Class 1 Lab*

Bioinformatics Databases and Key Online Resource

Barry Grant

2025-12-15

Instructions

Save this document to your computer and open it in a PDF viewer such as Preview (available on every mac) or Adobe Acrobat Reader (free for PC and Linux). Be sure to add your name and UC San Diego personal identification number (PID) and email below before answering all questions in the space provided.

Student Name	UCSD PID	UCSD Email

Learning Objectives

By the end of this lab you should be able to:

1. Use BLAST to annotate an unknown nucleotide sequence and interpret key metrics (E-value, coverage, percent identity);
2. Navigate between interconnected NCBI databases (GenBank, GENE, RefSeq, OMIM) to gather information about a gene of interest;
3. Perform sequence alignment and identify mutations at the nucleotide level;
4. Retrieve and visualize 3D protein structures from the PDB database;
5. Relate a single nucleotide mutation to its structural and phenotypic consequences at the protein level.

*http://thegrantlab.org/teaching/

Overview:

The purpose of this lab session is to introduce a range of bioinformatics databases and associated services available online whilst investigating the molecular basis of a common human disease.

Sections 1 and 2 deal with querying and searching GenBank, GENE and OMIM databases at **NCBI**. Sections 3 and 4 provide exposure to **EBI** resources for comparing sequences and **PDB** resources for visualizing protein structures.

> **Side-note:** The Web is a dynamic environment, where information is constantly added and removed. Servers "go down", links change without warning, etc. This can lead to "broken" links and results not being returned from services. Don't give up - give it a second go and try a search engine using terms related to the page you are trying to access.

Section 1

The following transcript was found to be abundant in a human patient's blood sample.

```
>example1
ATGGTGCATCTGACTCCTGTGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG
TTGGTGGTGAGGCCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGG
GGATCTGTCCACTCCTGATGCAGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGT
GCCTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACT
GTGACAAGCTGCACGTGGATCCTGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCA
TCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAAT
GCCCTGGCCCACAAGTATCACTAAGCTCGCTTTCTTGCTGTCCAATTT
```

The only information you are given is the above sequence so you must begin your investigation with a sequence search - for this example we will use NCBI's **BLAST** service at: http://blast.ncbi.nlm.nih.gov/

> **Side-note**: There are several different "basic BLAST" programs available at NCBI (including nucleotide BLAST (**BLASTn**), protein BLAST (**BLASTp**), and translated nucleotide nucleotide (**BLASTx**). We will explore using all of these in the coming sections.

Q1 Which BLAST program should we use in this case?

Searching against the **"Nucleotide collection" (NR database)** that includes GenBank is a good place to start your investigation of this sequence.

Q2 What are the names and accession numbers of the top four hits from your BLAST search?

Q3. What are the percent identities, coverage and E-values for these top few hits?

Tip

These three metrics (E-value, coverage and identities) are the most important for us to consider at this stage. I suggest you have a discussion with your neighbor and Barry to make sure you have a firm grasp of these concepts as you will need them later in your project.

To investigate these results further click on the **Alignments** section (tab) of your BLAST result page (Figure 1). This will give you more details on matched nucleotides and important links to "Related information" about a given *"subject sequence"*.

> **Side-note**: In BLAST terminology we talk about *query sequence* and *subject sequence*. The *query* being the input sequence you searched with and the *subject* being the identified hit sequence from the database.

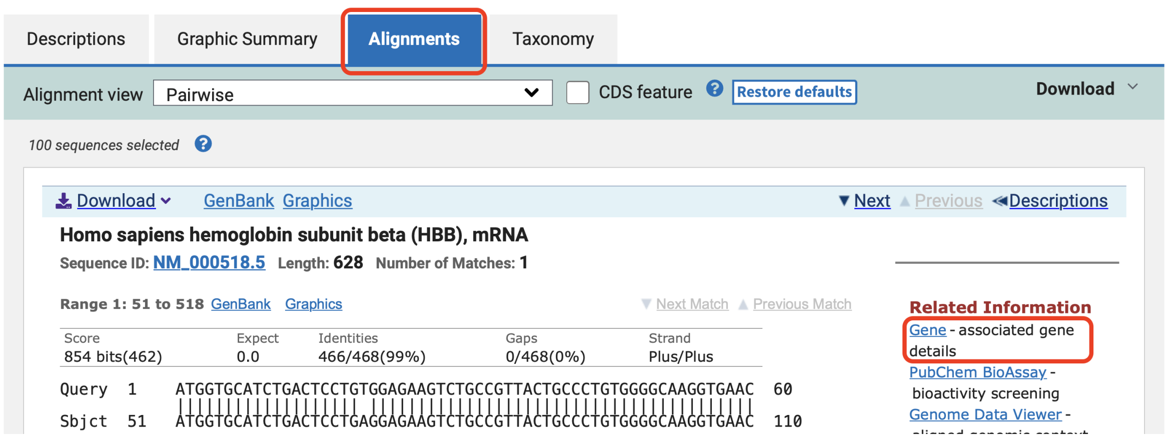


Figure 1: The BLAST *Alignments* tab contains more detailed information about your results and also a link to the **Gene** database that we will use for Q5

Tip

Scroll down to the end of the Alignments page to see lower ranked hits.

