Practical activity - part 2: Transcriptomic, cistromic and functional databases

In the first phase of this exercise, using public omic and functional databases, you have to characterize the expression and the regulatory elements of the *DSCAM* and *DSCAM-AS1* genes. In the second phase, you will analyze the functional role of *DSCAM* protein-coding gene in a neuronal disease phenotype.

Gene and protein expression databases

Access to Gene Expression Atlas (https://www.ebi.ac.uk/gxa/home) and using the FAQ section answer the following questions:

- What type of data is analyzed in this database?
- What are the expression units that are used in this database?
- What are the two main factors that are considered to compute this measure?
- What results are shown in the Baseline Expression section and how they are represented?
- What results are shown in the Differential Expression section and how they are represented?

Search DSCAM-AS1 in the database and the results from the Baseline Expression section.

- What are the tissues in which DSCAM-AS1 is expressed and in which level?
- Using the filter at the top of the heatmap sort the data by expression rank and identify the cell lines characterized by the highest expression of this gene.

Using the Differential Expression section identify two comparisons in which DSCAM-AS1 is down-regulated and two in which is up-regulated. Report the log2 fold change measured in these experiments. Selecting the results from the “Diseases” section, report the condition that is generally characterized by the highest DSCAM-AS1 expression.

From the home page access to the “Single Cell Expression Atlas” (top right). From the help, section understand which visualization and analysis technique is used for the data reported in this atlas.

From the “Gene search” section search the *ESR1* gene and select the dataset named “Single-cell RNA-seq of primary breast cancer cells and lymph node metastases from 11 patients representing the four subtypes of breast cancer: luminal A, luminal B, HER2, and triple negative breast cancer”.

- How many cells were analyzed in this study?
- *ESR1* is expressed prevalently in a specific cluster?
- By coloring the plot with the histology data, what are the breast cancer subtypes with the highest ESR1 expression?

Access to the Human Protein Atlas database (http://www.proteinatlas.org/). Using the “About / Introduction” section report the main characteristics of the Atlas contained in this database. Using the help section identify how the protein expression level is scored, then search for the DSCAM protein-coding gene.

- What are the cell lines and the tissues with the highest DSCAM RNA and protein expression?
- These results are consistent across the three different databases analyzed (HPA, GTex, and Fantom5)?

Cistromic data analysis

Since DSCAM-AS1 is highly expressed in MCF-7 compared to neuronal tissue in (which DSCAM protein-coding gene is expressed), it is pivotal to identify the regulatory elements that could be involved in the gene overexpression in breast cancer.

Access to Washu Genome Browser (http://epigenomegateway.wustl.edu/browser/) using the following section “Session bundle ID”:

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

click on “Restore” for the session called “ChIP-Seq”

Use the “-1” or “-⅓” bottom on the top to gain a clear view of upstream and downstream signals around the DSCAM-AS1 locus. Each track represents a ChIP-Seq coverage signal in which the intensity is proportional to the ChIP enrichment in a specific genomic position (expressed as a number of mapped ChIP-Seq reads). The number reported on the left for each track represents the max number of reads counted for that experiment considering the analyzed genomic region. By comparing these numbers, answer the following questions:

- Which epigenetic modifications have the highest signal on the DSCAM-AS1 locus?
- What is the modification with the higher signal at DSCAM-AS1 gene body?
- Which of the analyzed proteins bound at DSCAM-AS1 promoter? Estrogen Receptor alpha protein (coded from the *ESR1* gene) is detected at DSCAM-AS1 promoter?
- Comparing the intensity of activatory and inhibitory epigenetic modifications you can confirm the transcriptionally active status of the region?

Transcriptional regulation in complex organisms is mainly driven by distal regulatory elements including enhancers and insulators. To search for DSCAM-AS1 distal regulatory regions expand the analyzed genomic regions using two times the “-5” option at the top.

