Bioinformatics Practical

BS220 Medical Genetics

Vlad Teif (vteif@essex.ac.uk)

Objectives

Understand DNA sequence alignment and its applications for medical problems. Learn how to use online tools to map a DNA sequence to the human genome and to multiple bacterial genomes using BLAST. Familiarise yourself with the database Online Mendelian Inheritance in Man (OMIM).

Story plot

A patient is in a hospital in a critical condition. Medical doctors have extracted some pieces of DNA or RNA from patient's blood and have to decide, as a matter of life or death, what's going on with the patient. The patient is being treated for a genetic disease, cystic fibrosis, but the current symptoms cannot be simply explained by this diagnosis. In this situation, in addition to classic tests, a new test has been performed: all nucleic acids extracted from patient's blood plasma have been sequenced. You are provided with two sequences resulting from this experiment: sequence A and sequence B. You have two hours to analyse these and decide what these sequences mean for the patient's medical condition and how to save life.[i]

Introduction

The **Story Plot** and the **Plan of the Practical** above contain several important terms, and before we proceed let's make sure you understand their meaning:

Genetic disease is a genetic problem caused by one or more abnormalities formed in the genome. Some of them are caused by *Mendelian inheritance.*

Cell-free DNA consists of degraded DNA fragments released to the blood plasma (the liquid part of blood that does not include blood cells). cfDNA pieces can come from apoptotic (dead) cells from all parts of the body. Human blood plasma normally should not contain any foreign DNA, only the DNA from the dead cells of this organism. E.g. if bacterial or viral DNA or RNA are present in blood, it may indicate infection and even sepsis.

Sequencing is the experimental procedure of determining the nucleotide sequence in DNA.

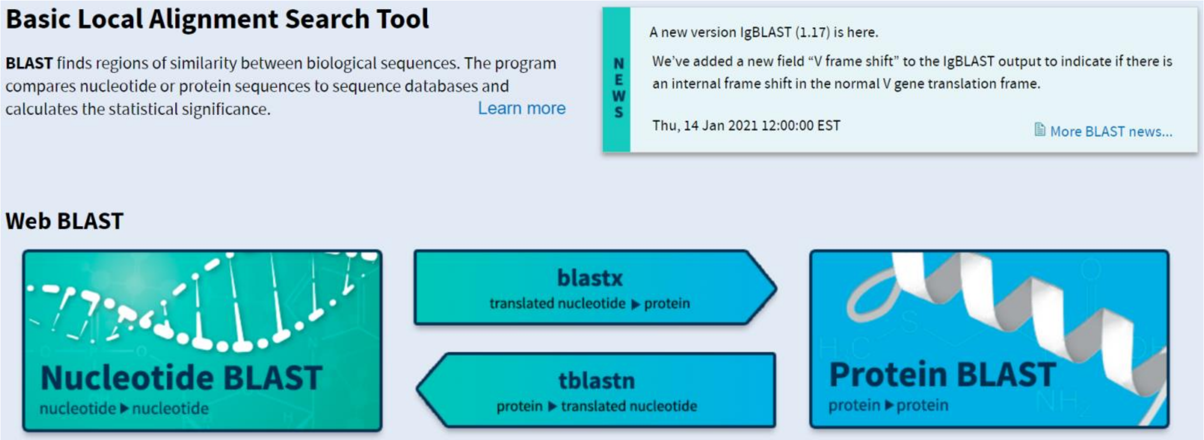
Alignment (mapping) is the process where you *align* (*map*) a given DNA sequence to some other DNA sequence. For example, you could compare a single short sequence to the long sequence of the human genome (~3 billion nucleotides), and ask a question, whether the human genome contains regions that have the same (or similar) sequences as our sequence of interest. If such region(s) exist in the human genome, then you can ask where these regions are located, and which of these regions better matches to our sequence of interest.

Task 1. Map sequence A to the human genome using BLAST

Here is "sequence A":

```
AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGA
TTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGATACAGAAGCGTC
AAGCATGCCAACTAGAAGAGGTAAGAAACTATGTGAAAACTTTTTGATTATGCATATGAAC
```

1.1. Let's go to the BLAST web site: https://blast.ncbi.nlm.nih.gov



Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

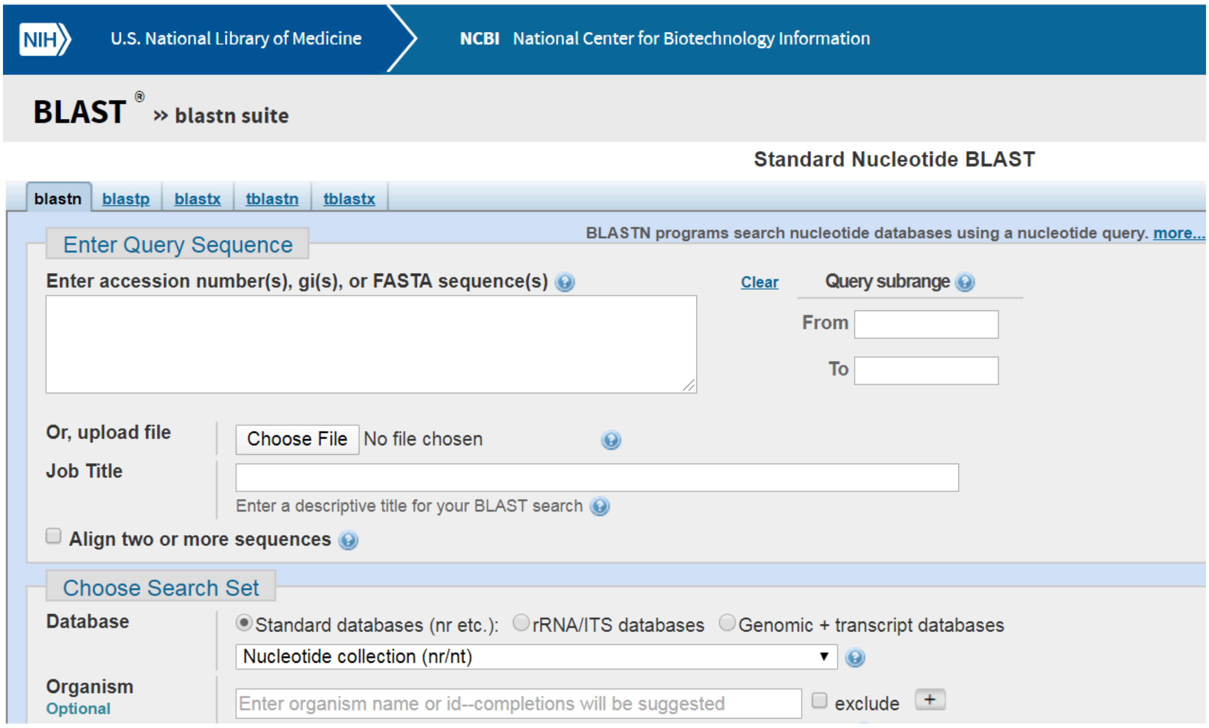
A new version IgBLAST (1.17) is here.

We've added a new field "V frame shift" to the IgBLAST output to indicate if there is an internal frame shift in the normal V gene translation frame.