Q4. How many identical and non identical nucleotides are there in your top hit compared to your last reported hit?

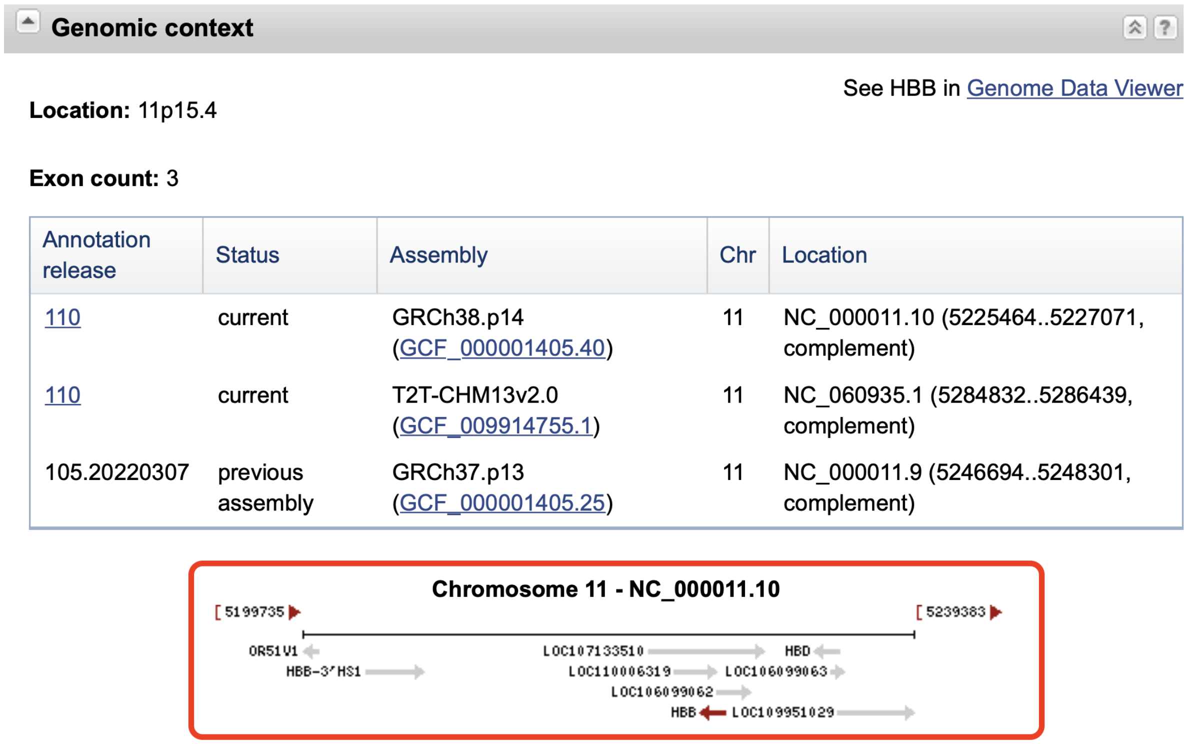
From the results of your BLAST search you can link to the **GENE** entry for one of your top hits.

Tip

Note that the **GENE** link is located under the "Related Information" heading at the right hand side of each displayed alignment on the "Alignments" tab (Figure 1).

Q5. What is the "Official Symbol" and "Official Full Name" for this gene?

Q6. What chromosome is this gene located on (see (Figure 2))?



Genomic context

See HBB in Genome Data Viewer

Location: 11p15.4

Exon count: 3

Annotation release	Status	Assembly	Chr	Location
110	current	GRCh38.p14 (GCF_000001405.40)	11	NC_000011.10 (5225464..5227071, complement)
110	current	T2T-CHM13v2.0 (GCF_009914755.1)	11	NC_060935.1 (5284832..5286439, complement)
105.20220307	previous assembly	GRCh37.p13 (GCF_000001405.25)	11	NC_000011.9 (5246694..5248301, complement)

Figure 2: NCBI GENEs Genomic context section indicating location, structure and neighboring genes

Q7. What are the names of neighboring genes on this chromosome?

Tip

Note that there is a schematic diagram of neighboring genes and their orientations in the "Genomic context" section (Figure 2). Our HBB gene is in maroon. All other gene arrows can be hovered over for full names and clicked on to link to that specific GENE page to find out more. We will explore more full featured *genome browsers* at ENSEMBE and UCSC in an upcoming lab but basically it is the same idea here in a more simplified form.

Q8. How many exons and introns are annotated for this gene?

In addition to reading the abstract like text on a given GENE entry I encourage you to explore the linked "**Gene Ontology**" information and discuss with your neighbor and Barry the advantages you think such controlled terms might have over free text?

Q9 What is the function of the encoded protein?

Scroll down to the "Phenotypes" section of the GENE entry page and also explore the link to the **OMIM** database

Q10. Does the protein have a role in human disease(s)? If so what diseases?

Section 2

By now you should be aware that there are a number of human diseases linked to particular variants of the beta-globing gene. In this case our example sequence corresponds to human sickle cell beta-globin mRNA with this disease resulting from a point mutation in the beta globin gene. In the following section, you will compare sickle cell and normal beta globin sequences to reveal the nature of the sickle cell mutation at the protein level.