- Do you see any other signals mapped upstream of the DSCAM-AS1 regions?
- Which epigenetic modifications and transcriptional regulators ChIP-Seq signals are enriched in these regions?
- Using the arrow at the top of the screen, approximately what is the distance between these elements and the DSCAM-AS1 gene?

Genomic experiment like ChIA-PET and Hi-C can identify long-range chromatin interactions between genomic regions. Using the public track hub (Tracks / Public data hubs) load the track “Long-range chromatin interaction experiments”. Using the “Track Facet Table” in the Tracks section load the tracks called “Long Range Interaction” obtained from the analysis of ChIA-PET experiment performed for the MCF-7 cancer cell lines. Among all the tracks select:

- GIS-Ruan_ChiaPet_MCF-7_Pol2-R1
- GIS-Ruan_ChiaPet_MCF-7_ERalpha_a-R1
- GIS-Ruan_ChiaPet_MCF-7_CTCF-R1

The new tracks were added at the bottom. By right-clicking on the track names change the Display mode from “Heatmap” to “Arc” and increase the height of the tracks.

- Since each purple arc represents an experimentally validated long-range interaction, the data confirm the interactions between the regulatory regions and the DSCAM-AS1 promoter?
- Finally, explore the DSCAM protein-coding gene promoter. Does the long-range chromatin interactions between these enhancers involve this regulatory region?

Cancer data integration databases

Since DSCAM-AS1 is a gene regulated by Estrogen Receptor alpha, would be interesting to investigate in which subtype of breast cancer the Estrogen Receptor alpha coding gene (*ESR1*) is over-expressed or altered by mutational or copy number variation events.

To analyze multiple genomic and transcriptomic data of primary tumors from large cohorts of patients, the website cBioPortal can be used. Access to cBioPortal (http://www.cbioportal.org/). How many studies are analyzed? How many of them concern to breast cancer disease?

Select the study “Breast Invasive Carcinoma (TCGA, Cell 2015)” and view the study summary by clicking on the pie chart on the right. How many patients were involved in this study? What is the gene that is more frequently mutated?

Return on the main page, select the same study and all the options in the field “Select Genomic Profiles”. Digit *ESR1* gene in the form on the bottom and then press on “Submit the query”. In how many patients *ESR1* gene is altered? What is the main alteration of the gene? There are mutations detected in the ESR1 gene? There is a difference in survival time between patients carrying or not an altered ESR1 gene (see the section “Survival”)?

Functional databases

Given a large amount of public biological experiments it is now easy to explore the function of a gene if belong to functionally annotated gene sets.

DSCAM was evidenced as altered in an individual with autism disorder. Access to GEO (https://www.ncbi.nlm.nih.gov/geo/) and using the GSE63524 identifier search the information about the experiment. In the “Summary” and “Overall design” section report the information about the tissues on which the experiment was performed. What is the name of the high-throughput technology that was used for this experiment (“Platform” section)?

From Moodle retrieve the list of the genes (Autism_genes.txt) was obtained by the analysis of this experiment which contains genes altered in stem cells from people with autism.

Access to Enrichr web tools (http://amp.pharm.mssm.edu/Enrichr/#). From “Help / Background information” section report the definition of the gene set, gene set library, enrichment analysis, and enrichment term. From the “Libraries” section report three examples of gene-set libraries collected in Enrichr. Copy this list of the gene into the search form of the Enrichr web tool to identify the enriched functional gene sets. Exploring the Enrichr results answer the following questions by clicking on the indicated gene set library and selecting the table section. Consider results associated with a p-value lower than 0.001.

- What is a candidate transcriptional regulator of these genes? (Transcription / ChEA 2016)
- What are the main biological processes involving a significant fraction of these genes? (Ontologies / GO Biological Process 2018)
- What are the brain regions in which a significant fraction of these genes are up-regulated and regions in which a significant fraction is down-regulated? (Cell Types / Allen Brain Atlas Up - Down)
- Excluding autism disorder, is there another neurological disease in which a significant number of these genes is down-regulated? (Crowd / Disease Perturbations from GEO down).