Tutorial 1. Introduction

The Chinese University of Hong Kong

BMEG3102 Bioinformatics

TA: Yizhen Chen

TA Info

- Yizhen CHEN (Channie)
- Office: SHB 1023
- Email: yzchen@cse.cuhk.edu.hk
- Consultation hour: Tue 14:00-16:00

Agenda

- Course Assessment
- Introduction to Bioinformatics
- Demonstration on how to do exercise related to Web tool
- Case Study: Mutation and disease (Sickle Cell Anemia)

Course assessment

- Assignments: 40%
- Class participation: 5%
- Reading presentation: 5%
- Mid-term examination: 20%
- Final examination: 30%

What is “Bioinformatics”?

- Biology + Informatics
- For Biology, you will learn / see:
  - Some biology terms
  - Basic but important biology concepts
  - Some medical problems (e.g. find genetic variants that are related to a disease)
- For Informatics, you will learn / see:
  - Some engineering terms
  - Large-scale data sets
  - Some systematic ways to solve problems and analyze results

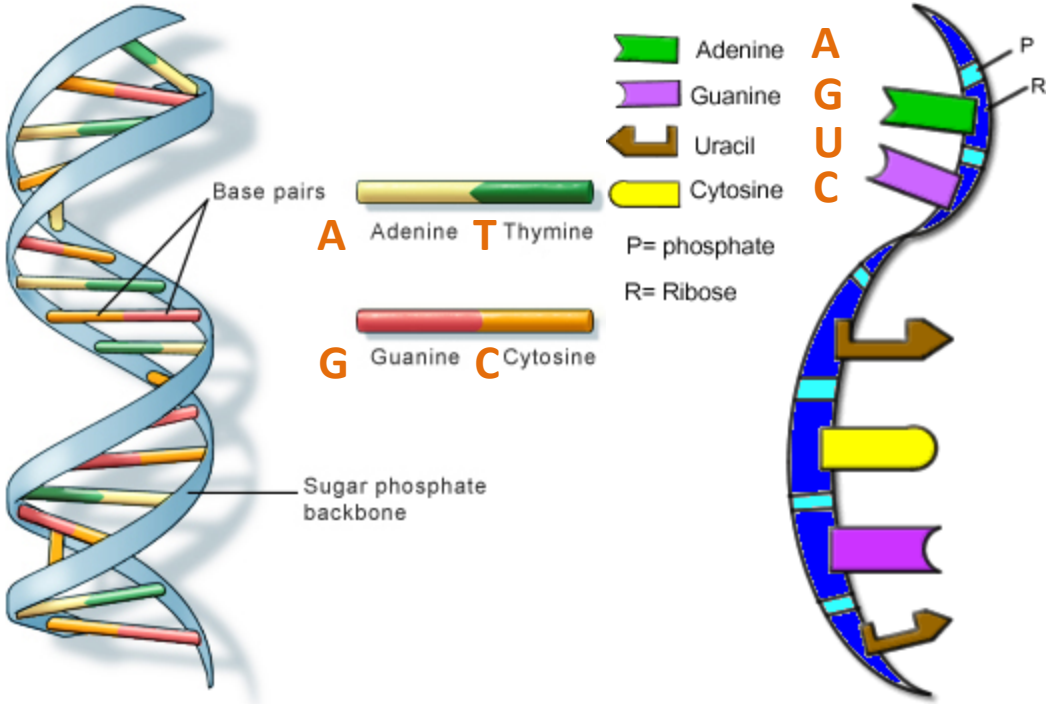
Basic terms in Biology

Term	Description
Bioinformatics	Application of computer science and information technology to the field of biology and medicine
Cell	Basic functional unit of life
Chromosome	Organized structure of DNA and protein found in cells
DNA	Carrying genetic information / double helix / A, T, C, G
RNA	Single strand / A, U, C, G
Protein	A strand of amino acids, folded into a particular structure
Amino acid	20 common types
DNA Replication	DNA → DNA (mitosis and meiosis)
Transcription	DNA → RNA
Translation	RNA → Protein
Codon	Specify which amino acid will be translated for each DNA triplet

Basic terms in Biology

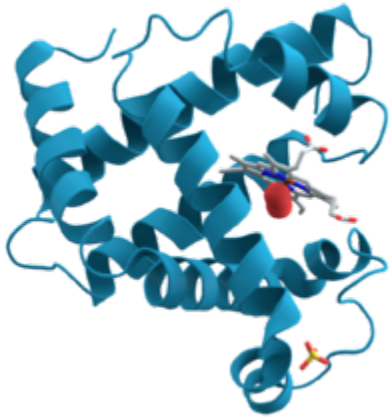
Term	Description
Coding Strand	Appears the same (except T → U) as the RNA transcribed
Template Strand	Used as the template for RNA synthesis
Mutation	Change of DNA sequence, e.g. substitution, insertion or deletion (indel) in some genomic regions
Transition	Nucleotide changes within purine (nucleotides with two carbon rings, i.e., A ↔ G) or within pyrimidine (nucleotides with one carbon ring, i.e., C ↔ T)
Transversion	Nucleotide changes other than A ↔ G and C ↔ T (more change in size, leading to more drastic consequences)
Synonymous Mutation	Different DNA sequences produce the same protein
Non-synonymous Mutation	Different DNA sequences produce different proteins
Missense Mutation	Change one amino acid to another (Non-synonymous)
Nonsense Mutation	Change one amino acid to stop codon (Non-synonymous)
Single Nucleotide Polymorphism (SNP)	a substitution of a single nucleotide at a specific position in the genome in a sufficiently large fraction of the population

Central Dogma



DNA

RNA



Protein

Image credit: http://ghr.nlm.nih.gov/handbook/basics/dna, http://www.biologycorner.com/bio1/DNA.html, wiki image

Central Dogma

Coding Strand: Appears the same (except T → U) as the RNA transcribed
Template Strand: Used as the template for RNA synthesis

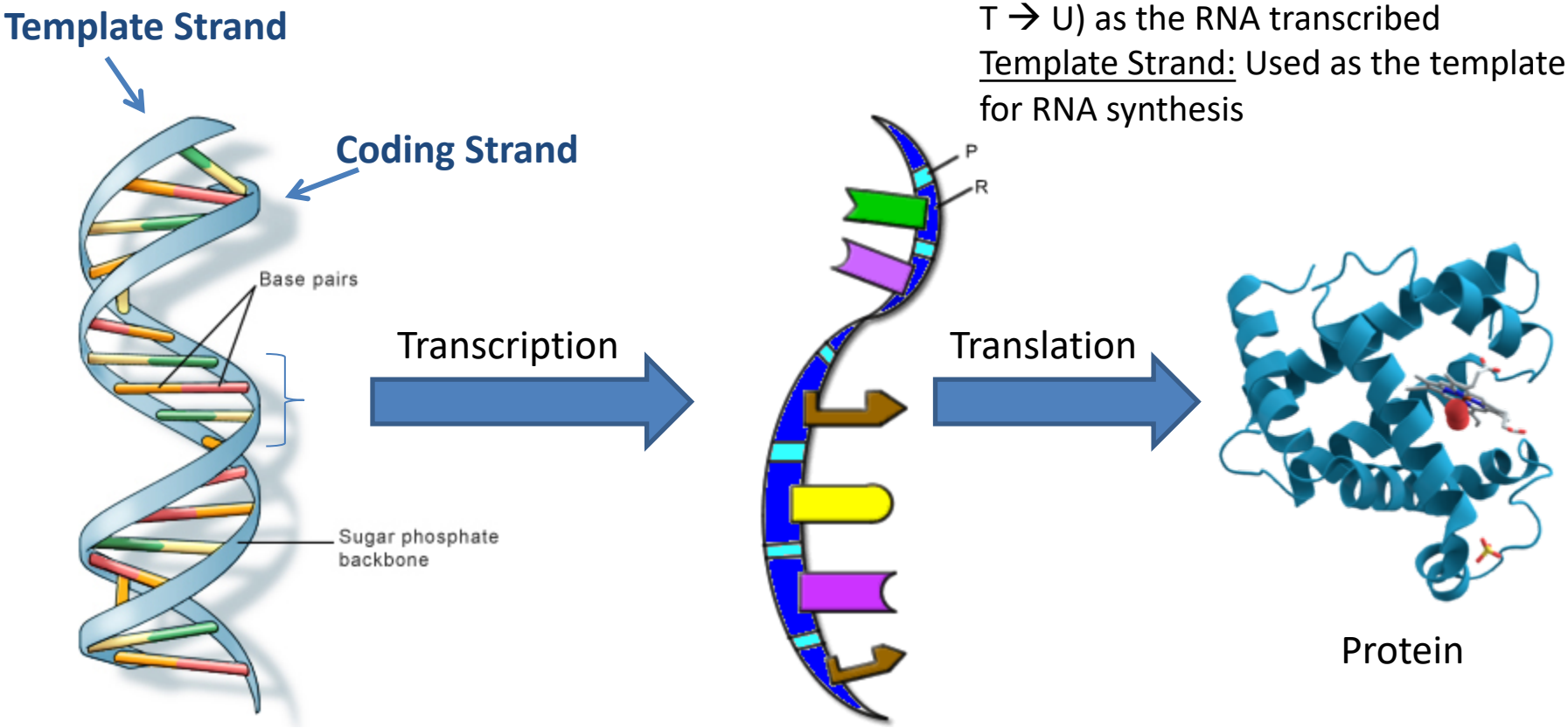


Image credit: http://ghr.nlm.nih.gov/handbook/basics/dna, http://www.biologycorner.com/bio1/DNA.html, wiki image

Some useful videos

DNA Replication: https://www.youtube.com/watch?v=TNKWgcFPHqw

Transcription: http://www.youtube.com/watch?v=WsofH466lqk

Translation: http://www.youtube.com/watch?v=5bLEDd-PSTQ

Mitosis: https://www.youtube.com/watch?v=f-ldPgEfAHI&t=32s

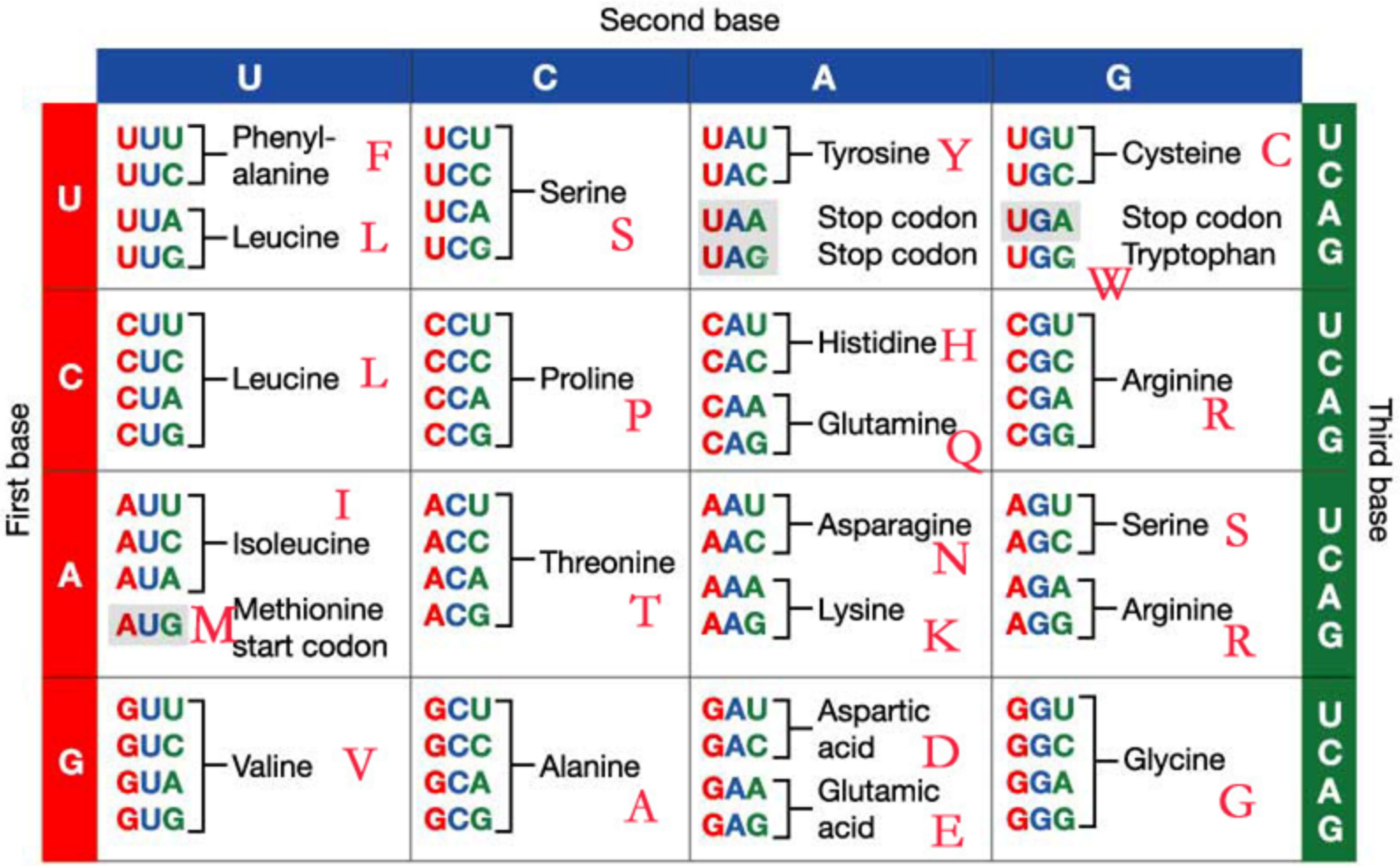
Meiosis: https://www.youtube.com/watch?v=VzDMG7ke69g

Youtube channel

Amoeba Sisters

https://www.youtube.com/user/AmoebaSisters

Codon table


Second base
U
C
A
G
First base
Third base
UUU Phenyl-alanine F
UUC
UUA Leucine L
UUG
UCU Serine S
UCC
UCA
UCG
UAU Tyrosine Y
UAC
UAA Stop codon
UAG Stop codon
UGU Cysteine C
UGC
UGA Stop codon
UGG Tryptophan W
CUU Leucine L
CUC
CUA
CUG
CCU Proline P
CCC
CCA
CCG
CAU Histidine H
CAC
CAA Glutamine Q
CAG
CGU Arginine R
CGC
CGA
CGG
AUU Isoleucine I
AUC
AUA
AUG Methionine start codon M
ACU Threonine T
ACC
ACA
ACG
AAU Asparagine N
AAC
AAA Lysine K
AAG
AGU Serine S
AGC
AGA Arginine R
AGG
GUU Valine V
GUC
GUA
GUG
GCU Alanine A
GCC
GCA
GCG
GAU Aspartic acid D
GAC
GAA Glutamic acid E
GAG
GGU Glycine G
GGC
GGA
GGG

Exercise: Transcription and translation

- Given the following coding strand of a DNA:

  ```
        12345678901234567890123456789 0
  5’-TGTTGCATGAGATATGTACAGAGGTAGTGA-3’
  ```

- Work out the pre-mRNA and mature mRNA, if position from 16 to 21 are intron, and the remaining are exons
- Deduce a protein sequence from the mature mRNA, and explain briefly how the sequence is deduced

Answer: Transcription and translation

- Pre-mRNA:
  - 5’-UGUUGCAUGAGAUAUGUACAGAGGUAGUGA-3’
- Mature mRNA:
  - 5’-UGUUGCAUGAGAUAU AGGUAGUGA-3’
- Protein:
  - NH_3- M R Y R *-COOH
- A coding sequence will always start with start codon (AUG) and end with stop codon (UAA, UAG or UGA).

