript.com/tools/codon-frequency-table. Note that you need to select "Human" from "Expression Host Organism".

(5): Write a script to obtain all protein sequences coded in the human genome. Your output should be in the multiple FASTA format, which looks like:

>ID1
Sequence 1…
>ID2
Sequence 2…

The ID field describes what the sequence is. You should use the concatenation (with colon ":" as the delimiter) of the RefSeq table name1 and name2 fields as the ID. For example, for the first record in the RefSeq table, the corresponding ID should be ">NM_001276352.2:Clorf141".

The sequence field simply records the corresponding sequence, all in one line. For example:
MVLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGS
AQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHC
LLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR

Submission:

Send your single FASTA file (gzipped) via the Blackboard system by Friday Mar 5th, 2021 11:59PM. **Please make sure you submission is properly titled and on time.**