NATIONAL UNIVERSITY

BIOINFORMATIC MODULE

WRITTEN BY: REEM SALAHELDIN KHALID HAMAD
REVISED BY: MOHMOUD KOKO

Table of content:

Lab 1 — EXPLORING NUCLEOTIDE DATABASES

Objectives :

By the end of Lab 1 (comprising the lab including its boxes, and the lecture), you should:

1. Know how to search for records at NCBI, both using search terms or identifiers and GQuery.

2. Know the difference between a GenBank accession number, a version number, and a GI number.

3. Understand the difference between the nucleotide sequence database part of GenBank and the protein sequence part of it.

4. Know the parts of a GenBank record and be able to switch between sequence formats (e.g., to FASTA format).

5. Be familiar with the interconnectedness of various NCBI databases and be able to call up linked records with ease.

6. Be able to use the Help function to address any question you may have with regard to the NCBI interface.

Part 1: NCBI

Software needed: web access

The National Center for Biotechnology Information (NCBI) maintained by the US National Library of Medicine and National Institutes of Health is one of the world's most important resources and repositories for biological data. This fantastic online resource provides an extensive network of databases cataloging an ever-growing wealth of genetic, medical, and biochemical information from all walks and crawls of life. Entire genomes, from viruses to humans, are compiled, organized, and cross-referenced within these networks, such that surfing the genome can be almost as easy as surfing the web.

But you have to know

a) What you're looking for, and

b) What you're looking at to get anything out of these databases.

This is what this first lab is going to help you do.

*Note that Google and other search engines typically do not index database-driven websites, which is why it cannot be used for searching for information that is stored at NCBI.

The primary portal for accessing data at NCBI is called GQuery. But first, let's start by visiting NCBI's website and examining the interface, which undergoes constant change.

-Open your Web browser and go to NCBI's homepage: www.ncbi.nlm.nih.gov. This page provides links to all of NCBI databases and resources. It's worth exploring here just to get a better idea of the scope of NCBI. If you click About the NCBI you will be taken to a page summarizing some of these resources. You can also check out the NCBI handbook for more information.

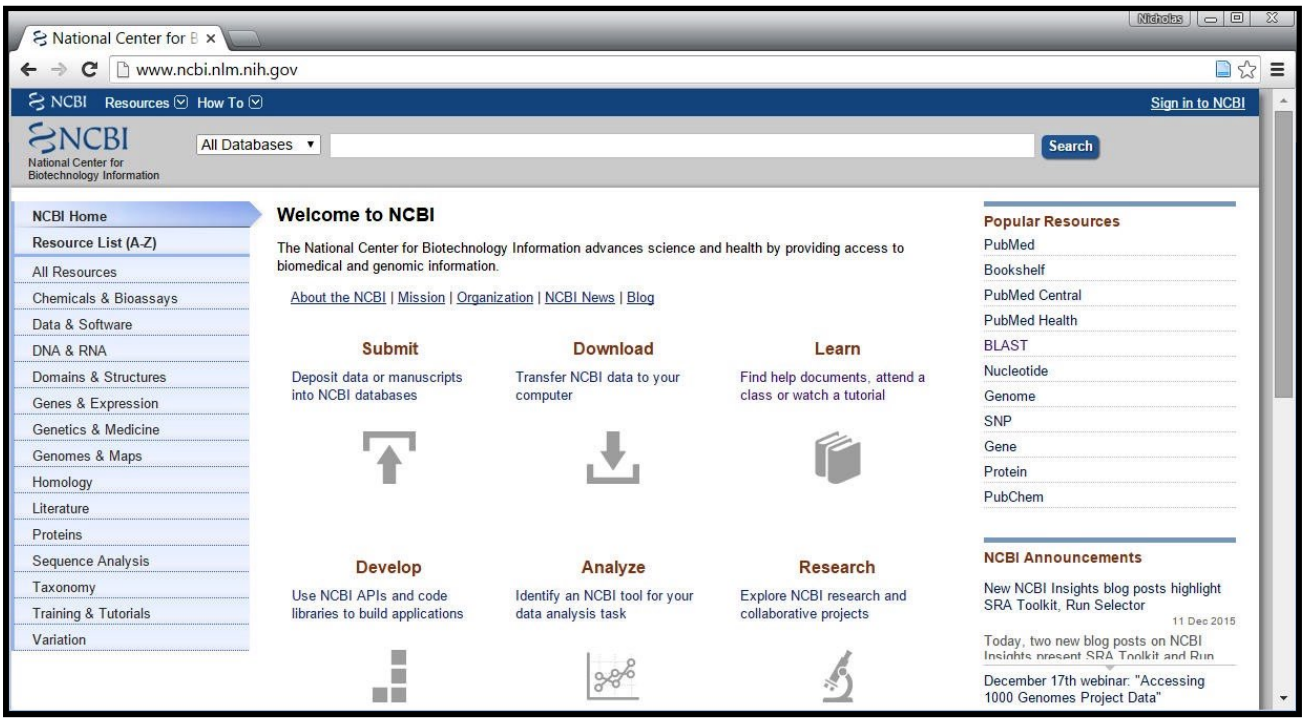


Figure (1): NCBI homepage.

- Let's move to the Search NCBI Databases (also known as GQuery) portal – select All Databases from the navigation bar at the top of the NCBI start page, by clicking "Search" on the empty field.

First, scan down the assortment of databases queried through this portal. You will notice there is everything from the biomedical literature at PubMed to nucleotide databases, taxonomy databases, protein structure databases, and expression profile databases.

Let's see what happens when you do an unguided search on the site.

In the "Search across databases" box, type in bacteria. The output is a summary page of the number of hits in each section. A search of bacteria gives millions of hits – not very helpful. We need specifics.

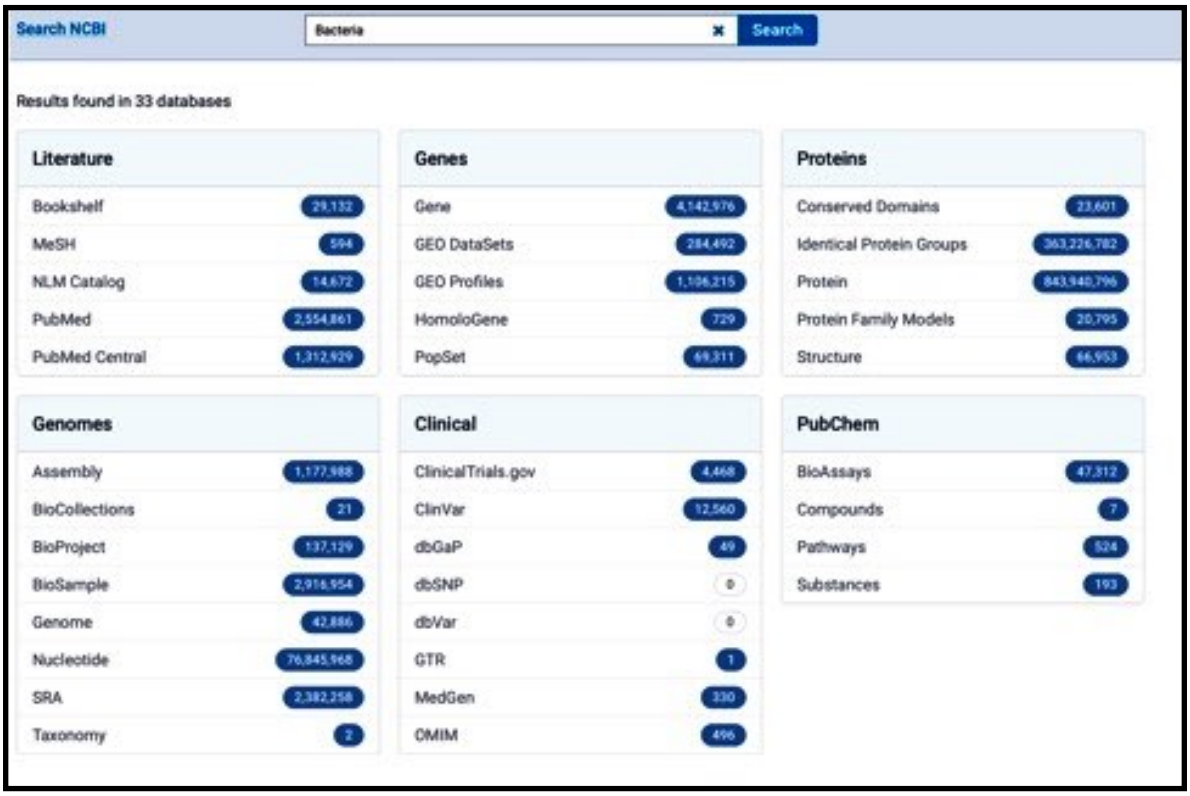


Figure (2): The Search NCBI Databases portal page with bacteria used as a search word.

- Usually when searching these databases, you have either a region of DNA or a protein (or protein function) of interest. For this lab you'll be using a gene from Human (BRCA1). The protein product of this gene is recorded under accession numberNP_009225.1, and it is a tumor suppressor.

-Go back to the NCBI GQuery portal page and try a more focused search. Use the search terms found associated with the gene sequence we'll be using with the GenBank Field Qualifiers shown below (a full list of qualifiers is presented in Appendix 1). Try the four different searches presented below and look at the number records, specifically "Protein" records, found:

- gene keywords
  e.g., BRCA1

- gene keyword AND organism
  e.g., BRCA1 AND Human

- gene keyword [PROT] AND organism [ORGN]
  e.g., BRCA 1 [PROT] AND Human [ORGN]

- accession or GI number e.g., NP_009225.1

  That narrowed things down significantly!

  Note that using parentheses can be very helpful in making sure you get exactly what you want. For example:

• BRCA1 AND (Mouse [ORGN] OR Human [ORGN]) is a very different search than
[SMC AND Mouse [ORGN] OR Human [ORGN]

Also, using quotation marks can also dramatically affect your search (ie,16s rRNA vs. "16s rRNA").

Finally, always capitalize the Boolean operators such as AND / OR / NOT. Ultimately, the most specific search items you can use are accession numbers.

Box 1. Accession numbers and Version numbers.

An Accession number is a unique identifier for a particular sequence record. An accession number is assigned to a specific record and stays with that record forever. In other words, Accession numbers track a particular record and do not change even if the information in the record is changed at the author's request (e.g., if a better annotation or more complete sequence is provided). Accession numbers are usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456).

Version numbers follow the Accession number and indicate the revision history of that entry starting with 1 and increasing with each revision. The standard format is Accession. Version.

Example: When a new entry was submitted to GenBank it was assigned an accession number (say AF000001). Since this is the first version the Accession would be appended with '.1', so it would look like AF000001.1. The updated record would keep the same Accession number but would increase in version number (AF000001.2). The new record would have been given a

- Search for our accession number of interest (e.g.NP_009225.1 from above) through the GQuery portal page. It should give you one protein sequence hit. Click on it (it is a hyperlink) so that you get its full GenBank description.

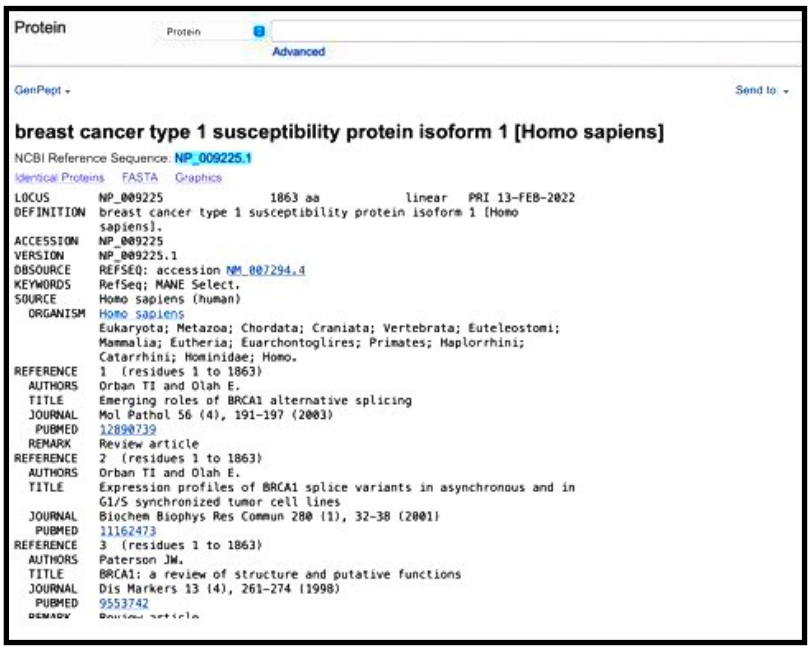


Figure (3): GenBank record for accession NP_009225.1, in GenPept format.

- Notice all the hyperlinks within the text. It looks messy but is in fact straightforward. For example, for taxonomic information, click on the SOURCE ORGANISM hyperlink. Some records have links to the primary publication where this sequence was originally cited in a PUBMED number hyperlink (not the case in the above example, but there is a PubMed reference for the sequence). Click around on different links and see what you find.

  - What is the taxonomic lineage of your organism?
  - Has the genome of this organism been sequenced, i.e., is there a Genome Project?
  - If so, can you find the accession for the full sequence or one of the chromosomes?

- To find out much more information on the structure of the Genbank file at http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html

- Go back to the GenBank record and click on the CDS link, just above the actual sequence.

  a. Where did this take you or what happened when you did this?

- Go back to the **Genbank** record and examine the **Related Information** section on the lower right. This gives you direct links to other databases with information on this query. Find the Gene link.

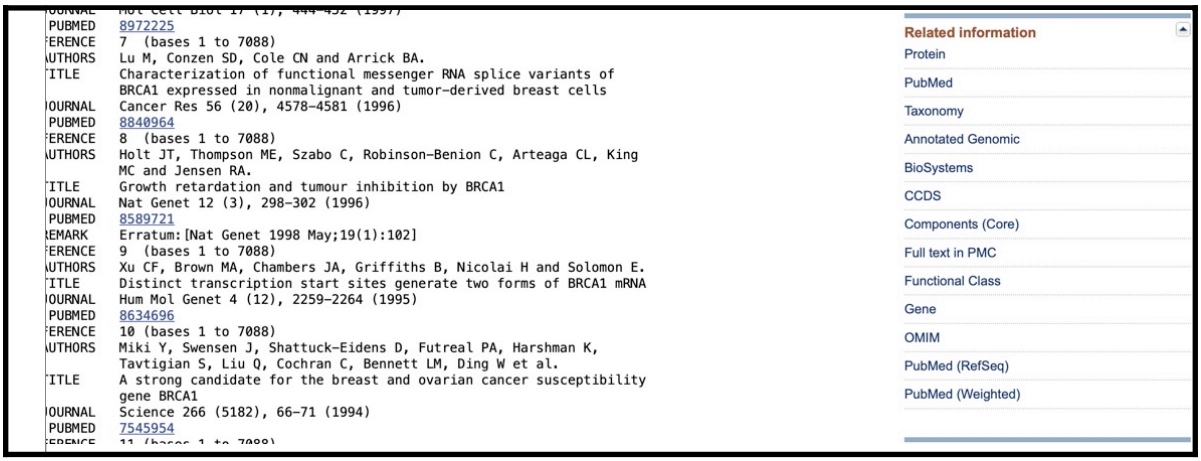


Figure (4): The Related Information menu for NM_007294.4, to the right of the record.

Select **Gene** from the Related Information menu. This is a great starter resource at NCBI. Scroll through the different sections. Use them to answer the following questions.

- Where is your gene's location in the genome? (Tip: hover with your cursor over the green bars in the "Genomic regions, transcripts, and products" section; the green bars represent the gene in the sequence viewer)

- How many exons do you see in this gene? Tip: how many green boxes are there?
- What are the names of the genes surrounding it (i.e. what is its "Genomic context")?
- Does it have any conserved domains? What are they called? (Tip: use the "Related Information" link to Conserved Domains on the right of the Gene page)
- After exploring conserved domains go back to the Gene page. What biological process (Gene Ontology terms) is this gene involved with (scroll down!)?

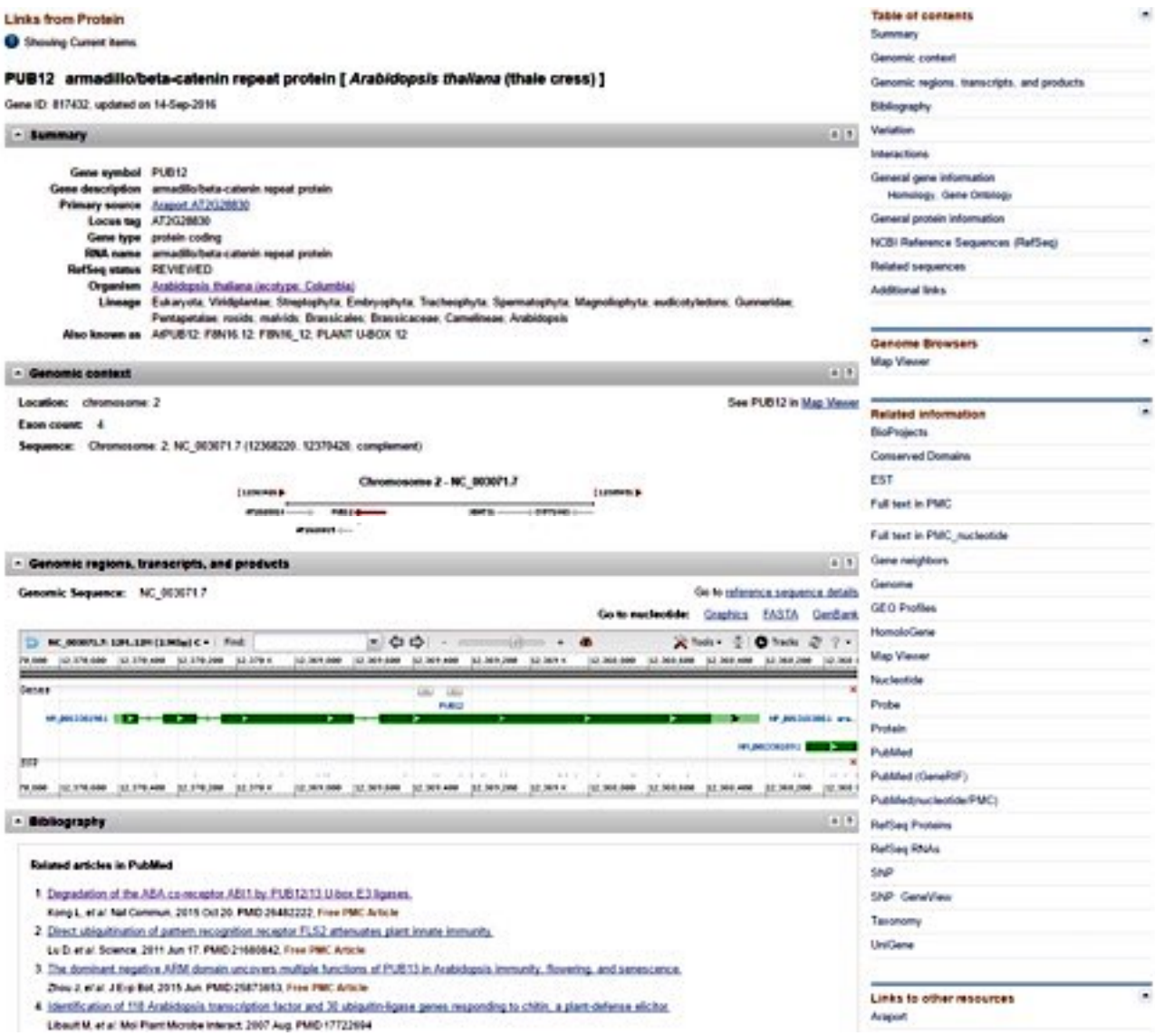


Figure (5): GenBank Gene page for At2g28830 (also known as PUB12), the gene that encodes NP_001318308.

- On the Gene page, there are also Additional links to examine a gene's structure, function and phylogenetic relationships further. The navigation sidebar on the right has an "Additional links" hyperlink which will take you to the bottom of the page, where they're found for most genes. Click [+] Gene Link Out to see them.

a. Click on Additional Links. What kind of information is in this section?

b.Click around and explore the variety of ways that data for BRCA1 are interconnected and displayed (don't worry, you can't break anything). Using the Related Information links can you find any publications associated with this gene? What about gene expression data? The next page shows the related "RefSeq" record for the corresponding mRNA (NCBI's RefSeq aims to provide canonical "reference" sequences – genomic, mRNA, CDS, protein etc. – for many model organisms).

c. Why is the length of the mRNA different from the value you can calculate from the start and stop positions in Question 9a?

Figure (6): RefSeq RNA linked from Gene page for At2g28830.

Box 3. Helpful Hints for NCBI searches

On most NCBI search pages (except, oddly, GQuery) click on "Save Search" below the search box. Register for an account and save your search. You can also combine previous searches using the History tab and the search numbers listed within it, as well as save your searches by

Part 2: ENSEMBL

Ensembl is a genome browser for vertebrate genomes that supports research in comparative genomics, evolution, sequence variation and transcriptional regulation. Ensembl annotates genes, computes multiple alignments, predicts regulatory function and collects disease data. Ensembl tools include BLAST, BLAT, BioMart and the Variant Effect Predictor (VEP) for all supported species.

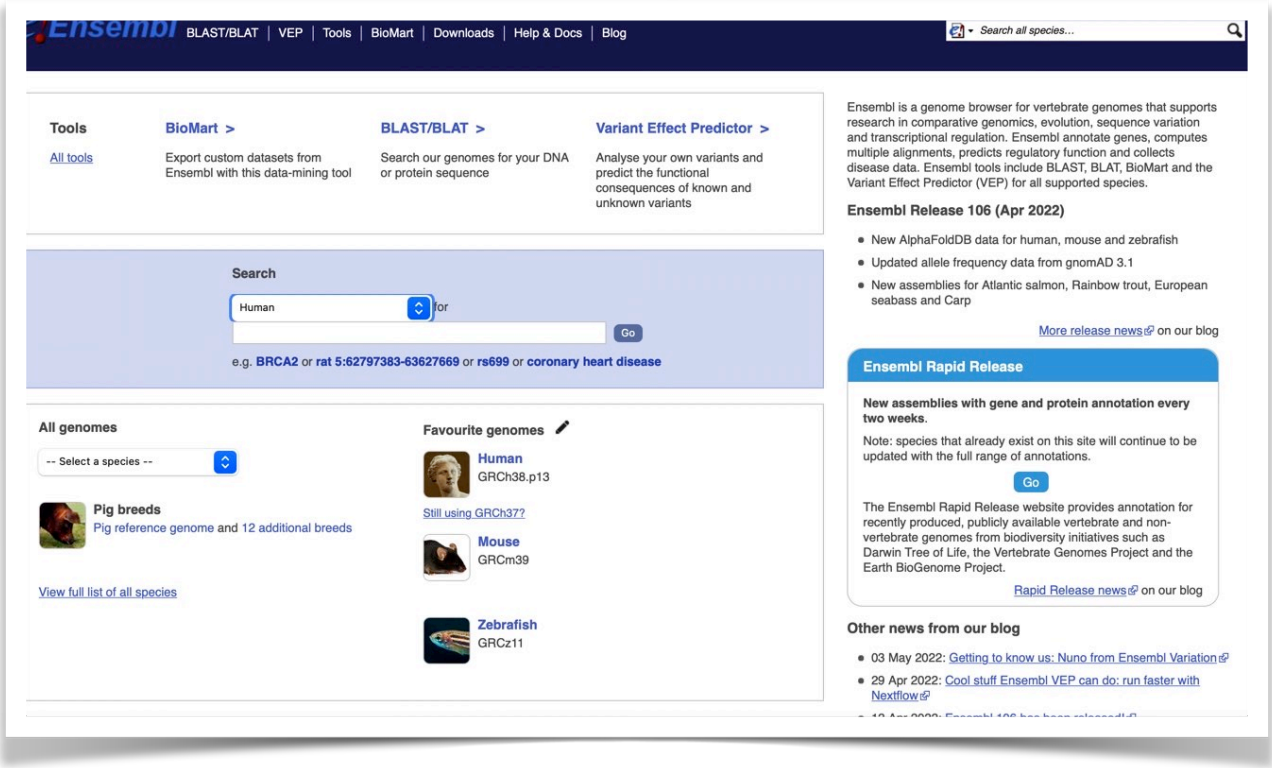


Figure (7): Ensemble start page

What is the content of the start page?

-type in the search box (human or species) and the second box BRCA1

Check what you have get after you press GO

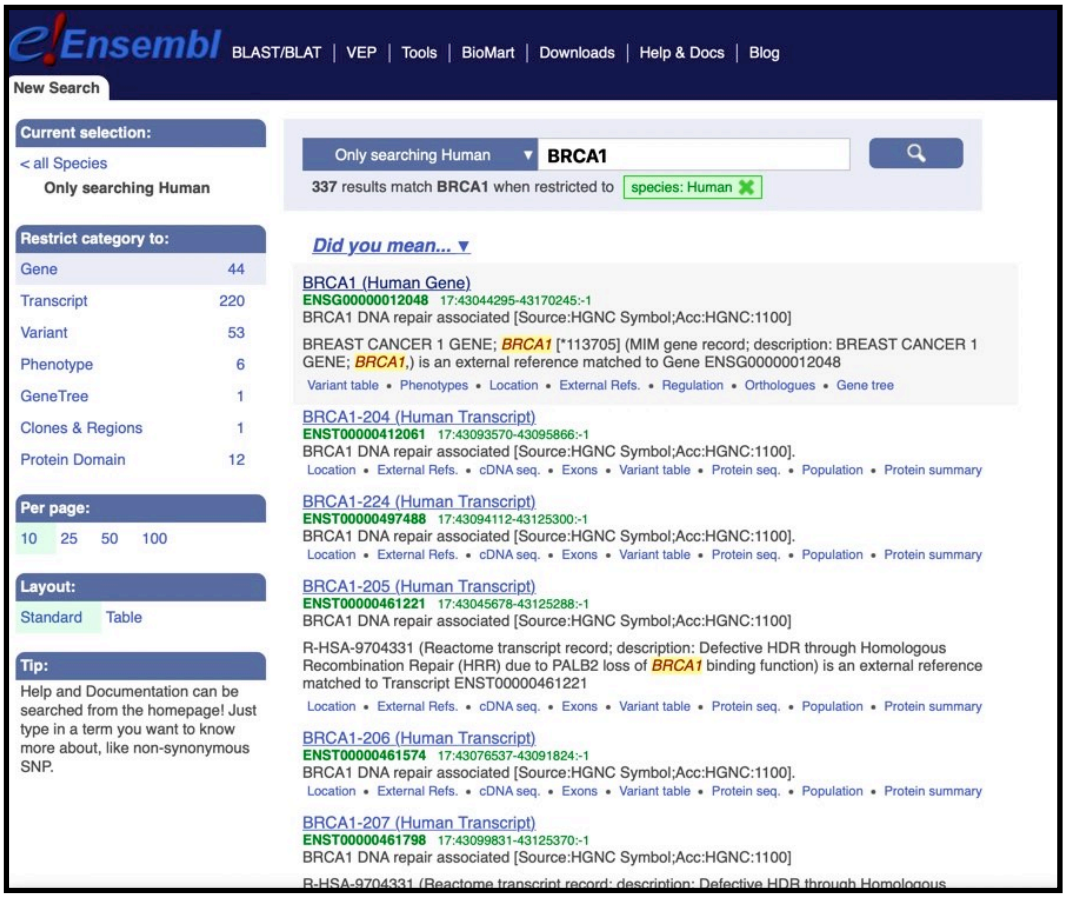


Figure (8): search result of BRCA1 on Ensembl.