Exercise: Six reading frames

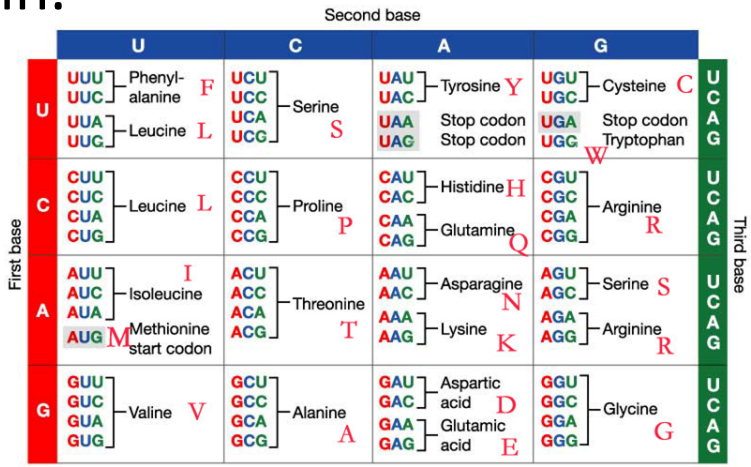
- You are provided with a DNA sequence within a coding region below:
  - Write down the sequence of reverse strand, and
  - Deduce ALL possible protein sequences (you may NOT worry about the start codon and stop codon)

- DNA sequence:

  5’-AACTTGTTCGTACA-3’

Six reading frames

- Given a sequence of DNA in coding region, we can have 6 ways to translate the DNA to protein.



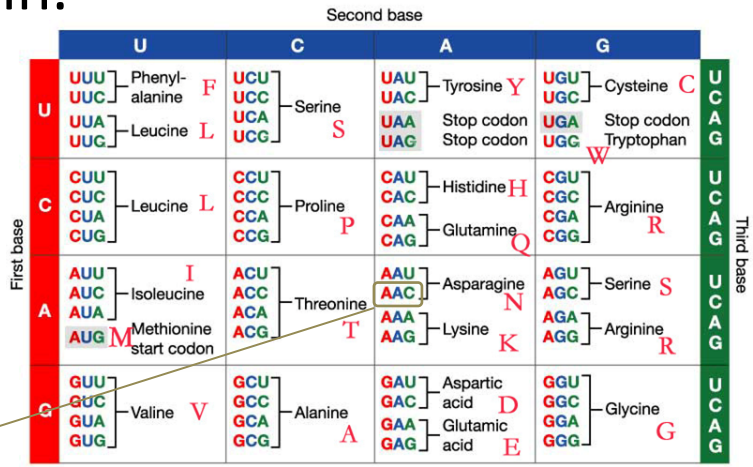
- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

```
5’-AACTTGTTCGTACA-3’
3’-TTGAACAAGCATGT-5’
```

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six reading frames

- Given a sequence of DNA in coding region, we can have 6 ways to translate the DNA to protein.



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

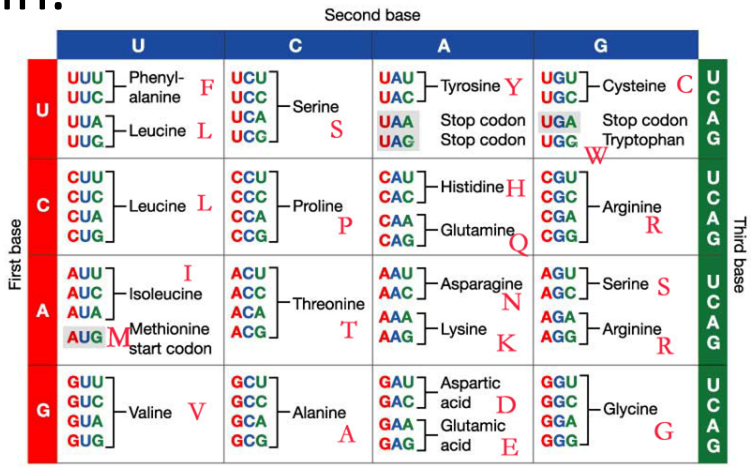
5’-AACTTGTTCGTACA-3’

3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six reading frames

- Given a sequence of DNA in coding region, we can have 6 ways to translate the DNA to protein.



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

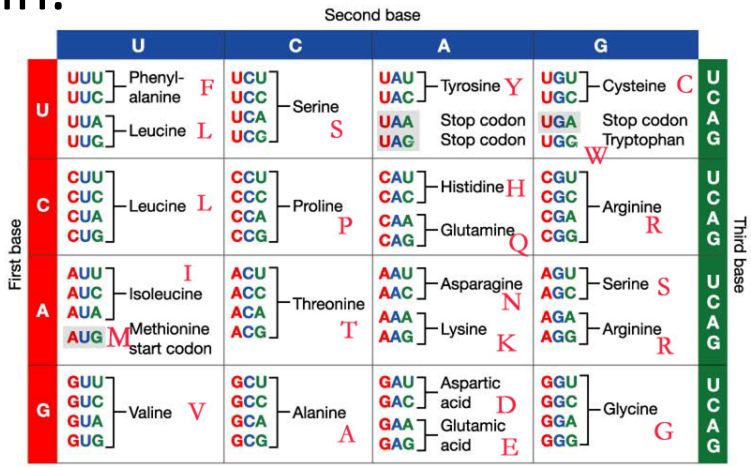
5’-AACTTGTTCGTACA-3’

3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six reading frames

- Given a sequence of DNA in coding region, we can have 6 ways to translate the DNA to protein.



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

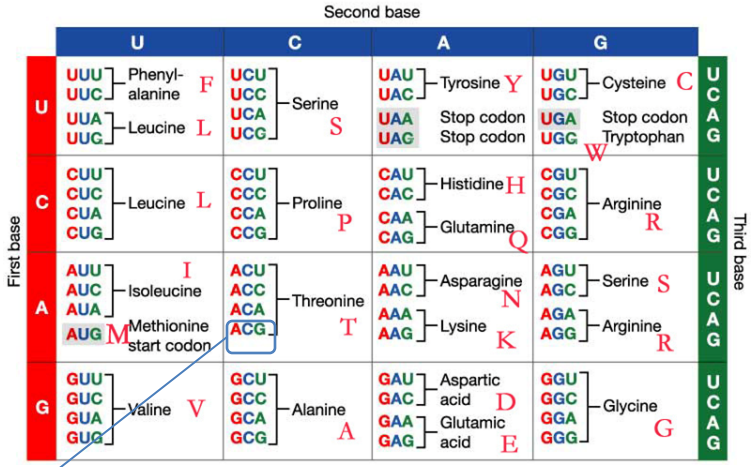
5’-AACTTGTTCGTACA-3’

3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six Reading Frames

- Given a sequence of DNA in coding region, we can have 6 ways to translated the DNA to protein



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

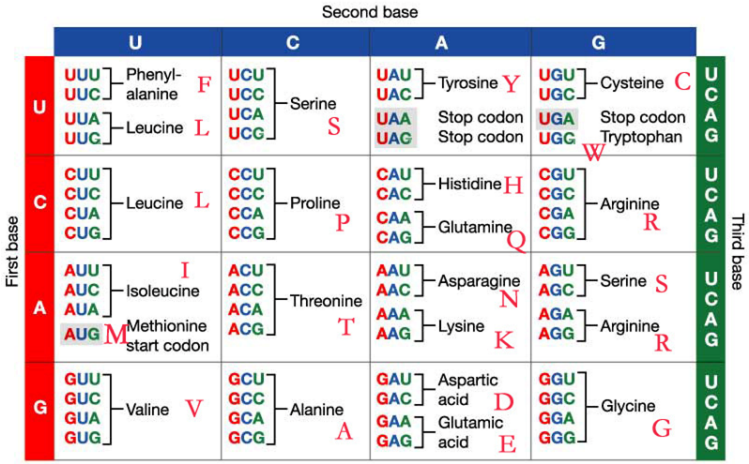
5’-AACTTGTTCGTACA-3’

3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six Reading Frames

- Given a sequence of DNA in coding region, we can have 6 ways to translated the DNA to protein



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

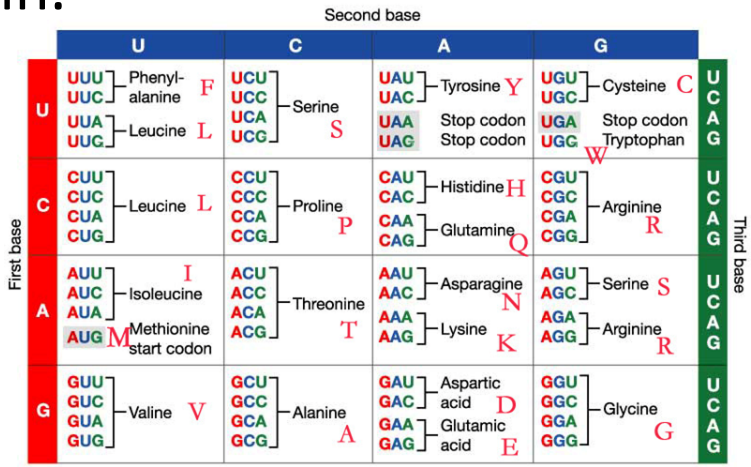
5’-AACTTGTTCGTACA-3’

3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Six reading frames

- Given a sequence of DNA in coding region, we can have 6 ways to translate the DNA to protein.



- Example:
- +3 L V R T
- +2 T C S Y
- +1 N L F V

5’-AACTTGTTCGTACA-3’

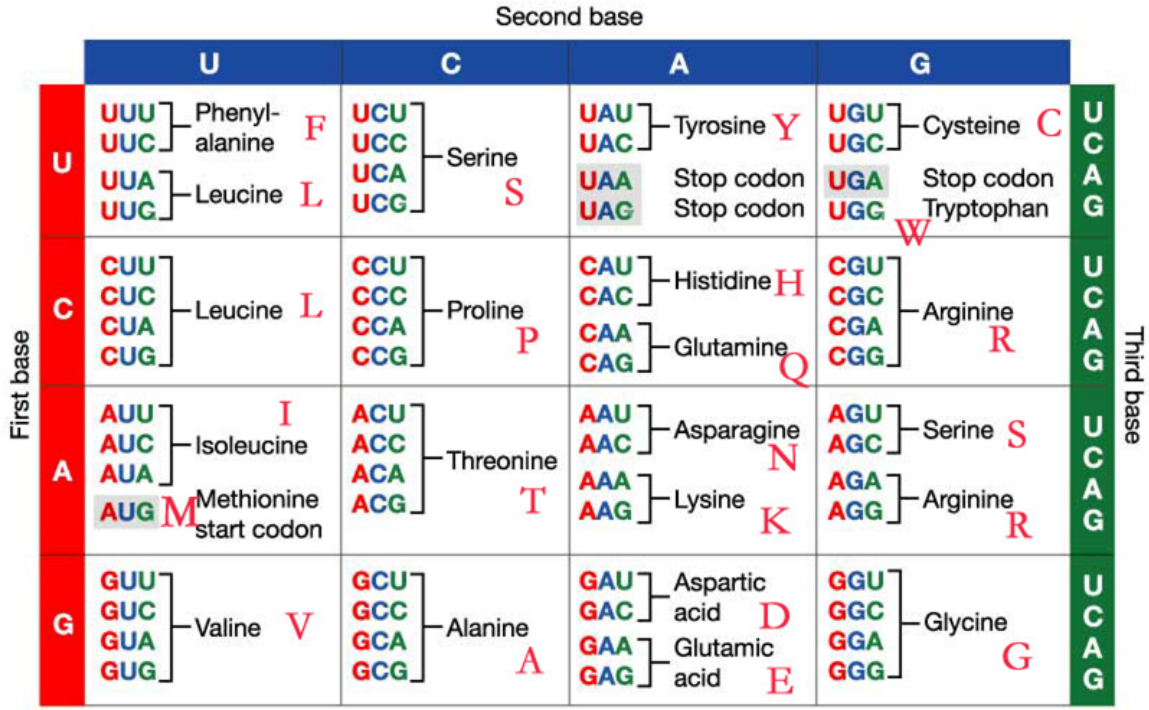
3’-TTGAACAAGCATGT-5’

