Practical activity - part 1: Primary and secondary genomic databases

From a genome-wide analysis of Estrogen Receptor alpha (ERα) chromatin binding performed in MCF-7 breast cancer cells, a research group identified a cluster of ERα binding sites mapped on chr21q22.2 region (Caroll et al., 2005). On this genomic region is mapped the *DSCAM* gene locus which could represent a candidate target of ERα activity in breast cancer. Using primary and secondary genomic databases you have to characterize the *DSCAM* gene locus and its expression in breast cancer cell lines.

- Connect to https://www.ebi.ac.uk/training/online/course/bioinformatics-terrified identify the main differences between flat file and relational databases and between primary and secondary genomic databases.
- Connect to the NCBI web site (https://www.ncbi.nlm.nih.gov/) and identify the list of database resources and the "how to" section.
- Search for *DSCAM* using the default search settings. Using the "Nucleotide" section of the results, identify the Homo Sapiens annotation for "*DSCAM* transcript variant 1".
    - What is the accession identifier of this sequence?
    - What is the first publication in which this DSCAM mRNA was cloned?
    - Identify the "Comment" section and identify the records from which this sequence was derived.
    - Using the "Feature" section, report the length of this *DSCAM* transcript variant?
    - In which nucleotide position the coding sequence starts?
    - Export the sequence in FASTA format.
- From the same page (using the "related information" module on the right), directly access to the *DSCAM* gene page and report:
    - The official full name
    - The genomic location
    - The functional role of *DSCAM* gene product

Using the "Genomic context" section identify the main features of *DSCAM* genomic locus:

- What are the genomic coordinates (location) of *DSCAM* genomic region in the human genome GRCh38 (hg38)?
- What is the length of the genomic locus?
- In which DNA strand *DSCAM* locus is annotated?

    - Are there any other gene annotations overlapping DSCAM locus?
    - By clicking on the gene name, identify the biotype of genes mapped in proximity of DSCAM. How many non coding genes are mapped in the region and in which strand?
    - How many RNA transcripts the *DSCAM* gene codes considering the Entrez Gene annotations?

At the page bottom, identify the protein accession ID of DSCAM and search to UniProtKB/Swiss-Prot database.

- What are the main biological processes involving DSCAM protein? In which part of the cell DSCAM protein is predicted to be located? Considering the DSCAM isoforms, what is the domain that is mainly affected by alternative splicing (compare the "topology" with the "alternative sequence" sections)?

Connect to Ensembl Genome Browser (http://www.ensembl.org) and access to the "Help & Docs" section. Using the information provided in the Ensembl Annotation section report:

    - The prefix used for gene and transcript annotations
    - The phases composing the standard Ensembl automatic annotation of coding genes

- Select the human genome GRCh38.
    - How many protein coding and noncoding genes are annotated in the genome primary assembly?
    - How many gene transcripts are annotated?
- Search for the *DSCAM* gene in the human genome
    - What is the Ensembl Gene ID for DSCAM?
    - How many transcripts the gene encodes?
    - What are the other gene loci overlapping the *DSCAM* gene? What is their gene biotype?
    - Is there any difference with the annotations reported in NCBI gene?
- Select the longest DSCAM isoform and using the "Exons" section, answer the following questions:
    - What is the length of the RNA and the protein sequence?

    - What is the absolute genomic position of the TSS of this isoform?
    - How many exons compose its sequence?
    - What is the length of the 5' and 3' UTR sequences?
    - What is the relative and absolute position of the ATG sequence?
- Considering the genomic variants overlapping the genes, indicate the position of the first variant generating a premature stop codon gain. What is the name of this variants?
- From the "Summary" section, extract the cDNA sequence of *DSCAM* longest transcript and store them in a FASTA file format. Separately, extract a region of 2,000 kbp upstream of the transcript TSS.
- From the "Orthologues" section, verify if *DSCAM* gene has an ortholog in mice? What is its gene identifier? Using the Genomic Alignment tools is there any sequence conservation in the DSCAM intronic and exonic regions?

Connect to Washu Epigenome Browser using the following link: http://epigenomegateway.wustl.edu/browser/

Insert the following code in the section "Session bundle ID" and click on "Retrieve session":

8b7e1f10-43de-11e9-8e53-ef34f4699812

Finally, click on "Restore" for the session called "DSCAM_RNA-Seq_MCF-7" to explore the expression of DSCAM in MCF-7 by the analysis of RNA-Seq data.

The top six tracks correspond to RNA-Seq reads mapped on plus strand while the bottom tracks to reads mapped on the minus strand. The tracks are relative to cytosolic (red), whole cell (purple), and nuclear (blue) long PolyA+ RNA-Seq experiments.

- Do RNA-Seq signals confirm the DSCAM expression in MCF7 cell line?
- Are there other genes overlapping DSCAM locus that are expressed in MCF-7?
- Considering the signal intensity, what is the higher expressed gene in this genome region?
- Using the information annotated to each track, explain the differences in the signal shape and distribution. Why are some signals spreaded on the gene body and other mapped only on gene exons?

- Search the highest expressed gene in NCBI gene. Is it involved in breast cancer pathology?