To do this you need to find at least one sequence representing the normal beta globin gene. Open a new window and visit the NCBI home page (http://www.ncbi.nlm.nih.gov) and select "Nucleotide" from the drop menu associated with the top search box. Then enter the search term: **HBB** (Figure 3).

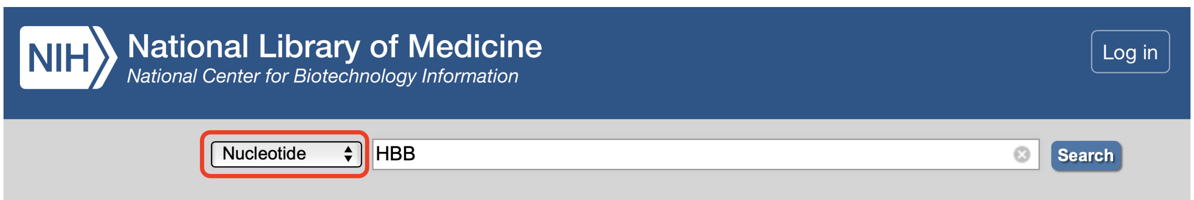


Figure 3: The main Google like search tool for searching accross all NCBI databases is called **Entrez**. Note that productive use often requires the use of addational "filters" as we will explore later.

Note that lots of often irrelevant results are returned so lets apply some "Filters" (available by clicking in the left-hand sidebar) to focus on **RefSeq** entries (under "Source databases") for **Homo sapiens** (under "Results by taxon" on the right-hand sidebar in this later case).

> **Side-note**: Boolean operators (NOT, AND, OR) as well as fielded queries (i.e. "HBB[Gene Name] AND Human[Organism]") can be used directly in ENTREZ searches to filter results for more efficient searching.

Remember that we are after mRNA so we can compare to the mRNA sequence from section 1 above.

Q11. What is the ACCESSION number of the "Homo sapiens hemoglobin, beta (HBB), mRNA" entry?

Select "Homo sapiens hemoglobin, beta (HBB), mRNA" from the results and scroll down to the *"FEATURES"* section to answer the following.

Tip

Note that that you can find much of the same information from either the *"GenBank"* format entry or by selecting the *"GRAPHICS"* display format and, for example, placing your mouse over the first exon (see Figure 4).

Q12. What are the numbers of the first and last base positions of exon 1 of this entry?

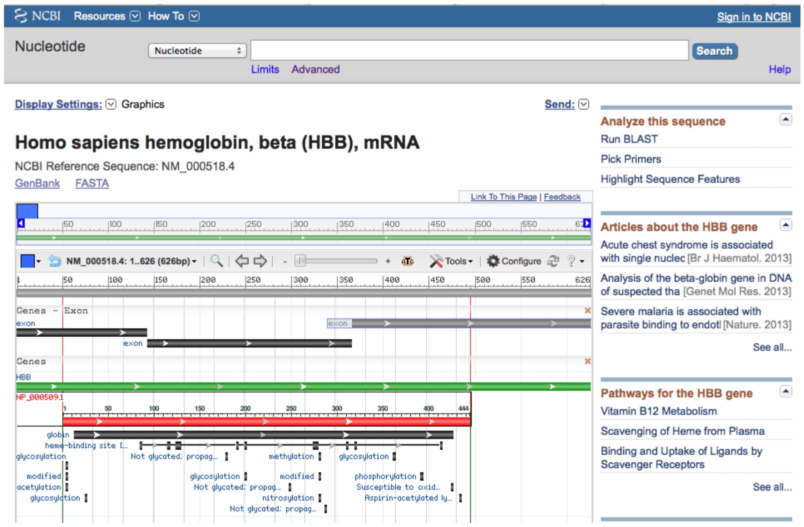


Figure 4: The GRAPHICS view of a GenBank entry can be more user friendly than the traditional text of the corresponding GenBank format display.

The red bar in the "*Graphics*" display (Figure 4) coresponds to the CDS (or "coding sequence"), which refers to the portion of a genomic DNA sequence that is translated, from the start codon to the stop codon. Successful translation of a CDS results in the synthesis of a protein.

Q13. What are the numbers of the first and last base positions of the CDS?

Section 3

Here we will compare the retrieved sequences by creating a sequence alignment. This will make the difference between the two sequences easy to spot.

To generate the alignment, we will use **MUSCLE** available on the EBI website at: https://www.ebi.ac.uk/jdispatcher/

Select the FASTA display for the "Homo sapiens hemoglobin, beta (HBB), mRNA" (NM_000518) entry from section 2.

Now copy-and-paste this FASTA format sequence and also the **example1** sequence from section 1 into the input box of the **MUSCLE** page. Then click the submit button (see red circle in Figure 5).