- -1 K N T C
- -2 S T R V
- -3 V Q E Y

Exercise: Six reading frames

- Write down the six reading frames of the following DNA strand:

  5’-CACTATCGCTTGTCC-3’



Answer: Six reading frames

- +3 L S L V
- +2 T I A C
- +1 H Y R L S

5’-CACTATCGCTTGTCC-3’

3’-GTGATAGCGAACAGG-5’

- -1 V I A Q G
- -2 * R K D
- -3 S D S T

Problems of previous exercise

- No splicing
- The 5’- and 3’-untranslated regions (UTR) are missing
- The protein does not start at “start codon”, and does not finish at “stop codon”
- The protein is too short (usually a protein should contain several hundred amino acids)

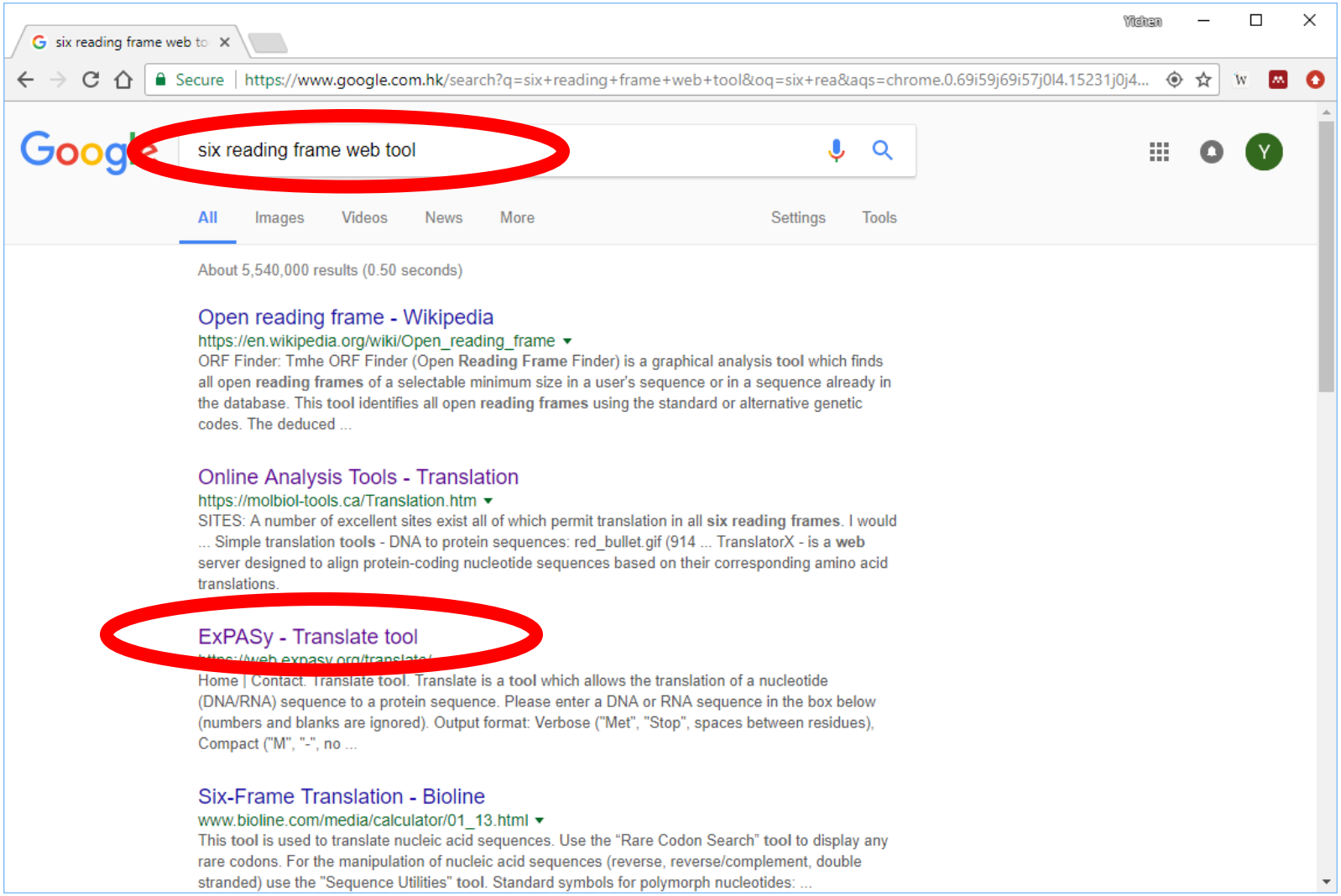
Demonstration: Web exercise

- Find a Web tool to perform six-frame translation on the sequence below:

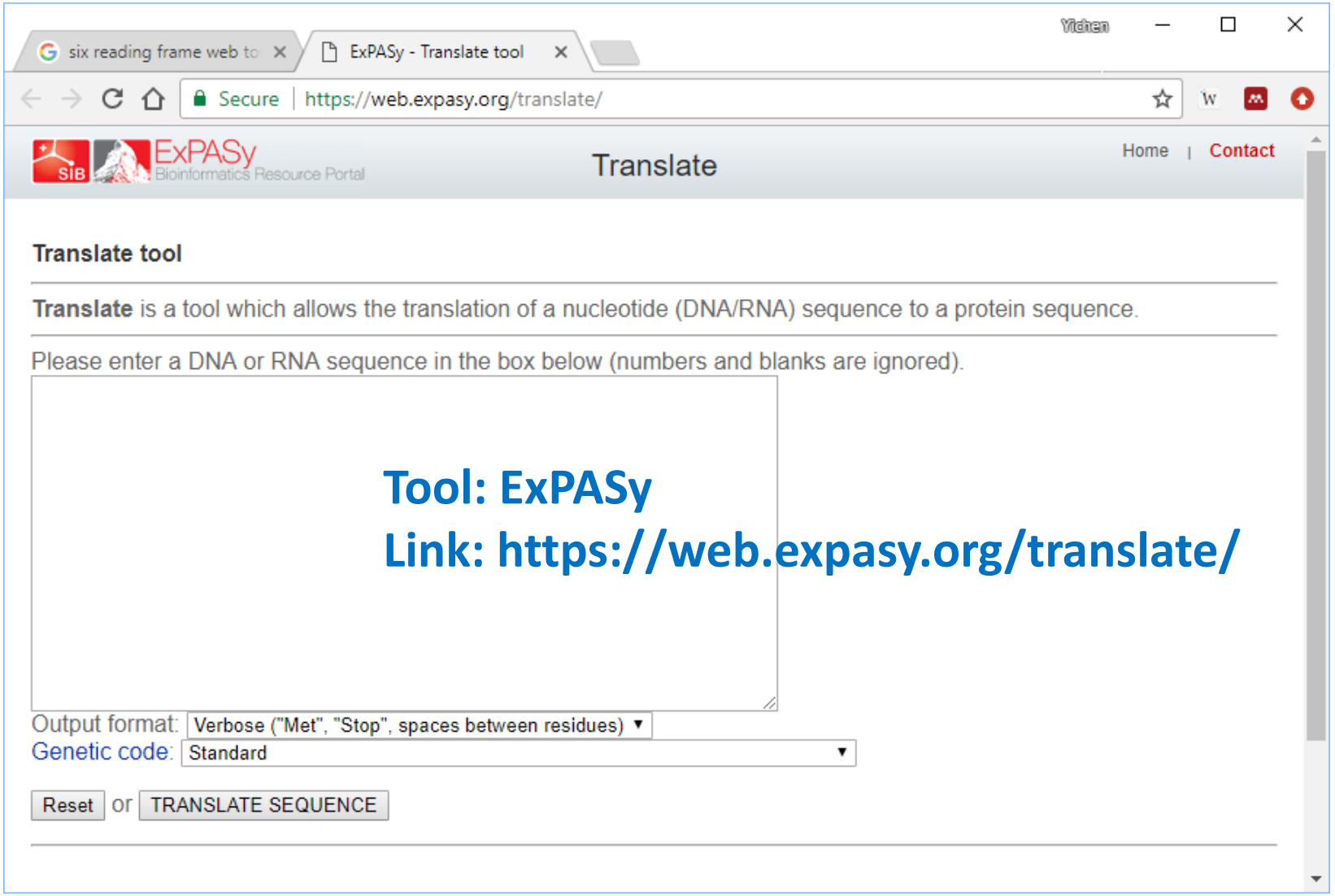
  5’-CACTATCGCTTGTCC-3’

- Remember to record details of tools you used including the input data, parameters used, and full screen shots of the outputs of every step.

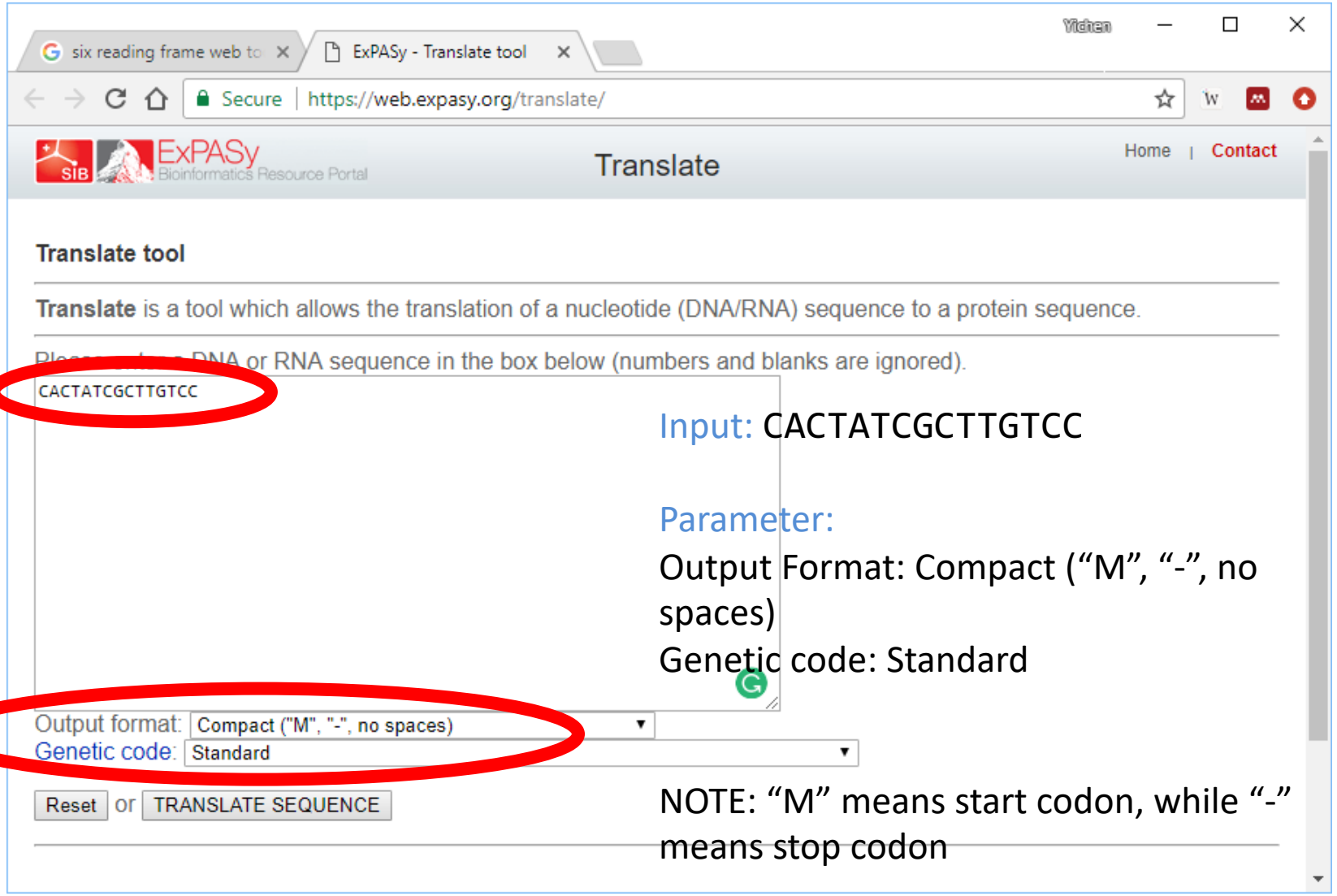
Step 1: Find tools from Google



Step 2: Describe the tool, ‘ExPASy’



