Tutorial

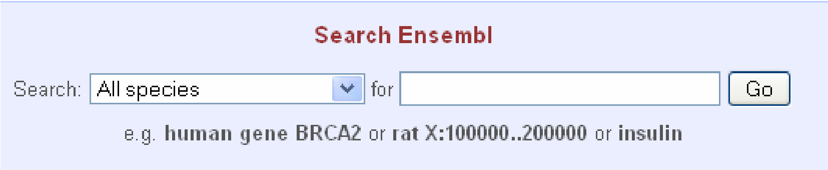
Module 2 – Search for a gene

You will learn about

- How to search for a gene
- How to search for a chromosomal region

Search for a gene in any species

To search for any gene in Ensembl, take advantage of the three descriptor fields. The first is species, the second is ‘view’ (gene or chromosome), and the third is the description (the gene name, or the base pair position). For example, type in the species name, the gene name, description or symbol, and the word ‘gene’. Let us try this with the IL2 gene, or ‘Interleukin-2 gene’ in human. To start, enter ‘**human IL2 gene**’ into the main search box at www.ensembl.org.

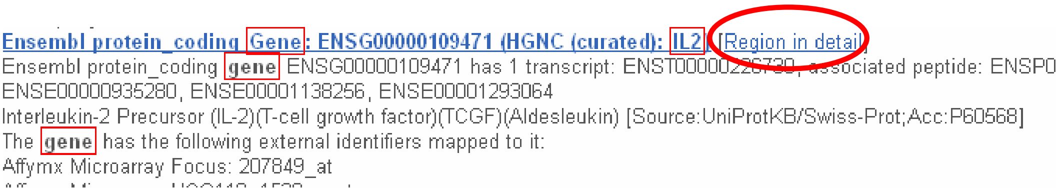


e.g. **human gene BRCA2** or **rat X:100000..200000** or **insulin**

(IL2 is the name given to this gene by the **HGNC**). Clicking on ‘Go’ yields the following results. It should be noted that this tutorial has been constructed using version 52 of Ensembl and therefore any updates to future versions of Ensembl may alter the search window and results. In order to emulate the example used in this pdf, you can use the archived version 52 of Ensembl which can be found here:

http://Dec2008.archive.ensembl.org/index.html

Here is the first result of the search. In this case it happens to be the correct entry but it should be noted that due to the method used by the search engine, the “proper” hit is not always the first one.

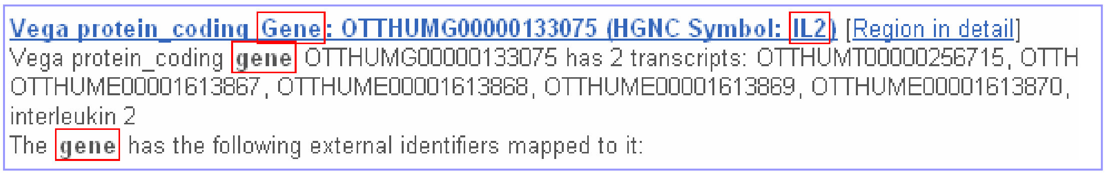


Ensembl protein_coding Gene: ENSG00000109471 (HGNC (curated): IL2) [Region in detail]
Ensembl protein_coding **gene** ENSG00000109471 has 1 transcript: ENST00000226730, associated peptide: ENSP0
ENSE00000935280, ENSE00001138256, ENSE00001293064
Interleukin-2 Precursor (IL-2)(T-cell growth factor)(TCGF)(Aldesleukin) [Source:UniProtKB/Swiss-Prot;Acc:P60568]
The **gene** has the following external identifiers mapped to it:
Affymx Microarray Focus: 207849_at

Clicking on the header for the IL2 entry will bring you to the **gene** page, where you can view more information about this gene (see **module 3**). The '**Region in Detail**' link at the top (circled in black) will bring you to a page to view a region of the chromosome corresponding to this gene (**module 4**).