Thu, 14 Jan 2021 12:00:00 EST

Web BLAST

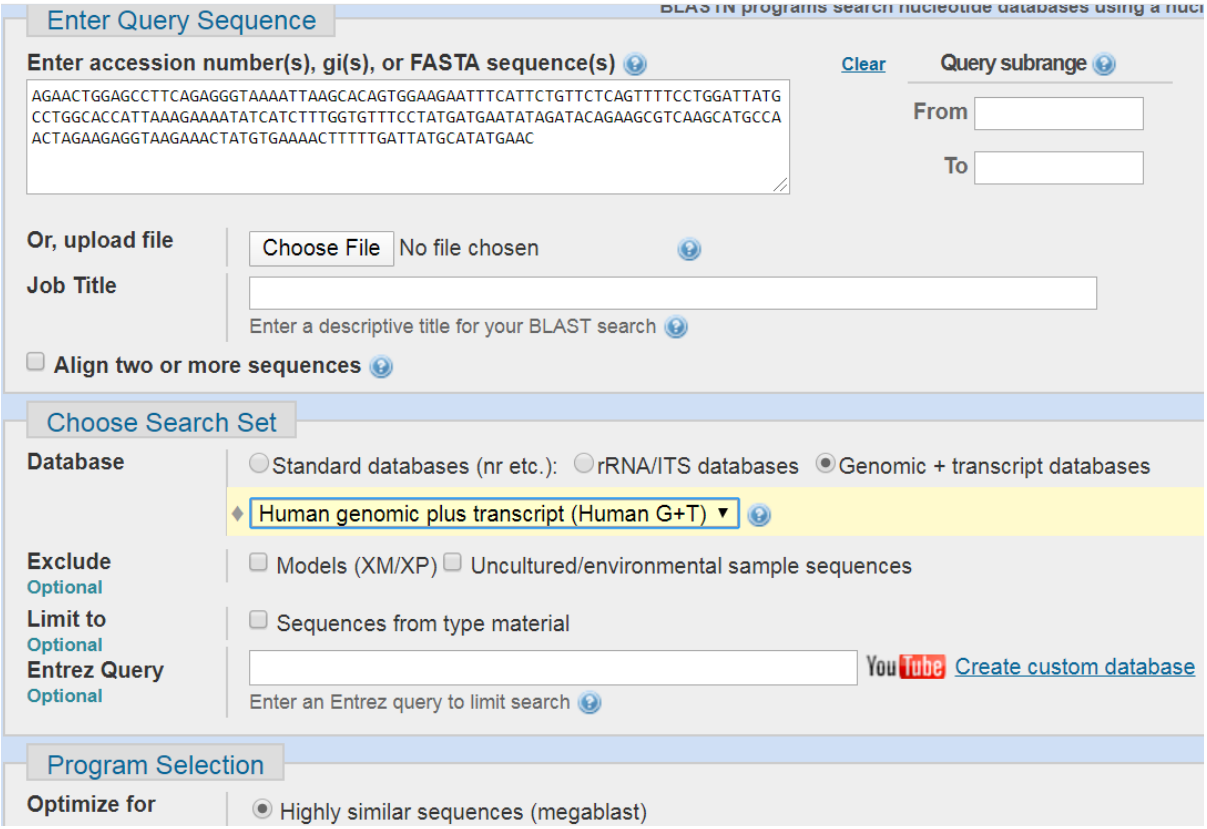
1.2. Select “Nucleotide BLAST”:


NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
BLAST® » blastn suite
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query. more...
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
□ Align two or more sequences
Choose Search Set
Database
● Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
□ exclude
+

1.3. Paste “Sequence A” in the form;

I the menu “Database” select “Genomic + transcript databases”;

In the drop-down menu under “Database” select “Human genomic plus transcript (G+T)”:



Enter accession number(s), gi(s), or FASTA sequence(s)

```
AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGATTATG
CCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGATACAGAAGCGTCAAGCATGCCA
ACTAGAAGAGGTAAGAAACTATGTGAAAACTTTTTGATTATGCATATGAAC
```

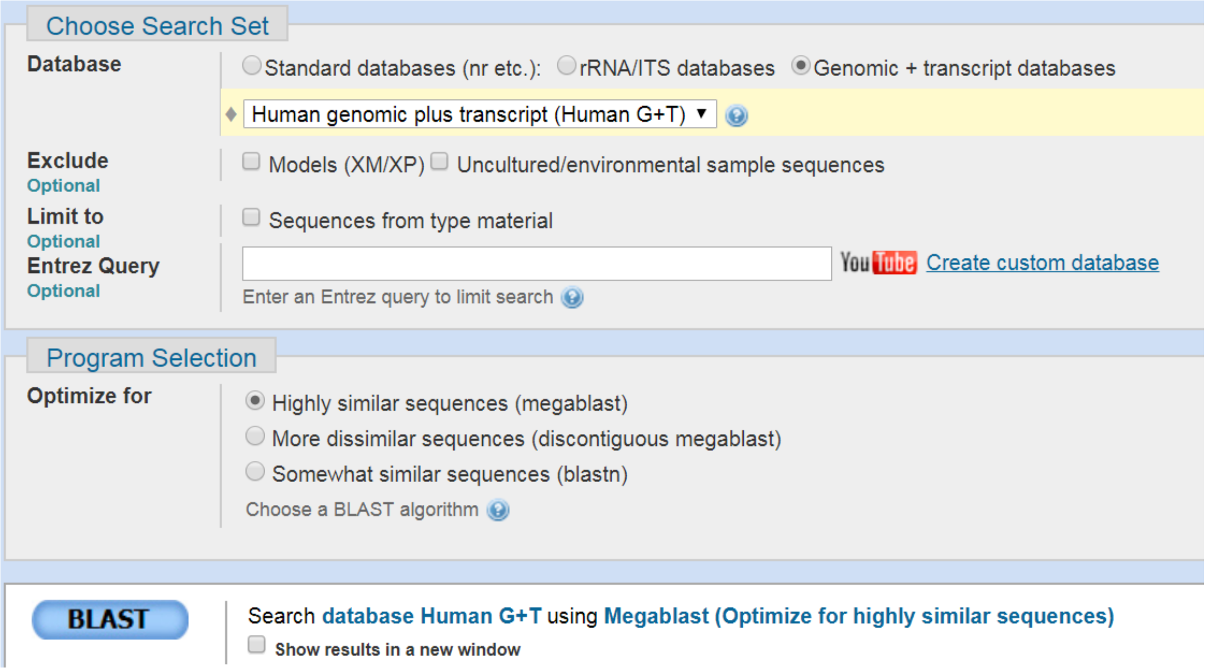
Database: ○ Standard databases (nr etc.): ○ rRNA/ITS databases ● Genomic + transcript databases

Human genomic plus transcript (Human G+T)

Optimize for: ● Highly similar sequences (megablast)

BLASTN programs search nucleotide databases using a nucl
Enter Query Sequence
Clear
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
Enter a descriptive title for your BLAST search
□ Align two or more sequences
Choose Search Set
Exclude
Optional
□ Models (XM/XP) □ Uncultured/environmental sample sequences
Limit to
Optional
□ Sequences from type material
Entrez Query
Optional
YouTube
Create custom database
Enter an Entrez query to limit search
Program Selection

1.4. In the section “Program selection”, select “Highly similar sequences (megablast)”:



Choose Search Set

Database: ○ Standard databases (nr etc.): ○ rRNA/ITS databases ● Genomic + transcript databases

Human genomic plus transcript (Human G+T) ▾

Exclude (Optional): □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to (Optional): □ Sequences from type material

Entrez Query (Optional): Enter an Entrez query to limit search

Program Selection

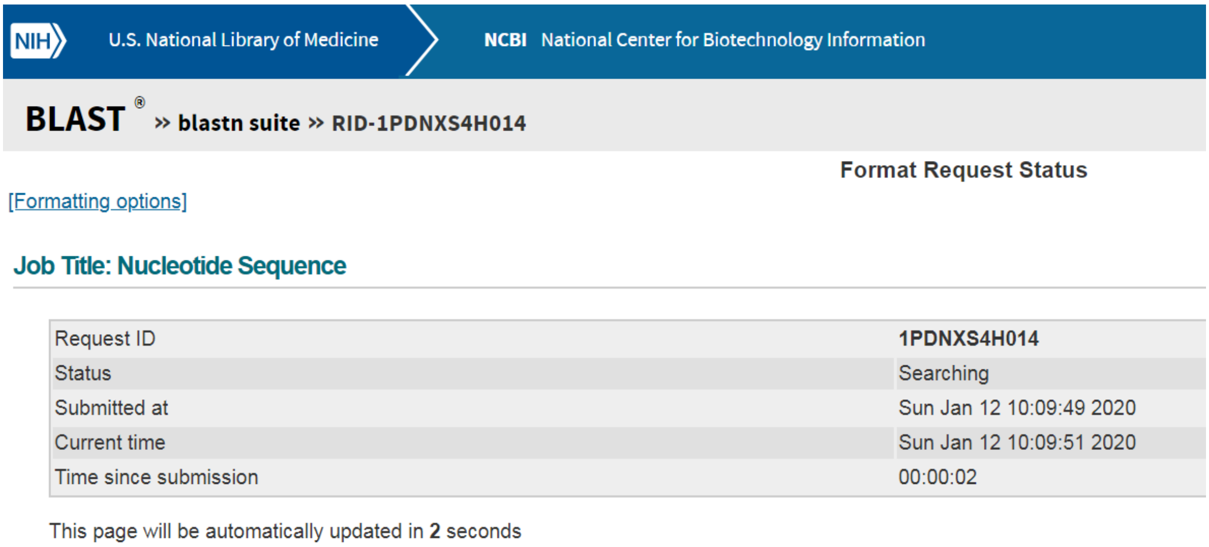
Optimize for:
- ● Highly similar sequences (megablast)
- ○ More dissimilar sequences (discontiguous megablast)
- ○ Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Search database Human G+T using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