Step 3: State the input and parameters



Input: CACTATCGCTTGTCC

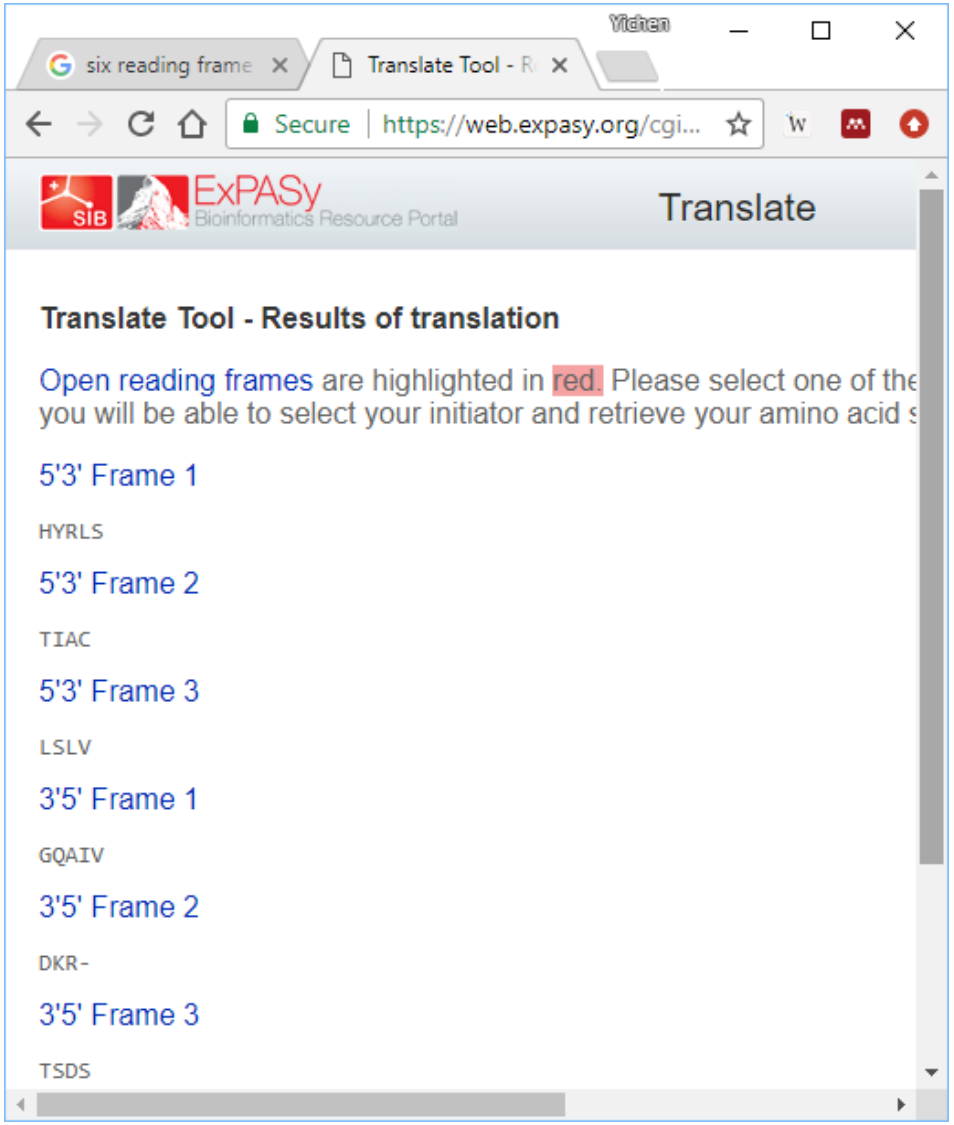
Parameter:

Output Format: Compact ("M", "-", no spaces)

Genetic code: Standard

NOTE: "M" means start codon, while "-" means stop codon

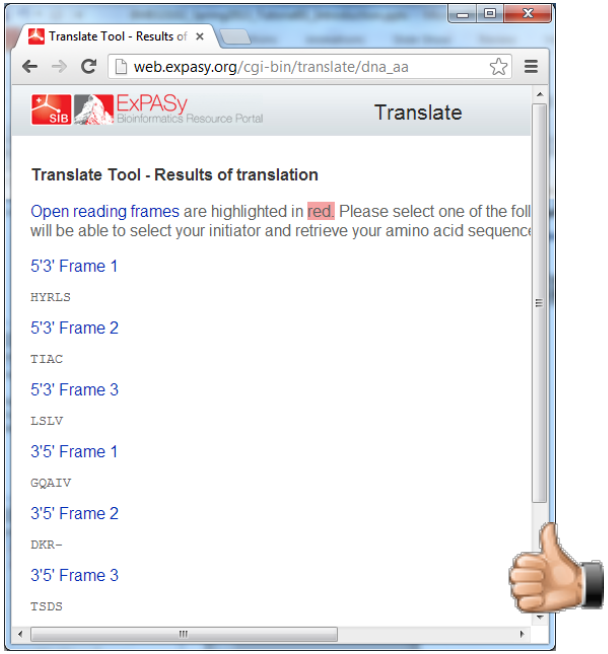
Step 4: Show the results

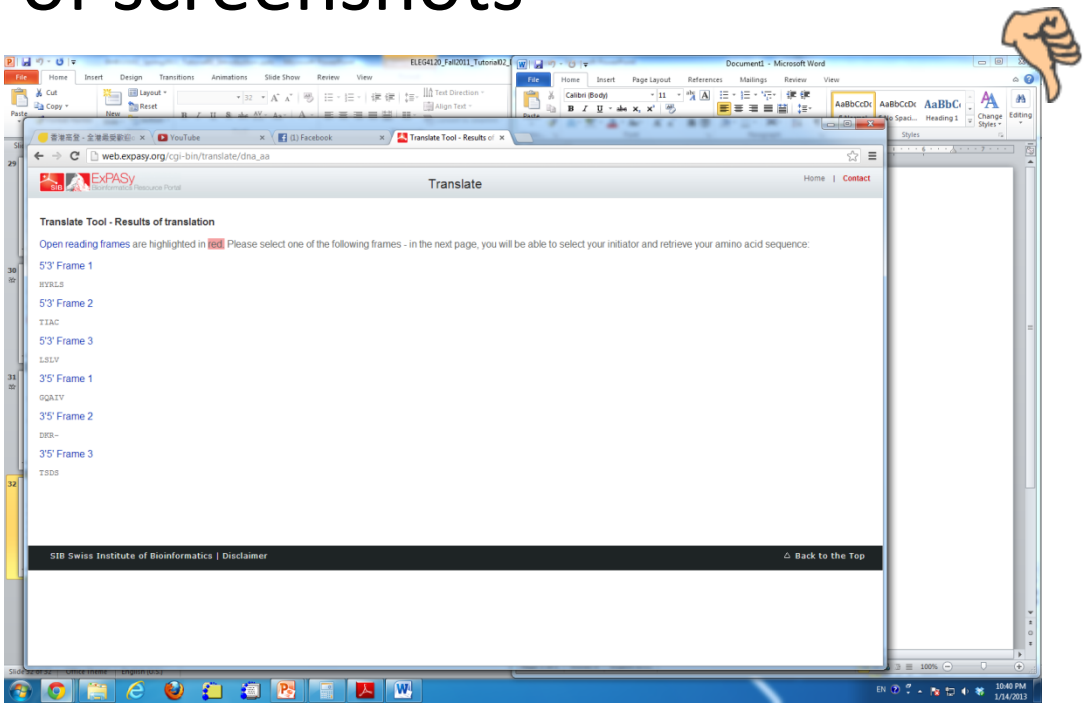


- Six reading frames:
  - +1 HYRLS
  - +2 TIAC
  - +3 LSLV
  - -1 GQAIV
  - -2 DKR-
  - -3 TSDS

Tips for making screenshot

- No extra background
- Good quality of words / figures
- Reasonable size of screenshots



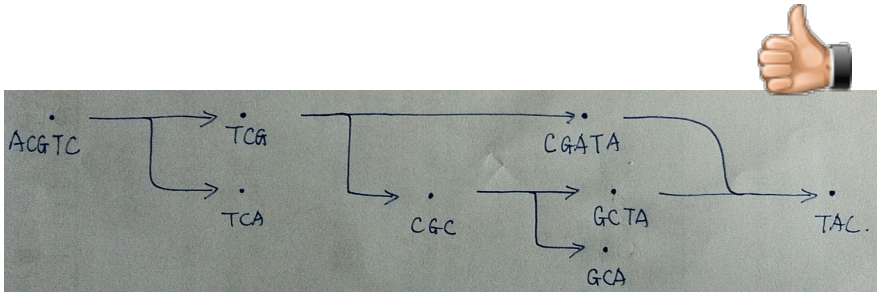


Tools for making screenshot

- For Windows users:
  - Press Alt-PrtScn, copy it to Word/PowerPoint, then crop (Format > Crop)
  - Use Snipping Tool to select the desired area
- For Mac users:
  - Press Shift+Command+4 to select the desired area
  - https://support.apple.com/en-hk/HT201361

Tips for submitting assignments

- Only accept softcopy for the assignment
- Submit your work using the Blackboard System
- When you are working with some *written problems*, you may:
  - Type and draw the figures in the computers, or
  - Write on paper and scan your work, or
  - Write on paper, take a photo and attach the photo in Word
- However, ……



Case study: Mutation and disease

- Classic example:

  Sickle-cell Anemia

  – Normal red blood cells are round and flexible, able to carry oxygen and travel through blood vessels

  – People with Sickle-cell Anemia have red blood cells that look like a sickle and get stuck in blood vessels

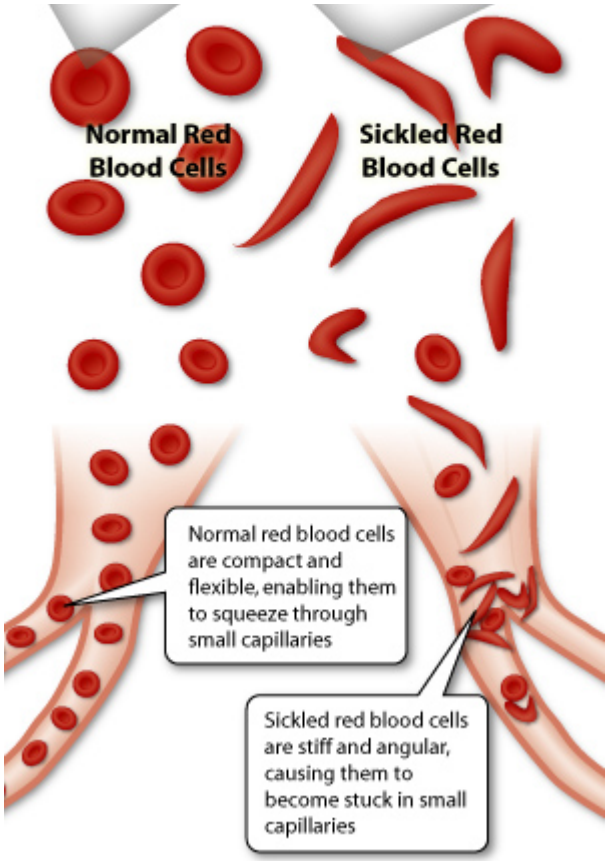


Image credit: http://learn.genetics.utah.edu/content/disorders/whataregd/sicklecell/images/sicklecell.jpg

Case study: Mutation and disease

- What went wrong?
  - Hemoglobin molecules, which are responsible for carrying oxygen, form a long chain in affected people
  - Why? There is a mutation in one of the hemoglobin genes on chromosome 11

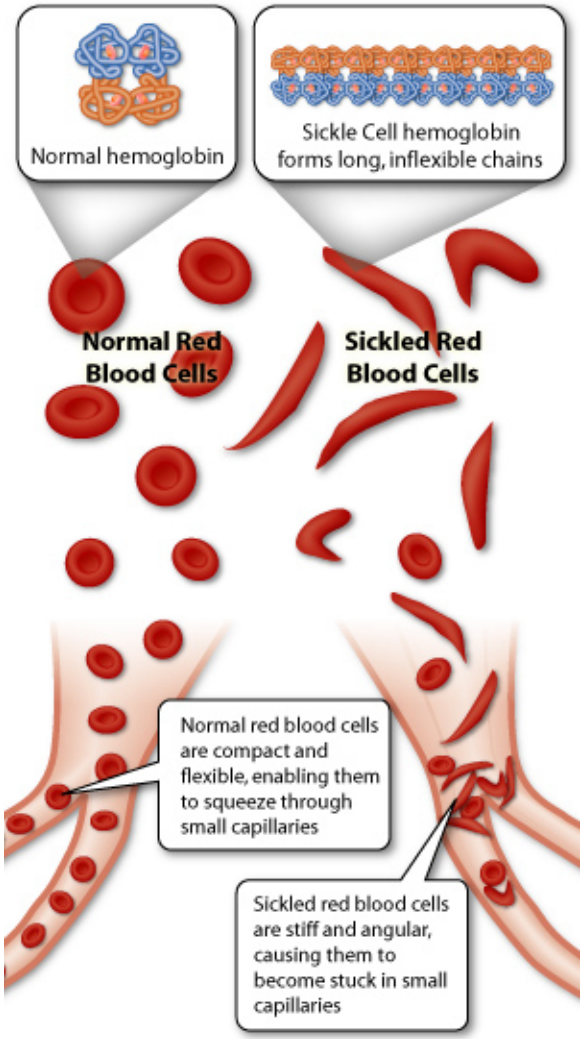


Image credit: http://learn.genetics.utah.edu/content/disorders/whataregd/sicklecell/images/sicklecell.jpg

Case study: Mutation and disease

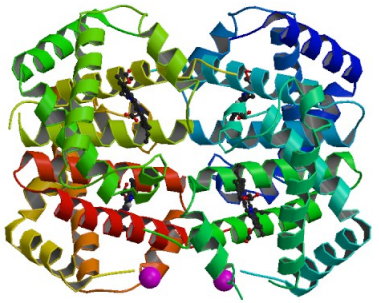
- Mutation in DNA causes change in protein structure

HBB Sequence in Normal Adult Hemoglobin (Hb A):

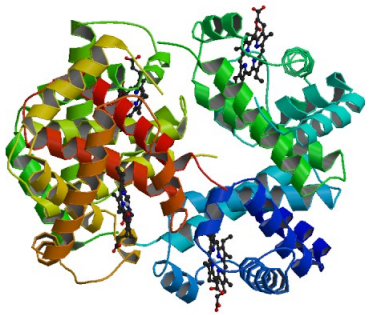
Nucleotide	CTG	ACT	CCT	GAG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Glu	Glu	Lys	Ser
	3			6			9

