**Module topic:** Genomics

# Practical Assignment

**Contact session title:** Comparative Genomics

**Trainer:** Fatma Guerfali

# Query Genes and Genomes

## Introduction

*The progress of sequencing technologies as well as the community efforts to release sequences* for genes and genomes are providing an unprecedented depth of biological information that is showing a constant increase. Exploiting the data available is rendered possible through the use of dedicated databases that provide now numerous tools to access, query or retrieve entire genomes or genes among other genomic features. We will navigate through the Ensembl database to retrieve information about genes and genomes.

## Tools used in this session

*Ensembl:* [*http://www.ensembl.org*](http://www.ensembl.org/)

## Please note

* **Hand-in information** If you are formally enrolled in the IBT course, please upload your completed practical assignment to the Vula ‘Practical Assignments’ tab. Take note of the final hand-in date for each practical assignment, which will be indicated on Vula.

## Task 1: Retrieve the CD74 gene in the Ensembl database and its corresponding Ensembl ID Task 1: instructions

*Open your browser and navigate to the Ensembl webpage (http://www.ensembl.org). Search* for the CD74 human gene.

*How many different hits are there? Which one would be the most relevant to choose? Indicate* its Ensembl ID.

## Task 2: Explore the related Gene information: about the sequence Task 2: instructions

*Click on the “CD74 (Human Gene)” link and have a look at the left-hand side panel. This* indicates options for the “Gene-based displays”. Use this panel to retrieve the sequence of this gene.

1. *How is the exon sequence indicated?*
2. *What is the format of this sequence?*
3. *What information is indicated in the header section of the sequence?*
4. *In what strand is this gene located?*
5. *How many transcripts are reported for this gene?*

## Task 3: Explore the related Gene information: about the variants Task 3: instructions

*Another option in the left-hand side panel is the “Configure this page” link. Use this link to* retrieve information about the annotated variants, by opening the link and setting the default option of the “Show variants” to the “Yes and show links” option (accept and close the configuration window in order to be able to see the changes in the sequence).

1. *Cite 3 types of variants highlighted in the sequence*
2. *Click on the link of the second inframe deletion present in the first exon (indicated GAG). What is its related rs number (dbSNP)?*
3. *Consider now the first variant detected in the first exon. Access it by clicking on the mutation itself, then click on the “Explore this variant” link on the new small window that opens. What information do you have about the mutation and the ancestral allele? What is the most severe consequence of this mutation ?*
4. *In the same previous page (rs1581262768) use the “Phylogenetic context” icon to retrieve information about this variant in different species. Select the Gorilla species to be compared in a pairwise alignment with Humans. Is this variant conserved between Human and gorilla?*

## Task 4: Explore the related Comparative Genomics information Task 4: instructions

*Back to the CD74 gene page, let’s explore and retrieve information from the “Comparative* Genomics” link (left-hand side options of the “Gene-based displays”).

1. *Click on the “Genomic alignments” link. In the “Alignments” box, click on “Select an Alignment” (or “Select another alignment” if the previous step has been performed) then select the “All alignments”, followed by the “Multiple” option and the “12 Primates EPO” option. Navigating through this page, could you tell what is the most evolutionarily distant one among these 12 species for this particular sequence?*
2. *Display the conservation regions and line numbering (through “Configure this page” then tick “Show conservation regions”, choose the most relevant option in the “Line Numbering”, and accept the modification by clicking on the upper-right icon). Display the full alignment. Which one of these species is showing no sequence conservation to the others for nucleotide positions around 5041-7321?*
3. *Select the “Orthologues” link (left-hand side options). In the list of orthologues, have a look at the Tetraodon nigroviridis region comparison. In which chromosome is the orthologuous gene located?*
4. *Go to the “Region in detail” using the corresponding link, then click on the synteny link (left-hand side options). What are the chromosomes in Zebrafish to which this Tetraodon chromosome is compared to by default?*
5. *Click on the location link indicated (bottom panel, see “1:524609-527222”) and explore the Variants reported. What are the splice region variants displayed?*

## Task 5: Explore the related Ontologies information Task 5: instructions

*Let’s explore and retrieve information from the “Ontologies” (left-had side options of the “Gene-* based displays”).

1. *Click on the “GO: Biological Process” link. What is the “GO:0002250” identifier referring to? What is the definition associated to that specific “vocabulary” ?*
2. *Use the “GO: Molecular Function” link to retrieve the different molecular functions of that gene. How many are there?*

## Task 6: Explore the related Genetic Variation information Task 6: instructions

*Let’s explore and retrieve information from the “Genetic Variation” (left-had side options of the* “Gene-based displays”). Click on the “Structural Variants” link. What large SV (Structural Variant) have been identified in this sequence?

*(for your information: you can click on the SV to view related publications)*