1.5. Now all parameters are selected, and we can press the “BLAST” button to start analysis:



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information

BLAST® » blastn suite » RID-1PDNXS4H014

Format Request Status

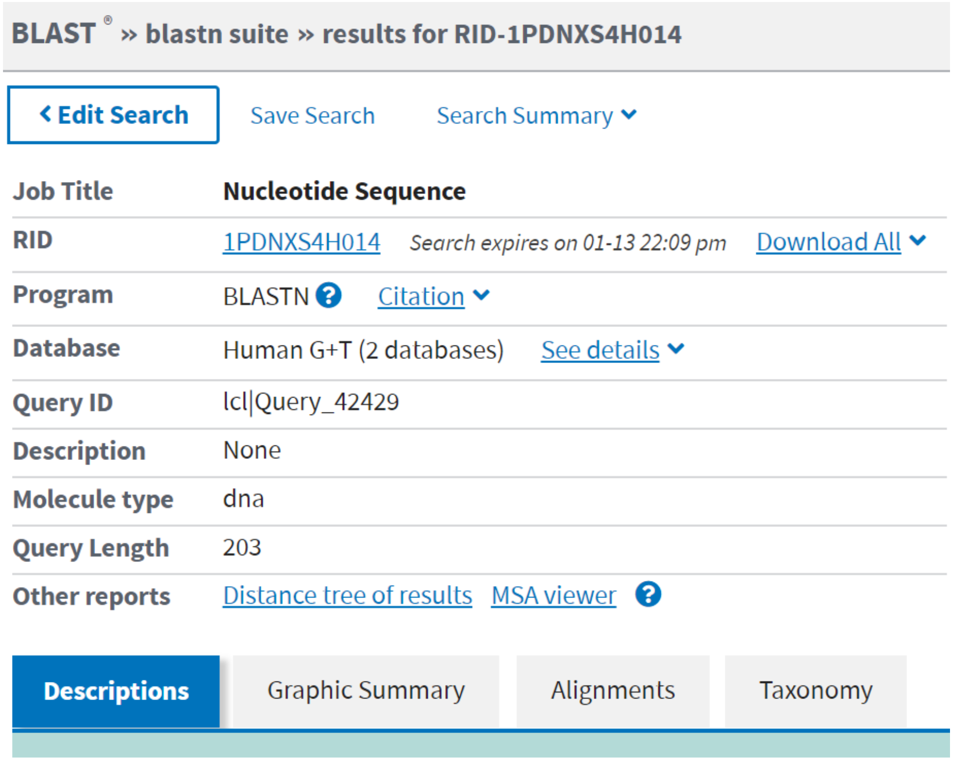
[Formatting options]

Job Title: Nucleotide Sequence

Request ID	1PDNXS4H014
Status	Searching
Submitted at	Sun Jan 12 10:09:49 2020
Current time	Sun Jan 12 10:09:51 2020
Time since submission	00:00:02

This page will be automatically updated in **2** seconds

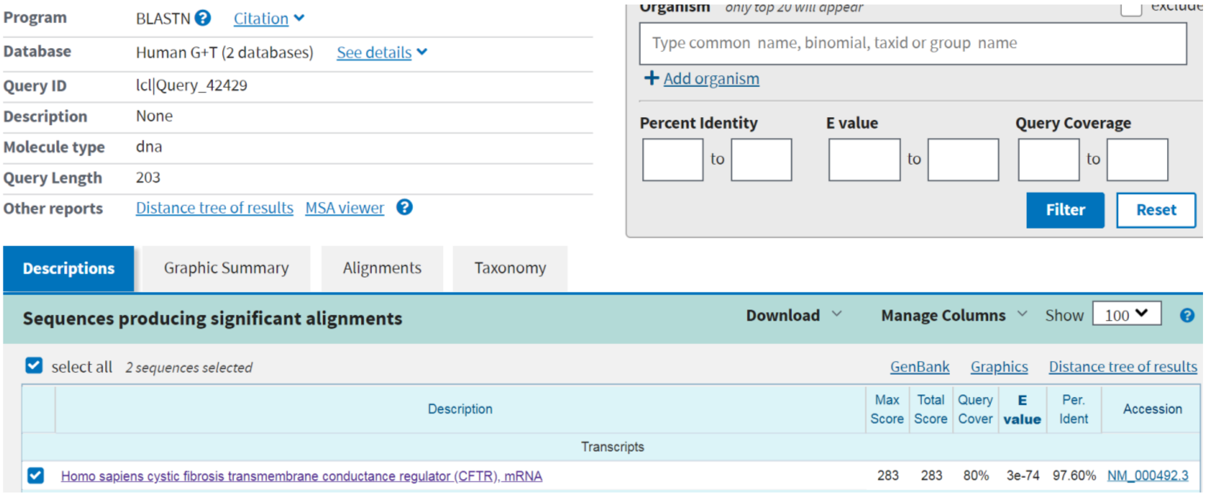
1.6. When the program finishes the analysis you will see the header like this:



BLAST® » blastn suite » results for RID-1PDNXS4H014

Job Title	Nucleotide Sequence
RID	1PDNXS4H014 *Search expires on 01-13 22:09 pm* Download All ⌄
Program	BLASTN ? Citation ⌄
Database	Human G+T (2 databases) See details ⌄
Query ID	lcl|Query_42429
Description	None
Molecule type	dna
Query Length	203
Other reports	Distance tree of results MSA viewer ?

Scroll down to the most important part of this page:



Program	BLASTN ? Citation ⌄
Database	Human G+T (2 databases) See details ⌄
Query ID	lcl|Query_42429
Description	None
Molecule type	dna
Query Length	203
Other reports	Distance tree of results MSA viewer ?

Sequences producing significant alignments

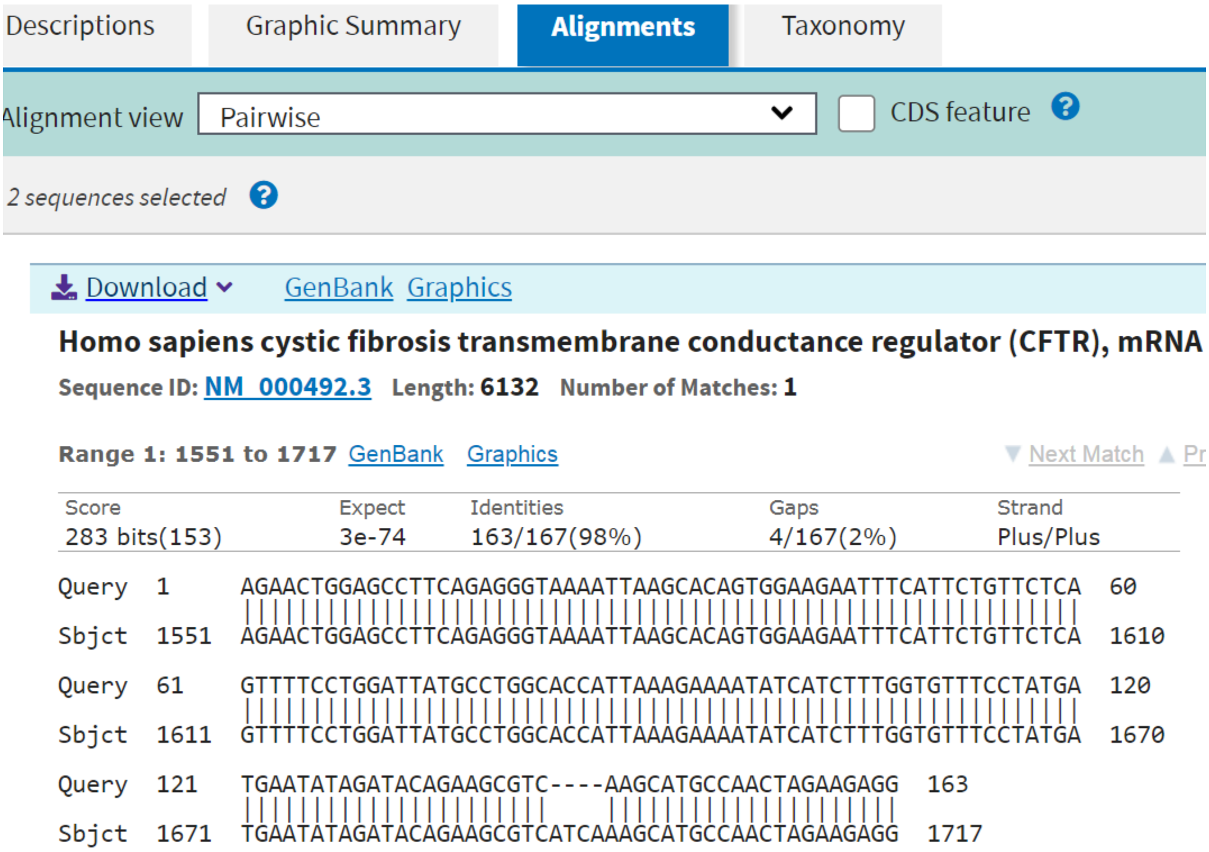
☑ select all *2 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
	Transcripts						
☑	Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR), mRNA	283	283	80%	3e-74	97.60%	NM_000492.3