HBB Sequence in Mutant Adult Hemoglobin (Hb S):

Nucleotide	CTG	ACT	CCT	GTG	GAG	AAG	TCT
Amino Acid	Leu	Thr	Pro	Val	Glu	Lys	Ser
	3			6			9



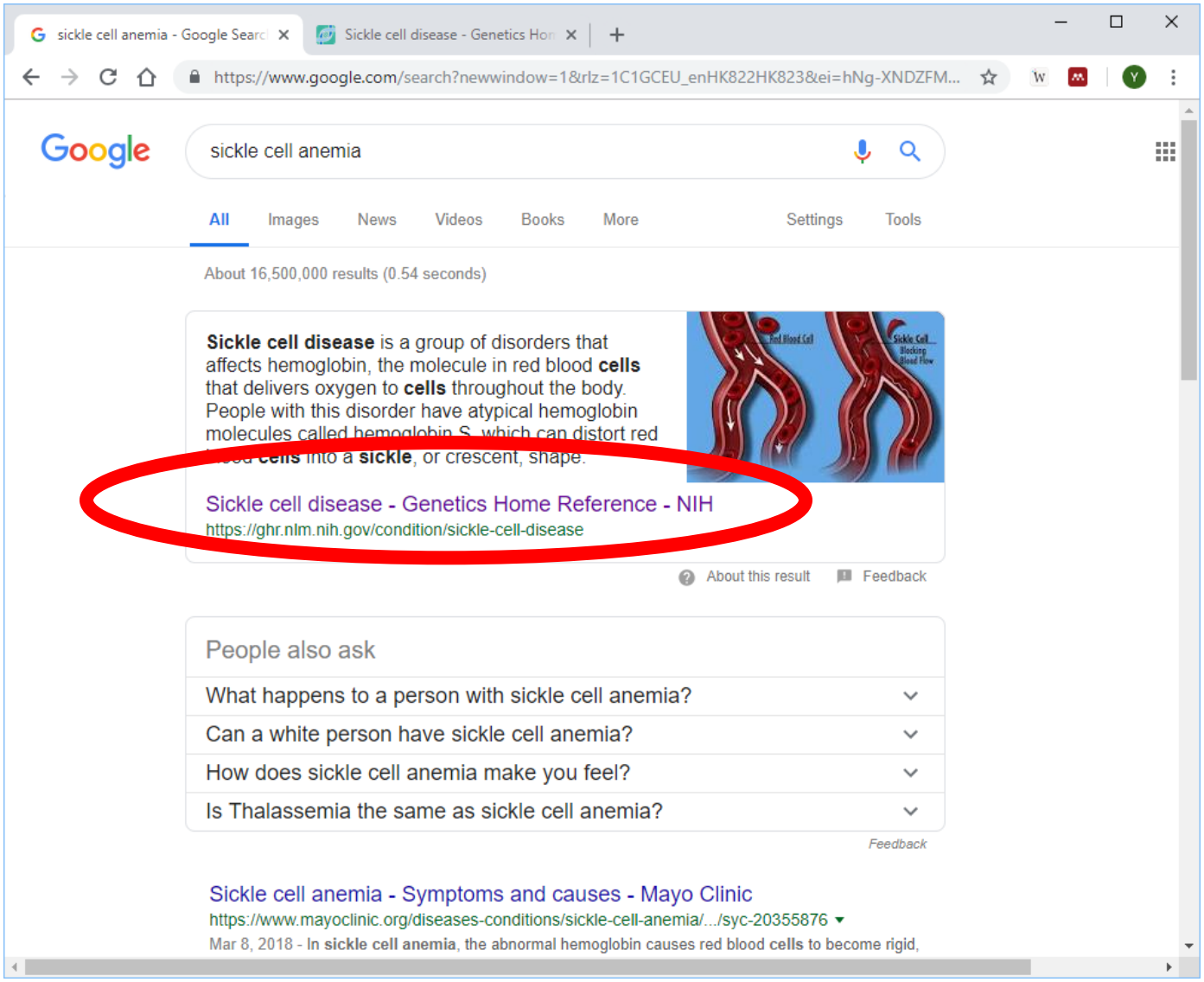
Normal human hemoglobin



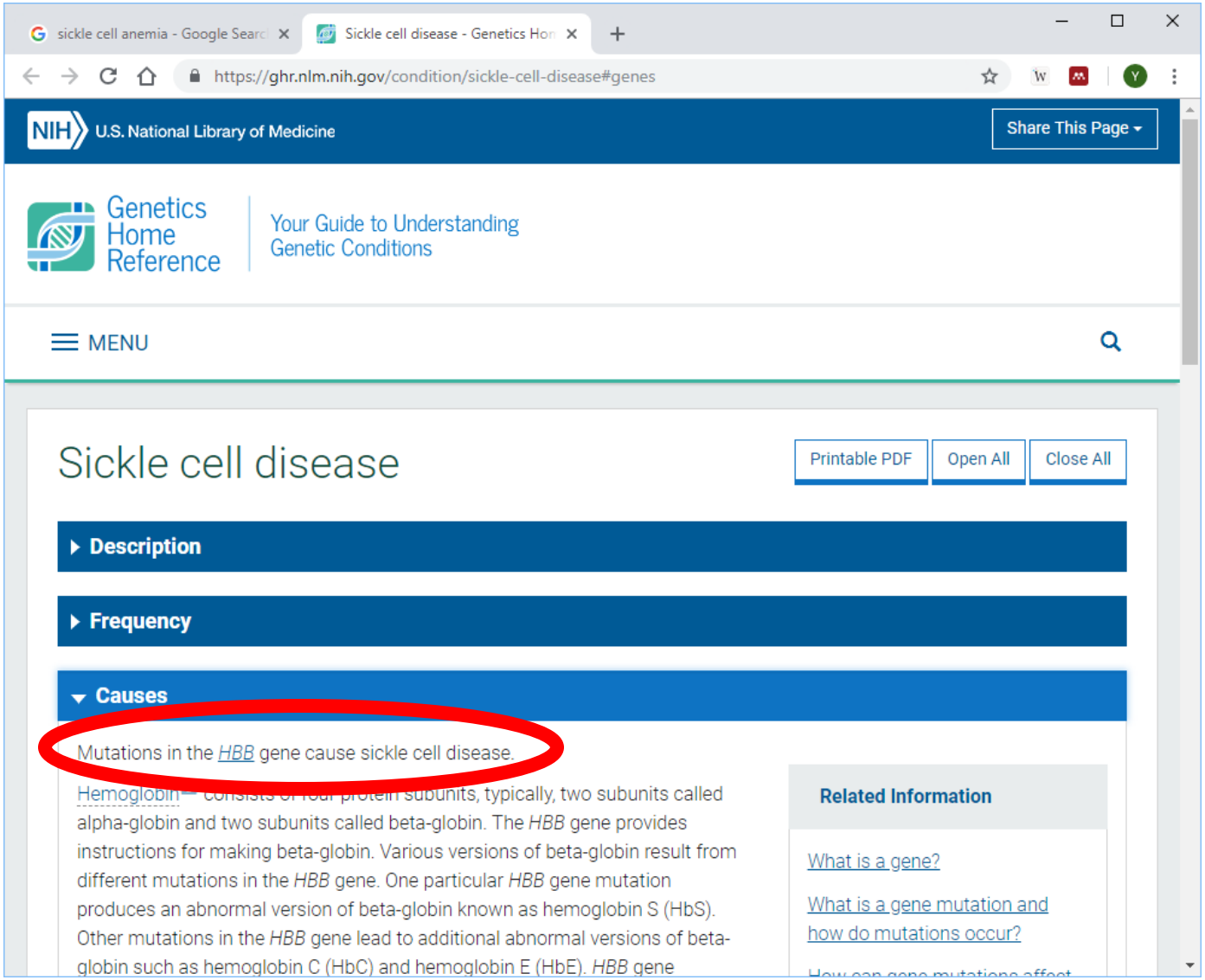
A common mutated form of human hemoglobin that causes Sickle-cell Anemia

Image credit: Protein Data Bank, entries 4HHB and 2HBS, http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/Gifs/HBBmutseq_2.gif

