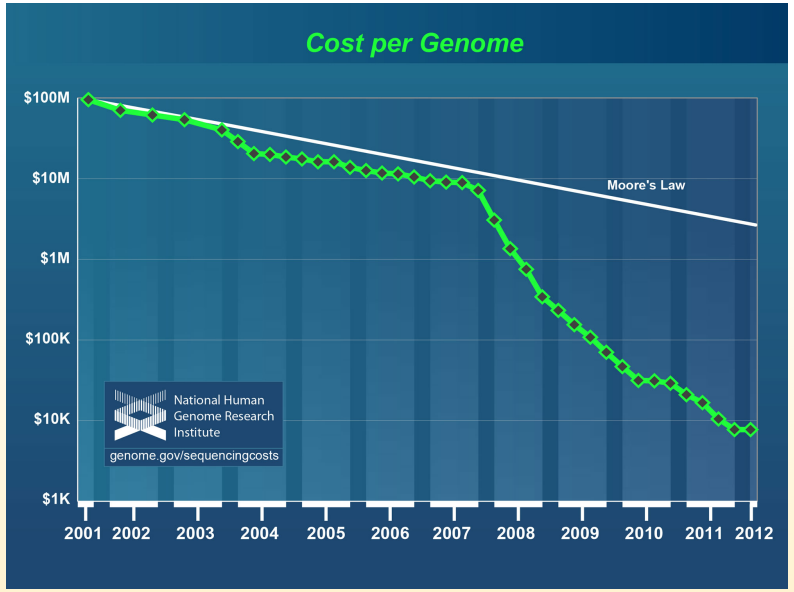
Project – main assignment

Every day from 15-17h

Project background

Whole genome sequencing in the clinic

- Sequencing costs low
- Screening for mutations causing diseases from whole genome sequencing data



Project background

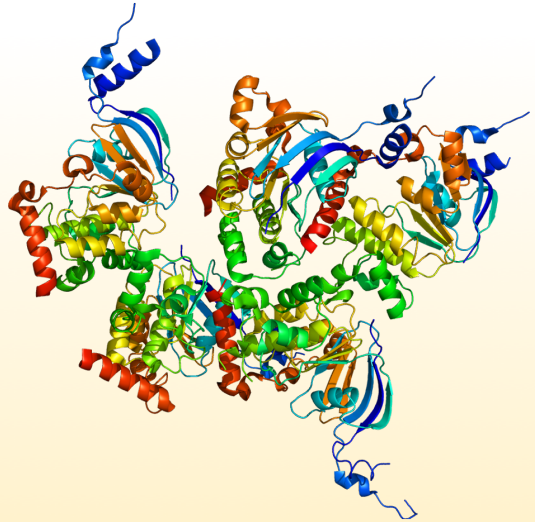
Cystic fibrosis (CF)

- Hereditary disease
- Affects lungs and digestive system
- Production of thick and sticky mucus
- No cure but symptom management

Project background

Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)

- Ion channel protein, acting in epithelial cells
- CFTR gene located on chromosome 7 of the human genome
- Over 1,500 mutations known to cause CF
- Non-synonymous (amino acid changing) mutations → premature stop codons



Goal

Write a python program that:

- Extracts the correct CFTR transcript from the human genome
- Translates it into its corresponding amino acid sequence
- Determines if one or more patients have a premature stop codon

→ You are guided step by step towards the final task

Available data

- Human reference genome
  - Chromosome 7 in fasta format
  - Gene annotations in GTF format
- Genome sequencing data from five patients
  - Chromosome 7 in fasta format

FASTA format

```
>MT dna:chromosome chromosome:GRCh38:MT:1:16569:1 REF
GATCACAGGTCTATCACCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTT
CGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTC
GCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATT
ACAGGCGAACATACTTACTAAAGTGTGTTAATTAATTAATGCTTGTAGGACATAATAATA
ACAATTGAATGTCTGCACAGCCACTTTCCACACAGACATCATAACAAAAAATTTCCACCA
AACCCCCCCTCCCCCGCTTCTGGCCACAGCACTTAAACACATCTCTGCCAAACCCCAAAA
ACAAAGAACCCTAACACCAGCCTAACCAGATTTCAAATTTTATCTTTTGGCGGTATGCAC
TTTTAACAGTCACCCCCCAACTAACACATTATTTTCCCCTCCCACTCCCATACTACTAAT
CTCATCAATACAACCCCCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATA
CCCCGAACCAACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAA
```

GTF format

- Gene transfer format
- Holds information about gene structure
- Tab-delimited
- Based on the general feature format (GFF), additional structure specific to genes

GTF format

Structure (per line):

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

GTF format

Structure (per line):

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

- seqname - name of the chromosome or scaffold; chromosome names without a 'chr' source - name of the program that generated this feature, or the data source (database or project name)
- feature - feature type name. Current allowed features are {gene, transcript, exon, CDS, Selenocysteine, start_codon, stop_codon and UTR}
- start - start position of the feature, with sequence numbering starting at 1.
- end - end position of the feature, with sequence numbering starting at 1.

GTF format

Structure (per line):

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

- score - a floating point value indiciating the score of a feature
- strand - defined as + (forward) or - (reverse)
- frame - reading frame, one of '0', '1' or '2'.
- attribute - a semicolon-separated list of tag-value pairs (separated by a space), providing additional information about each feature. A key can be repeated multiple times.

GTF format

Structure (per line):

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

Some attributes (always semi-colon separated key-value pairs):

- gene_id: The stable identifier for the gene
- gene_version: The stable identifier version for the gene
- gene_name: The official symbol of this gene
- gene_source: The annotation source for this gene
- transcript_id: The stable identifier for this transcript
- transcript_name: The symbold for this transcript derived from the gene name
- exon_id: The stable identifier for this exon

GTF format

Structure (per line):

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes]
```

```
1 havana gene 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene";
```

```
1 havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "processed_transcript"; tag "basic"; transcript_support_level "1";
```

```
1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "1"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene"; transcript_name "DDX11L1-202"; transcript_source "havana"; transcript_biotype "processed_transcript"; exon_id "ENSE00002234944"; exon_version "1"; tag "basic"; transcript_support_level "1";
```