# Practical session: Data and knowledge bases

In this practical session you will use several online-available data portals to find information on genes, transcripts, proteins, their function, and biological pathways they occur in.

Read the practical assignments carefully – you can write your answers below the questions. If you still have questions you can 1. Check the issue tracker on Github if there is already a question with answer. If not 2. Write down your question as a new issue. 3 If there are still problems or questions, ask an instructor.

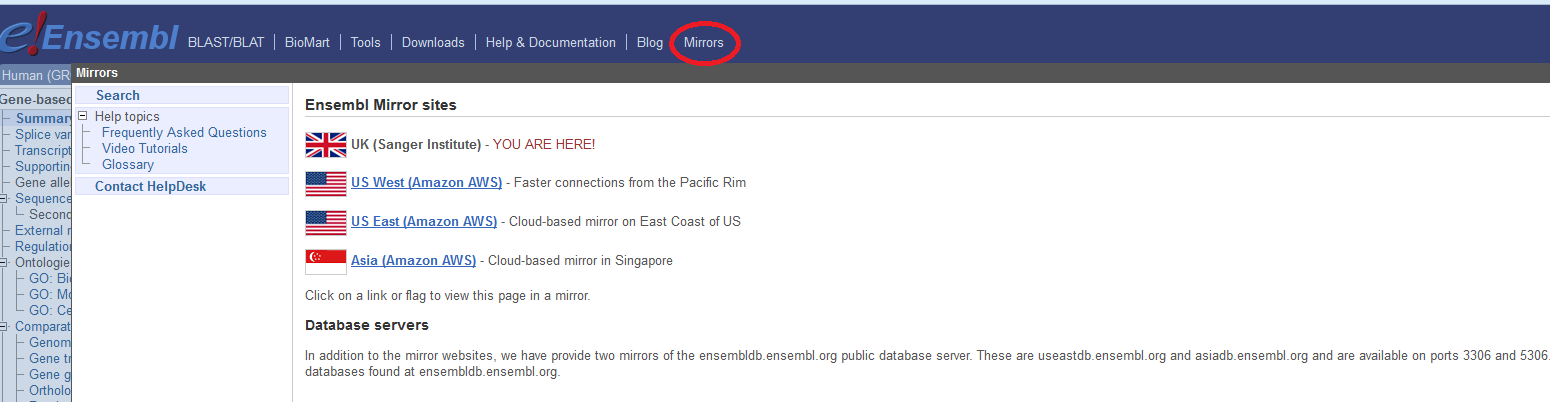
*Assignment 1: Introduction to Ensembl*

First you are going to learn how to use the Ensembl database. Together with the instructor you will walk through the Ensembl tutorial, see Tutorial\_Ensembl-walk\_through.pdf

Ensembl provides genes and other annotations such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on mRNA evidence in the NCBI RefSeq databases.

Go to the website: [***www.ensembl.org***](http://www.ensembl.org)

**!** Please note that if for some reason the ensembl website is slow you can choose a mirror in a different continent (see image below)!



*Assignment 2: Ensembl – do it yourself!*

You will learn how to obtain information on a specific gene in Ensembl ([www.ensembl.org](http://www.ensembl.org)). For this a member of the human superoxide dismutase family, the superoxide dismutase 1 (SOD1) was selected.

Superoxide dismutase (SOD) is an enzyme which plays an important role in handling oxidative stress in cells. It catalyzes the reaction of superoxide (O2-) to oxygen (O2) or hydrogen peroxide (H2O2). Superoxides are usually produced as a by-product of the oxygen metabolism but they are also produced in case of cellular stress. Superoxide can cause severe damage to cellular molecules leading to cell death. There are 3 main SOD subclasses known: SOD1, an enzyme which occurs in the cytoplasm, SOD2 which occurs in the mitochondria, and SOD3 which is released to the extracellular space.

🖰 Find the human SOD1 gene in Ensembl and go to the Gene tab.