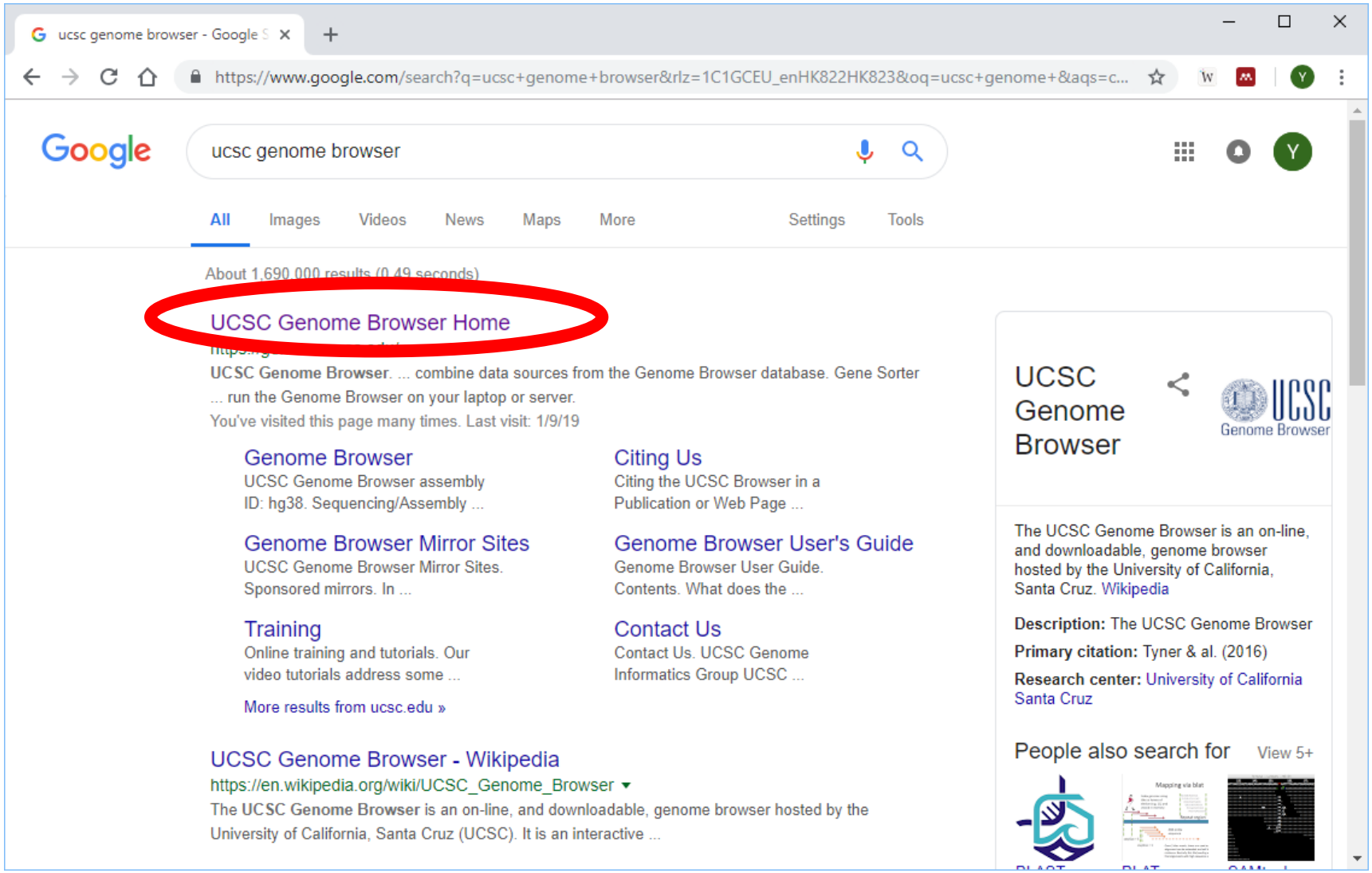
Find the gene that causes the disease



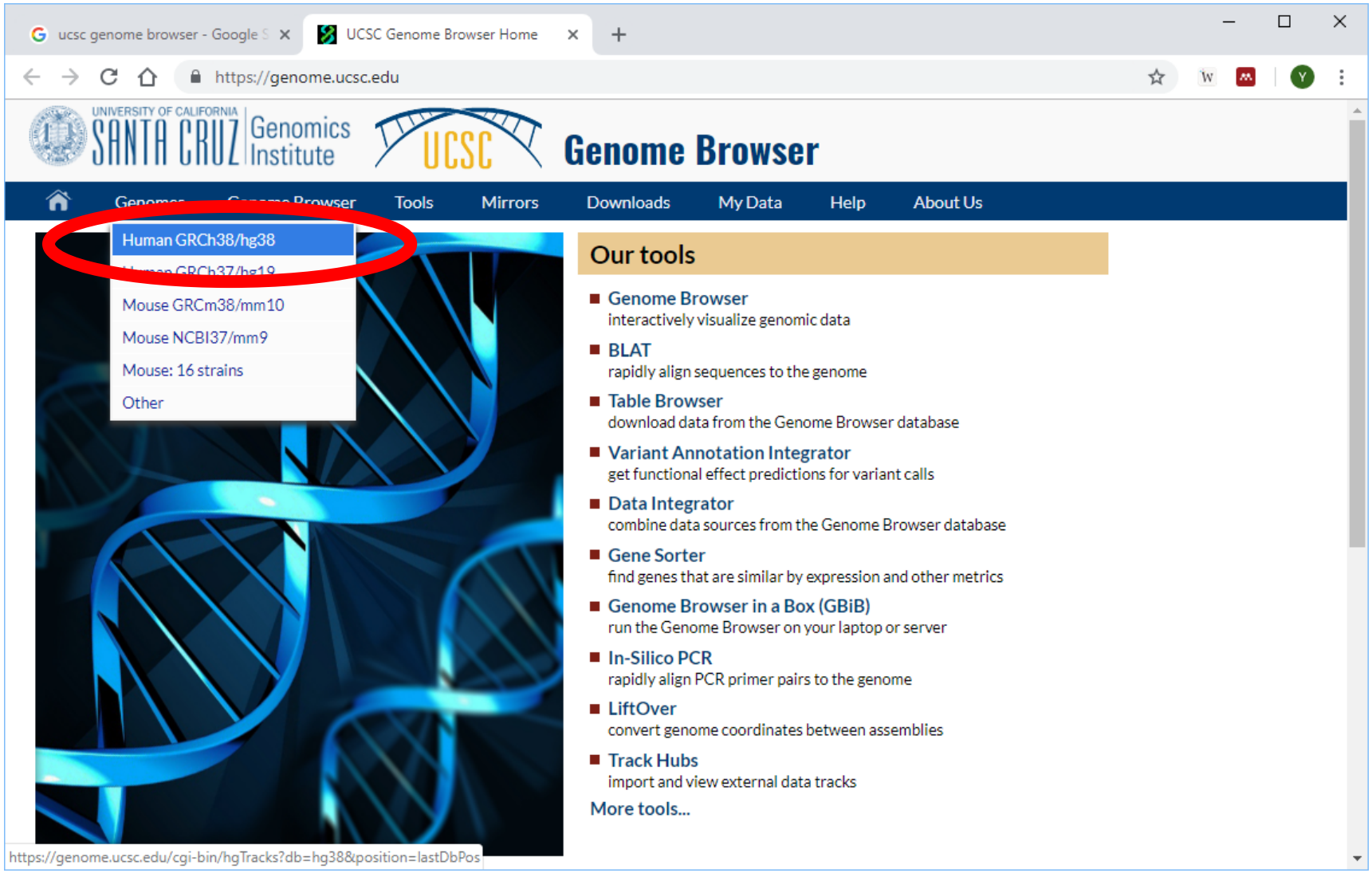
Find the gene that causes the disease



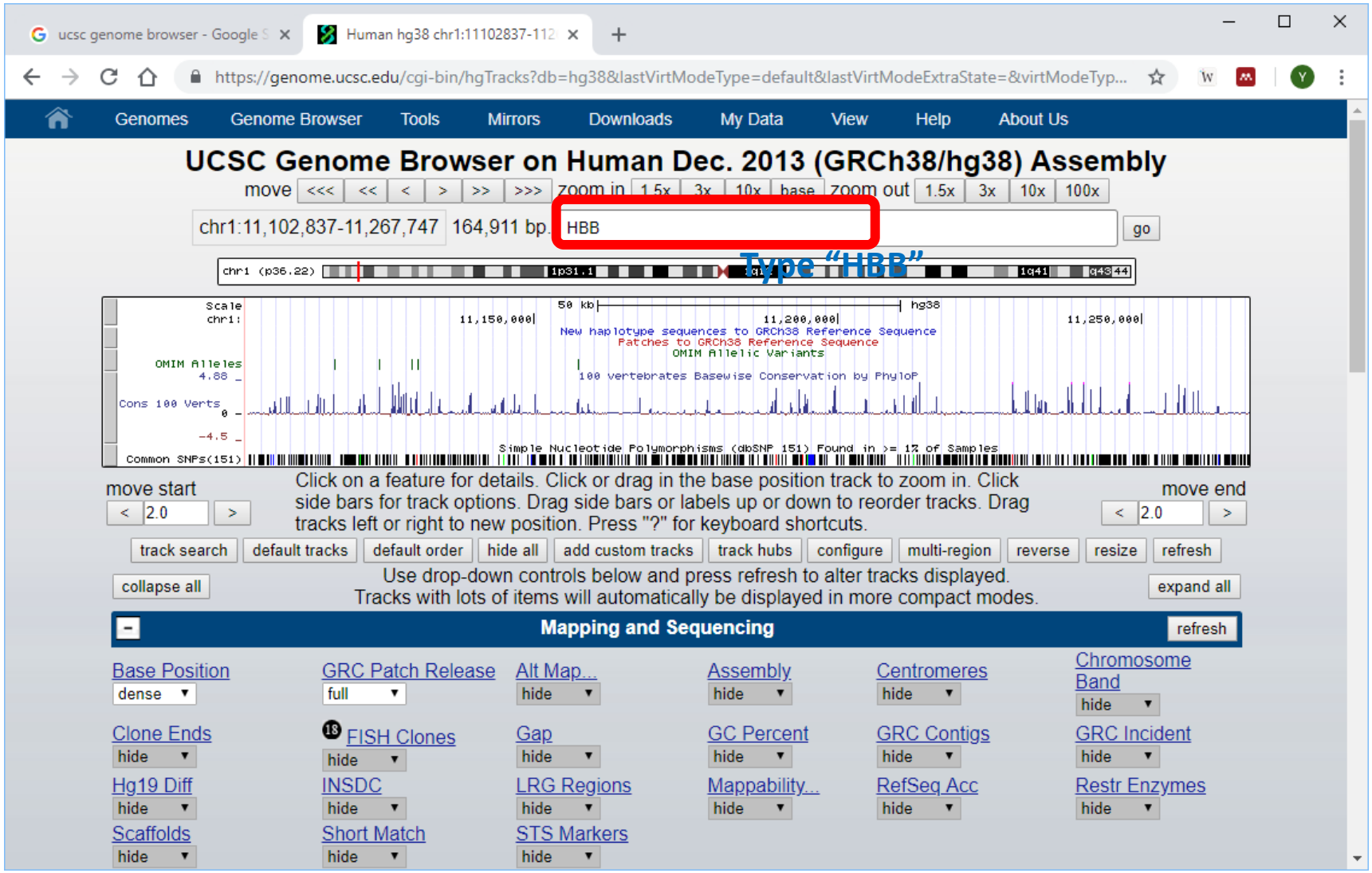
UCSC Genome Browser



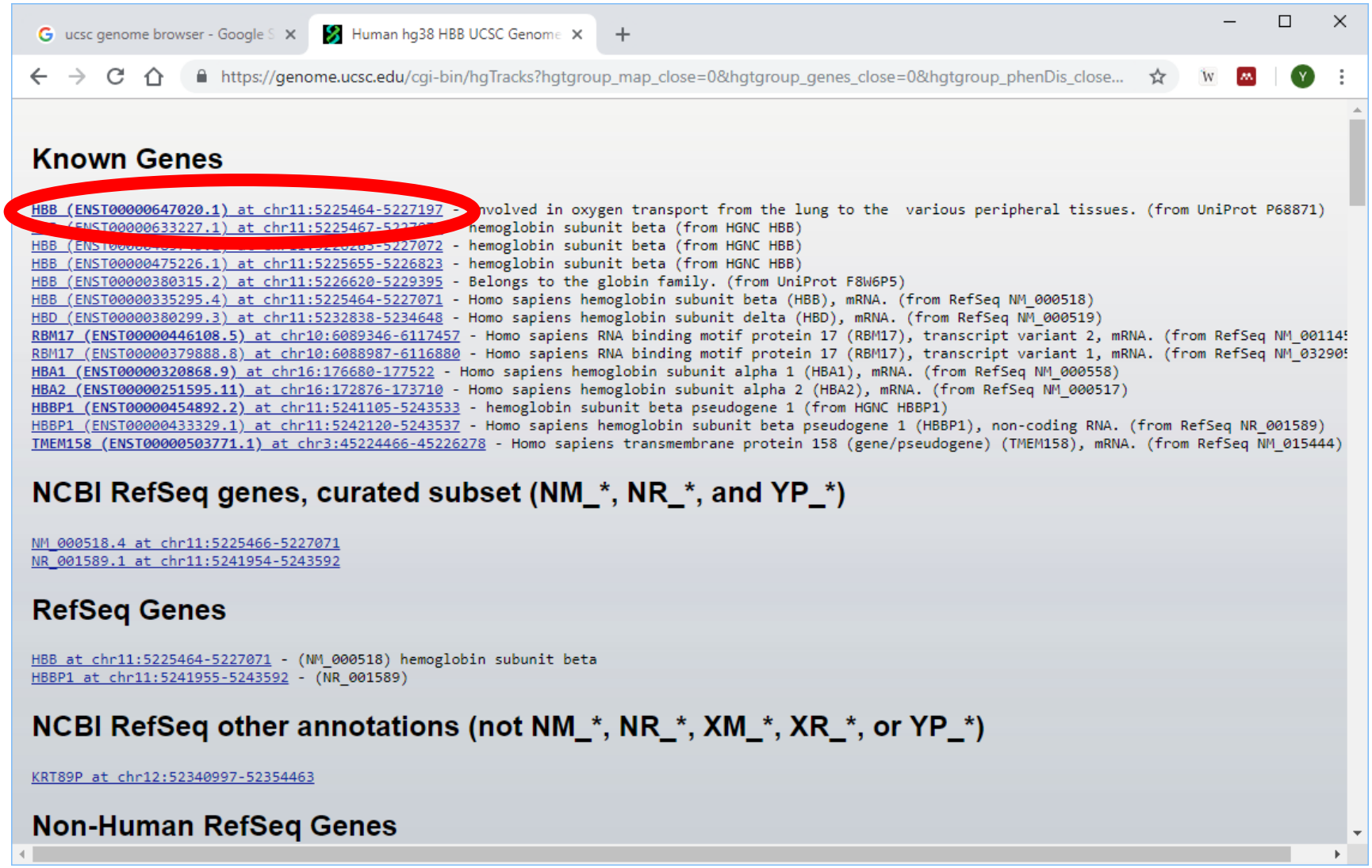
UCSC Genome Browser



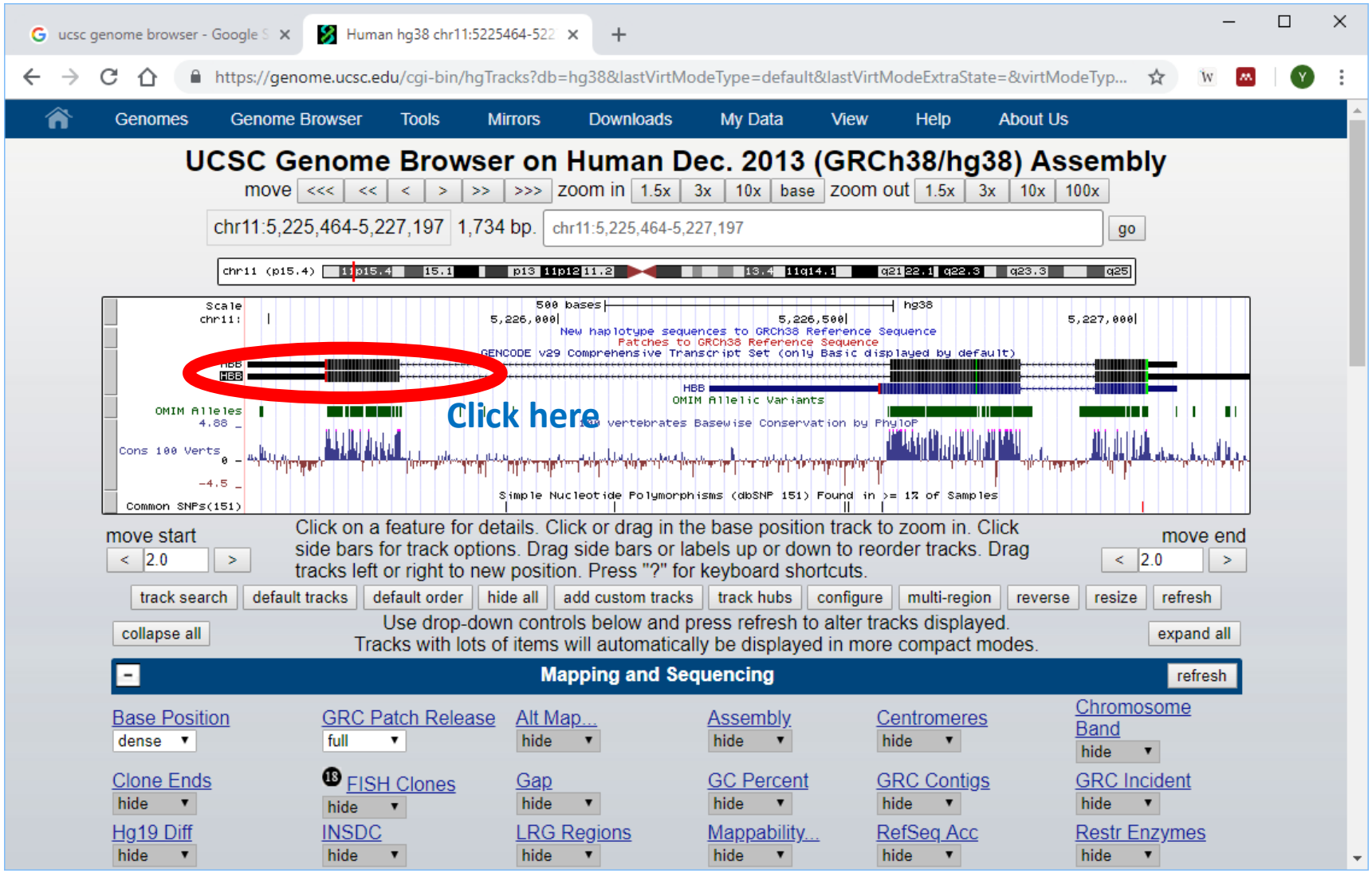
Search on UCSC Genome Browser



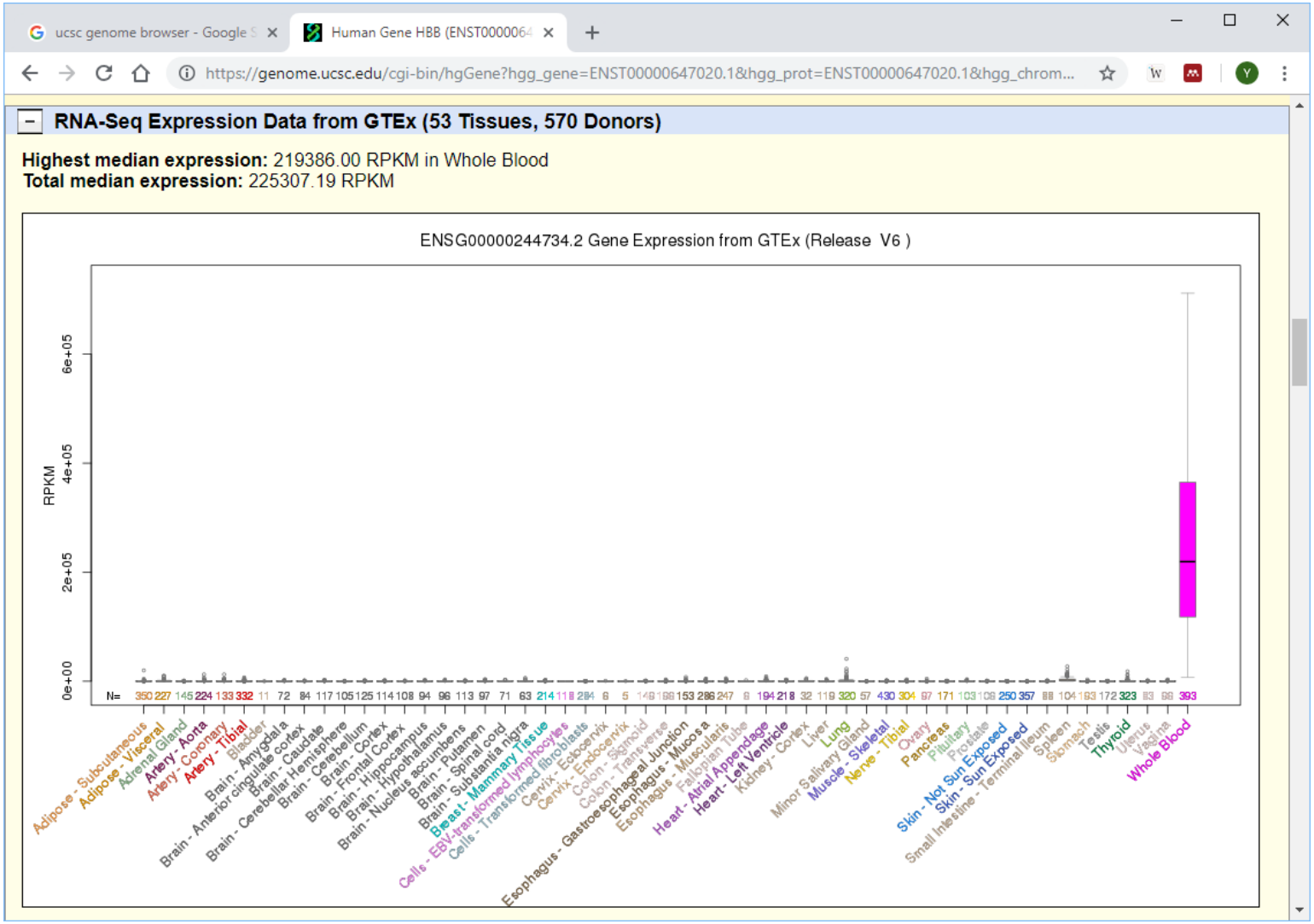