-Check the right bar.

-Click on (BRCA1 (Human Gene)).

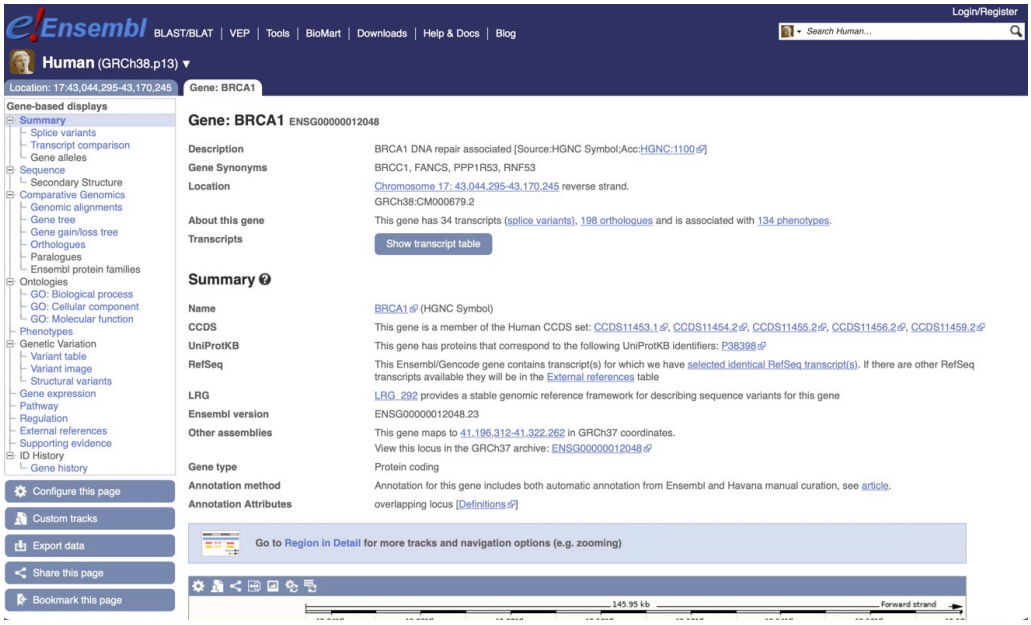


Figure (9): Ensemble record of Human BRCA1.

-What is Ensemble Accession number? Compare it with Genbank Accession Number.
-Look at Transcripts and Orthologues of BRCA 1(ENSG00000012048)

-Look at the Summary section.

Gene summary

-This page gives an overview of the information available at the gene level and it's composed of three sections.

-At the top, the page shows the gene name and Ensembl gene ID, the full description of the gene, its synonyms, its genomic location and strand, INSDC coordinates, and its number of transcripts.

-The following sections show the Transcript Table and the Summary with links to external databases, and a Gene Diagram.

TRANSCRIPT TABLE

-It shows each splice variant of a gene, i.e. protein-coding and non-coding transcripts, in addition to transcript and translation length, the transcript table displays information about biotype, mapped CCDS and RefSeq IDs as well as MANE, APPRIS and TSL flags. This table is hidden by default. Each transcript is given an Ensembl Transcript ID, which is unique and stable.

SUMMARY

-It provides additional information and links to external databases:

- Name - from official gene nomenclature committees such as HGNC (for human) and MGI (for mouse)
- CCDS - coding sequence IDs from the Consensus Coding Sequence Set
- UniProtKB - protein IDs from UniProtKB that match one of the translations of this gene
- RefSeq - Indicates if the gene has transcript(s) identified as MANE.
- LRG - IDs from the Locus Reference Genomic (LRG) project matching the Ensembl gene
- **Ensembl version - versioning of the Ensembl gene ID**
- **GRCh37 assembly - (for human only) with genomic coordinates and links to the Location and Gene views of the gene on the previous human assembly**
- Gene type - The gene type includes both status (e.g. known) and biotype (e.g. protein coding)

- **Annotation method** - It can be the Ensembl automatic, Havana manual or a merge between automatic and manual (for human, mouse, zebrafish, pig, and rat)
- **Alternative genes** - IDs from the HAVANA project that match the Ensembl gene

- **Scroll down. What you see?**

-Go to region in details

- Region in detail allows you to browse genes, variants, sequence conservation, and other annotation along the genome. There are three main images (or panels): **Chromosome, Overview** and **Region**

CHROMOSOME IMAGE

The first panel shows the chromosome of interest, marking any haplotypes or patches in red or green, respectively, and a cytogenetic banding pattern when available.

Explore what you have obtained

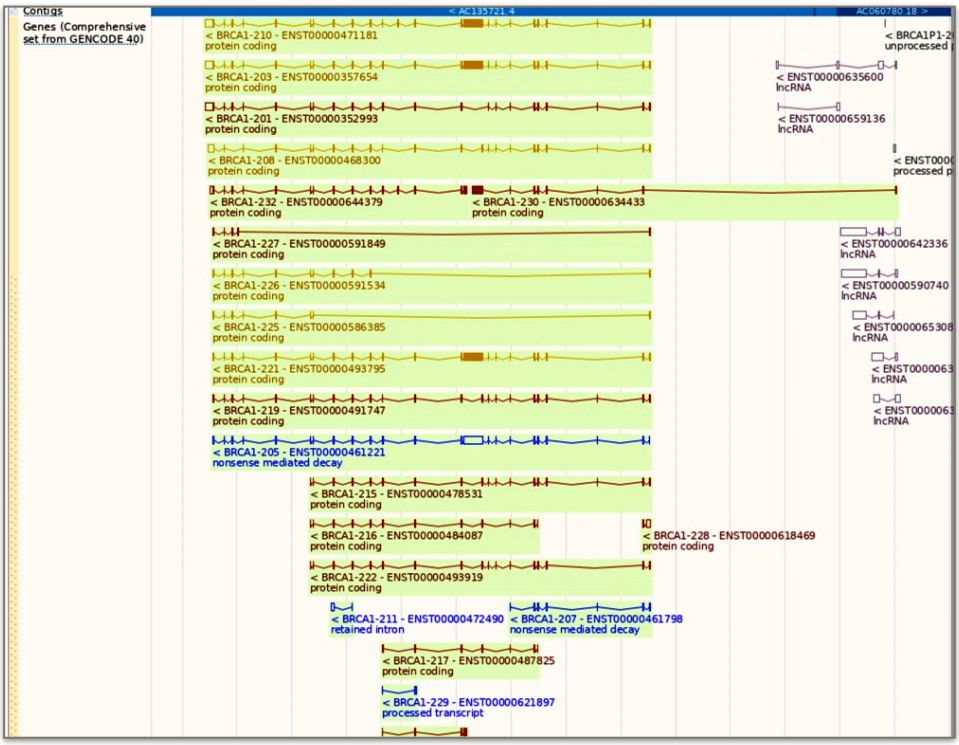


Figure (10): Ensemble transcript record of BRCA1.

GENE DIAGRAM

-It depicts the gene and all its transcripts in the context of the genome. The image can be configured to add or remove data tracks.

-Transcripts are drawn as **boxes** for **exons** and connecting **lines** for **introns**. Filled boxes show coding sequence, and empty boxes show UTRs (untranslated regions). Transcripts drawn above the blue bar (i.e the contig) are on the forward strand, whereas transcripts below are on the reverse strand.

-Transcripts are represented by different colours:

Blue, pink or **grey** transcripts are noncoding. Go to the transcript summary help page for more information

- **Red** or **gold** transcripts are protein coding. Gold transcripts are identical between the annotation from Ensembl automatic pipeline and the manual annotation from HAVANA

BIOTYPE

-it's an indicator of biological significance for genes.

-If a gene has been manually annotated (i.e., in human, mouse, zebrafish, pig, and rat), we use the -biotypes assigned by the HAVANA team.

-Biotypes can be grouped into protein coding, pseudogene, long noncoding and short noncoding.

- look at the right bar and then choose Sequence.
- explain what you see
- Compare with genbank records.
- How do you identify exons and introns?
- Now choose **Gene tree**

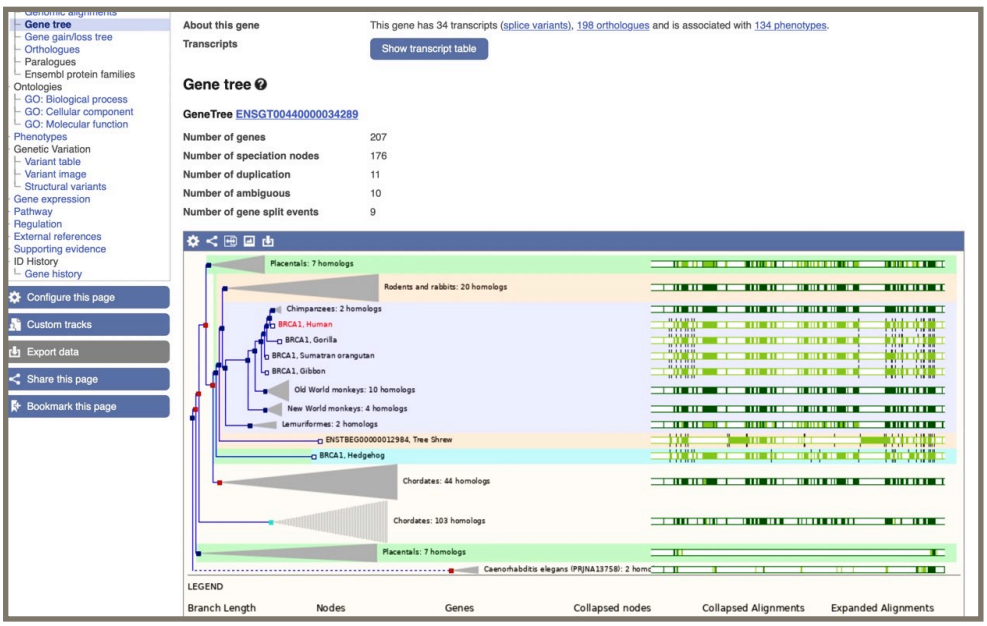


Figure (11): Ensemble Gene tree of BRCA1.
Explore the options on the right bar

Lab 2 — EXPLORING PROTEIN DATABASES

Objectives:

By the end of Lab 2 (comprising the lab including its boxes, and the lecture) You should:

1. To be able to explore UniProt Database.
2. To know different sections of UniProt.
3. To be able to explore and obtain secondary structure from Uniprot.
4. Know the advantages and disadvantages of representing structural elements in protein sequences as motifs or profiles.

UniProtKB:

The UniProt Knowledge-base (UniProtKB) is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation. In addition to capturing the core data mandatory for each UniProtKB entry (mainly, the amino acid sequence, protein name or description, taxonomic data, and citation information), as much annotation information as possible is added. This includes widely accepted biological ontologies, classifications and cross-references, and clear indications of the quality of annotation in the form of evidence attribution of experimental and computational data.

The UniProtKB consists of two sections:

A-UniProtKB/Swiss-Pro: Section containing manually annotated records with information extracted from literature and curator-evaluated computational analysis.

Reviewed, manually annotated.

B-UniProtKB/TrEMBL: Section with computationally analysed records that await full manual annotation.

Unreviewed, automatically annotated.

Where do the protein sequences come from?

-More than 95% of the protein sequences provided by UniProtKB are derived from the translation of the coding sequences (CDS) which have been submitted to the public nucleic acid databases, the EMBL-Bank/GenBank/DDBJ databases (INSDC). All these sequences, as well as the related data submitted by the authors, are automatically integrated into UniProtKB/TrEMBL.

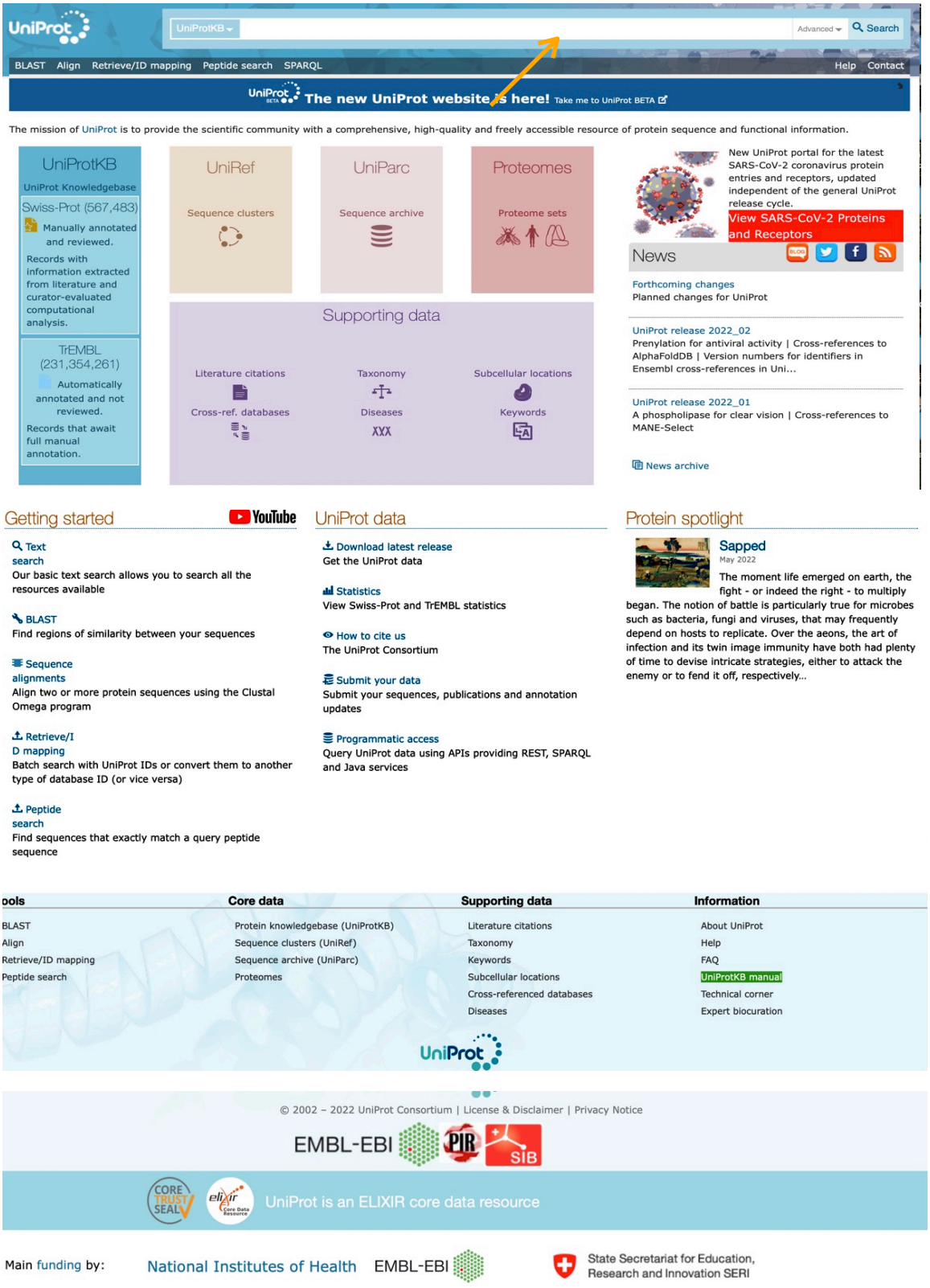


Figure (1): UniProtKB home page.

-Type *PMS1* in the search box and press search (according to orange arrow in Figure 1).

Figure (2): Result page of PMS1 search on UniProtKB.

Describe what the result page

- Notice the black arrow ↗, Inside the search box . Click on advanced search.

Explore the search parameter which appear.

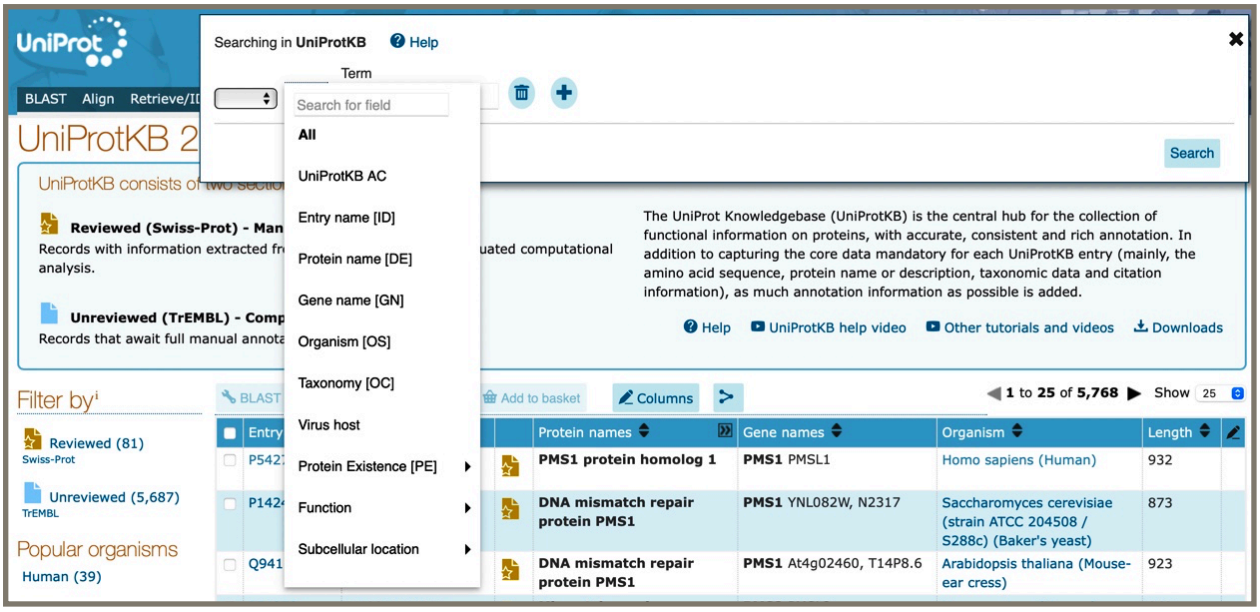


Figure (3): Advanced box choices on uniprot on search box.

What are the Golden and blue shapes referring to?

- Go to PMS 1 Human protein

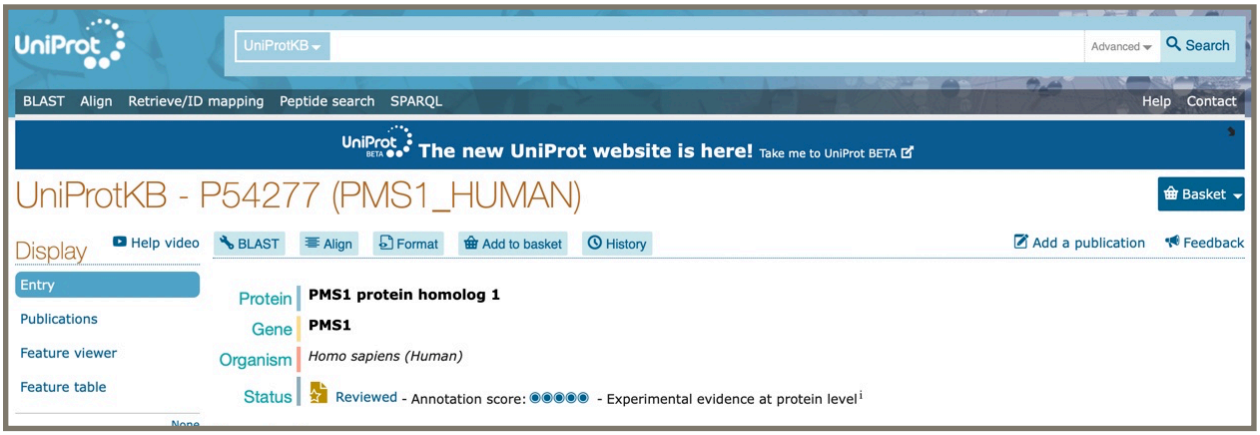
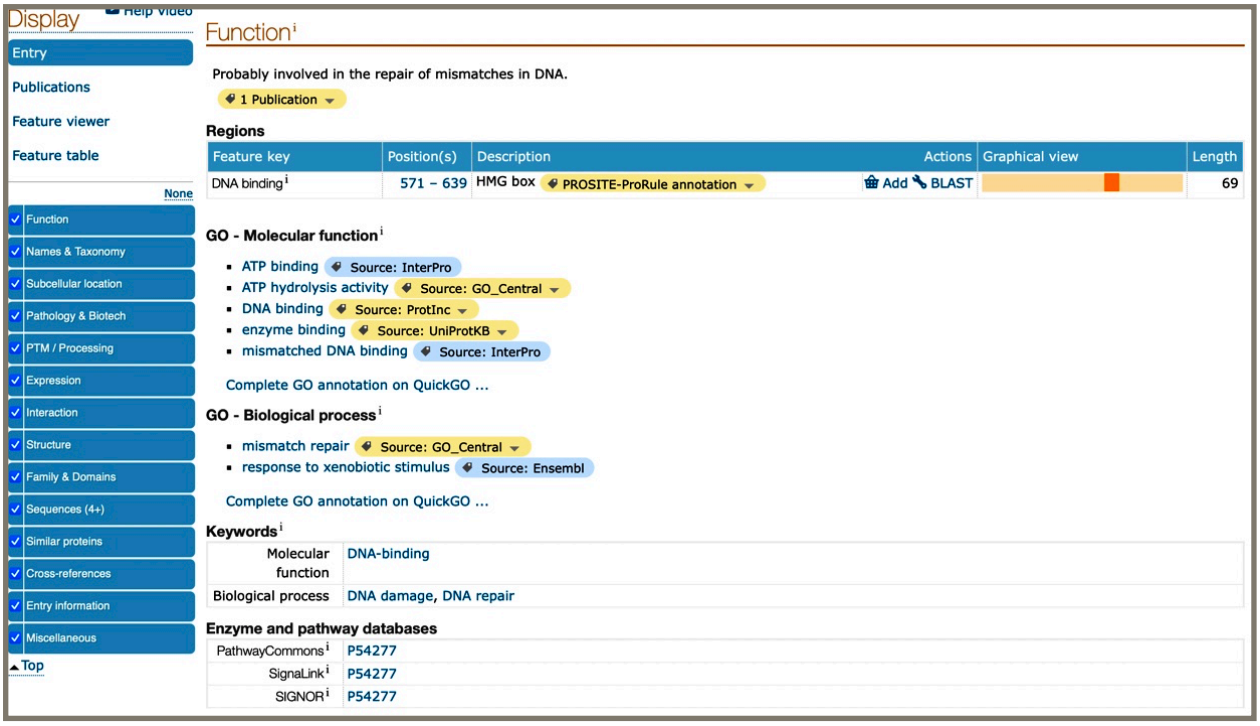


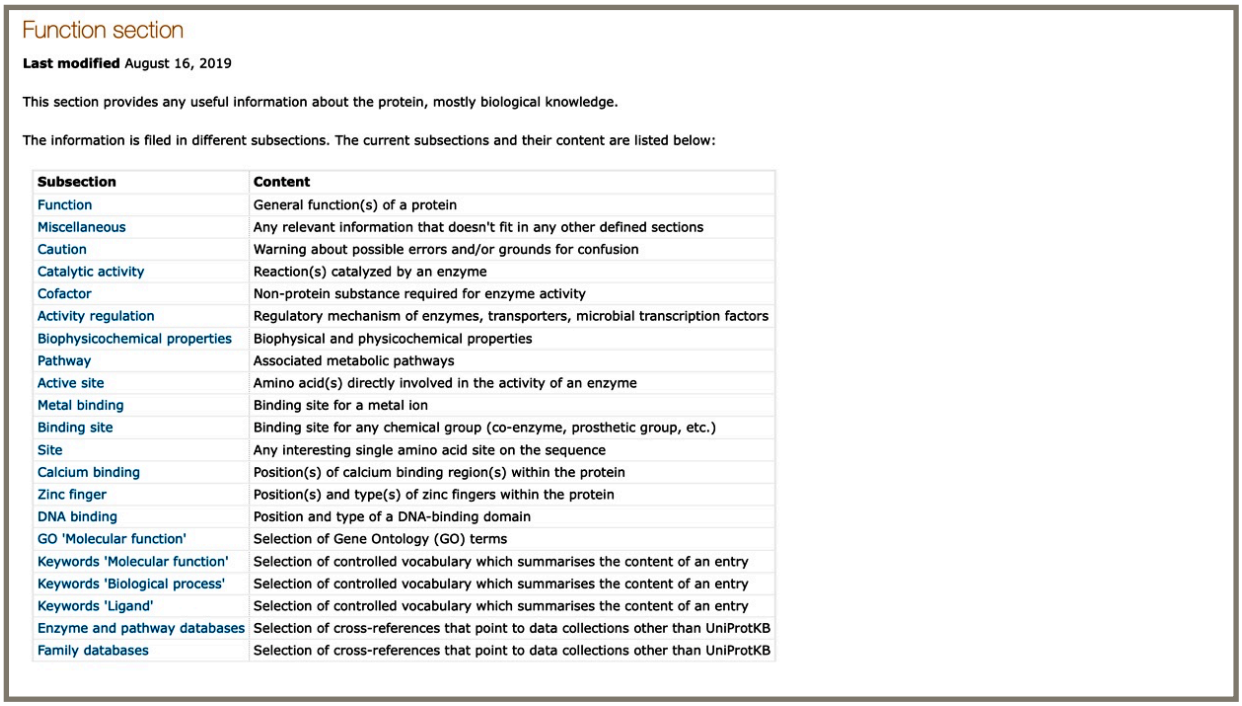
Figure (4): Summary Section for PMS1 on UniProt.

-Scroll Down to the Function Section

The Function section contains information about the molecular functions, biological processes and their related tools and websites.



Figure(5): Function Section for PMS1 on UniProt.



Function section

Last modified August 16, 2019

This section provides any useful information about the protein, mostly biological knowledge.

The information is filed in different subsections. The current subsections and their content are listed below:

Subsection	Content
Function	General function(s) of a protein
Miscellaneous	Any relevant information that doesn't fit in any other defined sections
Caution	Warning about possible errors and/or grounds for confusion
Catalytic activity	Reaction(s) catalyzed by an enzyme
Cofactor	Non-protein substance required for enzyme activity
Activity regulation	Regulatory mechanism of enzymes, transporters, microbial transcription factors
Biophysicochemical properties	Biophysical and physicochemical properties
Pathway	Associated metabolic pathways
Active site	Amino acid(s) directly involved in the activity of an enzyme
Metal binding	Binding site for a metal ion
Binding site	Binding site for any chemical group (co-enzyme, prosthetic group, etc.)
Site	Any interesting single amino acid site on the sequence
Calcium binding	Position(s) of calcium binding region(s) within the protein
Zinc finger	Position(s) and type(s) of zinc fingers within the protein
DNA binding	Position and type of a DNA-binding domain
GO 'Molecular function'	Selection of Gene Ontology (GO) terms
Keywords 'Molecular function'	Selection of controlled vocabulary which summarises the content of an entry
Keywords 'Biological process'	Selection of controlled vocabulary which summarises the content of an entry
Keywords 'Ligand'	Selection of controlled vocabulary which summarises the content of an entry
Enzyme and pathway databases	Selection of cross-references that point to data collections other than UniProtKB
Family databases	Selection of cross-references that point to data collections other than UniProtKB

Figure(6): Summary of Function section on UniProt.

-Now scroll down on the rest of the section which gives information about the protein. As you see on the right bar.

Box (1): Different Section of uniprot result page:

•**Names & Taxonomy:** This section provides information about the protein and gene name(s) and synonym(s) and about the organism that is the source of the protein sequence.

•**Subcellular location:** This section provides information on the location and the topology of the mature protein in the cell.

