**GMQL for the analysis of (epi)genomic data features**

Ex. 1 – Find the genomic regions of human genome GRCh38 covered by signal peaks of both H3K4me1 and H3K27ac histone marks, which are potential active enhancers, in the HCT116 human cell line. Use ENCODE *narrowpeak* data (dataset of November 2017) regarding *experiment target* H3K4me1 and H3K27ac.

Then, count the number of regions in the input data and in the result.

Report obtained sample and region numbers, and discuss the results.

Version 1

#1 Loading the data

H3K4me1 = SELECT(biosample\_term\_name == 'HCT116' AND experiment\_target == 'H3K4me1-human') GRCh38\_ENCODE\_NARROW\_NOV\_2017;

H3K27ac = SELECT(biosample\_term\_name == 'HCT116' AND experiment\_target == 'H3K27ac-human') GRCh38\_ENCODE\_NARROW\_NOV\_2017;

#2 Combining multiple replicates in a single sample

H3K4me1\_1 = COVER(1, ANY) H3K4me1;

H3K27ac\_1 = COVER(1, ANY) H3K27ac;

#3 Generating single input dataset

K4me1\_K27ac = UNION() H3K4me1\_1 H3K27ac\_1;

#4 Extracting the portion of intersecting regions

RES\_0 = COVER(2, ANY) K4me1\_K27ac;

#5 Counting output regions and saving the result

RES = EXTEND(\_RegionCountOutput AS COUNT()) RES\_0;

MATERIALIZE RES INTO res;

#6 Counting and saving input regions

HM\_0 = UNION() H3K4me1 H3K27ac;

HM = EXTEND(\_RegionCountInput AS COUNT()) HM\_0;

MATERIALIZE HM INTO HM;

**Execution time**: 00:03:53

RES samples: 1

RES regions: 93 562

HM samples: 16

HM regions: 1 490 710

Version 2

#1 Loading the data

H3K4me1 = SELECT(biosample\_term\_name == 'HCT116' AND experiment\_target == 'H3K4me1-human') GRCh38\_ENCODE\_NARROW\_NOV\_2017;

H3K27ac = SELECT(biosample\_term\_name == 'HCT116' AND experiment\_target == 'H3K27ac-human') GRCh38\_ENCODE\_NARROW\_NOV\_2017;

#2 Combining multiple replicates in a single sample

H3K4me1\_1 = COVER(1, ANY) H3K4me1;

H3K27ac\_1 = COVER(1, ANY) H3K27ac;

#3 Extracting the portion of intersecting regions

RES\_0 = JOIN(distance < 0; output: INT) H3K4me1\_1 H3K27ac\_1;

#4 Counting output regions and saving the result

RES = EXTEND(\_RegionCountOutput AS COUNT()) RES\_0;

MATERIALIZE RES INTO res;

#5 Counting and saving input regions

HM\_0 = UNION() H3K4me1 H3K27ac;

HM = EXTEND(\_RegionCountInput AS COUNT()) HM\_0;

MATERIALIZE HM INTO HM;

**Execution time**: 00:03:50 (00:04:33 without initial covers)

RES samples: 1 (64 without initial covers)

RES regions: 93 562 (3 326 076 without initial covers)

HM samples: 16

HM regions: 1 490 710

Ex. 2 – Identify HG19 DNA regions bound by the JUN transcription factor that overlap binding sites of FOS transcription factor, which could indicate co-binding events and provide information for finding functionally related factors. Use ENCODE *narrowpeak* data (dataset of November 2017) regarding *experiment target* JUN and FOS in the HeLa-S3 cell line; particularly, use only the single sample data produced by the *lab* '*ENCODE Consortium Analysis Working Group*'.

Then, count the number of regions in the input data and in the result.

Report obtained sample and region numbers, and discuss the results.

#1 Loading the data

FOS = SELECT(biosample\_term\_name == 'HeLa-S3' AND experiment\_target == 'FOS-human' AND lab == '*ENCODE Consortium Analysis Working Group*') HG19\_ENCODE\_NARROW\_NOV\_2017;

JUN = SELECT(biosample\_term\_name == 'HeLa-S3' AND experiment\_target == 'JUN-human' AND lab == '*ENCODE Consortium Analysis Working Group*') HG19\_ENCODE\_NARROW\_NOV\_2017;

#2 Identifying the JUN binding regions overlapping with FOS binding regions

RES\_0 = JOIN(distance < 0; output: LEFT\_DISTINCT) JUN FOS;

#3 Counting output regions and saving the result

RES = EXTEND(\_RegionCountOutput AS COUNT()) RES\_0;

MATERIALIZE RES INTO res;

#4 Counting and saving input regions

TF\_0 = UNION() FOS JUN;

TF = EXTEND(\_RegionCountInput AS COUNT()) TF\_0;

MATERIALIZE TF INTO TF;

**Execution time**: 00:02:06

RES samples: 1

RES regions: 7692

TF samples: 2

TF regions: 31 228

Ex. 3 - Identify DNA promotorial regions bound by the MYC transcription factor that present somatic mutations (if any) in breast cancer patients. Use narrowpeak data from the ChIP-seq *assay* of MYC binding sites for cell line MCF-7 from ENCODE (HG19 dataset of November 2016), DNA-seq data from TCGA breast cancer (BRCA *tumor tag*) patients with clinical *followup reason* *'additional new tumor event*', and genomic region annotations data (promoter, TSS, gene) from RefSeq (dataset HG19\_BED\_ANNOTATION).

Then, for each selected patient count the number of MYC bound promoters with patent’s mutations.

Report obtained sample and region numbers, and discuss the results.

Ex. 4 – Differential binding search. Answer the following questions using HG19 ENCODE *narrowpeak* data with *output type* ‘*optimal idr thresholded peaks*’regarding JUN *experiment target* from the ChIP-seq *assay* (dataset of November 2017) for cell line k562 (chronic myelogenous leukemia), with 1 control and 2 cases treated using interferon alpha or gamma, respectively, each of the latter ones with two replicas. Comment required GMQL statement parameters not specified in the exercise text.

Report obtained sample and region numbers, and discuss the results.

Question 1: Combining replicate samples (if existing) by considering only regions in at least two replicas and calculating the minimum pvalue of the combined regions, find those JUN binding regions in each treated cases where there are no JUN binding regions in the control case.

Question 2: Find the JUN binding regions that are differentially present between treated cases (e.g., in alpha and not in gamma interferon treated samples) and not in control.

Question 3: Find the RefSeq genes (dataset HG19\_BED\_ANNOTATION) that overlap with JUN binding regions either in control or treated samples.

Question 4: In each interferon alpha and gamma treated and control condition, find the overall average JUN ChIP-seq signal in gene promoters (areas surrounding the TSS) intersecting JUN binding regions either in control or treated samples.