View target gene on Genome Browser



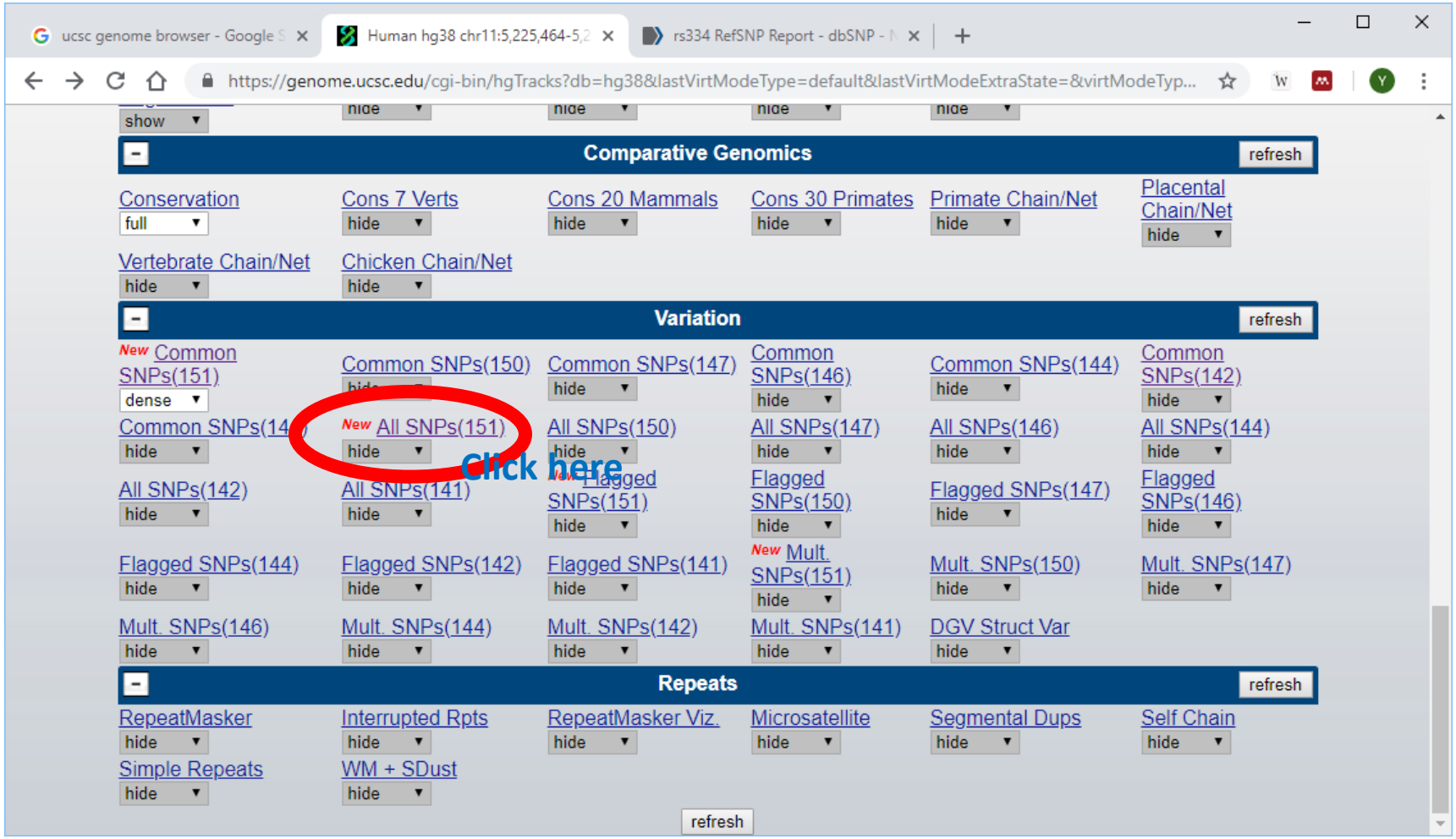
View target gene on Genome Browser



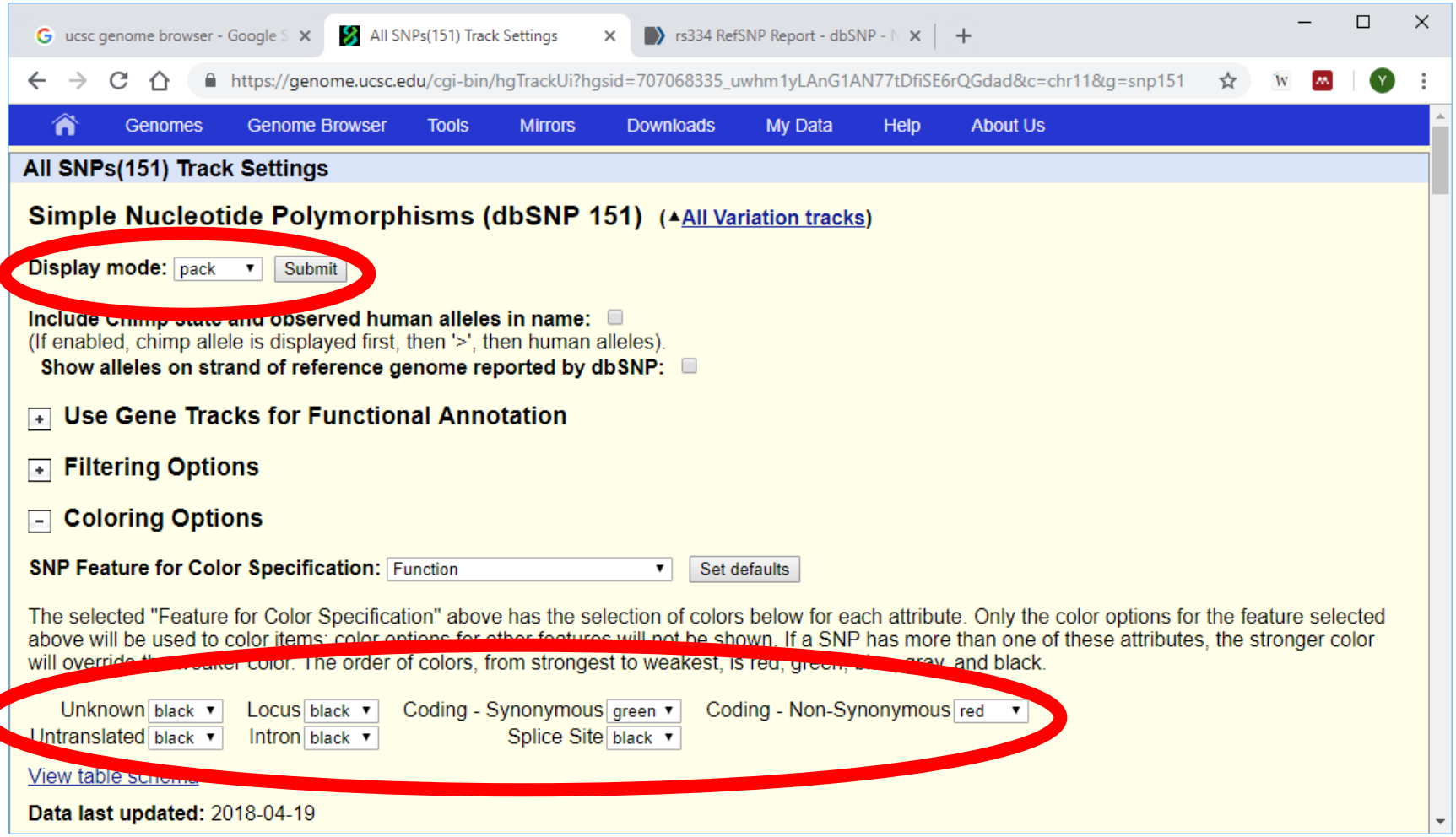
View RNA expression level



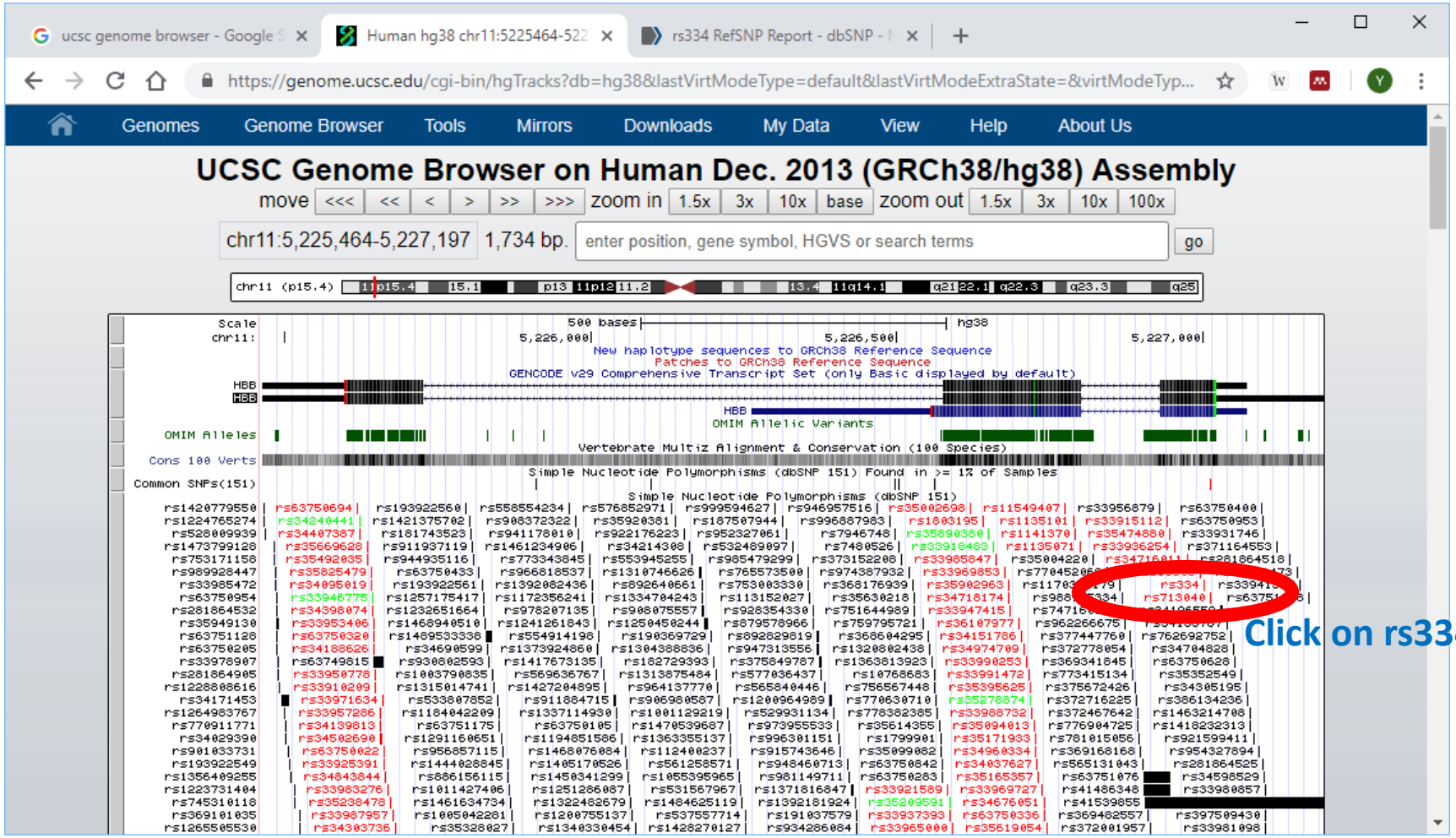
Change settings of Genome Browser



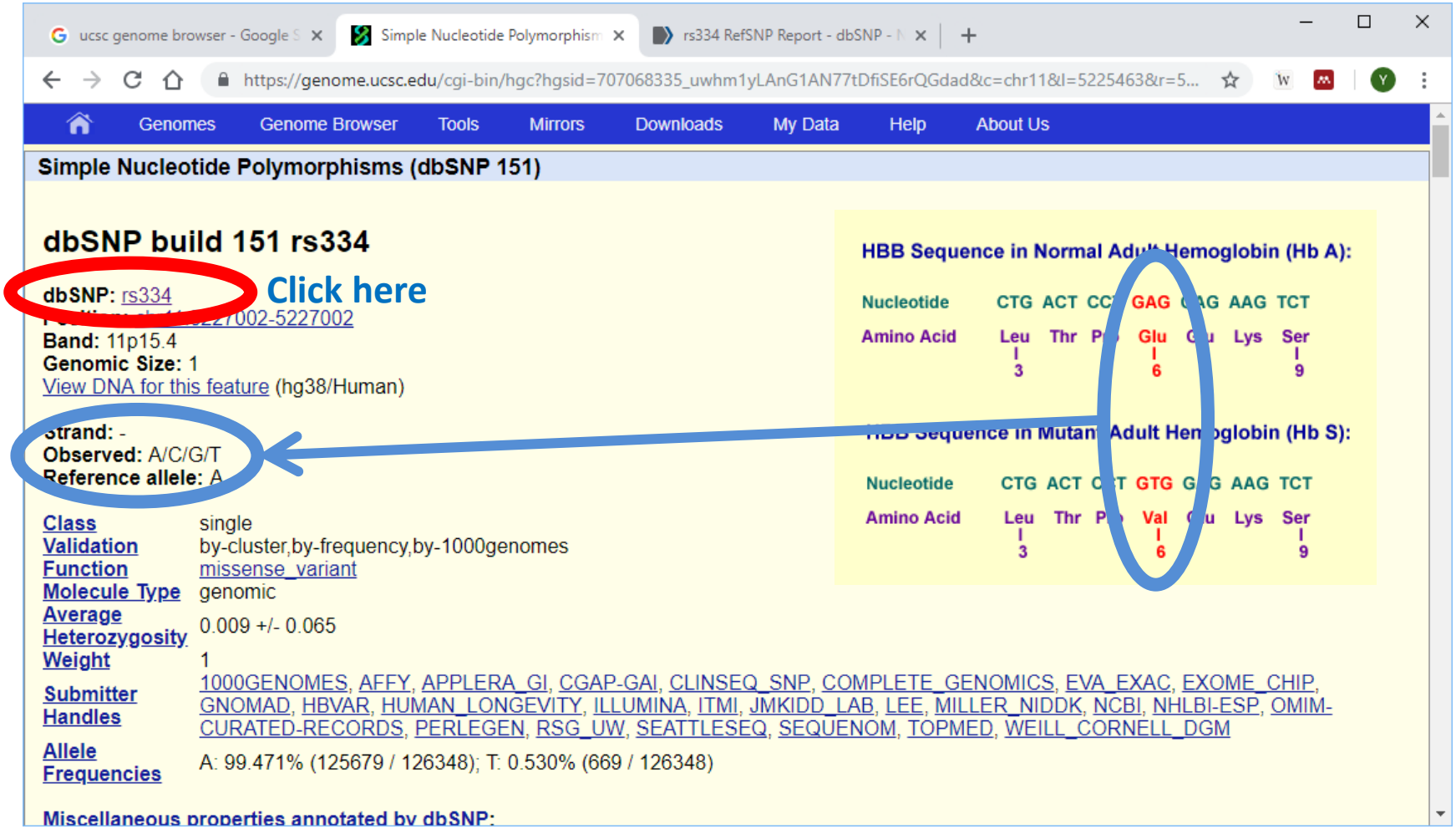
Change settings of Genome Browser



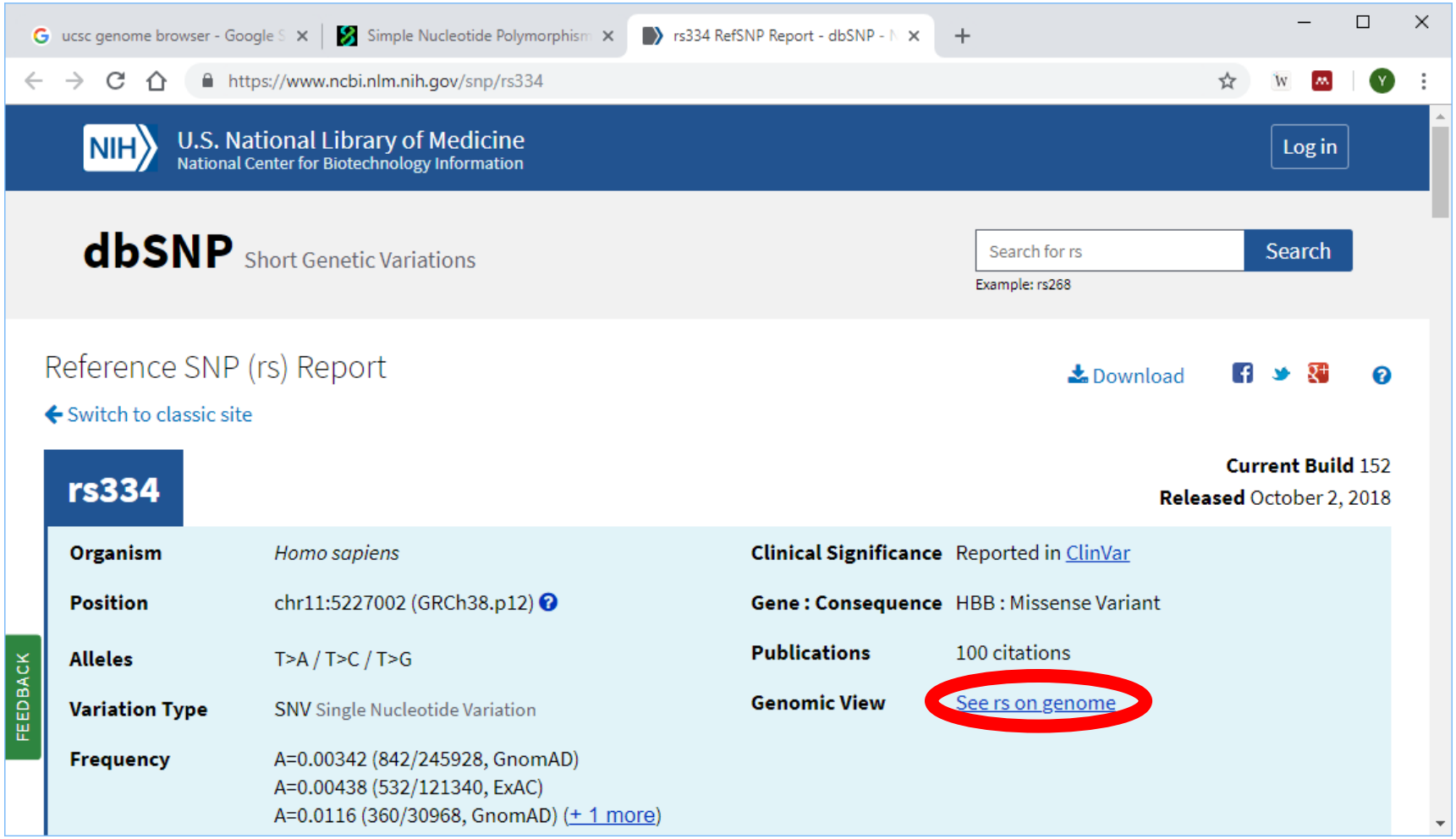
Show the non-synonymous mutations