•**Pathology & Biotech:** This section provides information on the disease(s) and phenotype(s) associated with a protein.

•**PTM / Processing:** This section describes post-translational modifications (PTMs) and/or processing events.

•**Expression:**This section provides information on the expression of a gene at the mRNA or protein level in cells or in tissues of multicellular organisms.

•**Interaction:** This section provides information on the quaternary structure of a protein and on interaction(s) with other proteins or protein complexes.

•**Structure:**This section provides information on the tertiary and secondary structure of a protein.

•**Family & Domains:** This section provides information on sequence similarities with other proteins and the domain(s) present in a protein.

•**Sequences (4+):**This section displays by default the canonical protein sequence and upon request all isoforms described in the entry. It also includes information pertinent to the sequence(s), including length and molecular weight. The information is filed in different subsections.

•**Similar proteins:**This section provides links to proteins that are similar to the protein sequence(s) described in this entry at different levels of sequence identity thresholds (100%, 90% and 50%) based on their membership in UniProt Reference Clusters (UniRef).

•**Cross-references:**This section is used to point to information related to entries and found in data collections other than UniProtKB.

•Entry information

•Miscellaneous

Finding Secondary Structure Information:

When examining the structure panel in UniProt, you can look at additional "Features", specifically, experimentally validated alpha helices and beta sheets (turns). UniProt collects these structures from other databases (mostly, the Protein Data Bank). You can find links to these resources as well. This is usually an easy way to find the main 2D features. Unfortunately, this might not be available for all proteins. Thankfully, there are dedicated structure databases that can be used instead.

The best way to find the secondary structure is probably to look for the tertiary structure. By definition, tertiary structure prediction means resolving the secondary structure as well. That means, you can use (most of) the resources providing tertiary structure to visualize the secondary structure. However, secondary structures can be resolved without necessarily predicting 3D (tertiary) folding.

Using Experimental Data:

Protein structure resources like RCSB PDB will give you the chance to examine the secondary structure. This is available in 2D in the **Sequence** panel of the Protein Feature View (you may need to hit expand) or otherwise in the **3D View** panel.

Using Predicted Structures:

Homology Modelling is predicting the structure of a single protein based on a (very) similar template (usually a homolog). The Swill-Model repository (https://swissmodel.expasy.org/repository/) is a database of structures created using homology modelling.

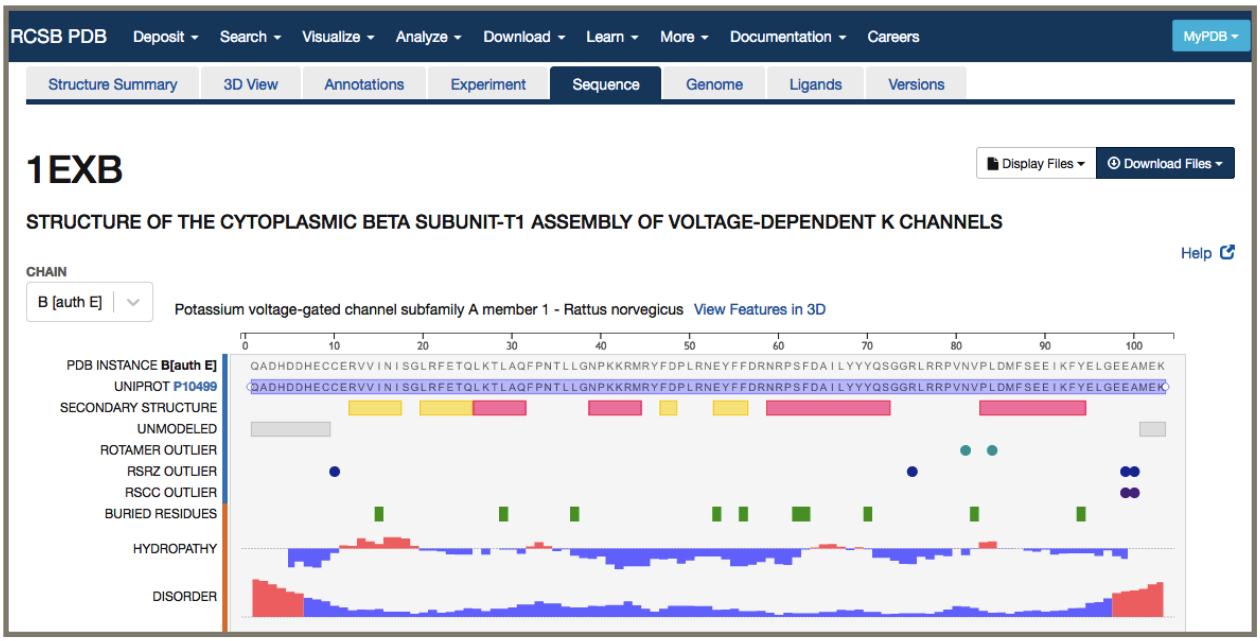
Recently, protein structure prediction has moved from using known structures as templates for homology modelling towards Artificial Intelligence and Machine Learning methods. The current state-of-the-art method is called Alpha Fold. AlphaFold Database (https://alphafold.ebi.ac.uk/) will give you the predicted secondary and tertiary structures of almost all known proteins (for sure all those in UniProt).

Structures from these databases can be visualised online on the same website (Alpha Fold, SwissModel) but the online visualisation is designed with 3D/tertiary structure viewing in mind (e.g., it might not be very easy to get the exact boundaries of helices & sheets).

However, structures can be downloaded as PDB files for local visualisation if one wants to highlight the secondary structure better. You can then use other online programs like PolyView (https://

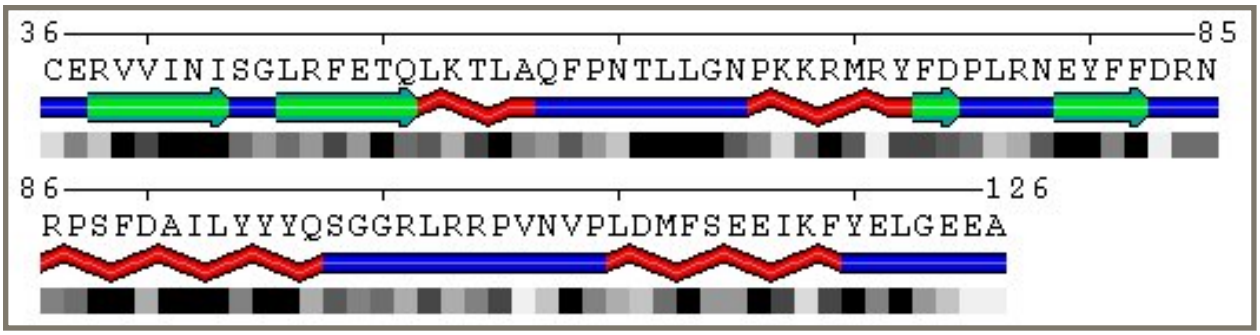
polyview.cchmc.org/) or some locally installed free programs (like Jalview, CLC Sequence Viewer, UniPRO UGene, Chimera).

As an exercise you can import a PDB structure to PolyView-2D and see how it looks compared to the original PDB Sequence view (https://www.rcsb.org/sequence/1EXB#E).

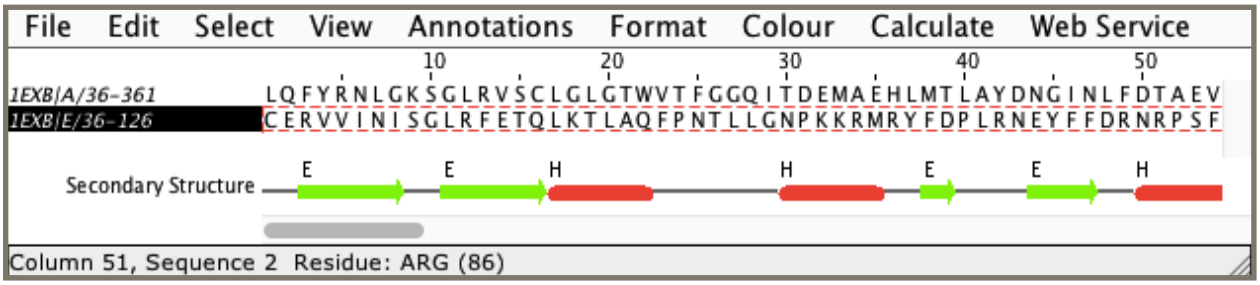


Figure(7): Example of a 2D structure of the T1 domain of the rat potassium channel $K_V1.1$ (from: https://www.rcsb.org/sequence/1EXB#E). Secondary structure is shown as panels of yellow and red blocks under the sequence. You can hover over one block to see if it's an alpha helix or a beta sheet.

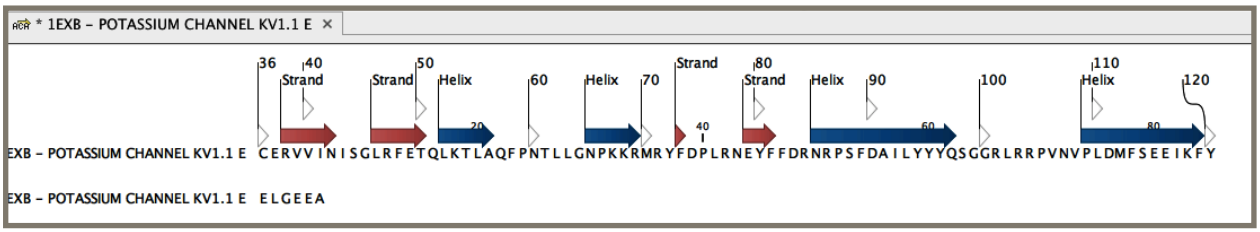
PolyView2D (Online)



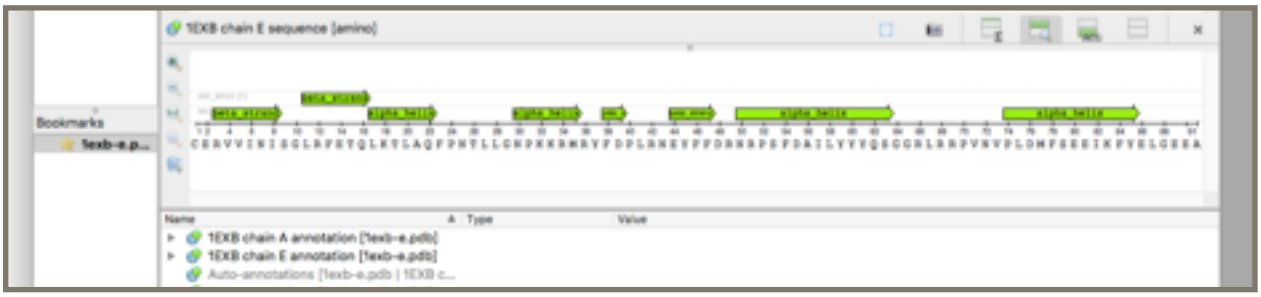
Jalview



CLC Viewer



UGene



Figure(8): Secondary structure visualisation of an experimentally obtained structure of the T1 domain of the potassium channel Kcna1. For the online visualisation, the structure was imported from RCSB PDB (PDB: 1EXB) to PolyView2D. For the local visualisation, the structure was downloaded and visualised using three different programs (Jalview, CLC Viewer, UniPro UGene).

These are the same structure features displayed in the previous figure from RSCB PDB. Note the slight difference in the boundaries between PolyView and the others.

Creating your own structure models:

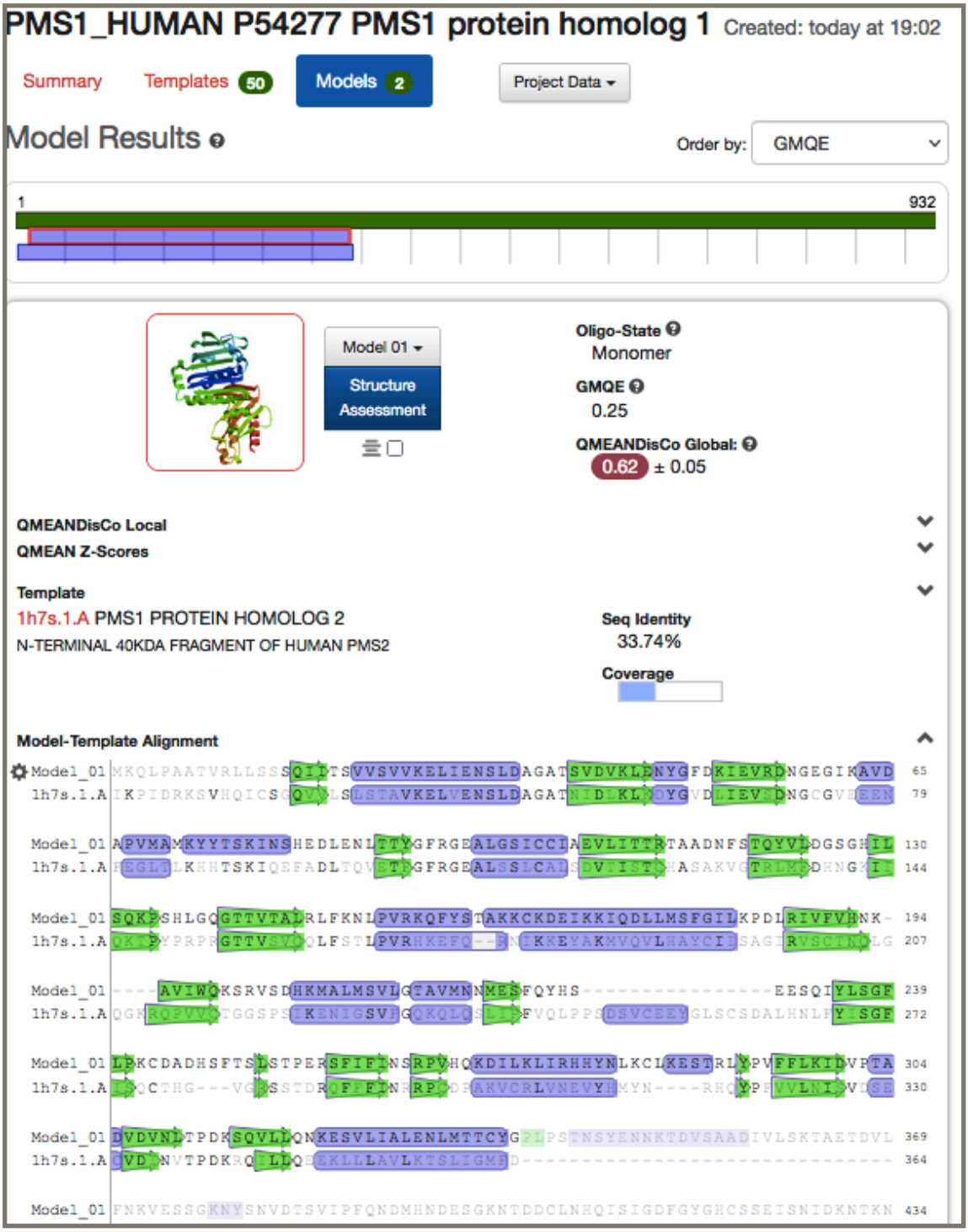
SwissModel:

Since predicting structures using complex and advanced methods like Alpha Fold is complex and computationally intensive, it might not be available for general users (no server where you can simply submit sequences to predict the structure). If your protein is not already on the Alpha Fold database, you may need to create your own model. Here, homology modelling comes handy as it is much simpler and faster. It is possible to search for templates with known structure and use these to predict the structure of a similar protein in SwillModel (https://swissmodel.expasy.org/interactive/).

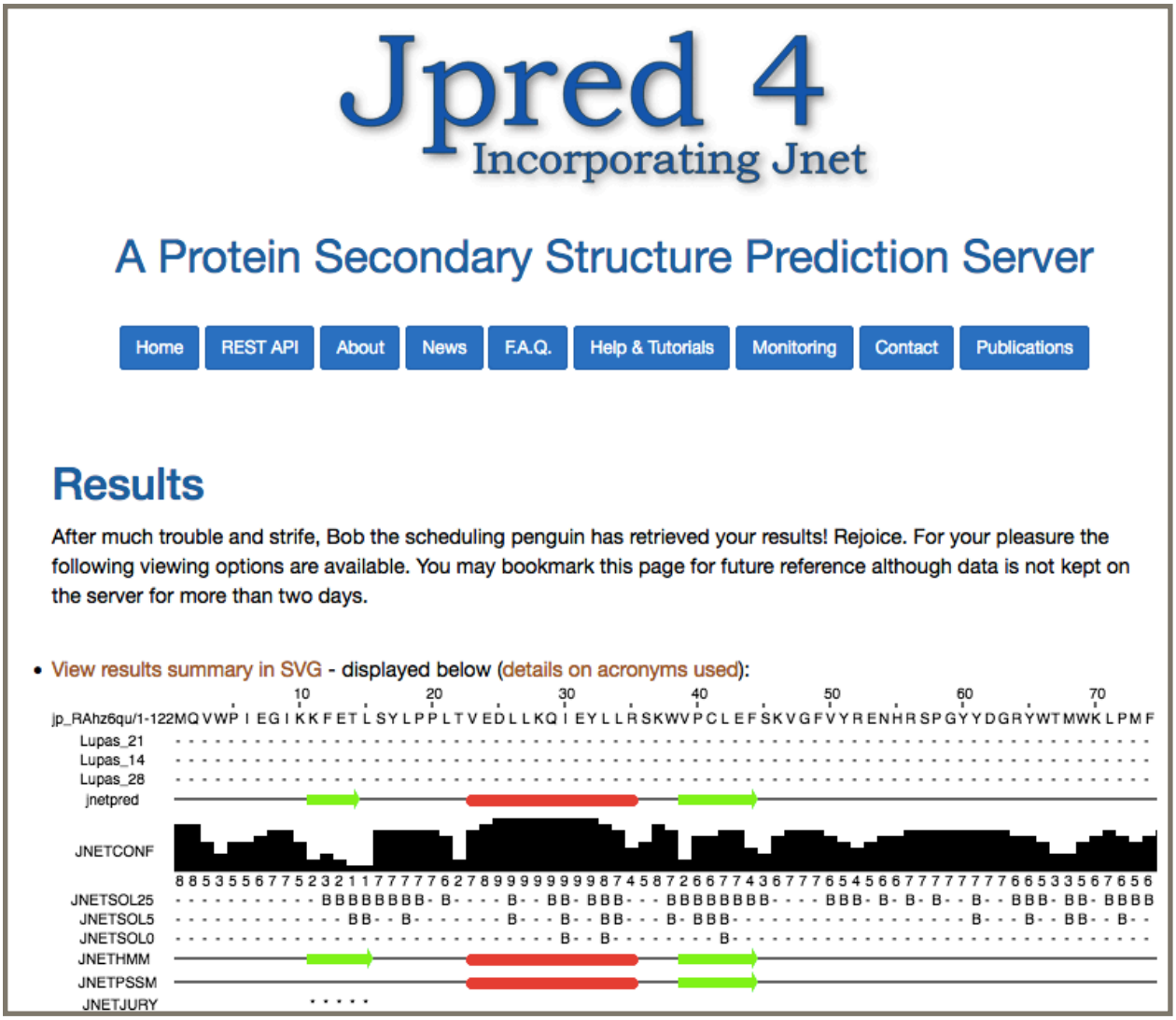
Take a protein accession number from uniprot and try to create a structure model using SwissModel (this may take 10-15 minutes if your sequence is not too long). You will get the 3D visualisation on the right-hand side and the 2D visualisation on the left-hand side (it is hidden by default; you will have to click on Model-Template Alignment). When looking at the 3D structure, click on the settings (the 'gear' icon) and see how different secondary structure prediction methods change the boundaries of the predicted helices/sheets.

JPred:

If you are only interested in the secondary structure and do not want to bother with 3D folding, there are some servers that are dedicated for this. JPred is one of those (https://www.compbio.dundee.ac.uk/jpred/). You can put a protein sequence and simply wait for the results.



Figure(9): 2D predicted secondary structure of the PMS1 protein structure using homology modelling (using PMS2 as a model). Helices (rectangles) and sheets (arrows) are shown in different colours.



Figure(10): predicted structure of the protein Ribulose Phosphatase using JPred. Compare this to the experimental structure found in RCSB PDB (PDB: 6kyi).

Lab 3—PROTEIN TERTIARY STRUCTURE VISUALISATION

Objective:

By the end of lab3 (comprising the lab including its boxes, and the lecture) you should know:

- Know the main methods for determining protein structure.
- Be familiar with Protein Database records and how to determine which method was used to ascertain a given protein's structure;

In this lab, we will visit the online protein structure repository, the Protein Data Bank (PDB), and will obtain models for the tertiary structure of several proteins.

Protein Data bank :

Since 1971, the Protein Data Bank archive (PDB) has served as the single repository of information about the 3D structures of proteins, nucleic acids, and complex assemblies.

The Worldwide PDB (wwPDB) organisation manages the PDB archive and ensures that the PDB is freely and publicly available to the global community.

PDBe (protein data bank Europe) is a founding member of the Worldwide Protein Data Bank which collects, organises and disseminates data on biological macromolecular structures. In collaboration with the other Worldwide Protein Data Bank (wwPDB) partners, we work to collate, maintain and provide access to the global repository of macromolecular structure models, the Protein Data Bank (PDB).

The PDB archive is a repository of atomic coordinates and other information describing proteins and other important biological macromolecules. Structural biologists use methods such as X-ray crystallography, NMR spectroscopy, and cryo-electron microscopy to determine the location of each atom relative to each other in the molecule. They then deposit this information, which is then annotated and publicly released into the archive by the wwPDB.

Figure(1): Protein Data Bank world wide.

Box1: wwPDB Partners

RCSB PDB

RCSB PDB (RCSB.org) is the US data centre for the global Protein Data Bank (PDB) archive of 3D structure data for large biological molecules (proteins, DNA, and RNA) essential for research and education in fundamental biology, health, energy, and biotechnology.

PDBe

Is the founding member of the Worldwide Protein Data Bank which collects, organises and disseminates data on biological macromolecular structures.

PDBj

PDBj (Protein Data Bank Japan) is a project team operating under the Joint Usage and Research activities of the Institute for Protein Research, Osaka University.

BMRB

Biological Magnetic Resonance Data Bank

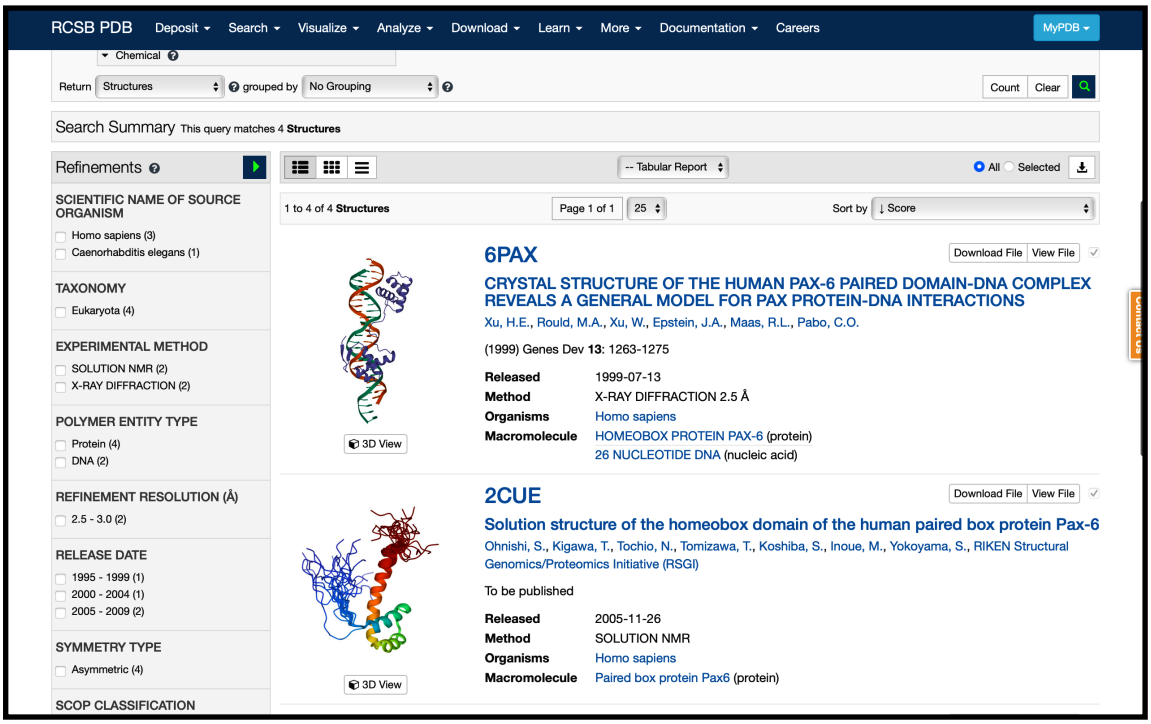
BMRB collects, annotates, archives, and disseminates spectral and quantitative data derived from NMR spectroscopic investigations of biological macromolecules and metabolites.

EMDB

The Electron Microscopy Data Bank is a public repository for electron cryo-microscopy maps and tomograms of macromolecular complexes and subcellular structures. It covers a variety of techniques, including single-particle analysis, electron tomography, sub-tomogram averaging, fibre diffraction and electron crystallography.

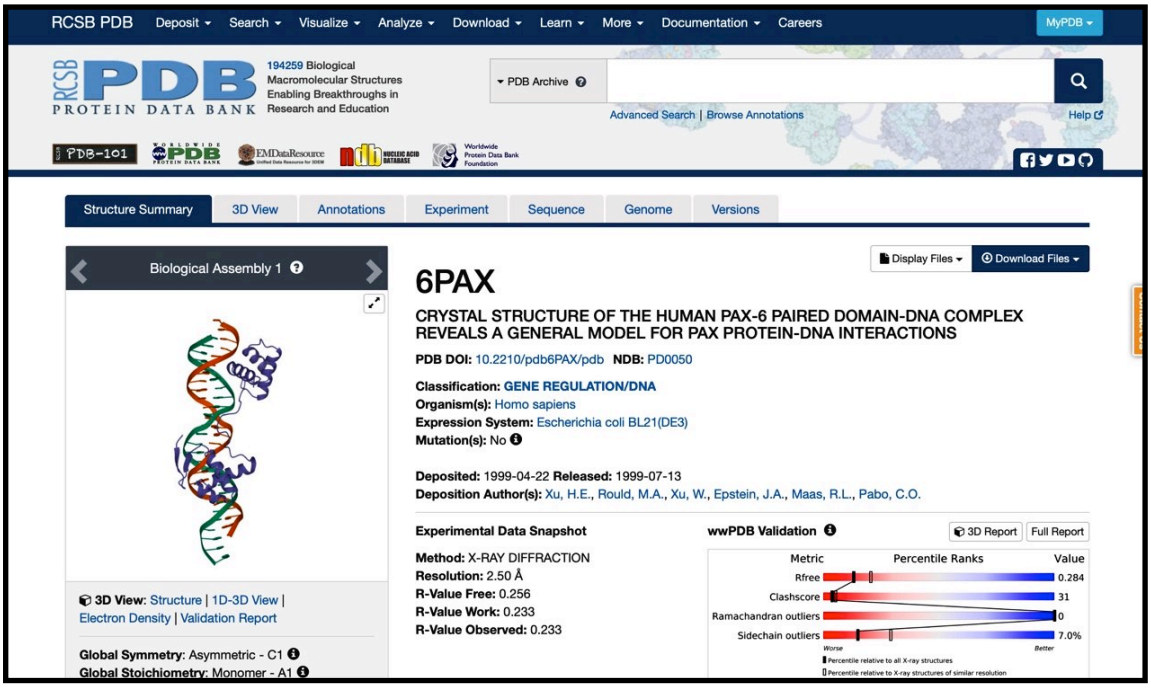
Figure(2) : Input page of RCSB PDB

-Connect to PDB through the RCSB portal (the black arrow): http://www.rcsb.org/pdb/home/home.do and type “PAX6” in the search field at the top of the page and click on Go.