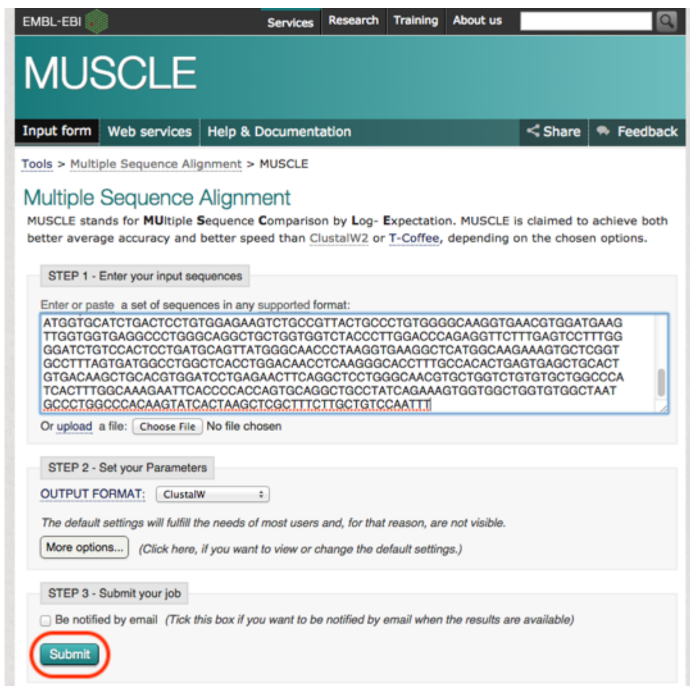


Figure 5: To use the EBI MUSCLE server you must paste multiple FASTA format sequences that include their ID lines.

The two sequences should now be aligned. Where the aligned sequences are identical, an * is placed under the alignment. Examine the results and note that your sequences are nearly identical. However, being much shorter, the sickle cell sequence has many padding gap characters (-----) to bring equivalent regions into the correct register (Figure 6).

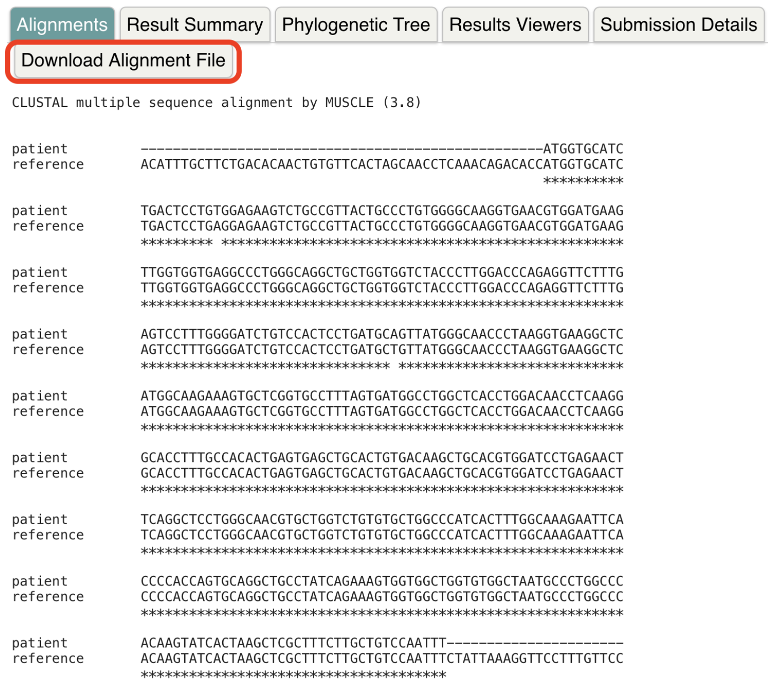


Figure 6: Alignment of patient and reference HBB sequence

When inspecting alignments (especially those with lots of sequences) it can be helpful to use a **g**raphical **u**ser **i**nterface (or GUI) to display colored, interactive and scrollable versions of your alignment. One such GUI is the **seaview** program.

From your muscle results web page click **Download Alignment File** (red highlight in Figure 6). Note that if a download does not automatically begin then you may need to save the resulting plain text page from your web browser via **File> Save As...**

Next download **seaview** for your computer from: http://doua.prabi.fr/software/seaview

Once downloaded open seaview by double clicking on it's icon (most likely in your Downloads folder) and then select **File > Open >** and select your muscle alignment results. A colored version of your alignment should now be displayed (Figure 7).

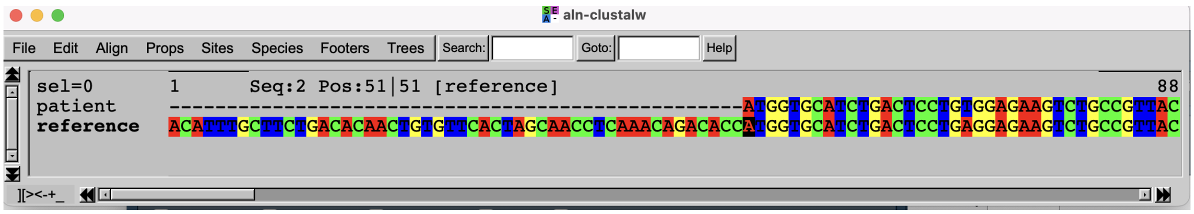


Figure 7: Programs like SEAVIEW are most useful when you have large many sequence alignments

See if you can now use seaview to answer the following 3 questions:

Q14. How many gap characters (-) are added to the beginning of the sickle cell beta-globin sequence in order to align it with the beta globin sequence? How might you have guessed this number from information you read in the GenBank annotation (See section 2, Q13)?