In this case, the DNA sequence A that we submitted has mapped to the human gene *CFTR* (cystic fibrosis transmembrane conductance regulator). The E-value is 3e-74 – the meaning of this parameter can be roughly understood as the probability to obtain the same result by chance (that is, if we would randomly construct a DNA sequence of 208 nucleotides, the probability that it would map to the same gene with the same similarity would be 3e-74). Thus, the chance that this result would be obtained by a random coincidence is very small, or in other words, the statistical significance of this result is very high.

2. If sequence A mapped to a known human gene, check in the BLAST output whether mutations are present in this gene.

Let’s click on the line “Homo sapience cystic fibrosis transmembrane conductance (CFTR), mRNA”:



Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR), mRNA

Sequence ID: NM_000492.3 Length: 6132 Number of Matches: 1

Range 1: 1551 to 1717 GenBank Graphics ▼ Next Match ▲ Pr

Score	Expect	Identities	Gaps	Strand
283 bits(153)	3e-74	163/167(98%)	4/167(2%)	Plus/Plus

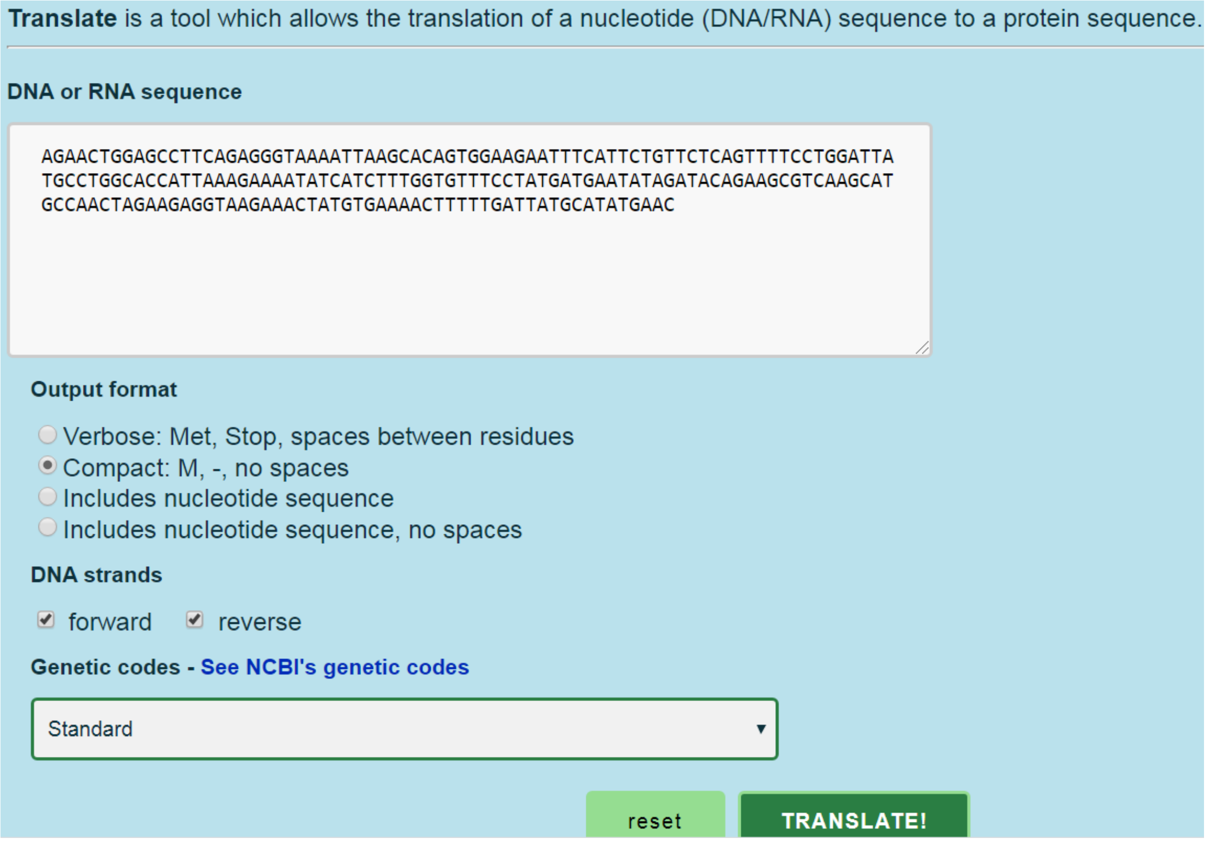
```
Query  1     AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1551  AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  1610

Query  61    GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1611  GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  1670

Query  121   TGAATATAGATACAGAAGCGTC----AAGCATGCCAACTAGAAGAGG  163
             ||||||||||||||||||||||    |||||||||||||||||||||
Sbjct  1671  TGAATATAGATACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGG  1717
```

This graph shows the alignment of “Sequence A” to the *CFTR* gene in the human genome. As you can see from this graph, only four nucleotides do not match (“Sequence A” has a deletion of these four nucleotides which appear in the reference human genome but do not appear in “Sequence A”). The overall identity between these two sequences is 98%, which is a very good match (highly unlikely to come up with such sequence randomly by chance). The four nucleotides which are missing represent a mutation (deletion).

3. Check how this DNA sequence translates into amino-acid sequence using ExPASy:

3.1. Go to the ExPASy web site: https://web.expasy.org/translate/

3.2. Paste the DNA sequence A in the form:



Translate is a tool which allows the translation of a nucleotide (DNA/RNA) sequence to a protein sequence.

DNA or RNA sequence

```
AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCAGTTTTCCTGGATTA
TGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGATGAATATAGATACAGAAGCGTCAAGCAT
GCCAACTAGAAGAGGTAAGAAACTATGTGAAAACTTTTTGATTATGCATATGAAC
```

Output format

○ Verbose: Met, Stop, spaces between residues
● Compact: M, -, no spaces
○ Includes nucleotide sequence
○ Includes nucleotide sequence, no spaces