Figure(3): RCSB PDB result of PAX6 gene.

- Choose the first option 6PAX and explore it.



Figure(4): 6PAX result summary of PDB.

-What is the Structure Summary page?

For any PDB entry, the Structure Summary page provides an overview of the structure. It presents information about the structure, what the structure looks like, which macromolecules and small molecule ligands it contains, which experimental method(s) were used to determine the structure, who solved the structure, what the quality of the structure is, which publications describe this structure, etc. All information pertaining to the structure in the PDB and in other bioinformatics data resources can be found here.

-Each section of the Structure Summary page is described here to describe

- What does the interface look like for this section?
- What can you learn about the structure from this section?
- How to explore the archive to find related structures?

-Let us explore the different section of PDB result .

-Move to 3D view section where you can see 3D visualisation.

Figure(5): 3D visualisation of PAX6 gene.

-What is the 3D View?

The real value of PDB data is the opportunity to visualise molecular structures and analyse them in three-dimensions (3D). Each PDB entry has a 3D View tab that can be used to upload the coordinate file(s) of the structure and display them for interactive analysis using Mol*. Detailed information about using the visualisation tool is available in the Mol* Documentation. Here we introduce the tool in the context of exploring a specific structure.

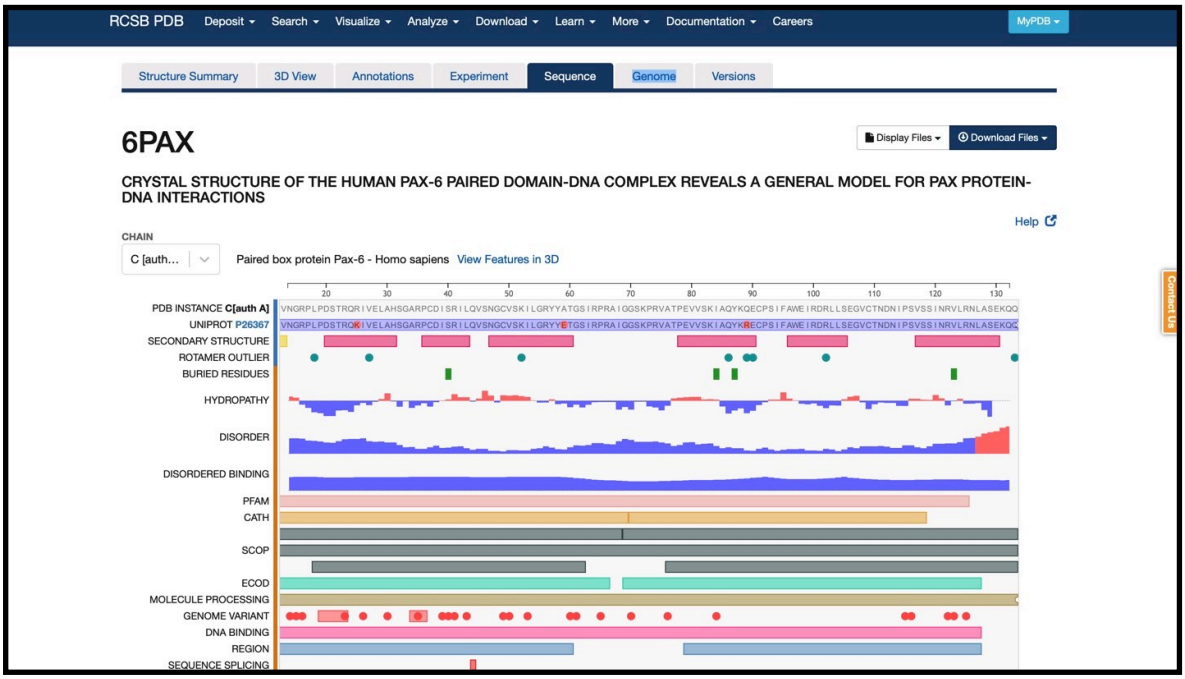
-The interface

The Mol* tool, used in the 3D View tab, simultaneously displays the molecules in the structure in 3D and the sequences of polymers present in the structure, as well as any ligands, ions, water molecules etc.

-The 3D canvas is where the molecule is displayed. Moving your mouse in this region of the screen allows you to move (rotate, translate, and zoom into) the structure.

-The sequence panel is marked with a horizontal box with a red outline. This can be used to click on any one or a group of amino acids to centre on them and zoom in to see the non-covalent interactions around it or them.

-The Controls panel provides options for you to display, hide, change representations, and color the polymer chains and ligands displayed.



Figure(6): Sequence view of PAX6.

-The Sequence Summary page is mostly used to integrate information about various aspects of the polymer being studied from various data resources. However, in the row displaying the UniProt sequence.

-The interactive sequence display provides a quick summary of the protein and nucleic acid polymers present in the structure. They help integrate information from a variety of resources and map them on the structure in an easily accessible format. Exploring this page can inform you about the structure and functions of the polymer - where the active site, binding site, etc. are located; where the hydrophilic and hydrophobic regions are located; where the mutations (if any) are present; and much more.

-The first row in the display lists the sequence of the protein from the PDB entry.

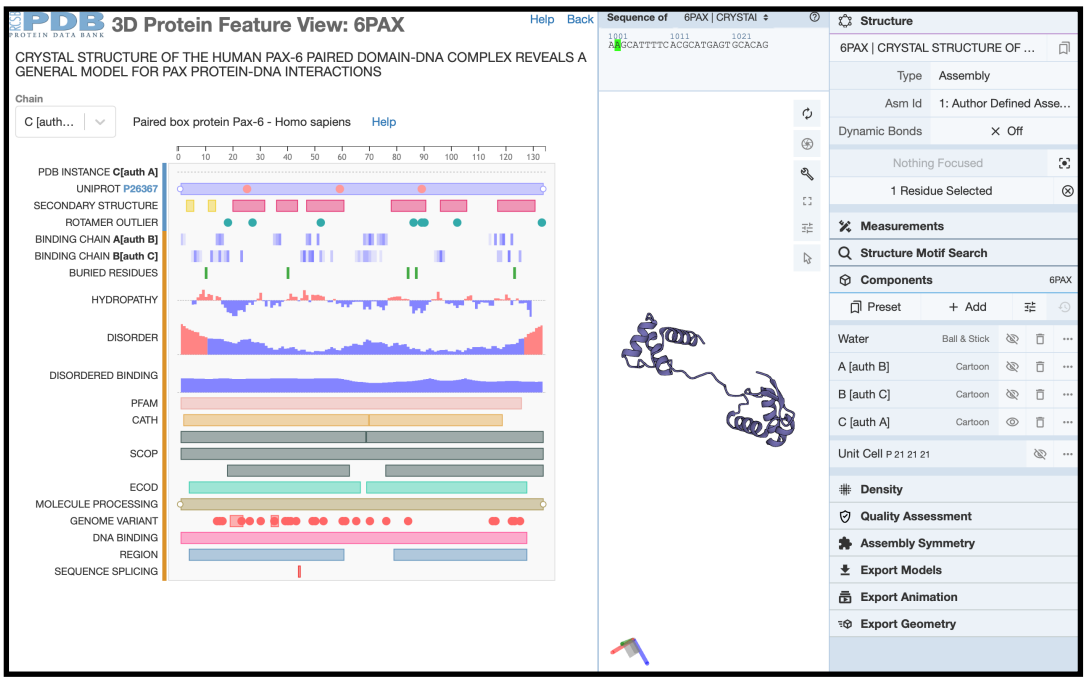
-The second row shows the reference sequence in a purple rectangle - e.g., UniProt.

-The data directly derived from the PDB or computed based on data from the PDB are marked with a blue line on the left of the display.

-Annotations integrated from various bioinformatics resources are marked with an orange line on the left of the display.

The display is interactive so you can zoom in and out to examine the sequence(s).

- In the sequence section, press view features in 3D and explore it



- **Figure (7):** 3D PAX6 protein Feature view.

- For more information about how to use PDB go the help link : https://www.rcsb.org/docs/exploring-a-pdb-entry/structure-summary-page.
- Other databases where you can visualise the 3D structure are:
    - NCBI-Structure database.
    - NCBI-conserved domain.
    - Uniprot structure section.
    - Alphafold.

Lab 4—BASIC BLAST

Objectives:

By the end of this lab (comprising the lab including its boxes, and the lecture) you should know:

1. Know how to use BLAST.

2. Be able to use nucleotide BLAST (Blastn) to search GenBank, and be able to interpret the output – what does the E-value tell you etc.

3. Understand the meaning of homologous, orthologous, and paralogous sequences.

Introduction:

The Basic Local Alignment and Search Tool **(BLAST)** is a very powerful approach to identifying database sequences that share local similarity to a query sequence (see below for definitions).

One of the most important bioinformatic strategies used for the functional annotation of genes and genomes is to predict the function of uncharacterised genes or proteins based on their similarity to sequences with better functional annotations**. BLAST** is perhaps the single most important tool for finding database sequences that are similar to a query sequence of interest.

Box 1. BLAST and Homology

There is a very important chain of assumptions used in biological research that is generally followed when using BLAST: Homologous genes share sequence similarity

• Orthologous genes have the highest similarity among multiple species

• Orthologous genes most likely have similar functions

• Consequently, sequences that are most similar between multiple species share similar functions

Note, it is very important to understand that these are only assumptions, and there are many reasons and instances where these assumptions prove to be false. Nevertheless, they are a reasonable starting place.

Definitions:

- **Similar sequences** – sequences that share a significant number of residues (nucleotides or amino acids). Sequences can be similar due to homology or simply by chance. The higher the similarity between sequences, the more likely they are to be homologous.

- **Homologous sequences** – sequences that are related through common ancestry. Homology is qualitative – two sequences either are, or are not related through common ancestry. Homologous sequences can vary greatly in their level of similarity – from 100% to 0%.

- **Orthologous sequences** – sequences that are related through a past speciation event. Orthologous sequences are assumed to share common functions.

- **Paralogous sequences** – sequences that are related through a past gene duplication event. Genes often diverge in function after duplicating; therefore, paralogous sequences are not assumed to share a common function.

- **Query sequence** – your sequence; the sequence you are interested in finding more about.

- **High Scoring Segment Pair (HSP) –** 'hits' to the database. A subsequence match between your query sequence and a database sequence returned by BLAST.

- **Local alignment** – a sequence alignment that extends only across part of the sequence.

- **Global alignment** – a sequence alignment that extends across the entire sequence (from end to end).

1. First, we need a query sequence for the search. Let's start with our given gene again, but this time we'll the nucleotide sequence corresponding to the protein sequence, not the protein sequence. First try finding the gene's DNA sequence using GQuery again.

• On the Search NCBI Databases (GQuery) Portal (All Databases) page, search for your given Gene sequence again using the search box Using the Gene (PMS1)

• The first page that comes up is the summary page. Once you're on this page you can move to the database of interest. In this case you probably don't have hits in too many databases since you had a very specific search.

• Choose the Gene link.

- Does the Gene page give you the gene sequence alone?

-What do you get instead?

Note the context specific link menus that pop up when you hover over the graphic of the gene with your mouse pointer. You can click on the green boxes denoting the exons of the gene to get links to various sequences and analyses associated with the gene. Note that the green track is a composite of the mRNA and CDS tracks – click on either the NM_ or NP_ number to see the deconvolution of the green track.

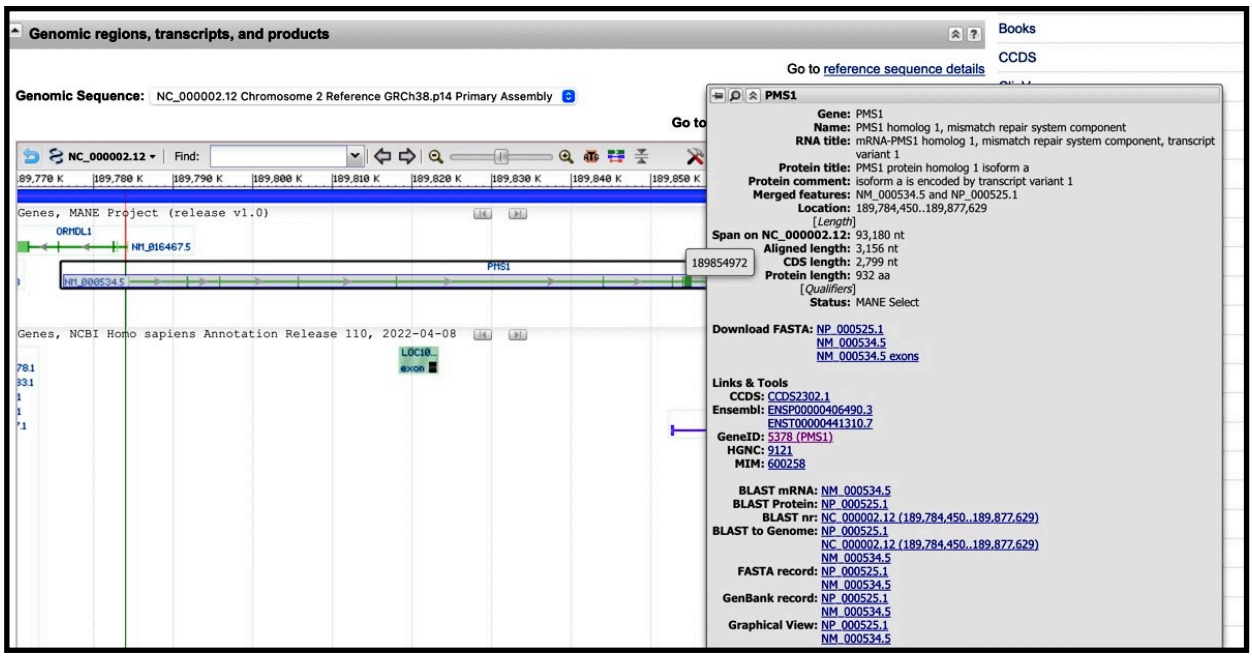


Figure (1): Part of the Gene page for PMS1 homologous 1, showing pop-up to sequence links. 1. Click the green bars to make mRNA and protein tracks appear; 2. Hover over the mRNA track to see info panel; 3. Click on NM_000534.5 link to see GenBank record.

• Click on the mRNA link (NM_000534.5 – the **"M"** in the accession number denotes mRNA – you may notice that this record is identical to the "RefSeq" record you accessed in a different way in Step 10 of the first lab) and select GenBank View (you may need to scroll to the right to access this link; see Figure 1). This takes you to the mRNA that encodes the protein you have been looking at. Notice the feature list in the record. One Feature in the GenBank record is gene, and corresponds to base position 1-3156 on this record. Another features is the coding sequence (CDS), which corresponds to base position165-2963.

a. Given your biology background knowledge, why do you think these are different?

• On the pop-up on the Gene page click on the Nucleotide Link NM_000534.5 , and select GenBank View. This takes you to the genomic region that encodes the mRNA you were just looking at. Notice how the gene feature corresponds to positions 1-3156, while the mRNA feature corresponds to positions 145..296, 297-479, 480-582, 583-746, and 747-863, 864-986, 987-1130, 1131-2020, 2021-2506, 2507-2637, 2638-2798, 2799-3156 and the CDS feature corresponds to positions 165..2963.

b. Again, why are these different? Tip: recall the Central Dogma of Molecular Biology.

- Let's return the mRNA record we were previously working with (NM_000534.5). Click on the CDS link. Now you are looking at the information for the coding sequence, as opposed to the whole gene or protein (highlighted in brown).

- Using the "Display: FASTA" option in the grey bar at the bottom of the page generate a FASTA-formatted version of the CDS.

- Now you have the sequence in the most basic and easily managed format – FASTA format. FASTA format is simply a header line that starts with a '>' followed by text describing the sequence, and then the actual sequence beginning on the next line. The sequence can be either DNA or protein, and may be continuous (scrolling off the page), or cut into more manageable lengths typically ranging between 60-80 residues.
  Figure 10. Sequence in FASTA text format.

2. Let's do some **BLASTing**! Use the "Run BLAST" link in the "Analyze This Sequence" part of the webpage. [Or open a new tab or window in your browser and go back to the NCBI home page (www.ncbi.nlm.nih.gov), then select BLAST from the Resources dropdown along the top, under the DNA&RNA subsection].

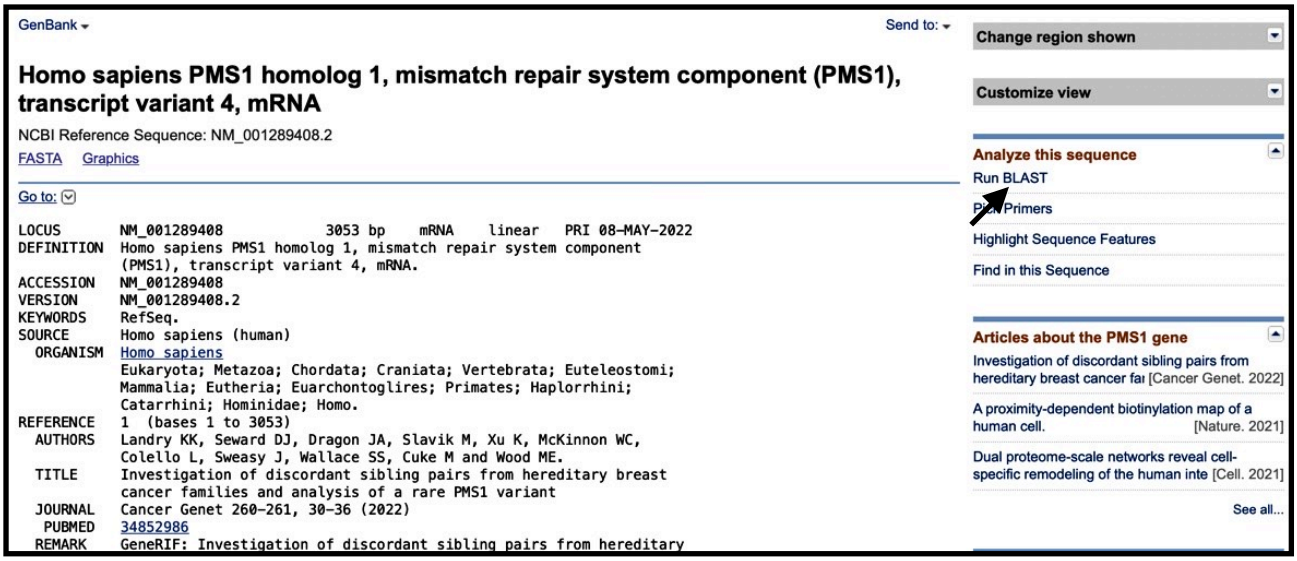
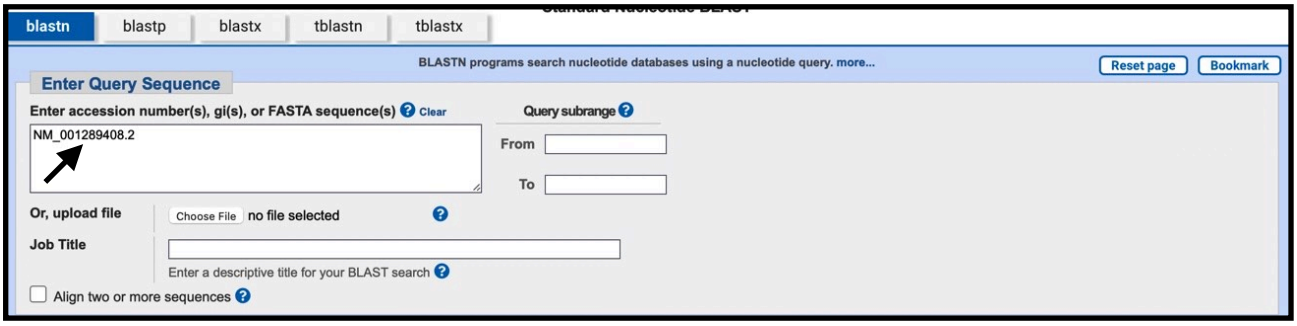


Figure (2): Run blast option on the left bar of gene database of NCBI

-There are lots of options here. Since our sequence is a nucleotide sequence, we want to do a nucleotide blast.



Figure(3): The blastn query page, with optimisation for "Somewhat similar sequences (blastn)" selected.

- On the BLAST page, note that under the Enter Query Sequence section, the NCBI system has automatically entered the accession number (but you can also enter a GI number, or FASTA sequence) and subrange (we'll be searching with just the coding sequence part of the mRNA sequence). You could also copy-and-paste the FASTA formatted CDS sequence you found as in without defining a subrange – you should be clear on the difference between an mRNA sequence and coding sequence at this point...

Box 2:A. Query Input and database selection

The query sequence(s) to be used for a BLAST search should be pasted in the **'Search'** text area. BLAST accepts a number of different types of input and automatically determines the format or the input. To allow this feature there are certain conventions required with regard to the input of identifiers (e.g., accessions or gi's). These are described in 3) below.

Accepted input types are FASTA, bare sequence, or sequence identifiers .

Upload file

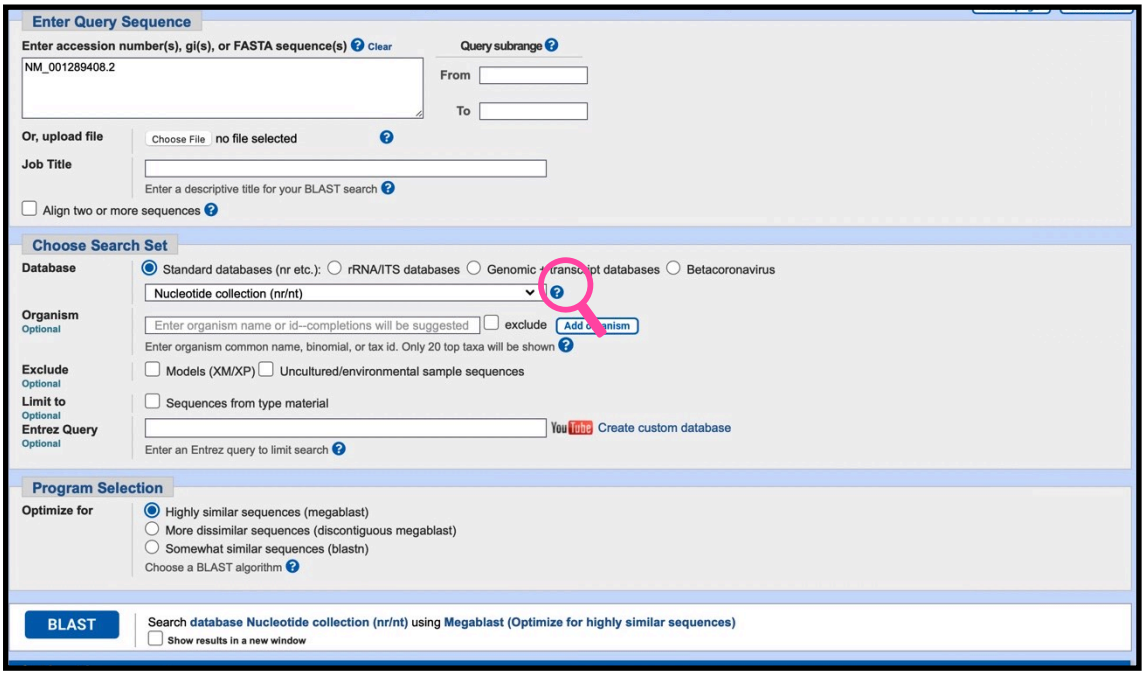
This function allows users to upload a text file containing queries formatted in FASTA format. The file can also contain sequence identifiers instead of FASTA sequences.

Query subrange

A segment of the query sequences can be used in BLAST searching. You can enter the range in the "Form" and "To" boxes provided under "Query subrange" to specify the position of this segment. For example to limit matches to the region from 24 to 200 of a query sequence, you would enter 24 in the "From" field and 200 in the "To" field. If one of the limits you enter is out of range, the intersection of the [From,To] and [1,length] intervals will be searched, where length is the length of the whole query sequence.

Query Genetic Code

Genetic code to be used in blastx and tblastx translation of the query. See list of Genetic Codes in Taxonomy.



Figure(4) : BLAST search parameters.

Box 3 :B. BLAST Search Parameters

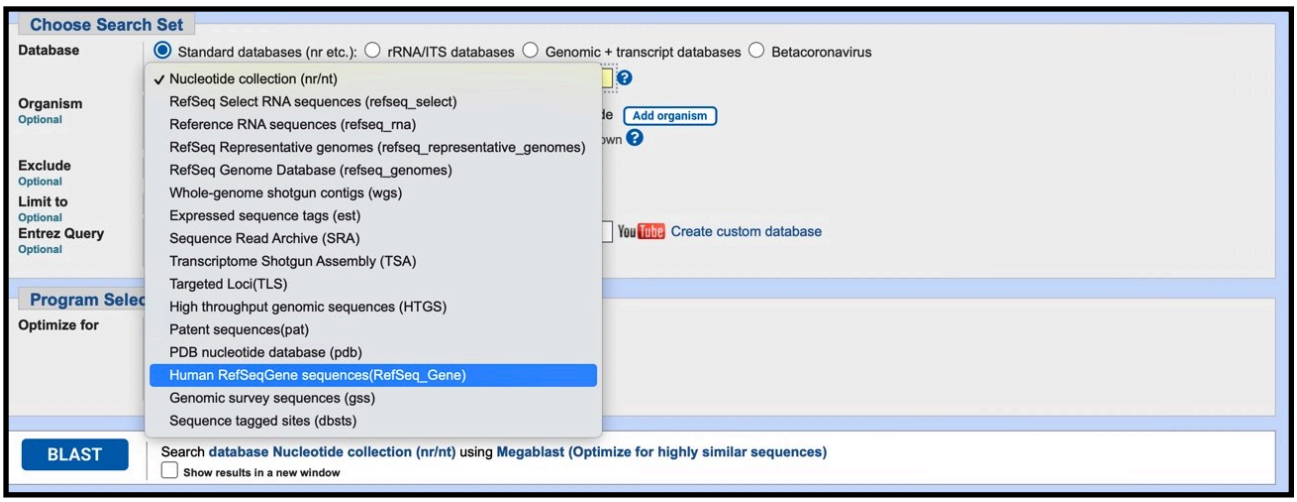
Limit by Organism

A BLAST search may be limited by organism. The entry field will suggest completions once a user starts typing. A checkbox will exclude rather than include the organism in the search.

Limit by Entrez Query

A BLAST search can be limited to the result of an Entrez query against the database chosen. This restricts the search to a subset of entries from that database fitting the requirement of the Entrez query. Terms normally accepted by Entrez nucleotide or protein searches are accepted here. Examples are given below.Scan the sections of the page. You have quite a bit of control over how the algorithm runs (particularly if you click [+] Algorithm parameters near the bottom.

-We want to query the full NCBI database; limit the search to human database .The nr database is the non-redundant collection of sequences in GenBank.



Figure(5): databases options.

- Change the Program Selected / Optimised for to Somewhat similar sequences (blastn).
- Note all the small question mark icons around the page pink search above. Click any one of these to find out more about the associated parameter. For example, by clicking the question mark in the Program Selection section you get a very brief summary of the different methods. By clicking more you jump to a new page with full documentation for the algorithms.