Q15. What are the nucleotide differences between the two sequences (note that there may be more than one)?

Q16. Which codon position from the start of the sickle cell sequence would this difference affect? Using the codon table below to help (Figure 8), what amino acid would the different codons encode in the two sequences?

Second Position

First Position	T	C	A	G	Third Position
T	TTT F Phe	TCT S Ser	TAT Y Tyr	TGT C Cys	T
T	TTC F Phe	TCC S Ser	TAC Y Tyr	TGC C Cys	C
T	TTA L Leu	TCA S Ser	TAA STOP	TGA STOP	A
T	TTG L Leu	TCG S Ser	TAG STOP	TGG W Trp	G
C	CTT L Leu	CCT P Pro	CAT H His	CGT R Arg	T
C	CTC L Leu	CCC P Pro	CAC H His	CGC R Arg	C
C	CTA L Leu	CCA P Pro	CAA Q Gln	CGA R Arg	A
C	CTG L Leu	CCG P Pro	CAG Q Gln	CGG R Arg	G
A	ATT I Ile	ACT T Thr	AAT N Asn	AGT S Ser	T
A	ATC I Ile	ACC T Thr	AAC N Asn	AGC S Ser	C
A	ATA I Ile	ACA T Thr	AAA K Lys	AGA R Arg	A
A	ATG M Met	ACG T Thr	AAG K Lys	AGG R Arg	G
G	GTT V Val	GCT A Ala	GAT D Asp	GGT G Gly	T
G	GTC V Val	GCC A Ala	GAC D Asp	GGC G Gly	C
G	GTA V Val	GCA A Ala	GAA E Glu	GGA G Gly	A
G	GTG V Val	GCG A Ala	GAG E Glu	GGG G Gly	G

Figure 8: The standard genetic code table. Codons are read by combining the first position (left column: T, C, A, G), second position (top row: T, C, A, G), and third position (right column). For example, ATG encodes Methionine (Met, M).

Section 4

In this section we will retrieve and visualize the 3D protein structure of sickle cell haemoglobin. The aim here is to ascertain how the Glu6 -> Val6 mutation might cause the mutant molecules to oligomerise into fibers, hence deforming erythrocytes. This will require you to examine the structural context of the mutation in the beta globin chains.

We could find sickle cell haemoglobin structures via a text search of main PDB website @ http://www.rcsb.org/. However, as we know the nucleotide sequence from our previous work, lets use BLASTx to search the PDB database from the NCBI site.

To do this visit http://blast.ncbi.nlm.nih.gov/ select the appropriate BLAST program and make sure the database you are searching against is set to "**Protein Data Bank (pdb)**".

Note the accession numbers and alignment statistics for the top few hits.

Q17. Are there any PDB structures with 100% identity to your *example1* query sequence? Give the PDB codes for these entries and note that there may be more than one.

To further examine these structures we will jump over to the main PDB database as it has more annotation data and more full featured **3D molecular viewers** than NCBI.

Visit http://www.rcsb.org/ and use the 4 character PDB accession code you found previously in your BLASTx search to pull up each PDB entry you listed in Q17.

From scrolling through this entry you can find out information about the "Experimental Data" (such as the resolution of the structure and quality of the data collected), "Literature" links (i.e. associated publications), "Macromolecules" (i.e. protein chains present) and "Small Molecules" (i.e. any ligands or co-factors that might be present).

From the "Macromolecules" section notice that the hemoglobin structure is composed of multiple alpha and beta globin molecules corresponding to gene names HBA and HBB.

Q18. Which four chain identifiers in the 2HBS structure represent beta globin?

At the top of the PDB entry page click the **Structure** tab to pull up an interactive 3D structure view (Figure 9).