The link to the **ENSG… ID** will bring you to the **gene** page. The search engine looks for a string of letters, so any **gene** page with the characters 'IL2' will be found in the search. Therefore beware… the search results may not be as simple as the above example. Let's look at the second hit for 'IL2'.

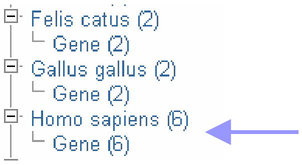
The second result:



Vega protein_coding Gene: OTTHUMG00000133075 (HGNC Symbol: IL2) [Region in detail]
Vega protein_coding gene OTTHUMG00000133075 has 2 transcripts: OTTHUMT00000256715, OTTH
OTTHUME00001613867, OTTHUME00001613868, OTTHUME00001613869, OTTHUME00001613870,
interleukin 2
The gene has the following external identifiers mapped to it:

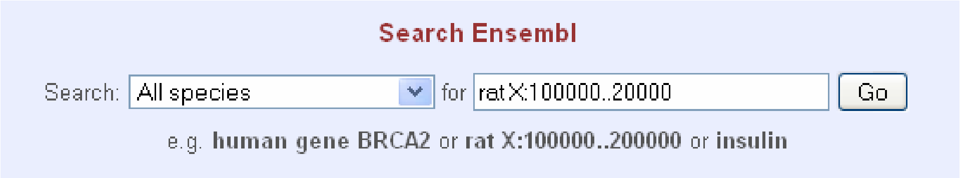
This shows a **VEGA** gene (see **manual curation** in the glossary at the end of this pdf), from the 'Vertebrate Genome Annotation' database. VEGA is maintained by a consortium of manual curators. These gene sets are determined by scientists who look at each individual case, in contrast to determining the genes all at once according to a set of rules, which is done by an **automatic annotation pipeline** such as that of Ensembl. Read more about Ensembl gene annotation here: www.ensembl.org/info/docs/genebuild/index.html. Part of the VEGA consortium, the HAVANA set (http://www.sanger.ac.uk/HGP/havana/havana.shtml a manually curated gene set at the Wellcome Trust Sanger Institute) is included into the Ensembl gene set for comparison. To select the gene that has come through the Ensembl annotation pipeline (rather than VEGA), we would choose the first search entry (ENSG00000109471) (note that the ID begins with 'ENS' for 'Ensembl'.

The search results may be filtered through the selections at the left hand of the page. For example, if we had searched for 'IL2' from the main page, without specifying 'human' and 'gene', we would have obtained families, protein domains, and genes across all Ensembl species. In order to narrow it down, 'human' or 'gene' could have been selected from the left hand side of the search results page, thus filtering the hits.



The above figure demonstrates the hits across species and types (in this case, markers and genes) resulting from a search for 'IL2'. To narrow down to one species, click on '*Homo sapiens*', in this example.

In addition to gene searches, entire chromosomal regions can be searched for using the format described below the search window. For example, rat X:10000..20000 would bring up a 'Region in Detail' display of the X chromosome in rat, specifically base pairs 100000 to 200000.


Search Ensembl

Search: All species for rat X:100000..20000

e.g. human gene BRCA2 or rat X:100000..200000 or insulin

We will examine the Region in Detail page in **module 4**.

Glossary

Annotation The assignment of genes and associated features to chromosome and base pair positions. The attachment of relevant information to a gene, such as its amino acid translation, single nucleotide polymorphisms, homologues, etc.

Automatic annotation pipeline The annotation of genes through a series of computer programs and algorithms in order to define a gene set all at once, rather than on a case-by-case basis.

HGNC symbol The gene name assigned by the HUGO Gene Nomenclature Committee (for human). http://www.genenames.org

Manual curation The annotation of genes by a team of scientists on a case-by-case basis using any evidence, including publications, literature, etc.

VEGA The manually curated gene set shown in Ensembl is the HAVANA set, part of the VEGA consortium. http://vega.sanger.ac.uk/

What to do next

For more about the gene build, read the article about how Ensembl annotates genes:

http://www.ensembl.org/info/docs/genebuild/index.html

Or, move on to **modules 3** and **4** to learn about the **Gene, Transcript** and **Region** pages.