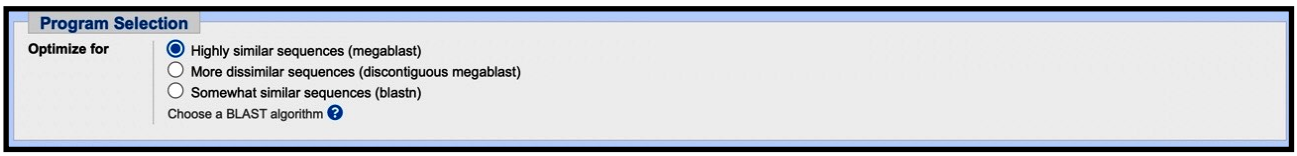


Figure (6):program selection parameters.

a. When would you want to use megaBLAST? What about discontinuous megaBLAST? (if you have time, try each to see how your results differ)

- Megablast is intended for comparing a query to closely related sequences and works best if the target percent identity is 95% or more but is very fast.
- Discontiguous megablast uses an initial seed that ignores some bases (allowing mismatches) and is intended for cross-species comparisons.
- BlastN is slow, but allows a word-size down to seven bases.

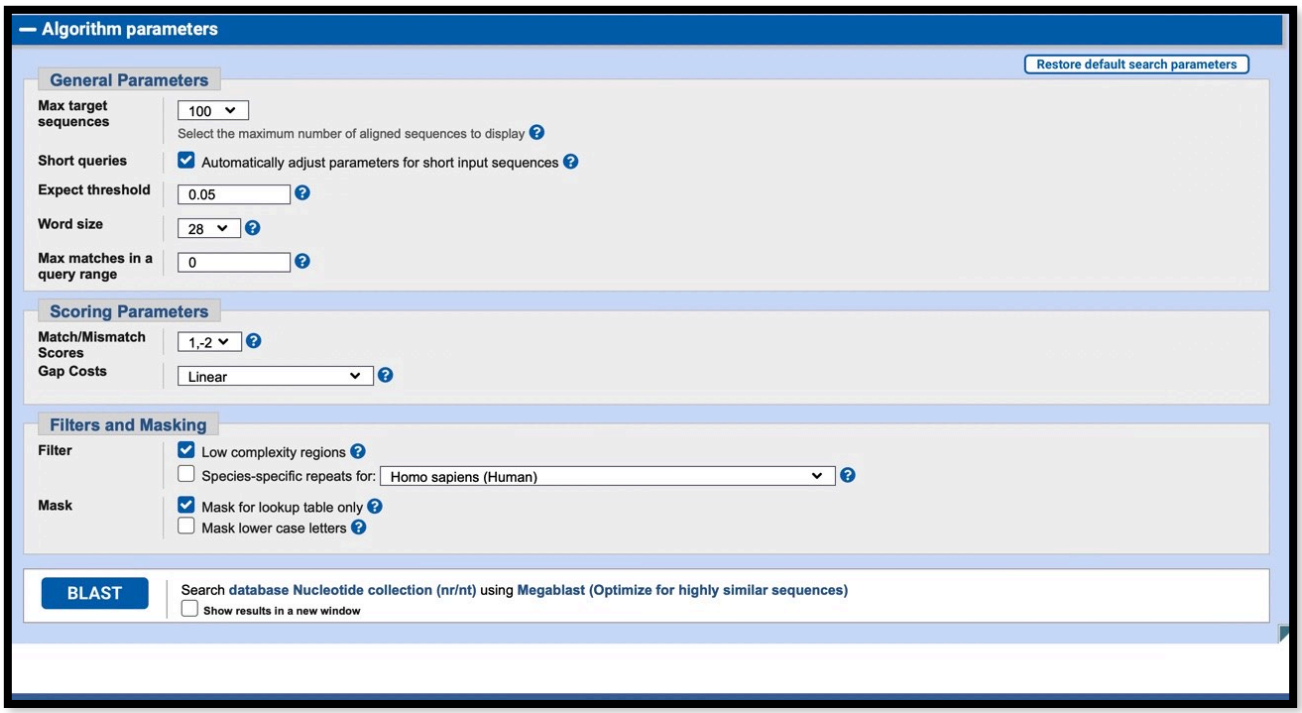
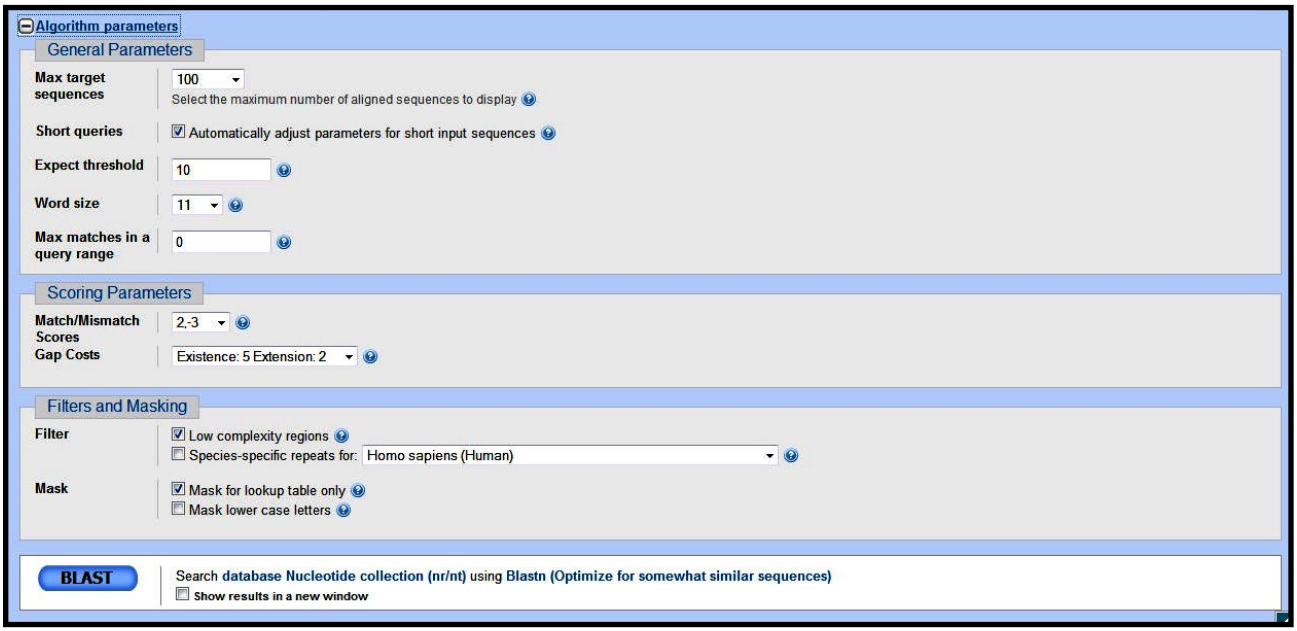


Figure (7): Algorithm parameters for blastn.

- Open the Algorithm Parameters near the bottom.
    - What is the Expect threshold?
    - What would happen if you decreased it? Increased it?
    - What would be the effect of increasing the Word size?
    - Why is there a Low complexity regions filter? Should we keep it on?

- Make sure you have your query sequence entered in the input box, and check the box next to Show results in a new window near the BLAST button. Now (finally) click the BLAST button.

- While BLAST is running or after the search is complete you can choose to adjust the format of the search results by clicking on the Format options link. We won't do this right now, as the defaults usually work fine.



Figure(8): custom Algorithm parameters.

Box 4:Word-size :

BLAST is a heuristic that works by finding word-matches between the query and database sequences. One may think of this process as finding "hot-spots" that BLAST can then use to initiate extensions that might eventually lead to full-blown alignments. For nucleotide-nucleotide searches (i.e., "blastn") an exact match of the entire word is required before an extension is initiated, so that one normally regulates the sensitivity and speed of the search by increasing or decreasing the word-size. For other BLAST searches non-exact word matches are taken into account based upon the similarity between words. The amount of similarity can be varied. The webpage allows the word-sizes 2, 3, and 6.

Box 5 : Algorithm parameters for BLAST:

Filter

- **Filter (Low-complexity)** This function mask off segments of the query sequence that have low compositional complexity, as determined by the SEG program of Wootton and Federhen (Computers and Chemistry, 1993) or, for BLASTN, by the DUST program of Tatusov and Lipman. Filtering can eliminate statistically significant but biologically uninteresting reports from the blast output (e.g., hits against common acidic-, basic-or proline-rich regions), leaving the more biologically interesting regions of the query sequence available for specific matching against database sequences.
  Filtering is only applied to the query sequence (or its translation products), not to database sequences. Default filtering is DUST for BLASTN, SEG for other programs.
  It is not unusual for nothing at all to be masked by SEG, when applied to sequences in SWISS-PROT or refseq, so filtering should not be expected to always yield an effect. Furthermore, in some cases, sequences are masked in their entirety, indicating that the statistical significance of any matches reported against the unfiltered query sequence should be suspect. This will also lead to search error when default setting is used.
- **Filter (Human repeats)** This option masks Human repeats (LINE's, SINE's, plus retroviral repeats) and is useful for human sequences that may contain these repeats. Filtering for repeats can increase the speed of a search especially with very long sequences (>100 kb) and against databases which contain large number of repeats (htgs). This filter should be checked for genomic queries to prevent potential problems that may arise from the numerous and often spurious matches to those repeat elements.
  For more information please see "Why does my search timeout on the BLAST servers?" in the BLAST Frequently Asked Questions.
- **Filter (Mask for lookup table only)** BLAST searches consist of two phases, finding hits based upon a lookup table and then extending them. This option masks only for purposes of constructing the lookup table used by BLAST so that no hits are found based upon low-complexity sequence or repeats (if repeat filter is checked). The BLAST extensions are performed without masking and so they can be extended through low-complexity sequence.
- **Mask Lower Case** With this option selected you can cut and paste a FASTA sequence in upper case characters and denote areas you would like filtered with lower case. This allows you to customise what is filtered from the sequence during the comparison to the BLAST databases.

One can use different combinations of the above filter options to achieve optimal search result.

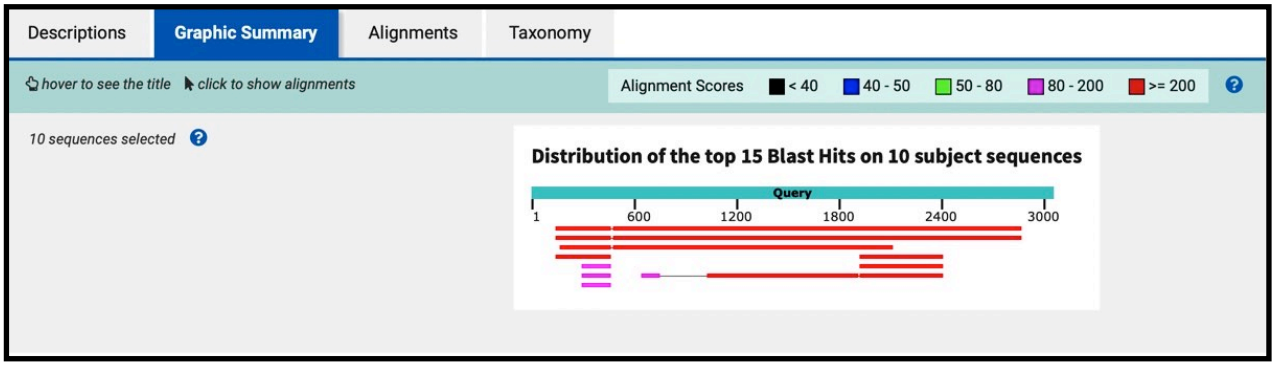
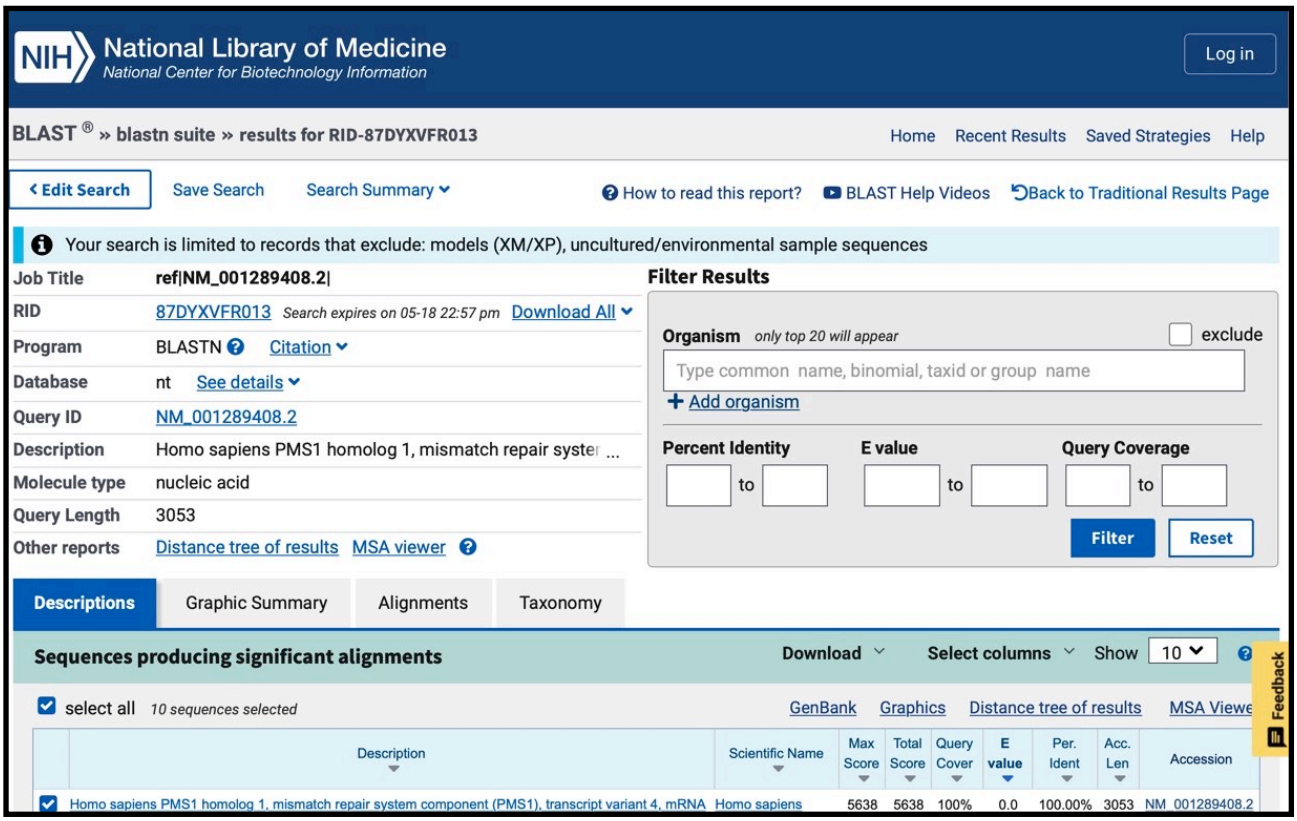


Figure (9): Graphic summary result of BLASTn PMS1.

At the very top is the job summary, which simply shows details about your query and the database searched. You can find more details about your search by clicking Search Summary.

- How many sequences are in the nt/nr database?
- What sequences are not included in the nt/nr database? (Trick question: this information is actually available by clicking on the question mark beside the Database option on the input page!)

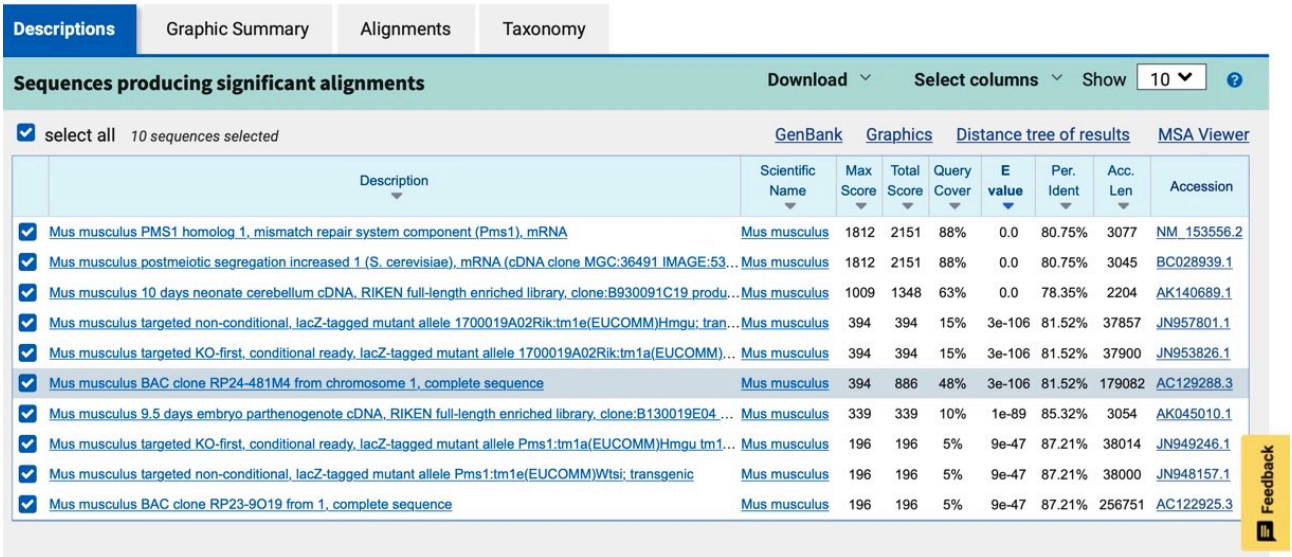


Figure(10): Summary page of BLASTn result.

•-Next is the Graphic Summary. Scroll your mouse over the coloured bars. c. What do the coloured bars mean?

- How does the colour code work?
- What information is displayed in the box near the top of the graphic summary?
- What do you notice about the significance values as you move down the graphical summary?
- What is the genus and species of the top (best) hit?
- What happens if you click on one of the entries?

• **The Descriptions section is next, listing:**



Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download Select columns Show 10

select all *10 sequences selected* GenBank Graphics Distance tree of results MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Mus musculus PMS1 homolog 1, mismatch repair system component (Pms1), mRNA	Mus musculus	1812	2151	88%	0.0	80.75%	3077	NM_153556.2
Mus musculus postmeiotic segregation increased 1 (S. cerevisiae), mRNA (cDNA clone MGC:36491 IMAGE:53...	Mus musculus	1812	2151	88%	0.0	80.75%	3045	BC028939.1
Mus musculus 10 days neonate cerebellum cDNA, RIKEN full-length enriched library, clone:B930091C19 produ...	Mus musculus	1009	1348	63%	0.0	78.35%	2204	AK140689.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele 1700019A02Rik:tm1e(EUCOMM)Hmgu; tran...	Mus musculus	394	394	15%	3e-106	81.52%	37857	JN957801.1
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele 1700019A02Rik:tm1a(EUCOMM)...	Mus musculus	394	394	15%	3e-106	81.52%	37900	JN953826.1
Mus musculus BAC clone RP24-481M4 from chromosome 1, complete sequence	Mus musculus	394	886	48%	3e-106	81.52%	179082	AC129288.3
Mus musculus 9.5 days embryo parthenogenote cDNA, RIKEN full-length enriched library, clone:B130019E04 ...	Mus musculus	339	339	10%	1e-89	85.32%	3054	AK045010.1
Mus musculus targeted KO-first, conditional ready, lacZ-tagged mutant allele Pms1:tm1a(EUCOMM)Hmgu tm1...	Mus musculus	196	196	5%	9e-47	87.21%	38014	JN949246.1
Mus musculus targeted non-conditional, lacZ-tagged mutant allele Pms1:tm1e(EUCOMM)Wtsi; transgenic	Mus musculus	196	196	5%	9e-47	87.21%	38000	JN948157.1
Mus musculus BAC clone RP23-9O19 from 1, complete sequence	Mus musculus	196	196	5%	9e-47	87.21%	256751	AC122925.3

Feedback

Figure(11): Blastn output descriptions

- Description [hyperlinked to corresponding Alignment(s) in Alignments section]
- Max Score – the alignment bit score
- Total Score – another alignment bit score which may differ from the Max Score if your query matched a single database entry in multiple regions.
- Query Coverage – what percent of the query had similarity to the database hit.

- E-value – probably the best measure of hit quality. Smaller numbers mean better hits, with 0.0 being the best value possible.
- Identity – the highest identity found between query and HSP.
- Accession – linked to the indicated sequence at NCBI

How many sequence matches are listed for this query sequence? How are they ordered? (you can sort these segments in other ways, like by identity, score, and query start position.)
What happens if you click the Accession hot-link?

What happens if you click the Alignments hot-link?

Figure (12): Blastn output alignments

Finally we get down to the actual HSP Alignments.

- Compare the information presented for the first HSP alignment to the first entry in the graphical summary and HSP summary.
- As you scroll down the alignments, you will see the alignment quality drop – that is, the e-value increases.

l. What do the vertical bars (|)represent between the Query and the Subject (database sequence)?

What does Strand=Plus/Plus, Strand=Plus/Minus mean? Hint: are genes always in the same direction on a piece of chromosomal DNA?

• Go back to the top of the page and click Formatting options. Change the Alignment View to Query-anchored with dots for identities. Click Reformat and score down to the HSP alignment section.

Describe the difference between this format and the previous format. Can you imagine cases where the different formats might be most useful?

o. Play with these format options to get a feel for what they mean.

• Return the formatting to the original Pairwise format. Go back to the graphical summary. If there are any low-scoring segments (i.e.: green or blue-coded blocks), click on one.

- What is its E-value?
- Does it have a high percent identity? If so, why would BLAST give it such a poor E-value?
- Do you think these hits are homologous? Why or why not?

Box 6:Alignment View

- **Pairwise:** The databases alignments are displayed as pairs of matches between query and subject sequence. A middle line between the query and subject sequence displays the status of a letter. For protein alignments (e.g, BLASTP/BLASTX/TBLASTN), identities present the letter, conservative substitutions present a "+", and nothing otherwise. For nucleotide alignments (e.g., BLASTN and megaBLAST) a "|" is shown for matches and nothing for mismatches. This is the default view.
- **Pairwise with dots for identities:** The databases alignments are anchored (shown in relation to) to the query sequence in pairwised fashion with mismatches colored in red. Subject will be in red and bold font if a line in the alignment contains mismatches. See example below.
- **Query-anchored with dots for identities:** The databases alignments are anchored (shown in relation to) to the query sequence. Identities are displayed as dots (**.**), with mismatches displayed as single letter abbreviations.
- **Query-anchored with letters for identities:** Identities are shown as single letter nucleotide abbreviations.
- **Flat Query-anchored with dots for identities:** The 'flat' display shows inserts as deletions on the query. Identities are displayed as dots (**.**), with mismatches displayed as single letter abbreviations.
- **Flat Query-anchored with letters for identities:** The 'flat' display shows inserts as deletions on the query. Identities are shown as single letter abbreviations.

Further Reading

Chapter 2 “Information Organization and Sequence Databases” in Concepts in Bioinformatics and Genomics by Jamil Momand and Alison McCurdy, Oxford University Press, 2017. pp 21-37.

SF Altschul , TL Madden , AA Schaffer , J Zhang , Z Zhang , W Miller , and DJ Lipman (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402.

NM Luscombe, D Greenbaum, M Gerstein (2001) What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics 2001:83.

CA Kerfeld, KM Scott (2011) Using BLAST to Teach “E-value-tionary” Concepts. PLoS Biol 9(2): e1001014. http://dx.doi.org/10.1371/journal.pbio.1001014.

Lab 5—MULTIPLE SEQUENCE ALIGNMENT

Objectives:

By the end of this lab (comprising the lab including its boxes, and the lecture)you should:

1-Understand How to use MUSCLE.

2-Understand how to use Clustal W.

3-Differentiate between different sequence alignment tool.

Clustal Omega:

Clustal Omega is a new multiple sequence alignment program that uses seeded guide trees and HMM profile-profile techniques to generate alignme58[1]nts between **three or more** sequences. For the alignment of two sequences please instead use our pairwise sequence alignment tools.

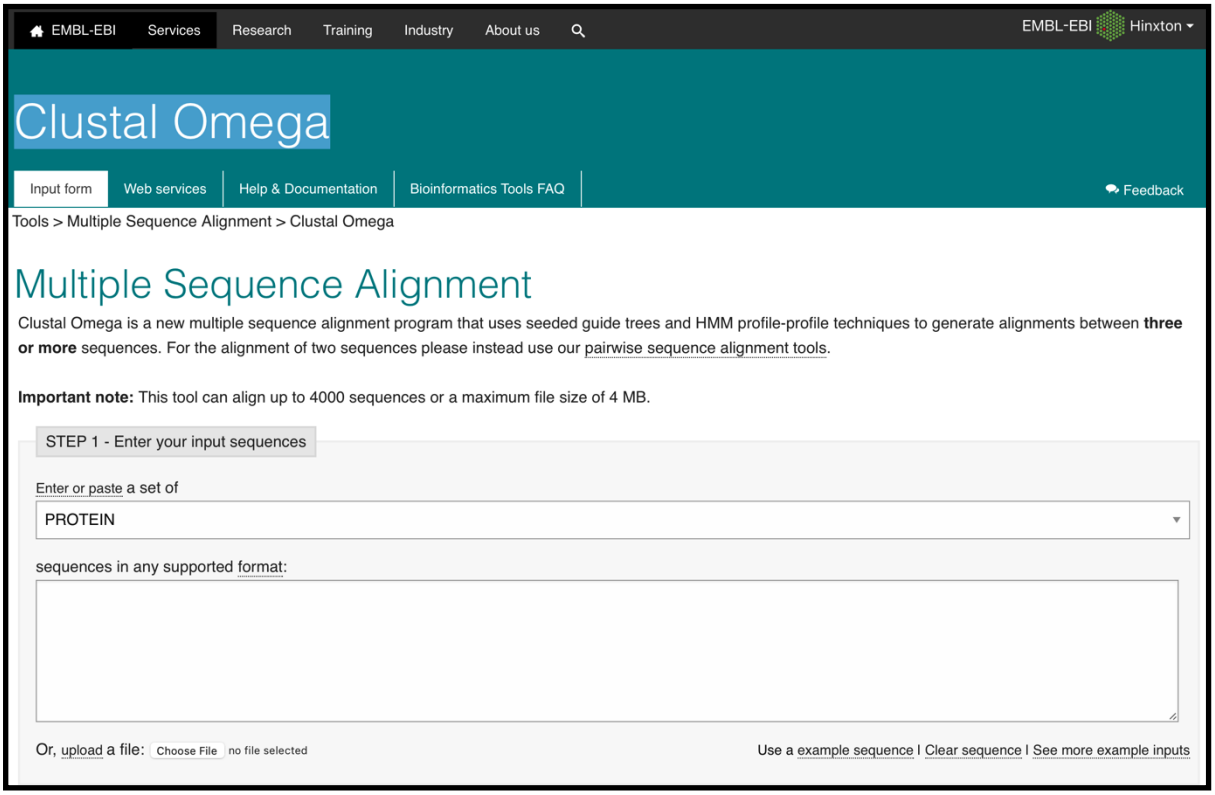


Figure (1): Clustal Omega Sequence Input Window.

Step 1 - Sequence

Sequence Input Window

Three or more sequences to be aligned can be entered directly into this box. Sequences can be in GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR/NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only) format.

Sequence File Upload

A file containing three or more valid sequences in any format (GCG, FASTA, EMBL (Nucleotide only), GenBank, PIR, NBRF, PHYLIP or UniProtKB/Swiss-Prot (Protein only)) can be uploaded and used as input for the multiple sequence alignment.