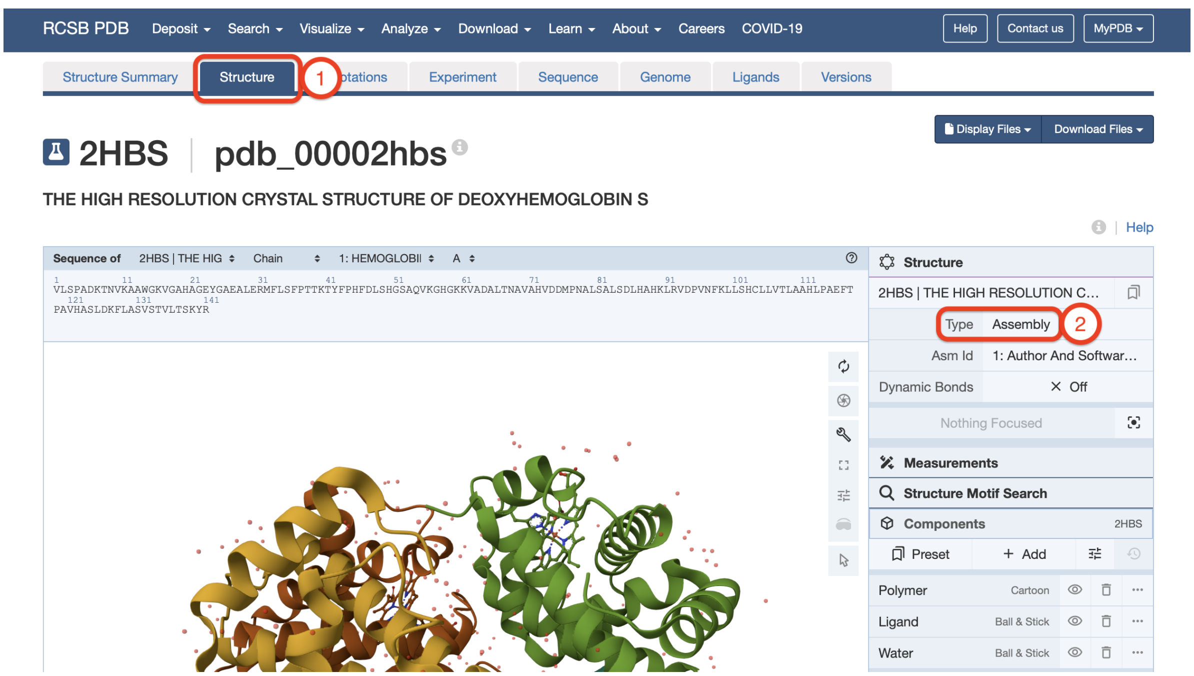


Figure 9: Using the molecular viewer at the PDB database

Under the *“Structure”* section of the right side control panel change the “Type” of view from “Assembly” to “Model” (see second red rectangle in Figure 9). This will now display the model observed in the asymmetric unit of the crystal (i.e. the packing of chains observed in the actual experiment) rather than the simplified default biological assembly view of the minimal functional form.

> **Side-note**: This is the relatively new Mol* 3D viewer. This viewer also features at a number of other major databases including UniProt. The user interface is still somewhat clunky and limited when compared to stand-alone software like PyMol, Chimera or VMD. However, these stand-alone viewes require considerable download and installion time so we will stick with Mol* for this lab. Essentially there are 3 major components to the Mol* viewer (Figure 10):

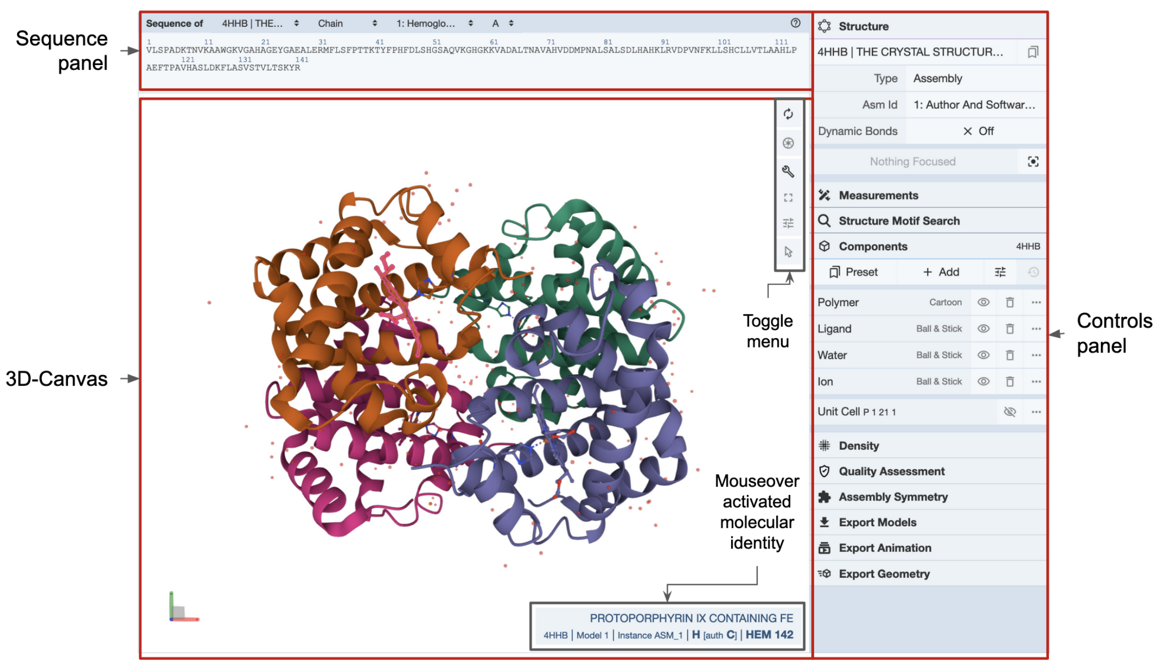


Figure 10: The three major components of the Mol* viewer include the Sequence panel, 3D-canvas and Control panel.

Notice by rotating the molecule in the 3D Canvas that there are now two hemoglobin molecules displayed rather than the previous single molecule. Notice that each is comprised of four distinctly colored chains with two alpha and two beta chains in each.

To highlight our beta globin amino acid of interest toggle the sequence display to the **beta chain** (Figure 11) and specifically chain **H** (Figure 12) in the top "Sequence panel".