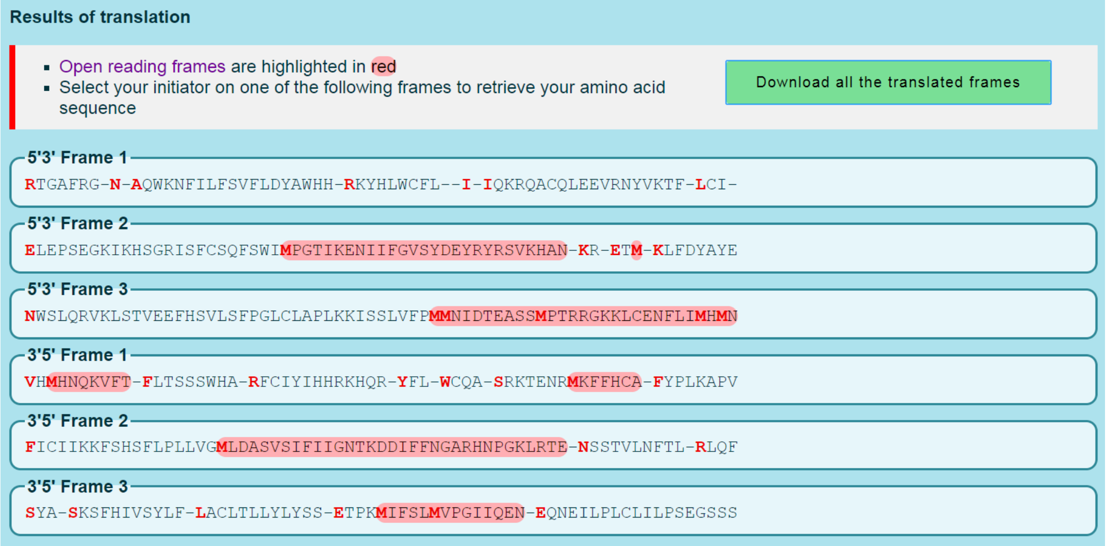
DNA strands

☑ forward ☑ reverse

Genetic codes - See NCBI's genetic codes

Standard ▾

3.3. Get resulting amino-acid sequences for different reading frames:



Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

5'3' Frame 1

```
RTGAFRG-N-AQWKNFILFSVFLDYAWHH-RKYHLWCFL--I-IQKRQACQLEEVRNYVKTF-LCI-
```

5'3' Frame 2

```
ELEPSEGKIKHSGRISFCSQFSWIMPGTIKENIIFGVSYDEYRYRSVKHAN-KR-ETM-KLFDYAYE
```

5'3' Frame 3

```
NWSLQRVKLSTVEEFHSVLSFPGLCLAPLKKISSLVFPMMNIDTEASSMPTRRGKKLCENFLIMHMN
```

3'5' Frame 1

```
VHMHNQKVFT-FLTSSSWHA-RFCIYIHHRKHQR-YFL-WCQA-SRKTENRMKFFHCA-FYPLKAPV
```

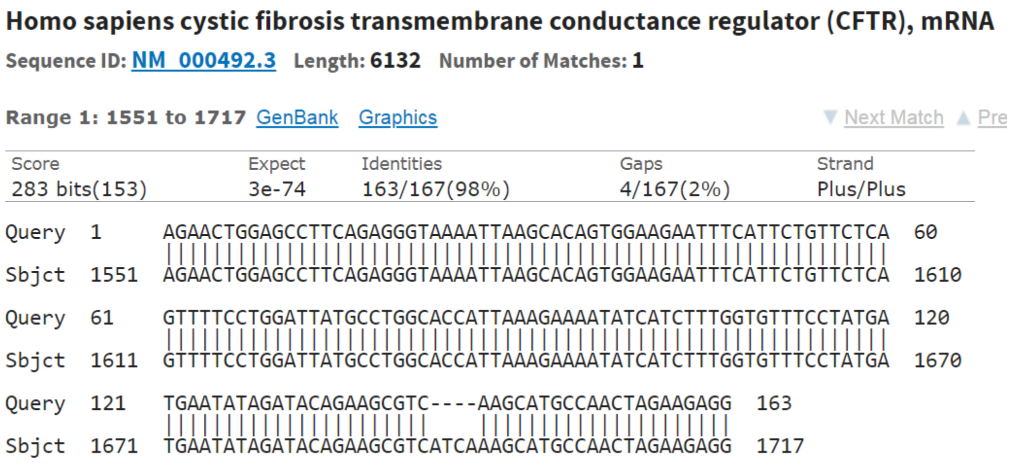
3'5' Frame 2

```
FICIIKKFSHSFLPLLVGMLDASVSIFIIGNTKDDIFFNGARHNPGKLRTE-NSSTVLNFTL-RLQF
```

3'5' Frame 3

```
SYA-SKSFHIVSYLF-LACLTLLYLYSS-ETPKMIFSLMVPGIIQEN-EQNEILPLCLILPSEGSSS
```

3.4. Now let's compare with the wild-type sequence in the reference DNA genome. You can get this sequence from the BLAST alignment output as shown in the figure below:



Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR), mRNA

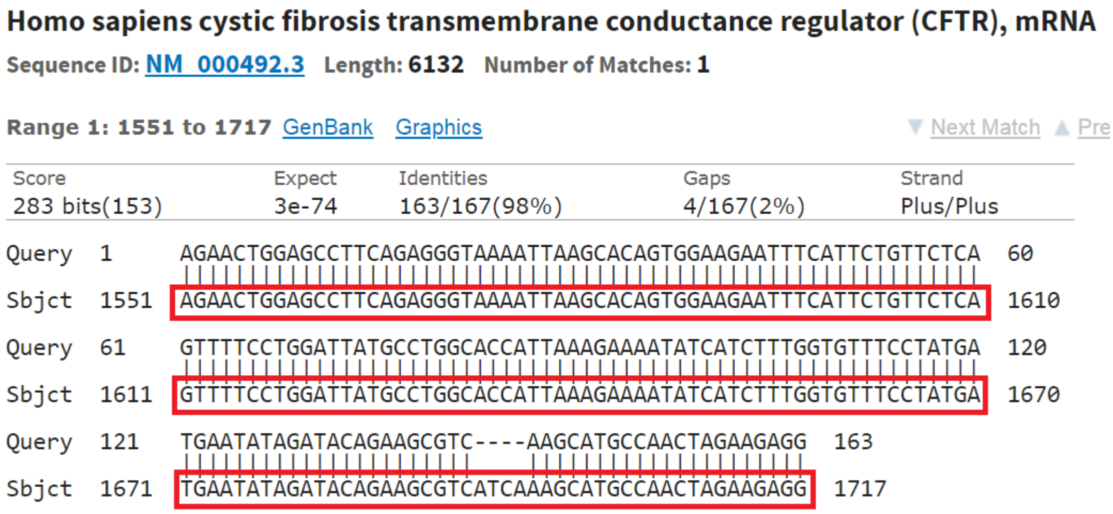
Sequence ID: NM_000492.3 Length: 6132 Number of Matches: 1

Range 1: 1551 to 1717 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
283 bits(153)	3e-74	163/167(98%)	4/167(2%)	Plus/Plus

```
Query  1     AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1551  AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  1610

Query  61    GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1611  GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  1670

Query  121   TGAATATAGATACAGAAGCGTC----AAGCATGCCAACTAGAAGAGG  163
             ||||||||||||||||||||||    |||||||||||||||||||||
Sbjct  1671  TGAATATAGATACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGG  1717
```

3.5. Copy the wild-type DNA sequence (it is shown in the red rectangle in the figure below):



Homo sapiens cystic fibrosis transmembrane conductance regulator (CFTR), mRNA

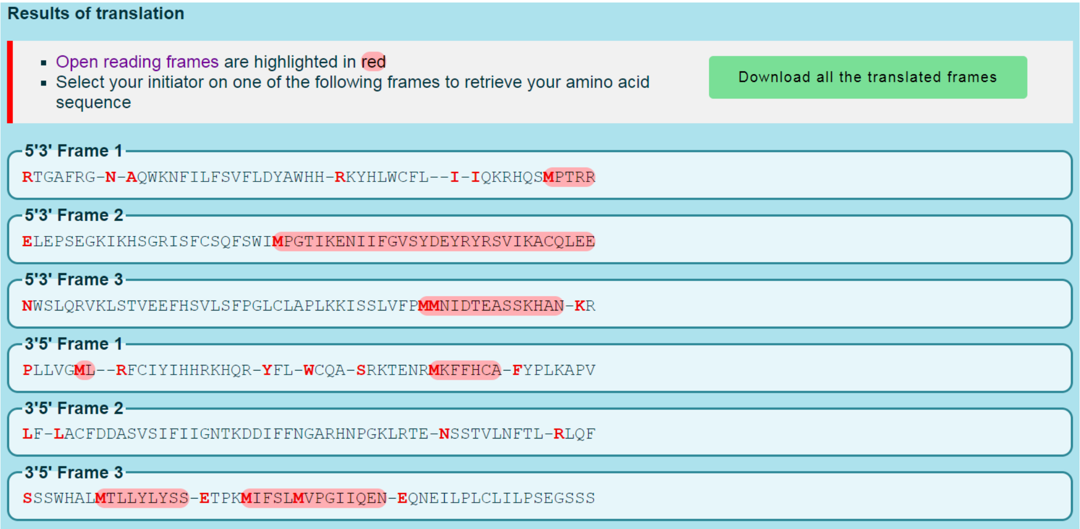
Sequence ID: NM_000492.3 Length: 6132 Number of Matches: 1