Sequence Type(PROTEIN, DNA, and RNA).

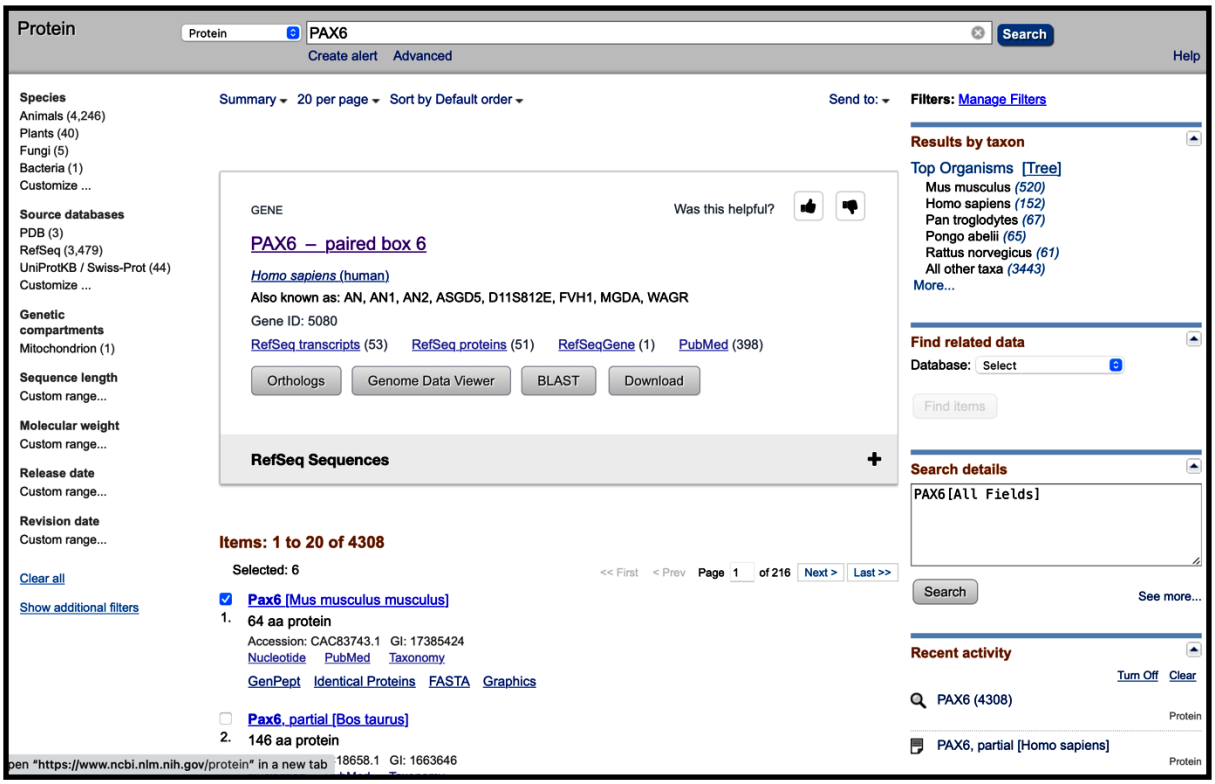


Figure (2): PAX6 search result on NCBI protein database.

-GO to NCBI protein database and search PAX6 gene, then download the FASTA for 6 different sequences.

-Then upload the file into the sequence upload window in the Clustal omega.

-After you upload your file then click on Submit.

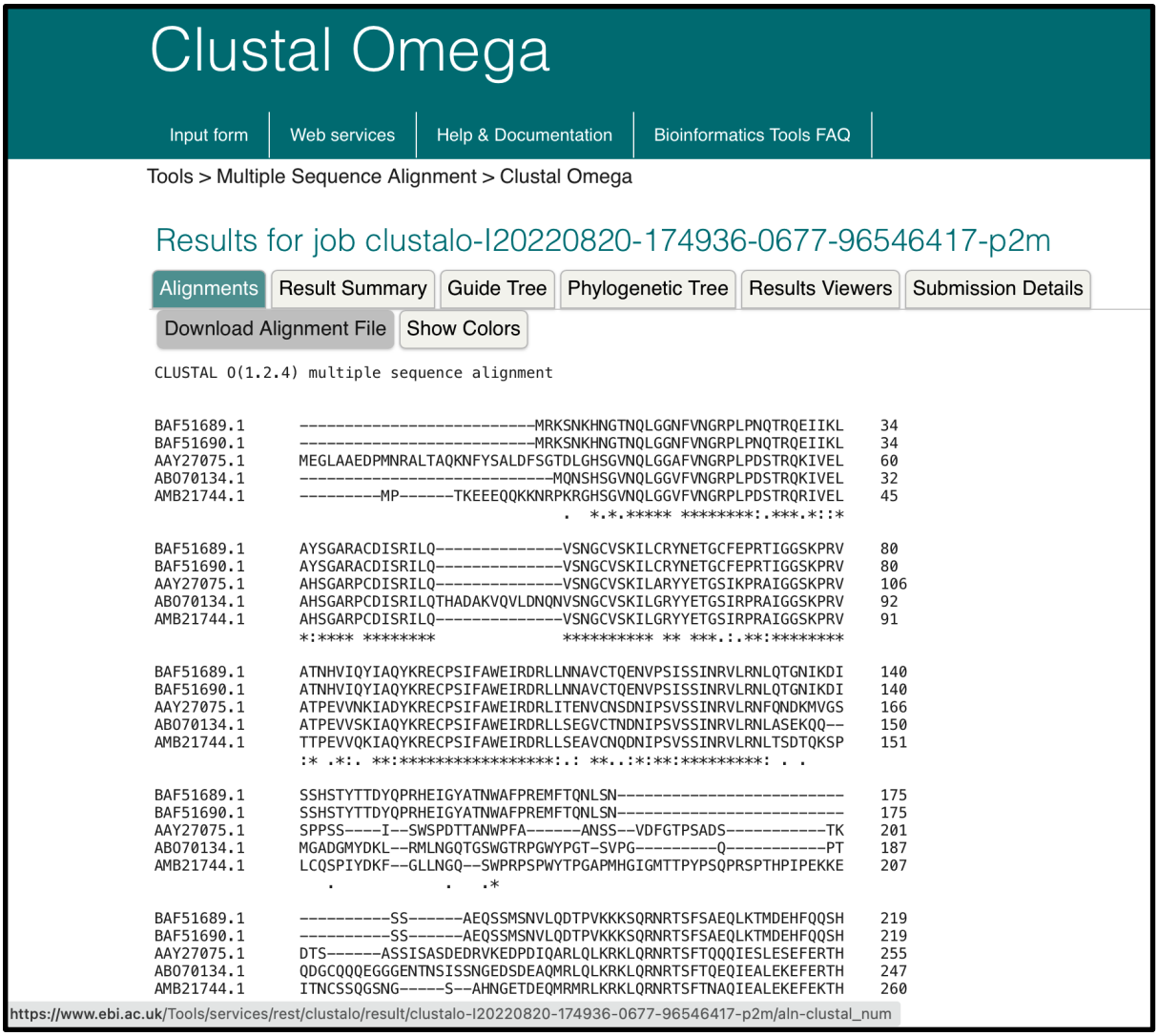


Figure (3): Alignment result of Clustal Omega.

What does (*, : , .) denotes?

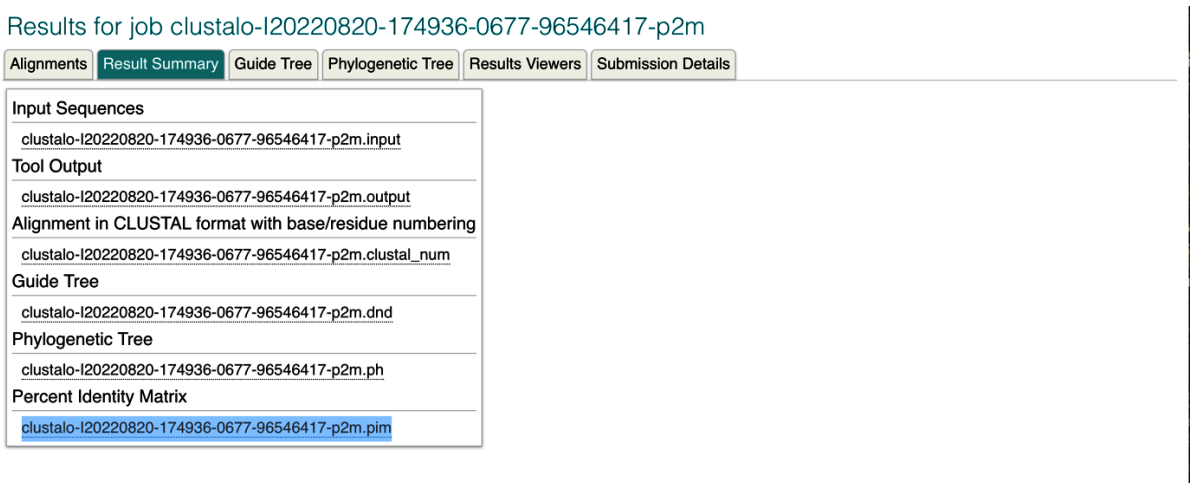


Figure (4): Result summary of Clustal Omega.

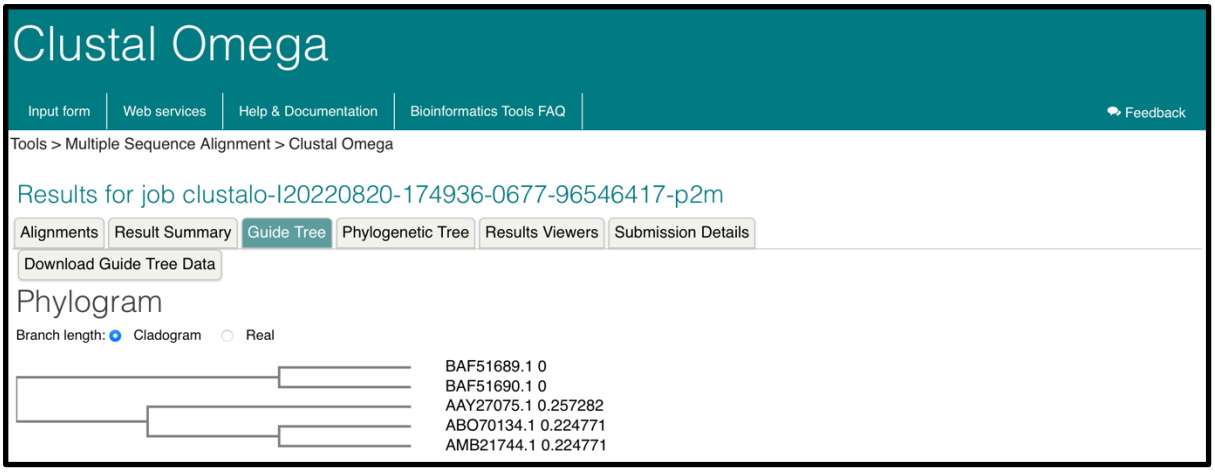


Figure (5): Guide tree of Clustal Omega.

BOX1: Multiple Sequence Alignment online tools:

COBALT (Constraint-based Multiple Alignment Tool) New

COBALT computes a multiple protein sequence alignment using conserved domain and local sequence similarity information.

ClustalW (everybody uses it),

MUSCLE (very fast)

BOX 2: Cabalistic signs:

(*) A **star** indicates an **entirely conserved** column.

(:) A **colon** indicates columns where all the residues have roughly the **same size** and the same **hydropathy**.

(.) A **period** indicates columns where the **size *OR*** the **hydropathy** has been **preserved** in the course of **evolution**.

MUSCLE:

MUSCLE stands for MUltiple Sequence Comparison by Log-Expectation.
MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.
MUSCLE enables high-throughput applications to achieve average accuracy comparable to the most accurate tools previously available, which is expected to be increasingly important in view of the continuing rapid growth in sequence data.
Multiple alignments of protein sequences are important in many applications, including phylogenetic tree estimation, secondary structure prediction and critical residue

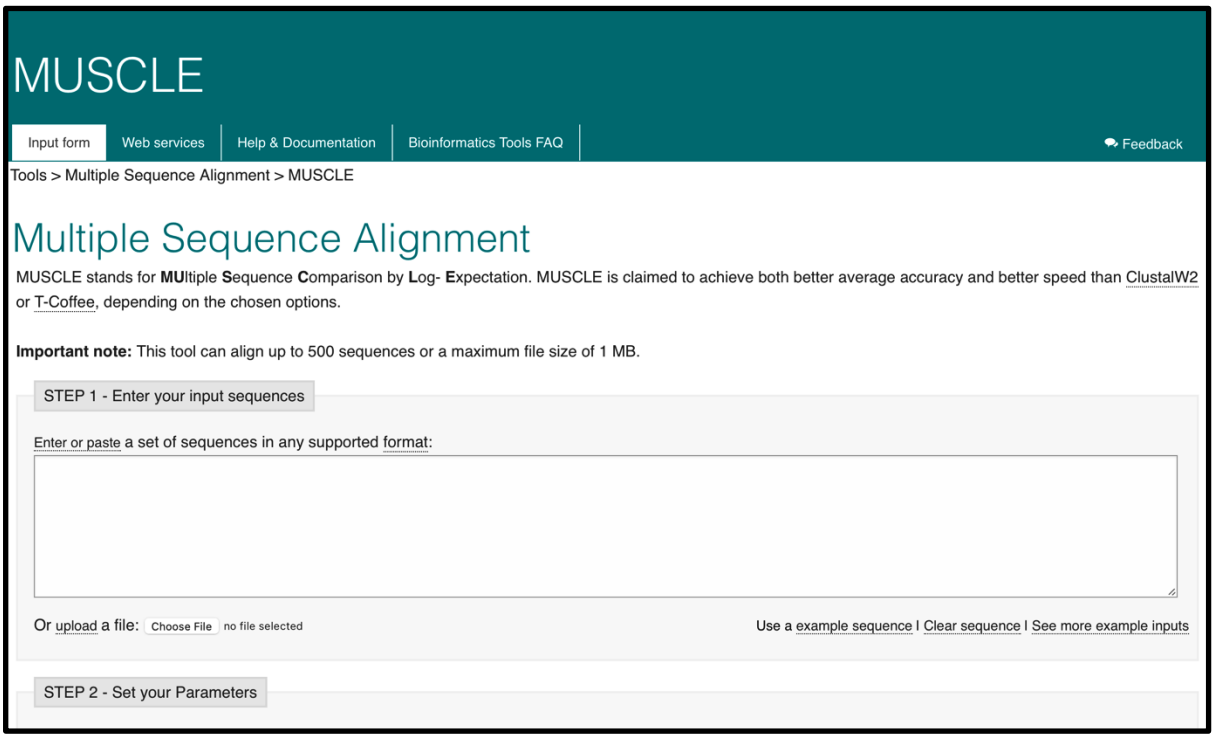


Figure (6): MUSCLE Sequence input box.

Upload the same file used in Clustal Omega. To compare.

Click submit and run the tool.

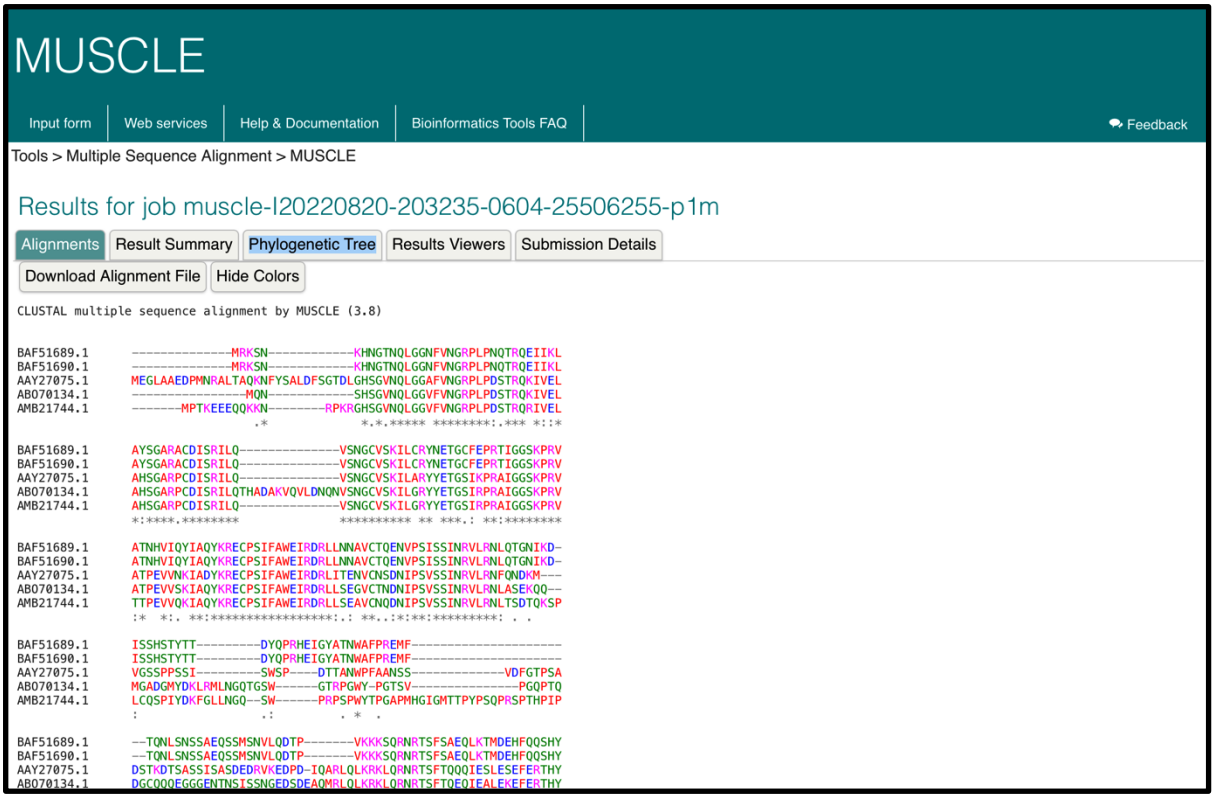


Figure (7): Colored sequence alignment of MUSCLE.

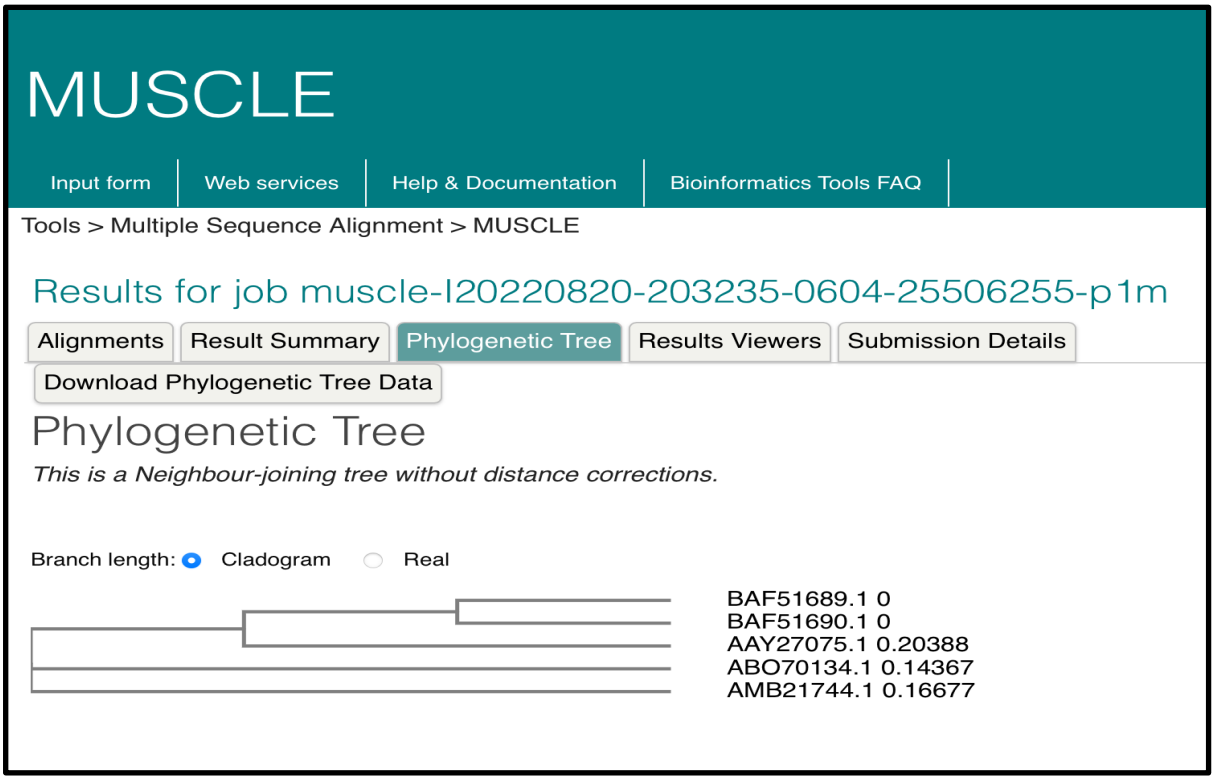


Figure (8): phylogenetic tree of MUSCLE.

NOTE:

-It is common to find conserved tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and is therefore difficult to mutate.

-It is common to find conserved columns with a glycine or a proline in a multiple alignment. These two amino acids often coincide with the extremities of well-structured beta strands or alpha helices.

-Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein domains and folds.

-Histidine and serine are often involved in catalytic sites, especially those of proteases. Conserved histidine or a conserved serine are good candidates for being part of an active site.

-K (Lysine), R (Arginine), D (Aspartic Acid), E (Glutamic Acid)These charged amino acids are often involved in ligand binding. Highly conserved columns can also indicate a salt bridge inside the core of the protein.

Lab 6 —VARIANT ANNOTATION AND SCORING:

Objectives:

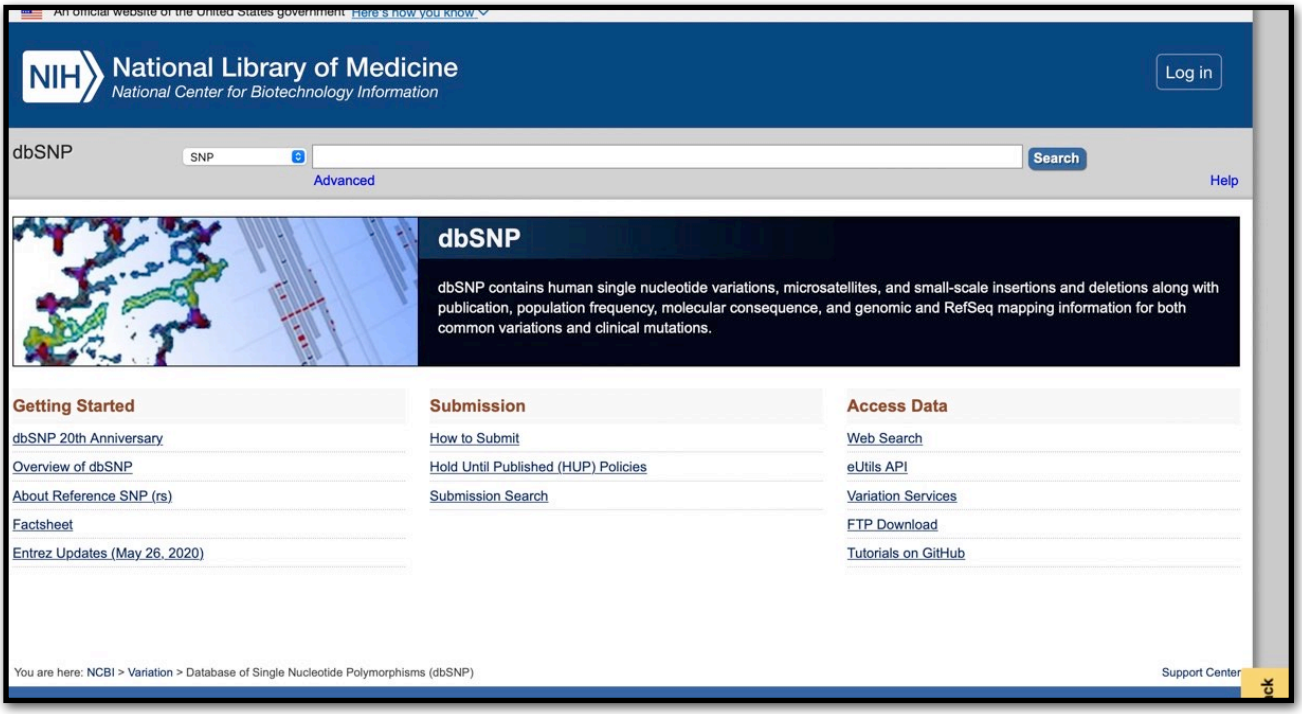
By the end of this lab (comprising the lab including its boxes, and the lecture)you should:

1. To Know how to use and interpret data on SNP database and ClinVar.
2. To be able to interpret variants according to ACMG guidelines.

SNP database:

The dbSNP has been designed to support submissions and research into a broad range of biological problems. These include physical mapping, functional analysis, pharmacogenomics, association studies, and evolutionary studies. Because dbSNP was developed to complement GenBank, it may contain nucleotide sequences from any organism.

dbSNP only assigned RefSNP for human organisms as an outcome of the recent collaborations with EMBL-EBI European Variation Archive (EVA). dbSNP Build 152 (November 2018) contains more than 650 million human RefSNP records, of which over 580 million records have population frequency data.



Figure(1):dbSNP start page.

In the search box write PMS1.

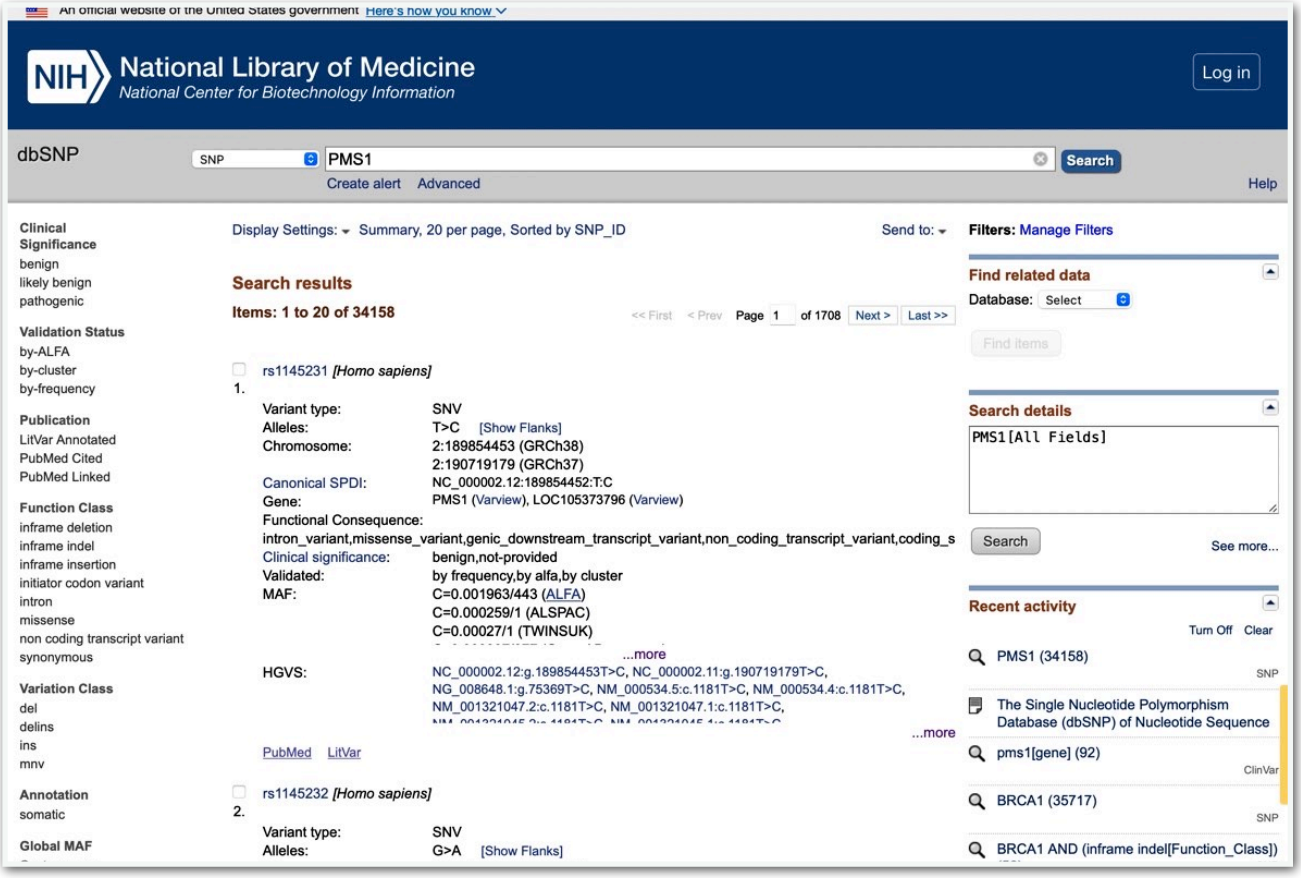
See the result page.