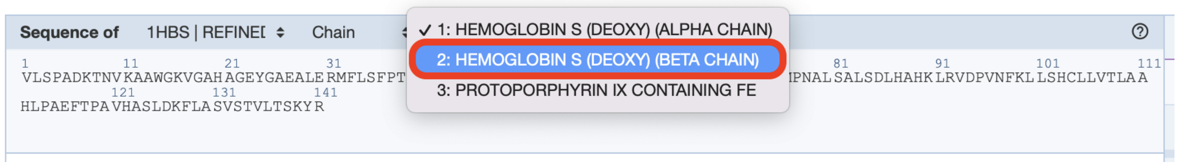


Figure 11: Display the sequence of beta globin chains in the PDB entry

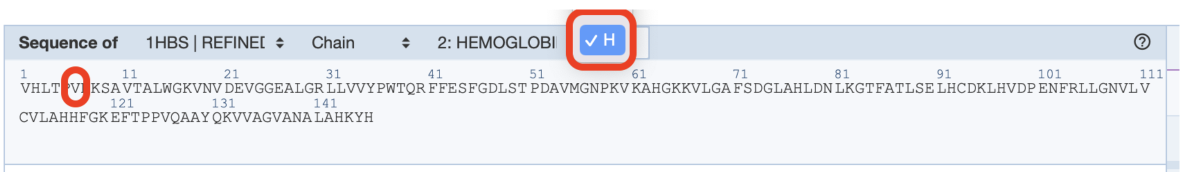


Figure 12: Focus on beta globin chain H

"Toggle selection mode" via the pointer/arrow icon in the *Toggle menu* of the 3D canvas (Figure 10 and step 1 in Figure 13).

With selection mode active find and click on amino acid V 6 in the sequence view to highlight and select this amino acid in both the sequence and 3D view (step 2 in Figure 13). Notice that this amino acid is now highlighted in green.

We can now add a new "*Representation*" to more clearly display the mutated Valine residue. Click the 3D cube icon (step 3 in Figure 13). Select **spacefill** from the dropdown list of "Representation" options (step 4 in Figure 13) and finally (step 5) click "*Create Component*".

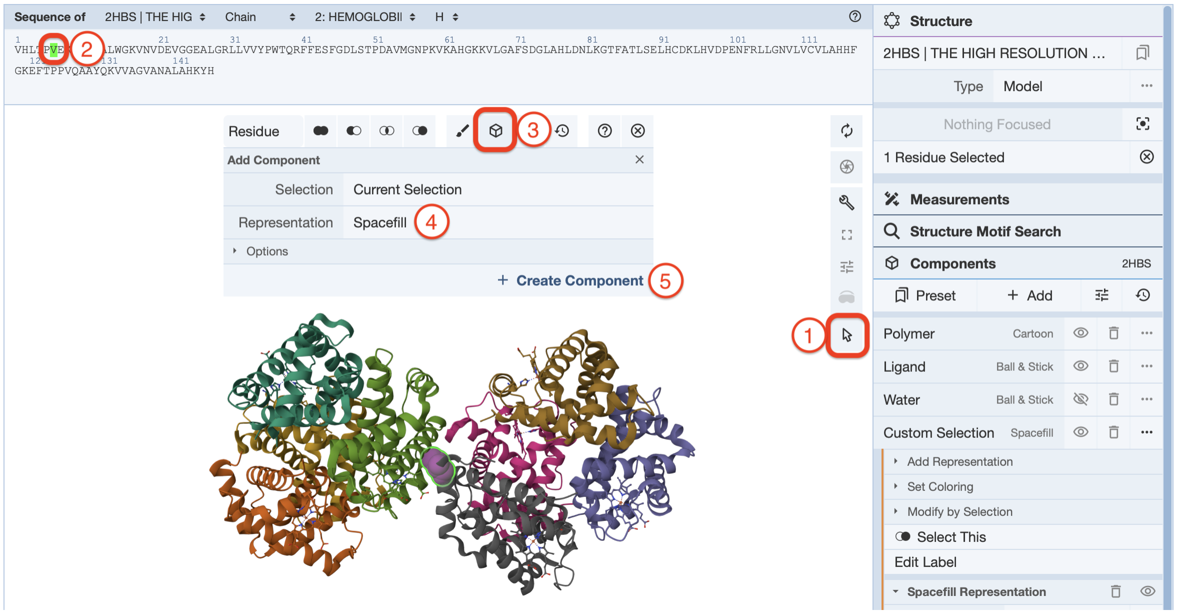


Figure 13: Adding a spacefill represention to our Val 6 residue. First toggle selection mode (see red highlight 1), Select V6 in beta globin chain H (2), Add a 3D component (3), using a spacefill represntation (4), and finally click create component.

This will result in all atoms of our mutant amino acid being displayed as so called "*spacefill spheres*" with different atom types in different colors (e.g. oxygens in red, carbons in gray etc.)

Play around with zoom level and the settings from the new spacefill "Custom component" menu item in the right hand side Control Panel until you have a reasonable feel for how the program works. Can you clearly see our mutated residue position?