Range 1: 1551 to 1717 GenBank Graphics

Score	Expect	Identities	Gaps	Strand
283 bits(153)	3e-74	163/167(98%)	4/167(2%)	Plus/Plus

```
Query  1     AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1551  AGAACTGGAGCCTTCAGAGGGTAAAATTAAGCACAGTGGAAGAATTTCATTCTGTTCTCA  1610

Query  61    GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  120
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1611  GTTTTCCTGGATTATGCCTGGCACCATTAAAGAAAATATCATCTTTGGTGTTTCCTATGA  1670

Query  121   TGAATATAGATACAGAAGCGTC----AAGCATGCCAACTAGAAGAGG  163
             ||||||||||||||||||||||    |||||||||||||||||||||
Sbjct  1671  TGAATATAGATACAGAAGCGTCATCAAAGCATGCCAACTAGAAGAGG  1717
```

3.6. Paste the wild-type DNA sequence to ExPASy (https://web.expasy.org/translate/) and translate it to the amino-acid sequence for all possible reading frames:



Results of translation

- Open reading frames are highlighted in red
- Select your initiator on one of the following frames to retrieve your amino acid sequence

5'3' Frame 1

```
RTGAFRG-N-AQWKNFILFSVFLDYAWHH-RKYHLWCFL--I-IQKRHQSMPTRR
```

5'3' Frame 2

```
ELEPSEGKIKHSGRISFCSQFSWIMPGTIKENIIFGVSYDEYRYRSVIKACQLEE
```

5'3' Frame 3

```
NWSLQRVKLSTVEEFHSVLSFPGLCLAPLKKISSLVFPMMNIDTEASSKHAN-KR
```

3'5' Frame 1

```
PLLVGML--RFCIYIHHRKHQR-YFL-WCQA-SRKTENRMKFFHCA-FYPLKAPV
```

3'5' Frame 2

```
LF-LACFDDASVSIFIIGNTKDDIFFNGARHNPGKLRTE-NSSTVLNFTL-RLQF
```

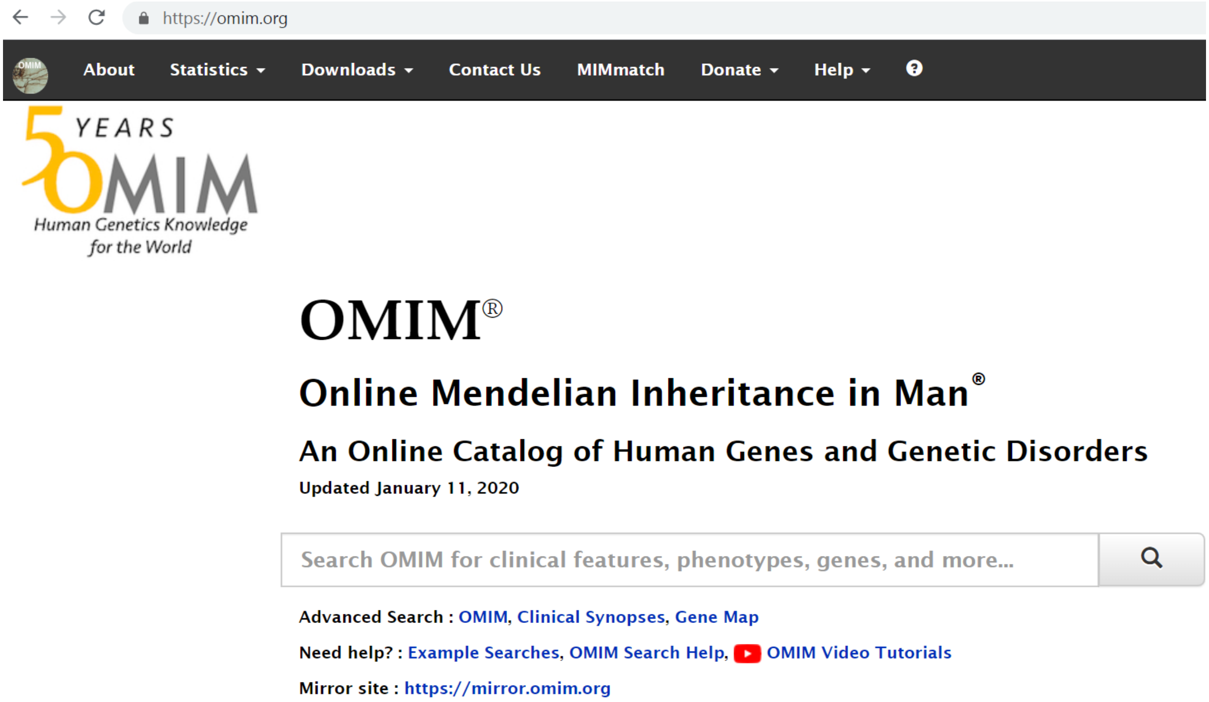
3'5' Frame 3

```
SSSWHALMTLLYLYSS-ETPKMIFSLMVPGIIQEN-EQNEILPLCLILPSEGSSS
```

3.7. Discuss: what amino-acid changes can be caused by this mutation?

4. Check in the OMIM database, which Mendelian disease is associated with this mutation.

4.1. Let's go to the OMIM database: https://omim.org



OMIM®

Online Mendelian Inheritance in Man®

An Online Catalog of Human Genes and Genetic Disorders

Updated January 11, 2020

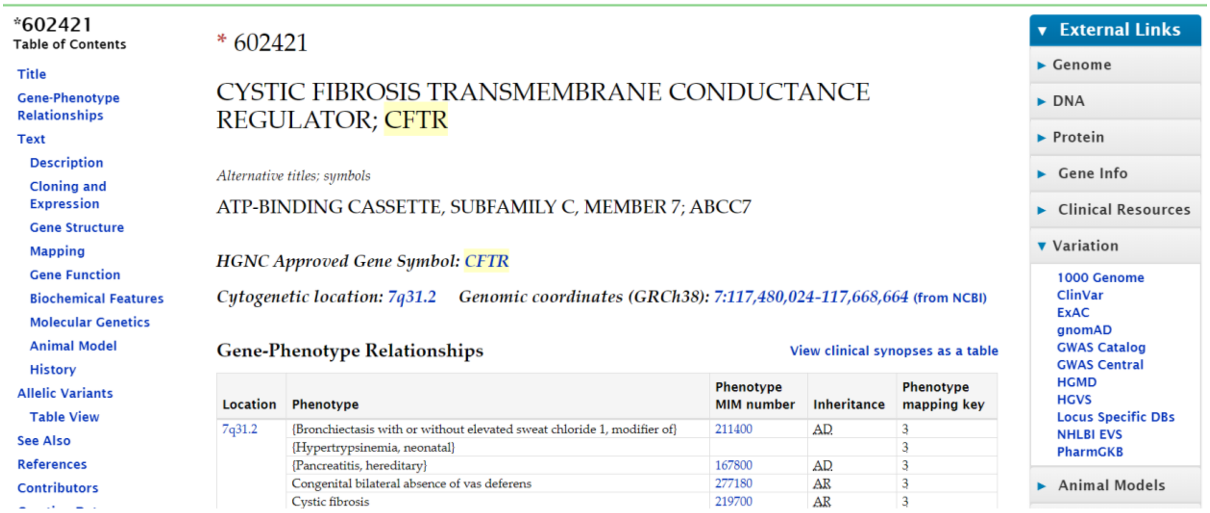
Search OMIM for clinical features, phenotypes, genes, and more...

Advanced Search : OMIM, Clinical Synopses, Gene Map

Need help? : Example Searches, OMIM Search Help, OMIM Video Tutorials

Mirror site : https://mirror.omim.org

4.2. Let's look for the *CFTR* gene in the OMIM database:



* 602421

CYSTIC FIBROSIS TRANSMEMBRANE CONDUCTANCE REGULATOR; CFTR

Alternative titles; symbols

ATP-BINDING CASSETTE, SUBFAMILY C, MEMBER 7; ABCC7

HGNC Approved Gene Symbol: CFTR

Cytogenetic location: 7q31.2 *Genomic coordinates (GRCh38): 7:117,480,024-117,668,664* (from NCBI)

Gene-Phenotype Relationships

View clinical synopses as a table

Location	Phenotype	Phenotype MIM number	Inheritance	Phenotype mapping key
7q31.2	{Bronchiectasis with or without elevated sweat chloride 1, modifier of}	211400	AD	3
	{Hypertrypsinemia, neonatal}			3
	{Pancreatitis, hereditary}	167800	AD	3
	Congenital bilateral absence of vas deferens	277180	AR	3
	Cystic fibrosis	219700	AR	3

4.3. Work independently with this page of the OMIM database to read about possible phenotypes associated with CFTR gene and associated medical information.