Answer the following questions

1. How many SNPs are returned?

Look at the right bar .What do you see?

Box(1):
You can access the dbSNP through the Entrez Gene page again, use the 'Links' menu on the right side to view the link out choices and select the 'SNP' option.
This will automatically query the Entrez SNP database for all SNPs in dbSNP for the any gene for species you are viewing (i.e., 'homo sapiens').



Figure(2) : dbSNP result page of PMS1.

2. Below the search box and tabbed menu choices (i.e., 'Advanced search,' etc), the 'Display' feature menu to show this list as a 'FASTA'. The page should automatically update when you make your selection.

3. In the 'Send To' drop down menu, select the 'Text' option. The page should Update the results in plain text format. This selection can be directly copied to a file on your computer.

4. Use the 'BACK' button on your browser. Alternatively this data can be "Sent To' a 'File' directly, that is saved on your computer.

Box(2): Definitions.

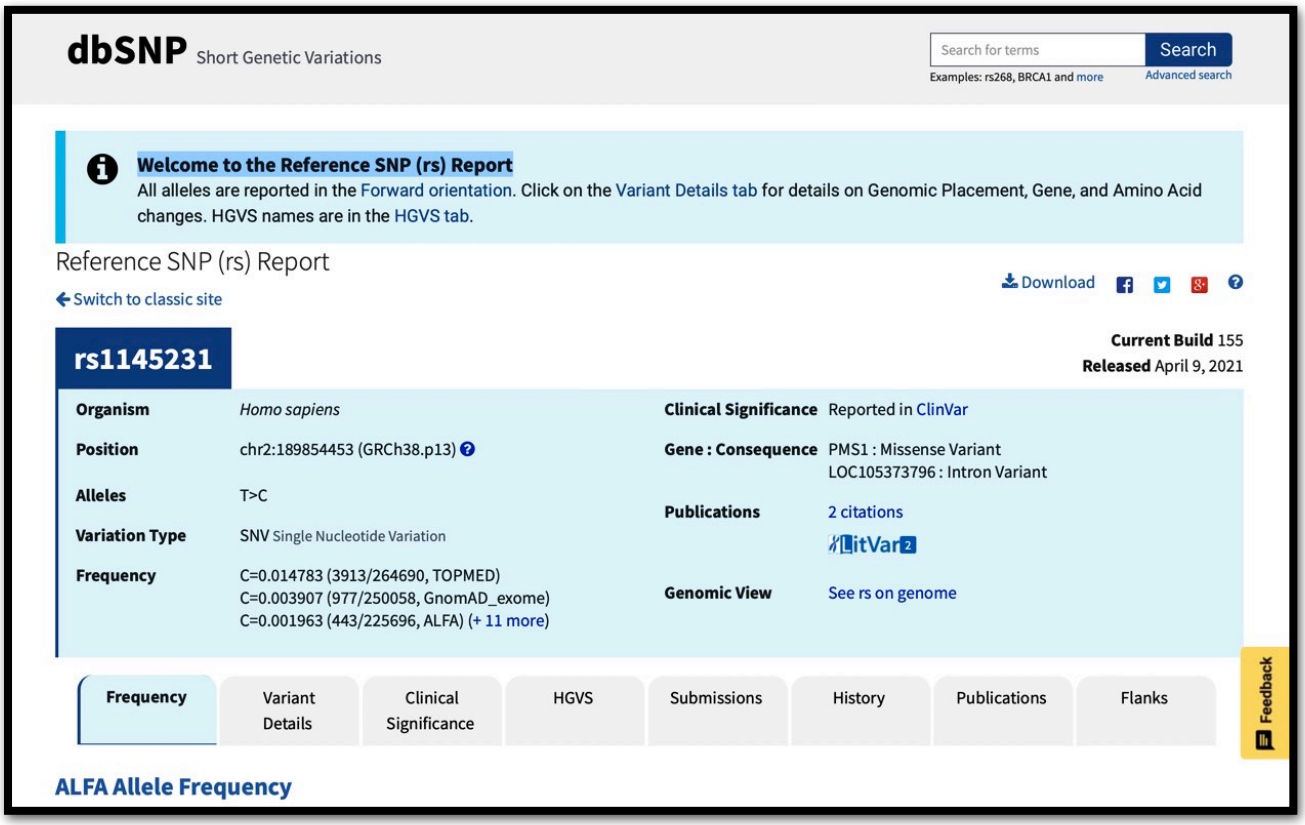
Variant type: **An alteration in the most common DNA nucleotide sequence. The term variant can be used to describe an alteration that may be benign, pathogenic, or of unknown significance.**

Canonical: **the longest transcript, though not necessarily the most biologically relevant**

MAF: **the frequency at which the second most common allele occurs in a given population.**

Press on any SNP rs:

Now explore the page which you encountered



Figure(3): Reference SNP report with rs1145231.

This page reports data for a single dbSNP Reference SNP variation (RefSNP or rs) from the new redesigned dbSNP build.

Top of the page reports a concise summary for the rs, with more specific details included in the corresponding tabs below.

All alleles are reported in the Forward orientation. Use the Genomic View to inspect the nucleotides flanking the variant, and it's neighbours.

For more information see Help documentation.

Download

Search:

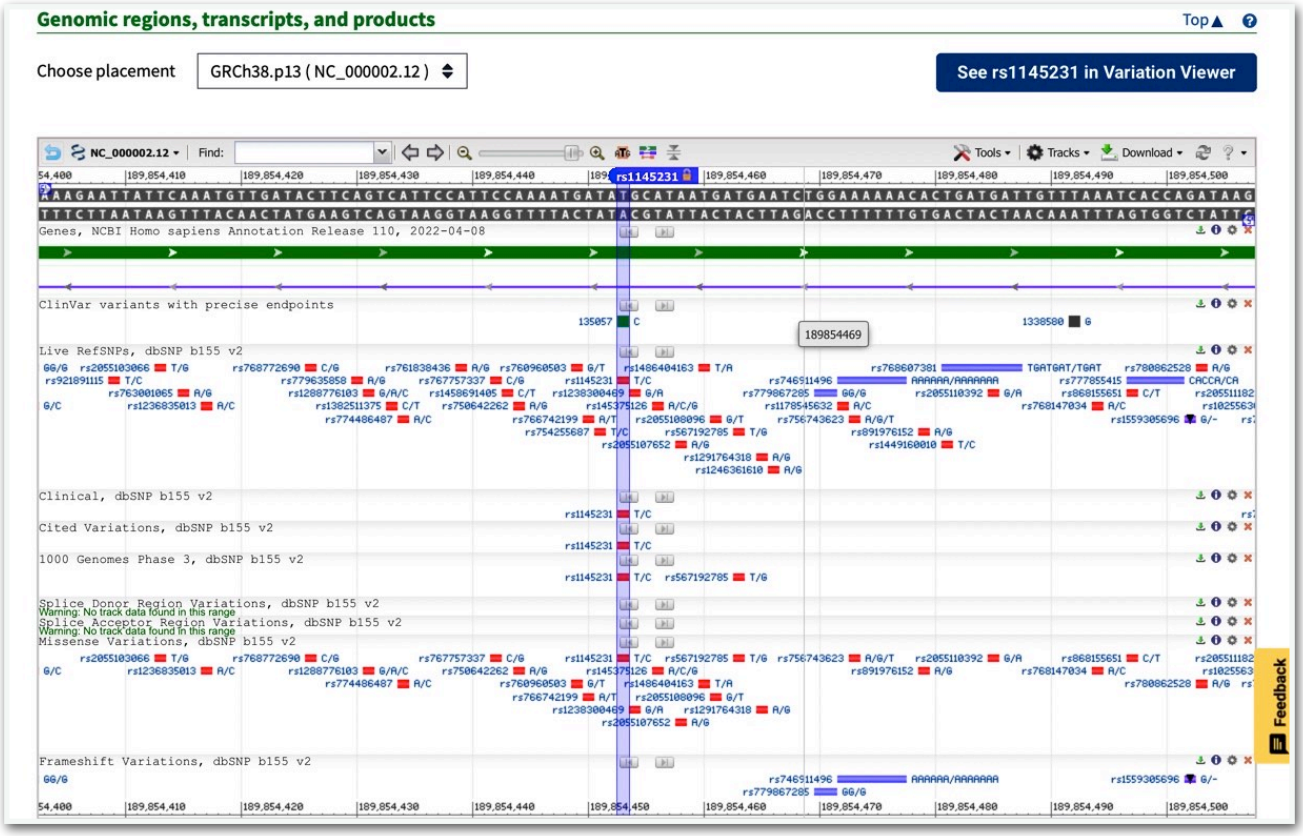
Study	Population	Group	Sample Size	Ref Allele	Alt Allele
The PAGE Study	PuertoRican	Sub	7918	T=0.9898	C=0.0102
The PAGE Study	NativeHawaiian	Sub	4534	T=0.9982	C=0.0018
The PAGE Study	Cuban	Sub	4230	T=0.9891	C=0.0109
The PAGE Study	Dominican	Sub	3828	T=0.9726	C=0.0274
The PAGE Study	CentralAmerican	Sub	2450	T=0.9910	C=0.0090
The PAGE Study	SouthAmerican	Sub	1982	T=0.9924	C=0.0076
The PAGE Study	NativeAmerican	Sub	1260	T=0.9929	C=0.0071
The PAGE Study	SouthAsian	Sub	856	T=0.999	C=0.001
TopMed	Global	Study-wide	264690	T=0.985217	C=0.014783
UK 10K study - Twins	TWIN COHORT	Study-wide	3708	T=0.9997	C=0.0003
South Asian	Sub		296	T=1.000	C=0.000

Figure(4): Allele frequency tables.

-Frequency tab displays a table of the reference and alternate allele frequencies reported by various studies and populations.

-Table lines, where Population="Global" refer to the entire study population, Whereas lines, where Group="Sub", refer to a study-specific population sub-groupings (i.e. AFR, CAU, etc.), if available.

-Frequency for the alternate allele (Alt Allele) is a ratio of samples observed-to-total, where the numerator (observed samples) is the number of chromosomes in the study with the minor allele present (found in "Sample size", where Group="Sub"), and the denominator (total samples) is the total number of all chromosomes in the study for the variant (found in "Sample size", where Group = "Study-wide" and Population = "Global").



Figure(5): Genomic region using NCBI Graphical Sequence Viewer.

-NCBI Graphical Sequence Viewer display of the genomic region, transcripts and protein products for the reported RefSNP (rs).

-Use the zoom option to view the nucleotides around the RefSNP and find other neighboring RefSNPs.

-Visit Sequence Viewer for help with navigating inside the display and modifying the selection of displayed data tracks.

What is a Reference SNP?

- The dbSNP Reference SNP (rs or RefSNP) number is a locus accession for a variant type assigned by dbSNP.

-The RefSNP catalog is a non-redundant collection of submitted variants which were clustered, integrated and annotated. RefSNP number is the stable accession regardless of the differences in genomic assemblies.

-RefSNP numbers facilitate large-scale studies in association genetics, medical genetics, functional and pharmacogenomics, population genetics and evolutionary biology, personal genomics, and

precision medicine. They provide a stable variant notation for mutation and polymorphism analysis, annotation, reporting, data mining, and data integration.

Distinguishing RefSNP Features

- Non-redundancy and globally unique accession series (1)
- Composed from over 2 billion Submitted SNP (ss) from thousands of submitters.
- More than 20 years of tracking histories for all assigned, merged, and deleted RefSNP.
- Annotated and linked to the latest human assembly and RefSNP nucleotide and protein sequences.
- Updates to reflect current knowledge of sequence data and biology
- Data validation.
- Ongoing curation and annotation by NCBI staff and collaborators.
- Searchable across variation and genomic databases
- Supported and reported in open-source and commercial software and tools.
- Over 400K RefSNP are in ClinVar
- Cited in over 51K publications with biological, functional, disease, and clinical information for variants across the genomes (2,3,4)
- Linked to many NCBI internal and external resources such as ClinVar, PubMed, PubMedCentral, RefSeq, UCSC, EBI, TopMed, and GnomAD.
- Supports consistent reporting and non-redundant variation annotations across related sequences including alternate haplotypes, GRC patches, and future graph genomes if the alignment or sequence relationship is known.

Variation Type

Despite its name, RefSNP is assigned to all variation types listed below with precise locations for both common and rare variations, including mutations. Most are typically small variations (<= 50bp).

- Single nucleotide variation (SNV)
- Short multi-nucleotide changes (MNV)
- Small deletions or insertions
- Small STR repeats

- retrotransposable element insertions

dbSNP Accession Types

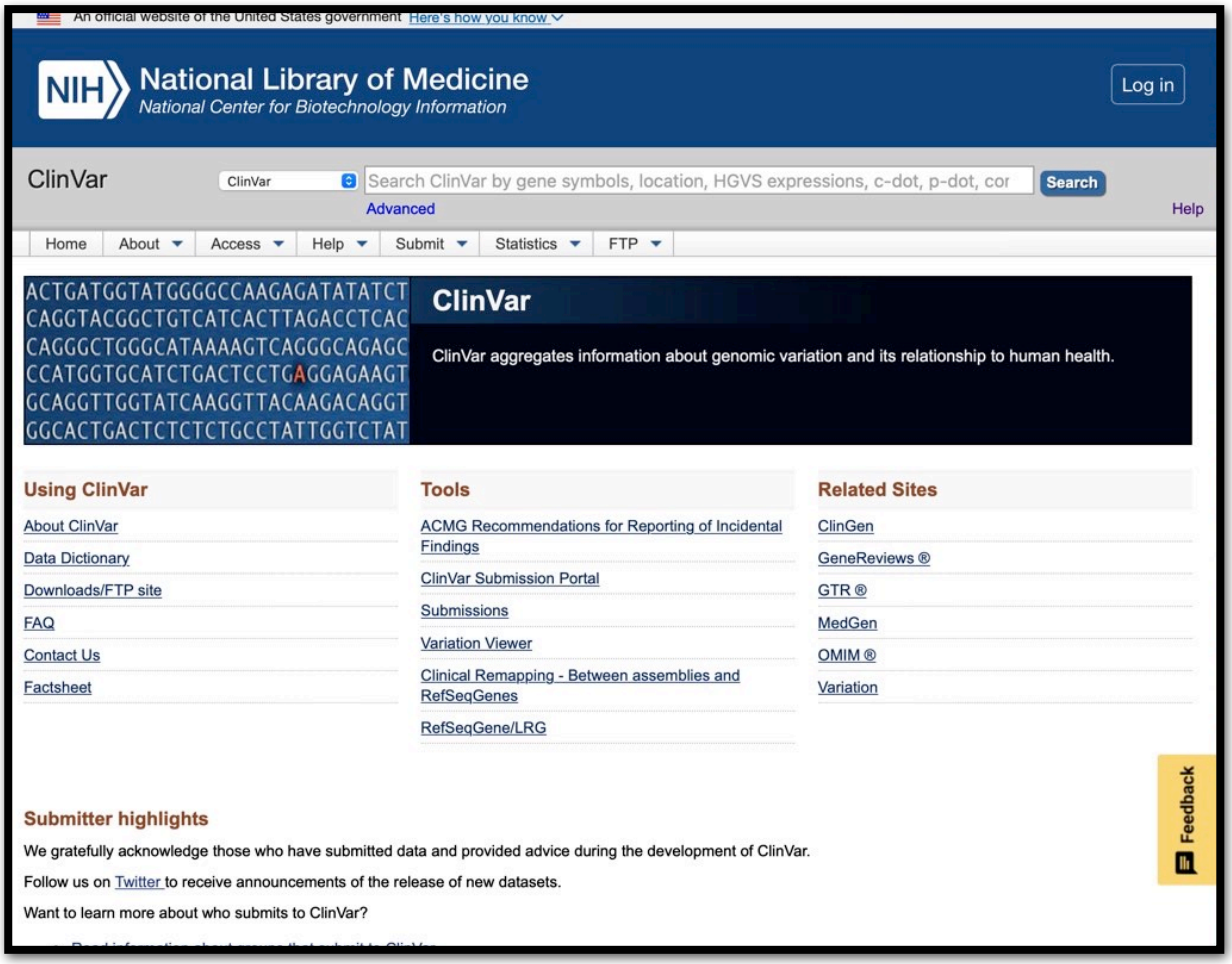
Submitted SNP (ss) – submitted variant based on asserted location or flanking sequences

Reference SNP (rs) - Non-redundant set of variations based on clustering of SS'es of same variant type and sequence position.

ClinVar:

ClinVar is a freely accessible, public archive of reports of the relationships among human variations and phenotypes, with supporting evidence. ClinVar thus facilitates access to and communication about the relationships asserted between human variation and observed health status, and the history of that interpretation.

ClinVar processes submissions reporting variants found in patient samples, assertions made regarding their clinical significance, information about the submitter, and other supporting data. The alleles described in submissions are mapped to reference sequences, and reported according to the HGVS standard. ClinVar then presents the data for interactive users as well as those wishing to use ClinVar in daily workflows and other local applications. ClinVar works in collaboration with interested organizations to meet the needs of the medical genetics community as efficiently and effectively as possible. Read more about using ClinVar.



Figure(6): ClinVar start page.

-ClinVar is an active partner of the ClinGen project, providing data for evaluation and archiving the results of interpretation by recognised expert panels and providers of practice guidelines. ClinVar archives and versions submissions which means that when submitters update their records, the previous version is retained for review. Read more about submitting data to ClinVar.

-ClinVar supports submissions of differing levels of complexity. The submission may be as simple as a representation of an allele and its interpretation (sometimes termed a variant-level submission), or as detailed as providing multiple types of structured observational (case-level) or experimental evidence about the effect of the variation on phenotype.

-A major goal is to support computational (re)evaluation, both of genotypes and assertions, and to enable the ongoing evolution and development of knowledge regarding variations and associated phenotypes.

Implementation

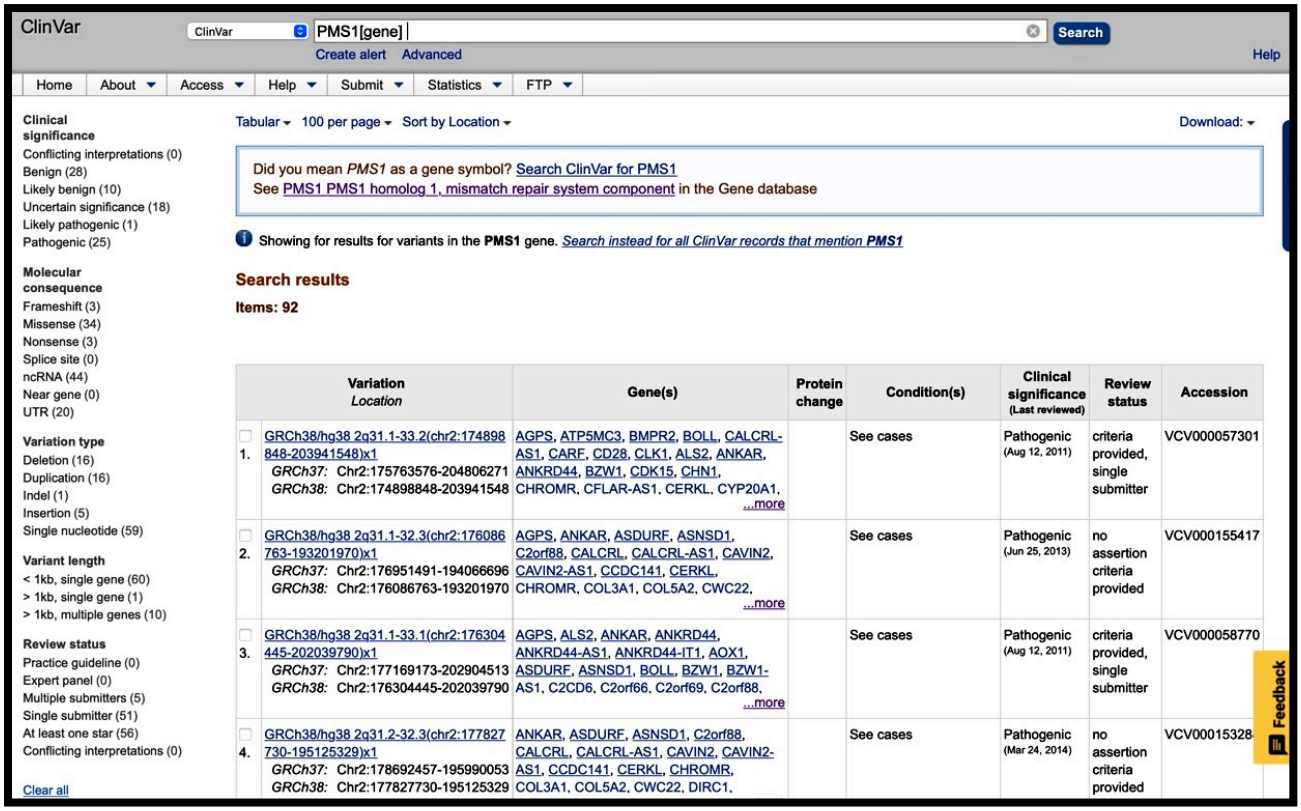
-A preliminary view of ClinVar was launched in 2012, with the first full public release in April 2013. The initial dataset included variations from OMIM, GeneReviews, some locus-specific databases (LSDB), contributing testing laboratories, and others. ClinVar is an active participant in the ClinGen project, leading to improved content and representation of that content. ClinVar continues to evolve in response to the needs of the clinical genetics community.

Scope

-ClinVar accepts variants in any part of the genome and interpreted for any type of condition.

ClinVar currently includes clinical assertions for variants identified through several methods of data collection, including clinical testing, research, and reports from the literature (literature only). See our documentation on submitting collection method for more details.

ClinVar currently does not include uncurated sets of data from GWAS studies, although variants that were identified through GWAS and have been individually curated to provide an interpretation of clinical significance are in scope.



Figure(7): Result Section of ClinVar.

-Check the right bar : Describe what you notice

-Click on one of the variant on the table

ClinVar Genomic variation as it relates to human health

Search by gene symbols, location, HGVS expressions, c-dot, p-dot, conditions, and mor | Search ClinVar

Advanced search

About Access Submit Stats FTP Help

Were new search queries using location, c-dot, and p-dot helpful?

Follow Print Download

NM_000534.5(PMS1):c.2766del (p.His923fs) Cite this record

Interpretation: Pathogenic

Review status: ☆☆☆☆ no assertion criteria provided
Submissions: 1 (Most recent: Feb 21, 2021)
Accession: VCV000998151.1
Variation ID: 998151
Description: 1bp deletion

Variant details
Conditions
Gene(s)

NM_000534.5(PMS1):c.2766del (p.His923fs)

Allele ID: 985851
Variant type: Deletion
Variant length: 1 bp
Cytogenetic location: 2q32.2
Genomic location: 2: 189877398 (GRCh38) GRCh38 UCSC
2: 190742124 (GRCh37) GRCh37 UCSC

HGVS:

Nucleotide	Protein	Molecular consequence
NM_000534.5:c.2766del MANE SELECT	NP_000525.1:p.His923fs	frameshift
NM_000534.5:c.2766delT MANE SELECT		
NM_001128143.2:c.2649del	NP_001121615.1:p.His884fs	frameshift

Canonical SPDI: NC_000002.12:189877397:TTTTTT:TTTTT
Functional consequence: -
Global minor allele frequency (GMAF): -
Allele frequency: -
Links: dbSNP: rs2057666079
VarSome

Feedback

Submitted interpretations and evidence

Interpretation (Last evaluated)	Review status (Assertion criteria)	Condition (Inheritance)	Submitter	More information
Pathogenic (-)	no assertion criteria provided Method: case-control	Colorectal cancer Affected status: yes Allele origin: germline	Genomic Center,National Cancer Institute Accession: SCV001481771.1 Submitted: (Feb 21, 2021)	

Functional evidence

There is no functional evidence in ClinVar for this variation. If you have generated functional data for this variation, please consider submitting that data to ClinVar.

Citations for this variant

There are no citations in ClinVar for this variation. If you know of citations for this variation, please consider submitting that information to ClinVar.

Text-mined citations for rs2057666079 none

These citations are identified by LitVar using the rs number, so they may include citations for more than one variant at this location. Please review the LitVar results carefully for your variant of interest.

Feedback

Figure(8): ClinVar description of variants.

A ClinVar record contains the following elements:

1-ClinVar Accession and version

1. Submission accession number/version number separated by a decimal (SCV000000000.0) assigned to each submitted record.
2. Reference accession number/version separated by a decimal (RCV000000000.0) assigned to sets of submitted records about the same variation/condition pair.
3. Variation accession number/version separated by a decimal (VCV000000000.0) assigned to sets of submitted records about the same variation.

2-Identifiers for each variant allele or allele set

1. HGVS expressions
2. Published allele names
3. Database identifiers

3-Attributes of each phenotype

1. Name
2. Descriptions
3. Defining features
4. Database identifiers

4-Description of the genotype/phenotype relationship

1. Review status of the asserted relationship
2. Submitter of the assertion
3. Clinical significance - see full documentation on clinical significance
4. Summary of the evidence for clinical significance
    1. Number of observations of genotype/allele in those with the phenotype
    2. Number of observations of genotype/allele in those without the phenotype
    3. Family studies
    4. Description of the population sampled
    5. In vitro studies
    6. In silico studies
    7. Animal models
5. **Mode of inheritance**

6. Study design
7. Citations, including URLs

Submission information

1. Submitter description
2. Dates submitted and updated
3. Data added by NCBI computation

Detailed descriptions of the data elements are available in the ClinVar Data Dictionary .

Box(4): ClinVar Accessions

Accessions, with the format SCV000000000.0, are assigned to each submitted record. If there are multiple submitted records about the same variation/condition pair, they are aggregated within ClinVar's data flow and reported as a reference accession with the format RCV000000000.0. ***Because of this model, one variant will be included in multiple RCV accessions whenever different conditions are reported for that variant***.

Submitted records for the same variation are also aggregated and reported as an accession with the format VCV000000000.0. This aggregation lets a user review all submitted data for a variant, regardless of the condition for which it was interpreted.