Try zooming (via scrolling up and down) and rotating (via clicking and moving your mouse) to get a better feel for the location of our Valine amino acid in the overall structure.

> **Side note**: You can always "reset" the view by clicking the reset like circular arows icon in the toggle menu. Also feel free to experiment with different settings and representations.

Q19. What do you notice about the location of the Val6 residue in chain H of the 2HBS structure in relation to porphyrin?

Tip

You can download a high resolution image with transparent background of your final representation using the iris like screenshot icon.

Q20: What one part of this exercise or associated lecture material is still confusing? If appropriate please also indicate the question number from this document and answer the question in the following anonymous form: **Mudy_Point_Assesment_Form** Your comments will let us know which material needs to be further clarified and will help us gain stronger control of the material in this course. Thank you!

Side note: Some folks have reported issues using the Mol* viewer with older versions of the Edge and Chrome browsers. The workaround is to use a different web browser. If, the structure is still not displayed correctly for you, download its coordinates from the **PDB** database at: http://www.rcsb.org/ and **ask for assistance**.

Discussion

The original paper discussing the 1HBS and 2HBS crystal structures is available online:

http://www.sciencedirect.com/science/article/pii/S0022283697912535

In this article, Figure 3 demonstrates how the Glu6->Val6 mutation could result in the characteristic "sickle" phenotype. The charged Glu6 mutating to Val6 creates a superficial hydrophobic patch on one HbS molecule that interacts with hydrophobic surface residues of another. The molecules thus polymerize, creating extended fibers that distort the shape of the red blood cell.

Assessment of the disparate biochemical properties of normal and sickle haemgolobin, together with microscopy studies showing long crystal fibres inside sickle cells, led Linus Pauling (1949) to (correctly) predict the morphological effects of these changes. The abnormal sickle form causes the cells to clump together, hampering their passage through blood vessels, depriving tissues of oxygen. See this YouTube video for an illustration: http://www.youtube.com/watch?v=Qd0HrY2NlwY

The sickled blood cells have a short lifetime and cannot be replaced fast enough, leading to chronic anaemia. Sickle cell anemia was one of the first diseases to be linked to a defect at the molecular level, providing a clear demonstration that a single base mutation can change a single amino acid, which in turn can result in a defective protein.

Bluebird Bio as well as Vertex Pharmaceuticals and CRISPR Therapeutics have ongoing **sickle-cell gene therapy trials**. As noted in this excellent recent easy in the New Yorker: "There is something of a paradox in the fact that patients with sickle-cell disease — a population that has faced extraordinary levels of bias, neglect, and marginalization - may be among the first to have their illnesses transformed by the most cutting-edge of medical technologies".

Appendix

```
>gi|179408|gb|M25079.1|HUMBETGLA Human sickle cell beta-globin mRNA
ATGGTNCAYYTNACNCCNGTGGAGAAGTCYGCYGTNACNGCNCT
NTGGGGYAAGGTNAAYGTGGATGAAGYYGGYGGYGAGGCCCTGG
GCAGNCTGCTNGTGGTCTACCCTTGGACCCAGAGGTTCTTNGAN
TCNTTYGGGGATCTGNNNACNCCNGANGCAGTTATGGGCAACCC
TAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTG
ATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACA
CTGAGTGAGCTGCACTGTGACAAGCTNCAYGTGGATCCTGAGAA
CTTCAGGCTNCTNGGCAACGTGYTNGTCTGYGTGCTGGCCCATC
ACTTTGGCAAAGAATTCACCCCACCAGTGCANGCNGCCTATCAG
AAAGTGGTNGCTGGTGTNGCTAATGCCCTGGCCCACAAGTATCA
CTAAGCTNGCYTTYTTGYTGTCCAATTT
```

```
>gi|28302128|ref|NM_000518.4| Homo sapiens hemoglobin, beta (HBB), mRNA
ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACA
GACACCATGGTGCATCTGACTCCTGAGGAGAAGTCTGCCGTTAC
TGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGG
CCCTGGGCAGGCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTC
TTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTGTTATGGG
CAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCT
TTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTT
GCCACACTGAGTGAGCTGCACTGTGACAAGCTGCACGTGGATCC
TGAGAACTTCAGGCTCCTGGGCAACGTGCTGGTCTGTGTGCTGG
CCCATCACTTTGGCAAAGAATTCACCCCACCAGTGCAGGCTGCC
TATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAA
GTATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGG
TTCCTTTGTTCCCTAAGTCCAACTACTAAACTGGGGGATATTAT
GAAGGGCCTTGAGCATCTGGATTCTGCCTAATAAAAAACATTTA
TTTTCATTGC
```

http://www.rcsb.org/pdb/files/2hbs.pdb

The mutation causing sickle cell anemia is a single nucleotide substitution (A to T) in the codon for amino acid 6. The change converts a glutamic acid codon (GAG) to a valine codon (GTG). Changing a hydrophilic amino acid to a hydrophobic one, see http://themedicalbiochemistrypage.org/sicklecellanemia.php

Note there is also a T -> A difference at position 162 (162/3 => codon 54 GCT -> GCA). This is in the third position of the codon and hence does not change the corresponding amino-acid.