1. As explained in today’s lecture, in Ensembl all the genes have a unique identifier. **What is the Ensemble ID of the SOD1 gene?**
2. **On which chromosome and which strand the gene is located?**
3. Genes that are located on the same chromosome are linked genes. This means that they have the tendency to be inherited together. Therefore, it is of interest to know which protein coding genes are in close proximity. **What is the next known protein coding gene on the sequence to the right of the SOD1 gene? Which strand?** (Hint: go to the Location tab)
4. Genes can have more than one transcript due to alternative splicing. These transcripts can give rise to proteins which differ in their sequences and often in their activities or can cause changes in regulatory elements such as translation enhancers or RNA stability domains, which may have a dramatic effect on the level of protein expression. **How many transcripts have been predicted for this gene?**
5. **How many of these transcripts are protein coding?**
6. **How many base pairs has the longest protein coding transcript and what is the length of the protein (the number of amino acids) it encodes?**

*Assignment 3: NCBI*

In assignment 2 you used the Ensembl website to gather information on the human SOD1 gene. In this exercise you will use several NBCI-Entrez databases to find more information about this gene. NCBI offers a training and tutorial page at: <http://www.ncbi.nlm.nih.gov/guide/training-tutorials/>

Start with opening the **NCBI Gene page** for human SOD1 and answer the following questions. (NCBI homepage <http://www.ncbi.nlm.nih.gov/>, NCBI Gene page <http://www.ncbi.nlm.nih.gov/gene>). Hint: apart from information directly on this page, you may also want to use links to other web pages (for example at the right side of the page).

1. NCBI gene also uses specific identifiers. **What is the NCBI gene ID of the human SOD1 gene?**
   1. **!** Hint: Background information on the *Gene database* is available at <http://www.ncbi.nlm.nih.gov/books/NBK21085/> !
2. Human genes are given an official name by the HUGO Gene Nomenclature Committee (HGNC). However, one gene can have other names which are not official. **Give two alternative names (not IDs) of the human SOD1 gene.**
3. In the NCBI gene entry the NCBI reference sequences are given. **What is the RefSeq ID of the mRNA of this gene and of the protein?**
4. **What are the NCBI Gene IDs of the homologues of human SOD1 in mouse and rat?** **How/where did you find this?**
   1. **!** Hint: You find this information looking for HomoloGene (under the box on the right side Related Information) or select Orthologues: all in the summary.

*Assignment 4: Find protein function information in UniProt*

UniProt is a database about protein and their related information.

1. Type SOD1 in the search field and select human SOD1 from the result table. **How many amino acids does SOD1 have?**
2. Have a look at the catalytic activity. Any proteins with known catalytic function the catalyzed reaction is listed in UniProt and in other reaction databases, e.g. Rhea. **Which cofactors are required for SOD1 catalytic activity?**
3. **In which “other”cellular components can SOD1 be found?** (Hint: click on the square blue box of “other cellular components” to get the list.)
4. Have a look at the 3D protein structure. **From which database is this information?**
5. Go to the “Variant viewer”. The overview table allows filtering for variant consequence and provenance (references). Search for the VAR\_013518. **Which amino acid is changed? What is the clinical significance?**

*Assignment 5: Find gene – disease interactions in DisGeNET*

DisGeNET <https://www.disgenet.org/> is a gene-disease association database. It collects knowledge on gene and genetic variants that can cause diseases from different resources like curated databases, literature, animal models and inferred data (details here: <https://www.disgenet.org/dbinfo>).

1. Go to the search page of DisGeNET. You can search for diseases, genes and genetic variants. Check the “Gene” search and search for disease associations of SOD1 -> Summary of Gene-Disease Associations. You can sort the results by column by hitting the arrow in the header. Which disease has the highest GDA score?
2. How many literature references (PMIDs) are available for this?
3. How is the GDA and the VDA score calculated? (check <https://www.disgenet.org/dbinfo>)