e!Ensembl Tutorial

Ensembl provides genes and other annotation such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in UniProtKB and NCBI+ RefSeq+ databases, along with manual annotation from the VEGA/Havana group.

This tutorial aims at explotring different aspects and information from ensembl.

EXPLORING A REGION

1. Go to ensembl home page (or ensembl verterbrates) and select the human genome. Go to the region from bp 52000000 to 53000000 on human chromosome 4 by typing **4:52000000-53000000** on the search bar. Can you identify the contigs that lie within this region? They are indicated by alternating light and dark blue bands.
2. Zoom into the SGCB and LRRC66 genes, including some upstream and downstream sequences. This is done either by adjusting the location number on the search box, or dragging the red box around the region.
3. There is more information that is not shown by default. For example. CpG islands are genomic regions that contain high frequency of CG dinucleotides and are often locate d near promoter of mammalian genes; and are not shown by default. To activate, go to the side bar, click on **configure this page** and select **simple features** and select **CpG islands.**
    a. What wEmboss programs can you use to calculate the percentage of the CpG rich sequences?
4. To save the sequences, click on export data and follow the guide.

EXPLORING A GENE

1. Search for the human gene TAC1 by selecting the human genome and on the dropbox, select gene.
    a. On which chromosome is this gene located?
    b. How many splice variants are there for this gene? You can click on the **show splice variants** to show the transcript table.
    c. What is the transcript Id for the longest transcript? How long is it (base pairs, and amino acid residues)?
    d. How many coding exons does the gene contain?
    e. Browse through the Gene ontology terms for the transcript.
2. Are there any orthologues for the gene TAC1 in Mice?
    a. Click on the **orthologues navigation** tab on the side menu. How many orthologues are present? What is the ensembl identifier?

VARIATION

SNPs within a coding sequence do **not** necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. **SNPs** in the coding region are of two types, **synonymous** and **nonsynonymous SNPs**.

1. For the human TAC1, how many non-synomous SNP have been discovered for the protein encoded by transcript ENST00000319273?

Select the transcript (ENST00000319273). On the transcript page, go to the side bar and click on **protein information – variation.**

CONSERVED DOMAINS

Go to **Protein Information, Domain & features.**
How many different types of domain types have been identified?