**Mutation Mapping Design:**

* Given:
  + VCF file of HGMD mutations
  + gtf file with information from all chromosomes
  + hg38 fasta (sequence of each chromosome)
  + Transcript fasta
  + Protein fasta
* From gtf file:
  + Filter so that left only with “CDS” lines
    - **Function**: something similar to “CDS\_list.py”
  + Filter so that only missense and nonsense mutations, no indels
    - **Function**: position 7 and 8, “insertion” and “deletion” should be “NULL”, and position 6 “amino” should not be “NULL”
  + Filter gtf further: Search for gene name, then scan the VCF to see if gene is there.
    - If a gene in the gtf is not found in the VCF, remove that line from the gtf.
    - **Function:** split file by “gene\_name: “…if gene name in VCF, put that gtf line into a list. Export to new file.
* For each gene in VCF, find the position of each mutation
  + Concatenate the CDS sequence’s for each transcript
    - **Function**: something similar to “concatenate\_CDS\_3.py”
    - Save the concatenated sequence as a string
  + The VCF file gives position in terms of codon # and position within the codon (1, 2, or 3)

If position\_within\_codon == 3:

Nucleotide\_position = (codon#)\*3

Elif position\_within\_codon == 2:

Nucleotide\_position = (codon#)\*3 – 1

Elif position\_within\_codon == 1:

Nucleotide\_position == (codon#)\*3 – 2

* These values give the position within the coding sequence, starting at position 1
* From the nucleotide\_position collected, put that as the index (adjust!) in the string of the concatenated sequence
  + Convert that position to the chromosome coordinate!
* Make a table in an excel file:
  + Transcript: ENST Accession (from gtf)
  + Gene: ENSG Accession (from gtf)
  + Gene Name: gene name (from VCF)
  + Chromosome: # (from gtf)
  + Coordinate of mutation: coordinate corresponding to the chromosome
  + Mutation: A 🡪 T (from VCF)
  + Amino acid change: Cys-Tyr (from VCF)
  + Codon: # (from VCF)
  + Base: # (1, 2, or 3) (from VCF)
  + Coding sequence: (put entire coding sequence, with mutation coordinate marked)