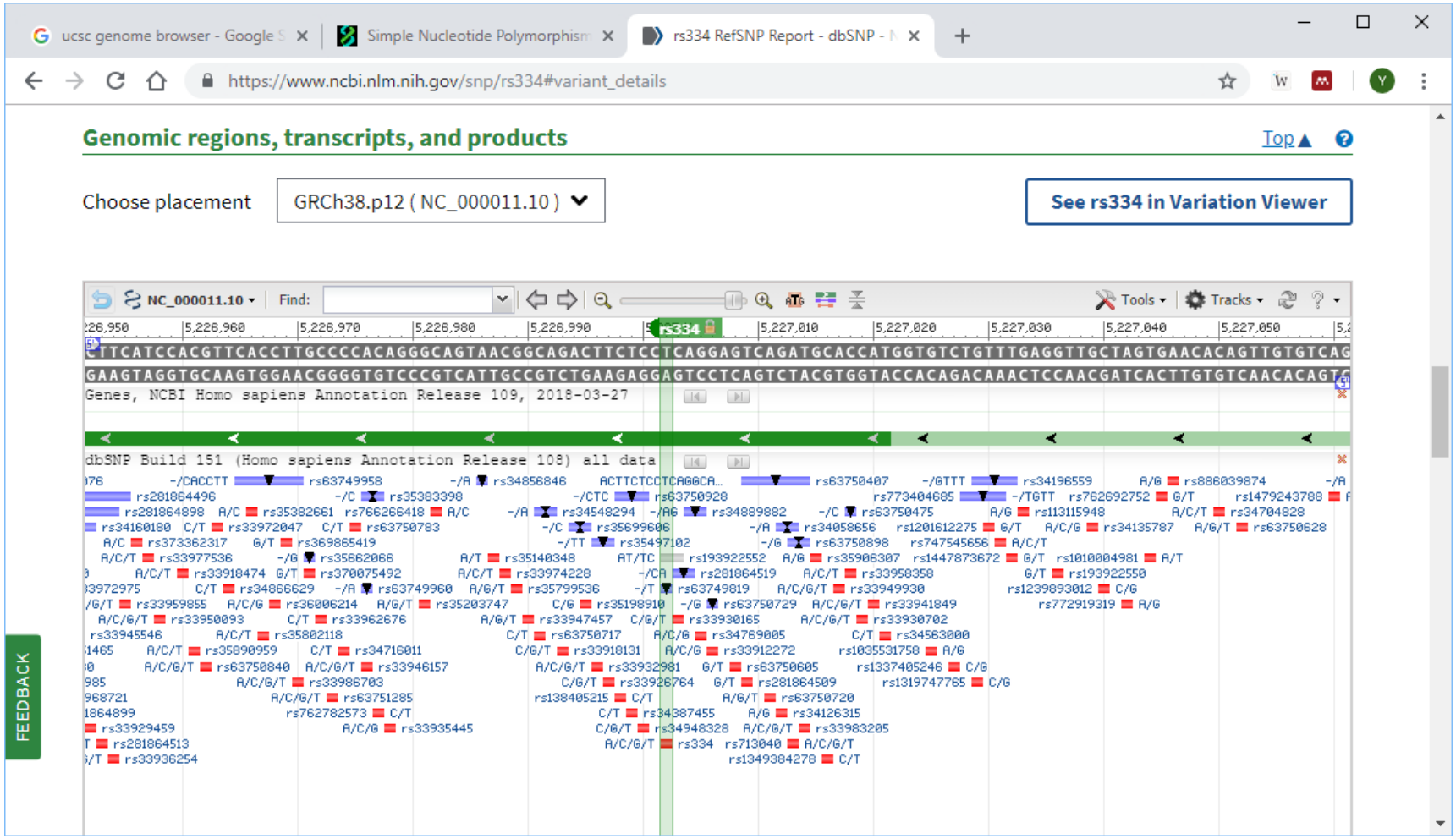
More about the SNP rs334



Go to dbSNP to view the SNP



Go to dbSNP to view the SNP



Check List

- What is “bioinformatics”?
- Describe a problem in bioinformatics, with relevant data, databases, and algorithms.
- What is Central Dogma?
- What is RNA splicing?
- What is the unit of sequence data?
- How can we perform six-frame translation with a given DNA sequence?