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.

- Connecting 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 how to section.
- Search for *DSCAM* using the default search settings. Using the Nucleotide section of the results identify a Genbank 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, 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 of *DSCAM* genomic region in the human genome GRCh38 (hg38)?
    - How long is the genomic locus?

    - In which DNA strand *DSCAM* locus is annotated?
    - There are any other gene annotations overlapping DSCAM locus?
    - 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?
- 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?
- Connect to Ensembl Genome Browser (http://www.ensembl.org) and access to the documentation. Using the information provided in the Ensembl Annotation section report:
    - The prefix used for gene and transcript annotations
    - The phases composing the standard Ensembl annotation process
    - The difference between known, novel or merged gene status.
- Select the human genome GRCh38.
    - How many protein coding and noncoding genes are annotated in the genome primary assembly?
    - How many gene transcript are annotated?
- Search for the *DSCAM* gene in the human genome
    - What is the Ensembl Gene ID for DSCAM?
    - How many transcripts the gene codes?
    - What are the other gene loci overlapping the *DSCAM* gene? What is their gene biotype?
    - There is any difference with the annotations reported in NCBI gene?
- Select the longest DSCAM isoform and answer the following questions:
    - How long is the RNA and the protein sequence?
    - What is the absolute genomic position of the TSS of this isoform?
    - How many exons compose its sequence?

    - How long are 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?
- 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.
- *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 (http://epigenomegateway.wustl.edu/browser/), select hg19 genome and import the ENCODE datasets public hub “Encyclopedia of DNA Elements”. Use the following session to analyse ENCODE data about RNA-Seq experiments performed in MCF-7 cell lines:

http://epigenomegateway.wustl.edu/browser/?genome=hg19&session=gKJuhc2Nyr&statusId=1727519239

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

- The RNA-Seq signals confirm the DSCAM expression in this cell line?
- There are other genes overlapping DSCAM locus that are expressed in MCF-7?
- Considering the signal intensity what is the higher expressed gene in this genome regions?
- Using the information annotated to each track, explain the differences in the signal shape and distribution. Why some signals are spreaded on the gene body and other mapped only on gene exons?
- Search the highest expressed gene in NCBI gene. It is involved in breast cancer pathology?