*ClinVar archives submitted information, and adds identifiers and other data that may be available about a variant or condition from other public resources. However ClinVar neither curates content nor modifies interpretations independent of an explicit submission. If you have data that differs from what is currently represented in ClinVar, we encourage you to submit your data and the evidence supporting your interpretation. There is a submission wizard to guide you through that process.

Represents medical phenotypes

-ClinVar aggregates the names of medical conditions with a genetic basis from such sources as SNOMED CT, GeneReviews, Genetic Home Reference, Office of Rare Diseases, MeSH, and OMIM®. ClinVar also aggregates descriptions of associated traits from Human Phenotype Ontology (HPO), OMIM, and other sources. Each source of information is tracked, and can be used in queries.

Represents variations

-Human variations are reported to the user as sequence changes relative to an mRNA, genomic and protein reference sequence (if appropriate), according to the HGVS standard. The defaults are as ‘c.’ and any protein sequence change. Genomic sequences are represented in RefSeqGene/LRG

coordinates, as well as locations on chromosomes (as versioned accessions and per assembly name, such as NCBI36/hg18 and GRCh37/hg19). Novel variations are accessioned in NCBI's variation databases (dbSNP and dbVar).

Represents the relationships among phenotypes and variations

ClinVar is designed to support the evolution of our understanding of the relationship between genotypes and medically important phenotypes. By aggregating information about variations observed in individuals with or without a phenotype, ClinVar supports establishment of the clinical validity of human variation.

HOPE(Have (y)Our Protein Explained)

HOPE is an easy-to-use web service that analyses the structural effects of a point mutation in a protein sequence. Input your protein sequence and the mutation and HOPE will collect and combine available information from a series of web services and databases and will produce a report, complete with results, figures and animations.

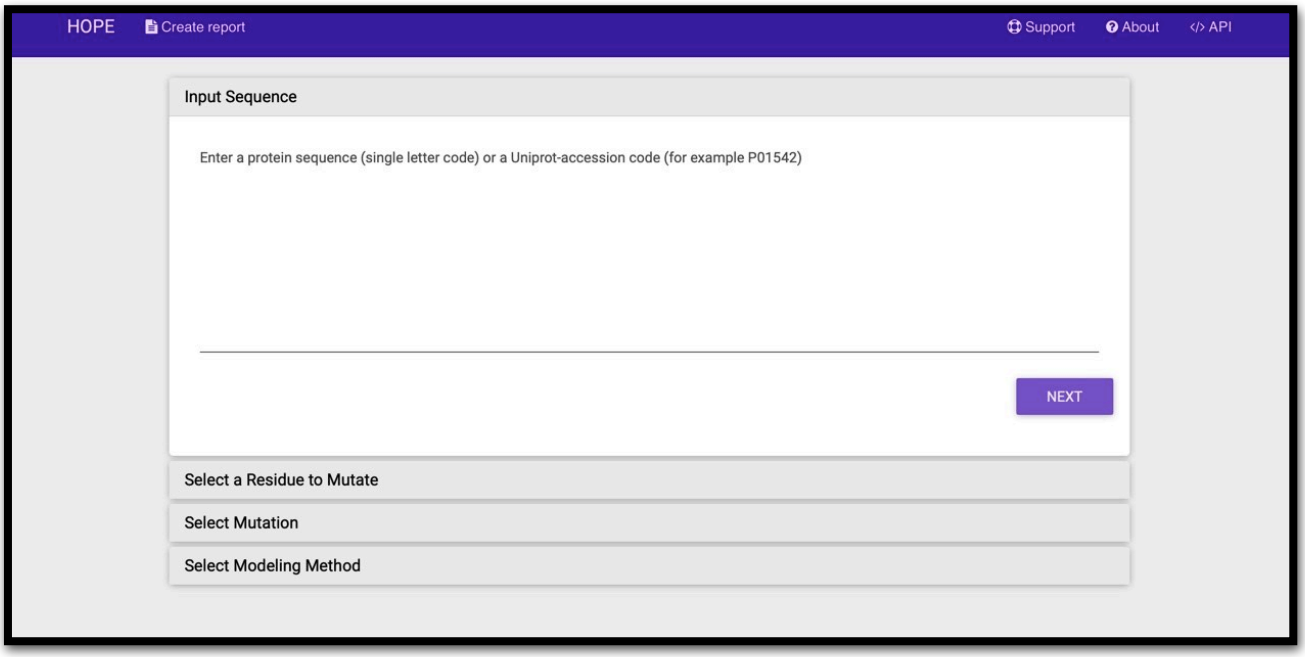
-To explain the molecular origin of a disease related phenotype caused by mutations in human proteins. In this aspect HOPE resembles the aforementioned systems (PolyPhen, SIFT, ALAMUT).

HOPE we have takes the logical next step in the e-Science era in that the data gathering is done using Web services and DAS servers.

-HOPE takes a protein 3D structure centred approach. HOPE collects information from data sources such as the protein's 3D structure and the UniProt database of well-annotated protein sequences.

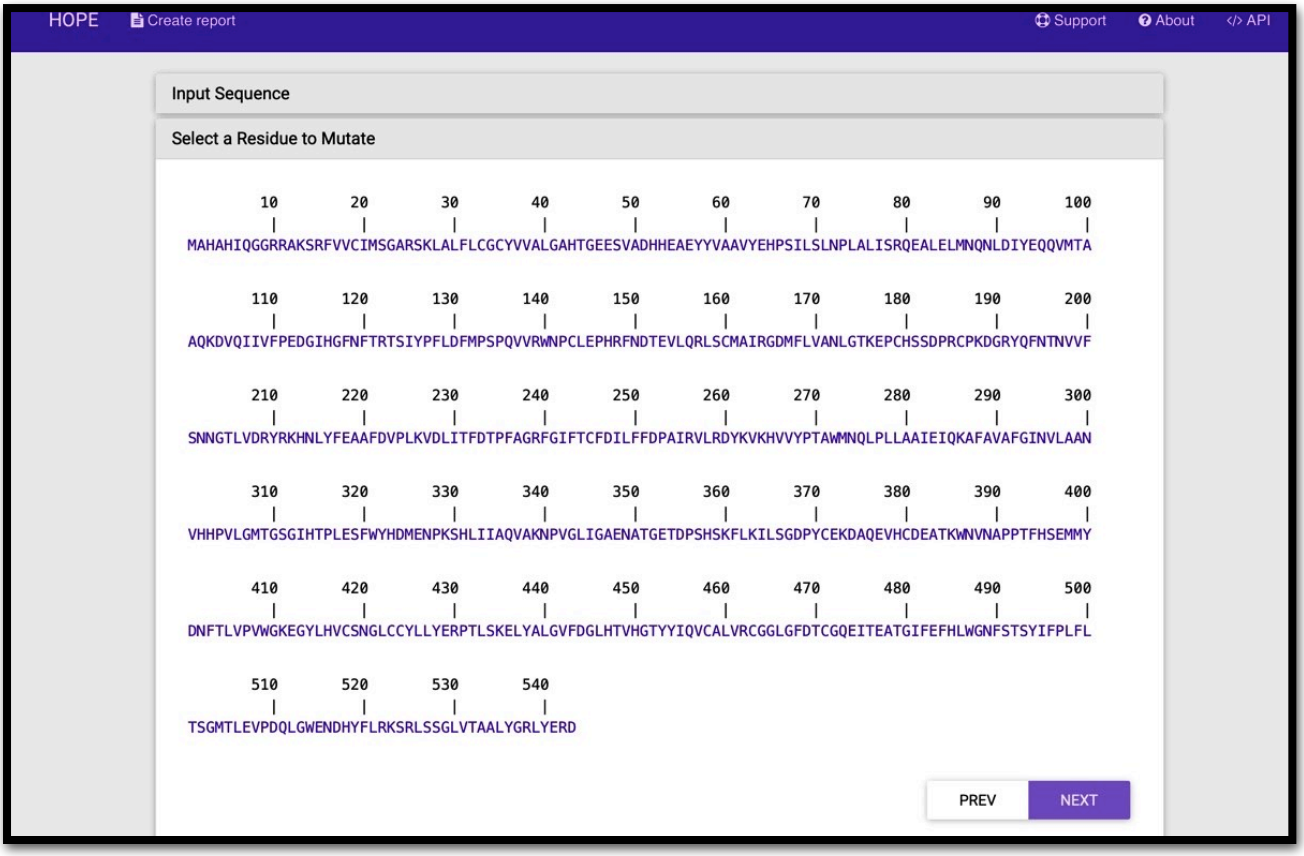
-A life-scientist friendly report is produced that explains and illustrates the effects of the mutation.

-This report is presented using an interface that is designed specifically for the intended user community of human genetics researchers. The report is enriched with figures that illustrate the effects of the mutation, while any residual bioinformatics jargon is linked to our in-house, online dictionary of bioinformatics jargon. The conclusions are drawn in the report .



Figure(9): Sequence deposition window.

Go to UniProt and choose the BTD gene and copy the uniprot ID P43251 to the sequence input box.



Figure(10): Select residue to mutation box.

Now select the amino acid at position 444, which is Aspartic acid.

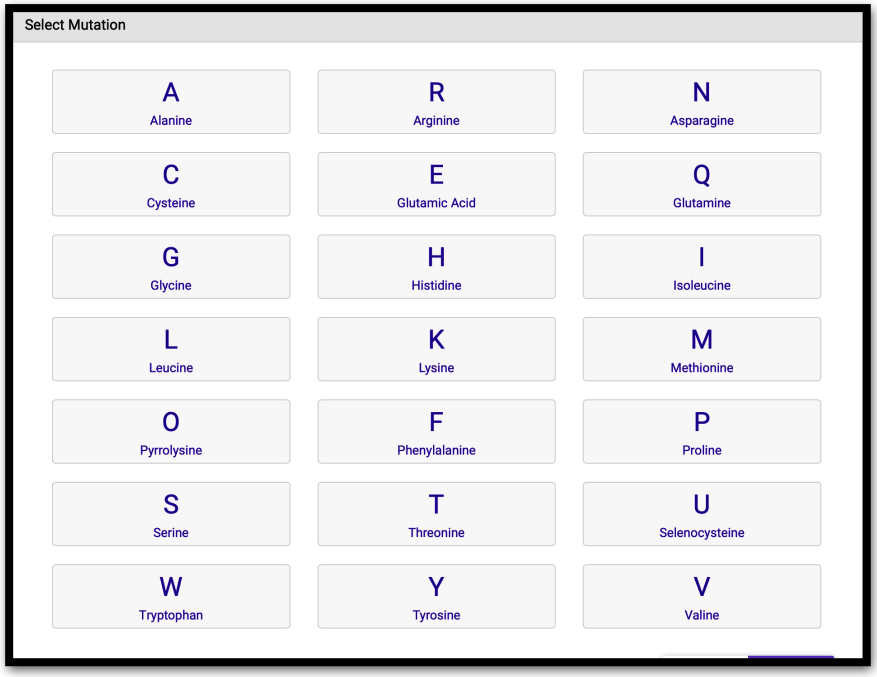
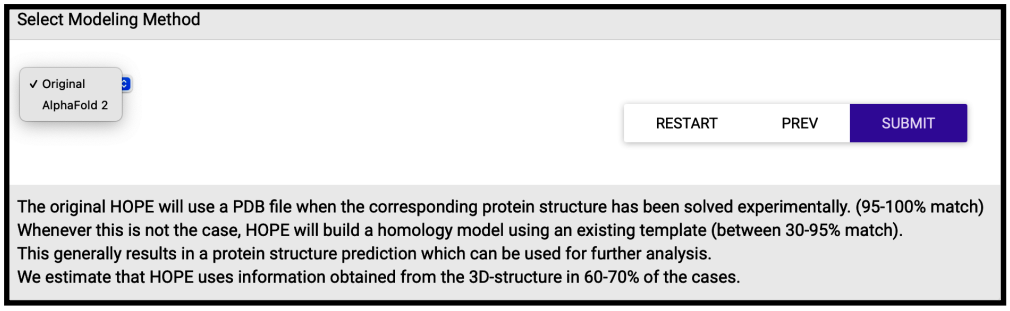


Figure (11): Selection of the substitution amino acid.

-Choose Histidine as the substitution amino acid.

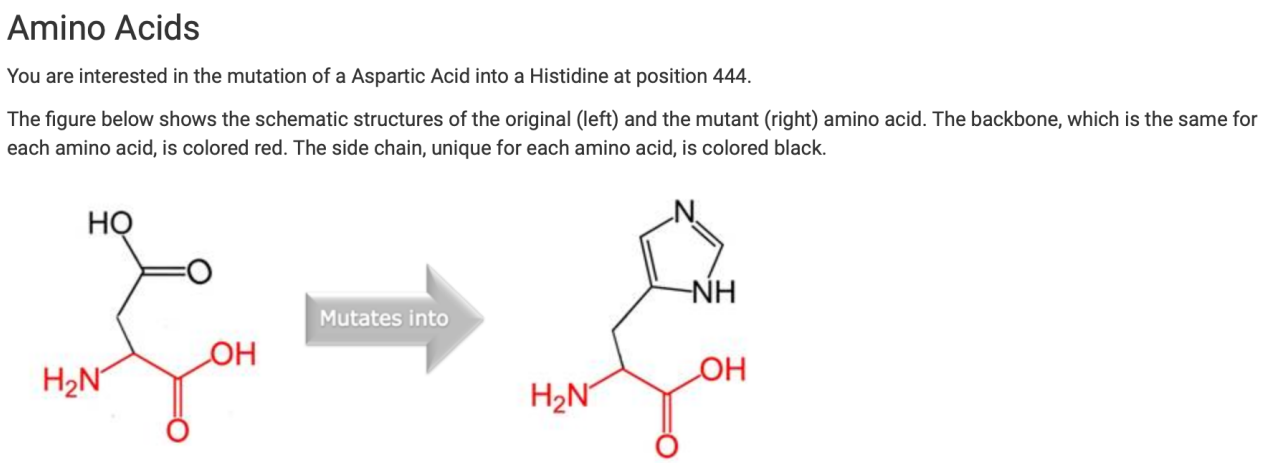


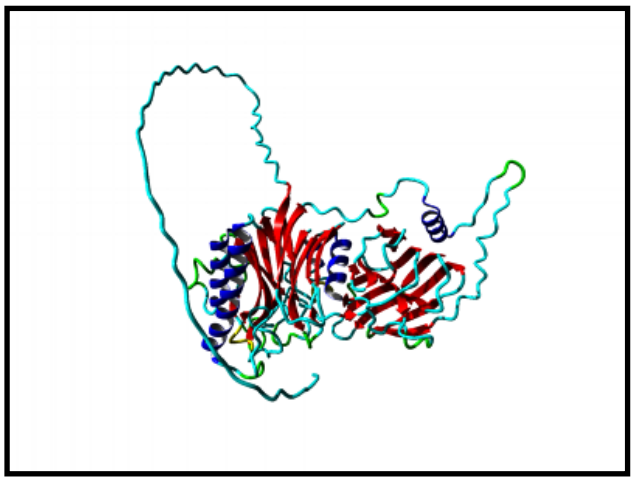
Figure(12): Modelling Method selection box.

This is new section has been added last year.

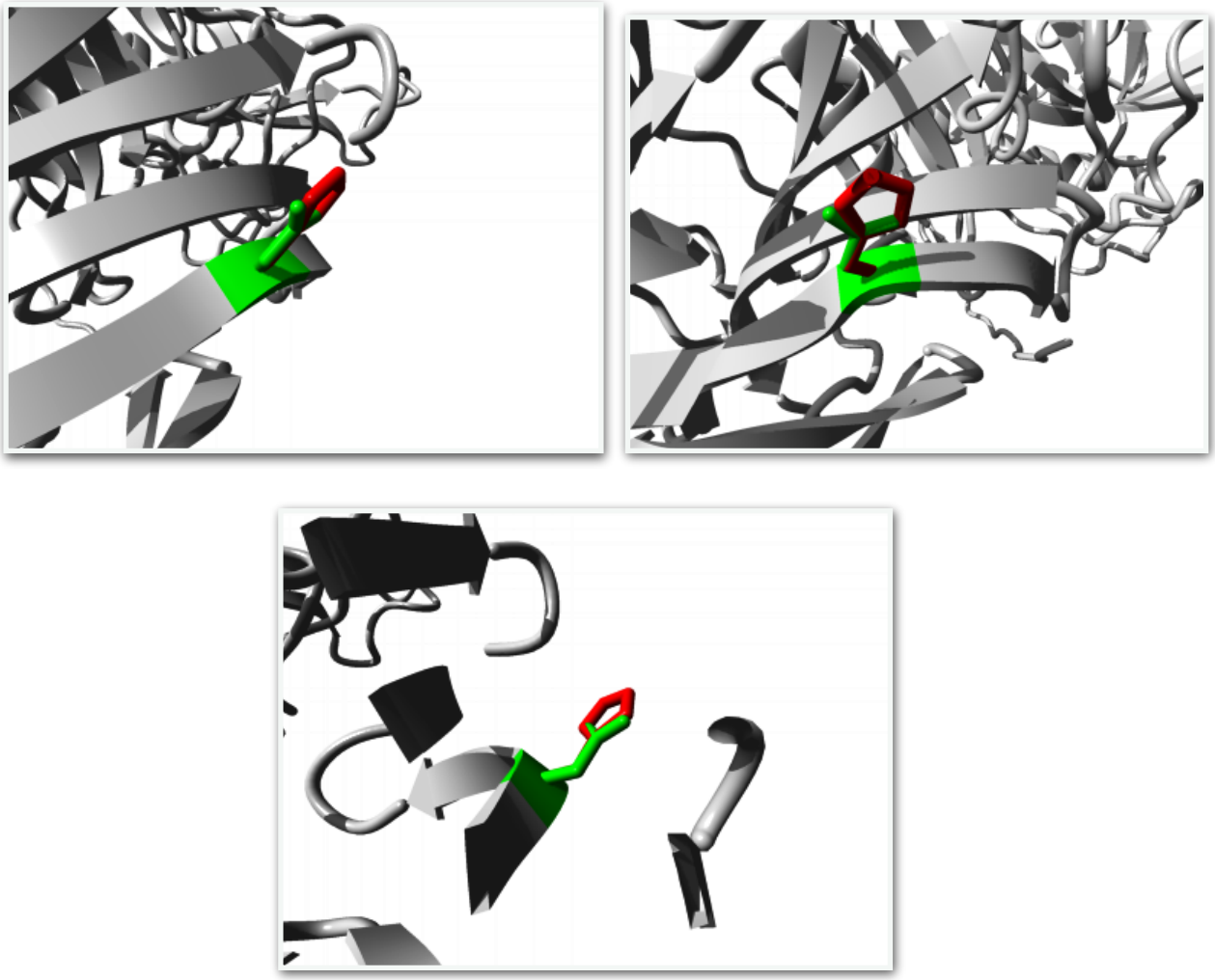
Choose the AlphaFold2 method and clic submit.

- **Now explore the result report.**





Figure(12): Overview of the protein in ribbon-presentation. The protein is coloured by element; α-helix=blue, β-strand = red, turn=green, 3/10 helix=yellow and random coil=cyan. Other molecules in the complex are coloured grey when present.

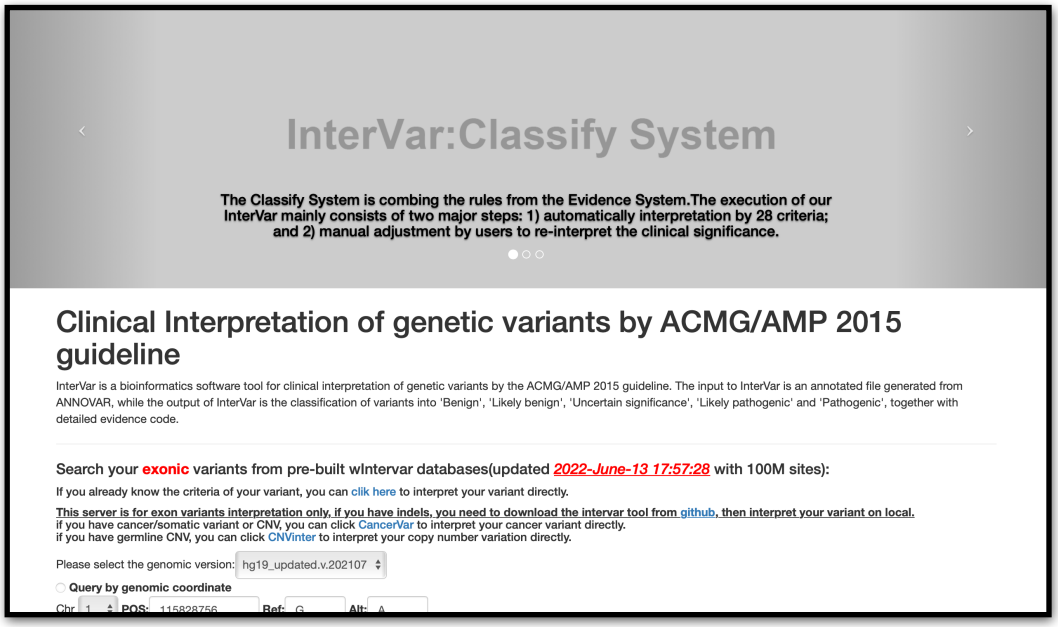


Figure(13): Close-up of the mutation. The protein is coloured grey, the side chains of both the wild-type and the mutant residue are shown and coloured green and red respectively.

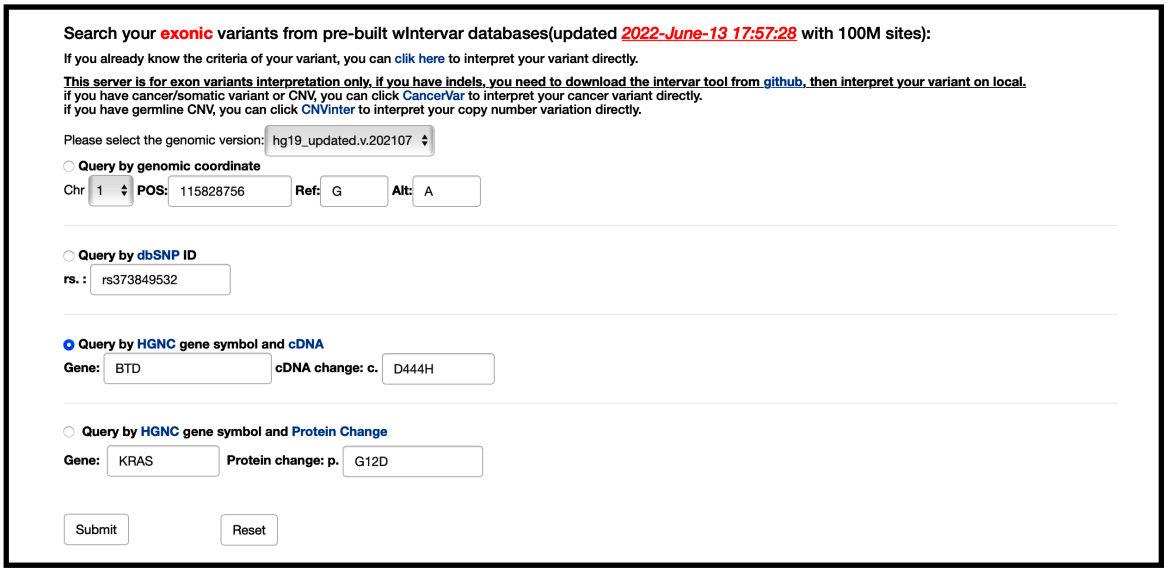
InterVar:

Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline.

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

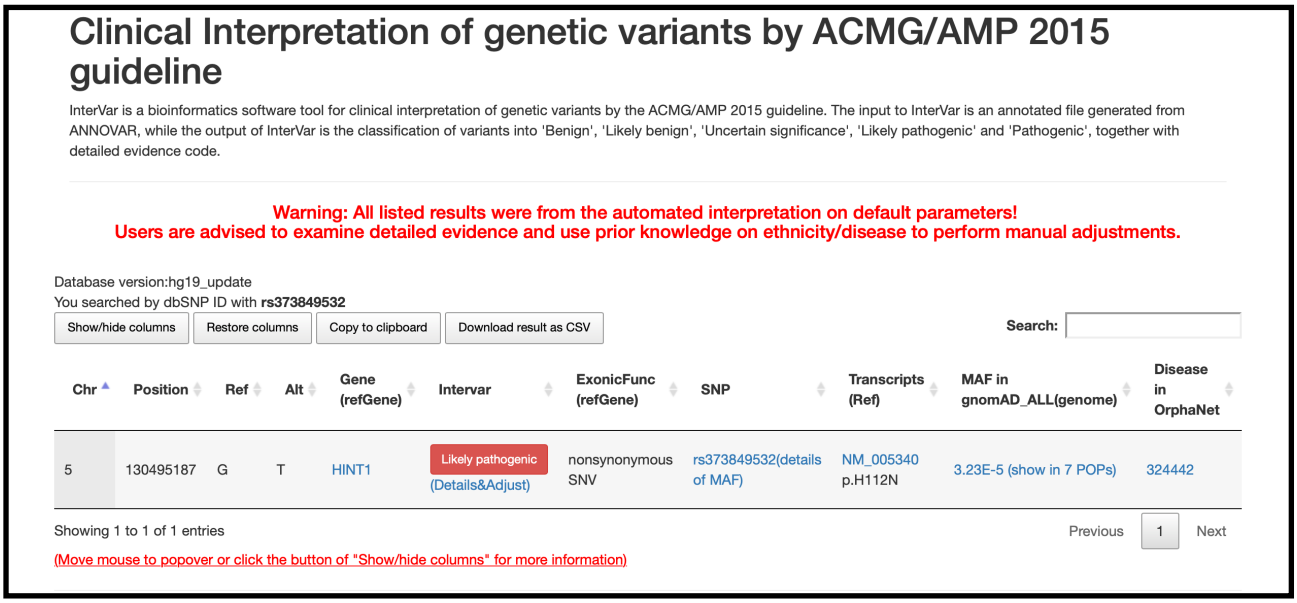


Figure(14): Home page of InterVar.



Figure(15): Input section of InterVar.

Now insert any mutation you want to examine.



Clinical Interpretation of genetic variants by ACMG/AMP 2015 guideline

InterVar is a bioinformatics software tool for clinical interpretation of genetic variants by the ACMG/AMP 2015 guideline. The input to InterVar is an annotated file generated from ANNOVAR, while the output of InterVar is the classification of variants into 'Benign', 'Likely benign', 'Uncertain significance', 'Likely pathogenic' and 'Pathogenic', together with detailed evidence code.

Warning: All listed results were from the automated interpretation on default parameters!
Users are advised to examine detailed evidence and use prior knowledge on ethnicity/disease to perform manual adjustments.

Database version:hg19_update
You searched by dbSNP ID with **rs373849532**

Show/hide columns | Restore columns | Copy to clipboard | Download result as CSV

Search:

Chr	Position	Ref	Alt	Gene (refGene)	Intervar	ExonicFunc (refGene)	SNP	Transcripts (Ref)	MAF in gnomAD_ALL(genome)	Disease in OrphaNet
5	130495187	G	T	HINT1	Likely pathogenic (Details&Adjust)	nonsynonymous SNV	rs373849532(details of MAF)	NM_005340 p.H112N	3.23E-5 (show in 7 POPs)	324442

Showing 1 to 1 of 1 entries

Previous 1 Next

(Move mouse to popover or click the button of "Show/hide columns" for more information)

Figure(16): Result section of InterVar.

References:

NCBI About section.

Ensemble About section.

UniProt document Section.

The dbSNP help section.

ClinVar help section.

HOPE about section.

BLAST help section.

PDB documentation section.