Discuss: is cystic fibrosis a recessive or dominant disease? If this patient has a piece of DNA with mutation in CFTR, does it mean she/he has cystic fibrosis? What are its molecular mechanisms?

5*. Try to map sequence B to the human genome using BLAST

Repeat steps 1.1-1.6 to map “Sequence B” to the human genome.

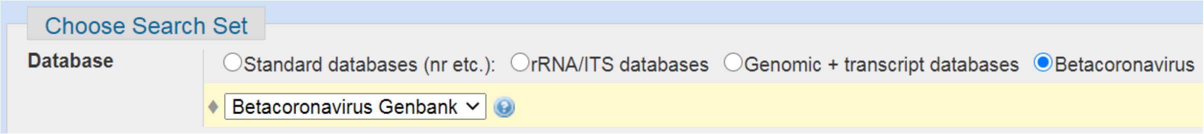
Here is “sequence B”:

```
ATTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAAGGAGGAG
CTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCTGCCTGGTAGCGGGGGATAAC
TATTGGAAACGATAGCTAATACCGCATAAGAGTGGATGTTGCATGACATTTGCTTAAAAGGTGCACTTGC
ATCACTACCAGATGGACCTGCGTTGTATTAGCTAGTTGGTGGGGTAACGGCTCACCAAGGCGACGATACA
TAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGKCCCAGACTCCTACGGGAGGCAGCA
```

Discuss: Did you manage to map “Sequence B” to the human genome? Why?

6. Try to map sequence B to the coronavirus genome using BLAST

Repeat steps 1.1-1.6, but select “Betacoronovirus” as the database:



Choose Search Set

Database ○ Standard databases (nr etc.): ○ rRNA/ITS databases ○ Genomic + transcript databases ● Betacoronavirus

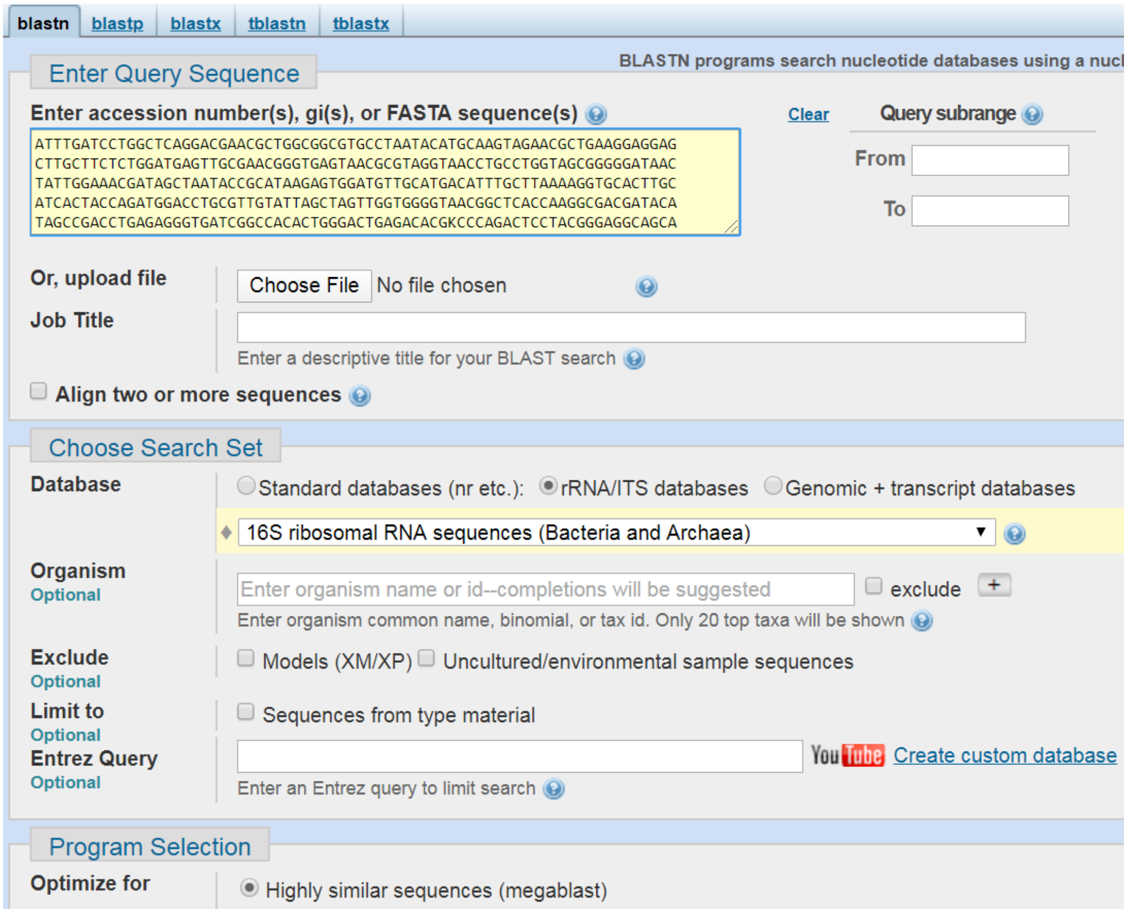
Betacoronavirus Genbank

Discuss: Did you manage to map “Sequence B” to the coronovirus genome? Why?

BTW, it’s possible to say that it’s not RNA from virus even without this analysis! Why?

7. Map “Sequence B” to bacterial and fungal genomes

7.1. Open BLAST (https://blast.ncbi.nlm.nih.gov) and in the section “Database” select “rRNA/ITS databases”. In the drop-down menu select “16S ribosomal RNA sequences (bacteria and fungi)[ii]:



BLASTN programs search nucleotide databases using a nucl

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s)

```
ATTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCTGAAGGAGGAG
CTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCTGCCTGGTAGCGGGGGATAAC
TATTGGAAACGATAGCTAATACCGCATAAGAGTGGATGTTGCATGACATTTGCTTAAAAGGTGCACTTGC
ATCACTACCAGATGGACCTGCGTTGTATTAGCTAGTTGGTGGGGTAACGGCTCACCAAGGCGACGATACA
TAGCCGACCTGAGAGGGTGATCGGCCACACTGGGACTGAGACACGKCCCAGACTCCTACGGGAGGCAGCA
```

Query subrange

From

To

Or, upload file: Choose File No file chosen

Job Title:

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: ○ Standard databases (nr etc.): ● rRNA/ITS databases ○ Genomic + transcript databases

♦ 16S ribosomal RNA sequences (Bacteria and Archaea)

Organism Optional: Enter organism name or id--completions will be suggested □ exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Limit to Optional: □ Sequences from type material

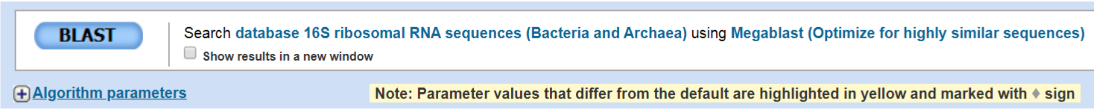
Entrez Query Optional:

Enter an Entrez query to limit search

Program Selection

Optimize for: ● Highly similar sequences (megablast)

Click button “BLAST”:

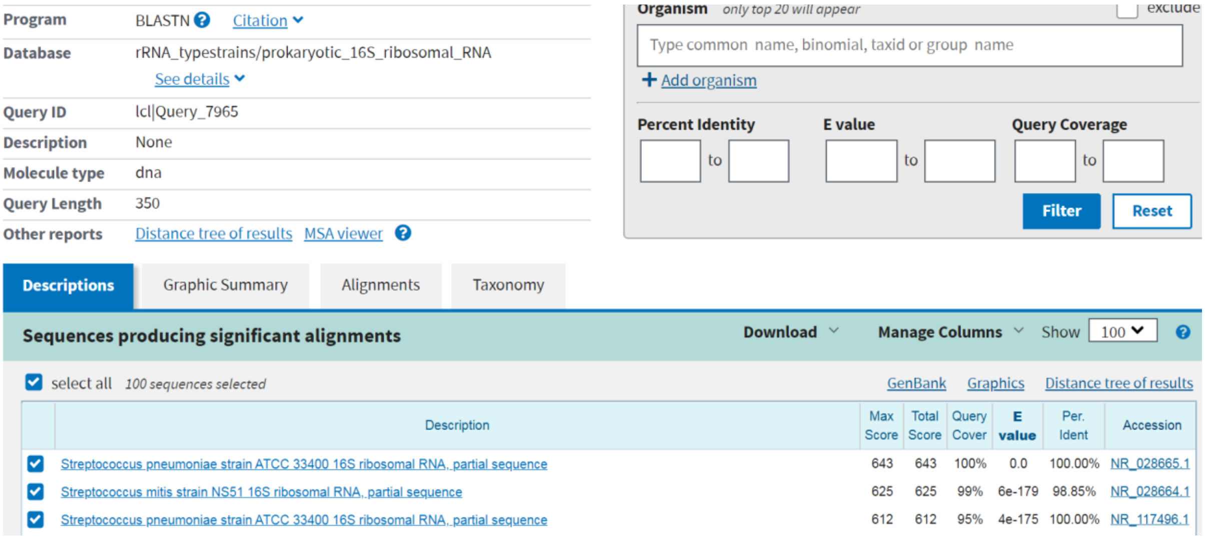


Search database 16S ribosomal RNA sequences (Bacteria and Archaea) using Megablast (Optimize for highly similar sequences)

□ Show results in a new window

Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

7.2. After the program finishes calculations, you will get the following results:



Program	BLASTN Citation
Database	rRNA_typestrains/prokaryotic_16S_ribosomal_RNA See details
Query ID	lcl|Query_7965
Description	None
Molecule type	dna
Query Length	350
Other reports	Distance tree of results MSA viewer

Organism *only top 20 will appear* □ exclude

Type common name, binomial, taxid or group name

Percent Identity ___ to ___ E value ___ to ___ Query Coverage ___ to ___

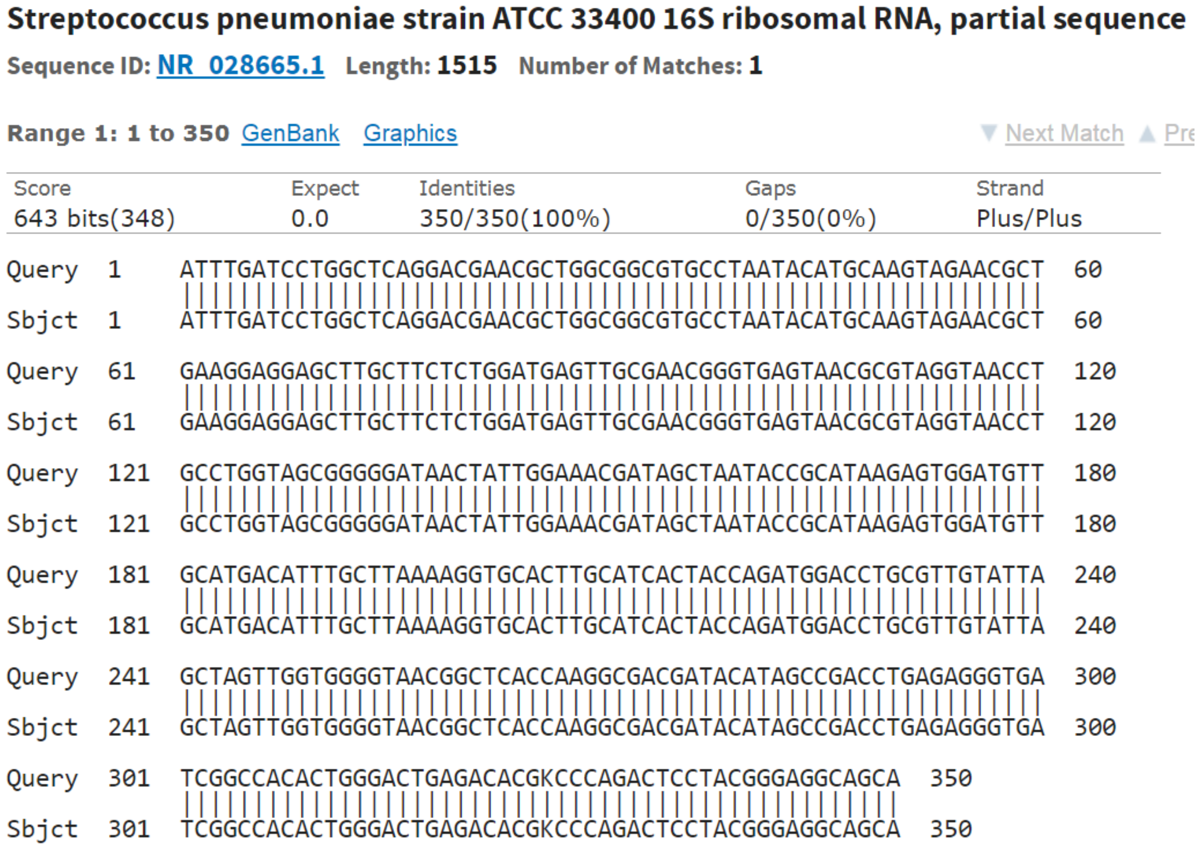
Sequences producing significant alignments

Show 100

☑ select all *100 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Streptococcus pneumoniae strain ATCC 33400 16S ribosomal RNA, partial sequence	643	643	100%	0.0	100.00%	NR_028665.1
☑	Streptococcus mitis strain NS51 16S ribosomal RNA, partial sequence	625	625	99%	6e-179	98.85%	NR_028664.1
☑	Streptococcus pneumoniae strain ATCC 33400 16S ribosomal RNA, partial sequence	612	612	95%	4e-175	100.00%	NR_117496.1

7.3. Click on the top match to see how “Sequence B” aligned with it:



Streptococcus pneumoniae strain ATCC 33400 16S ribosomal RNA, partial sequence

Sequence ID: NR_028665.1 Length: 1515 Number of Matches: 1

Range 1: 1 to 350 GenBank Graphics

```
Score           Expect  Identities     Gaps         Strand
643 bits(348)   0.0     350/350(100%)  0/350(0%)    Plus/Plus

Query  1    ATTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCT  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    ATTTGATCCTGGCTCAGGACGAACGCTGGCGGCGTGCCTAATACATGCAAGTAGAACGCT  60

Query  61   GAAGGAGGAGCTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   GAAGGAGGAGCTTGCTTCTCTGGATGAGTTGCGAACGGGTGAGTAACGCGTAGGTAACCT  120

Query  121  GCCTGGTAGCGGGGGATAACTATTGGAAACGATAGCTAATACCGCATAAGAGTGGATGTT  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  GCCTGGTAGCGGGGGATAACTATTGGAAACGATAGCTAATACCGCATAAGAGTGGATGTT  180

Query  181  GCATGACATTTGCTTAAAAGGTGCACTTGCATCACTACCAGATGGACCTGCGTTGTATTA  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  181  GCATGACATTTGCTTAAAAGGTGCACTTGCATCACTACCAGATGGACCTGCGTTGTATTA  240

Query  241  GCTAGTTGGTGGGGTAACGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGA  300
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  241  GCTAGTTGGTGGGGTAACGGCTCACCAAGGCGACGATACATAGCCGACCTGAGAGGGTGA  300

Query  301  TCGGCCACACTGGGACTGAGACACGKCCCAGACTCCTACGGGAGGCAGCA  350
            ||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  301  TCGGCCACACTGGGACTGAGACACGKCCCAGACTCCTACGGGAGGCAGCA  350
```

8. As you can see from this output, the piece of DNA extracted from the blood of the patient belongs to Streptococcus pneumoniae strain ATCC 33400

8.1. Discuss: How can it be that the cell-free DNA fraction in the blood plasma contains this piece of DNA that maps to a pathogenic bacterium?

8.2. Discuss: What would you advise to medical doctors?

[i] In a real situation sequencing of cell-free DNA would return millions of short pieces of DNA and we would need to do more advanced analysis, but for the purpose of this practical for simplicity only two DNA sequences were reported. For example, this could be the result of targeted amplification of DNA sequences of interest.

[ii] rRNA genes are extremely conserved across many bacterial and fungal species, therefore rRNA is frequently used in cross-species mapping. While rRNA is very conserved, there are differences between different species, so if a given sequence of rRNA is compared to rRNA from each known species of bacteria and fungi it is possible to identify the best match. Thus, we can uniquely identify the bacteria or fungi to which a given piece of DNA belongs.