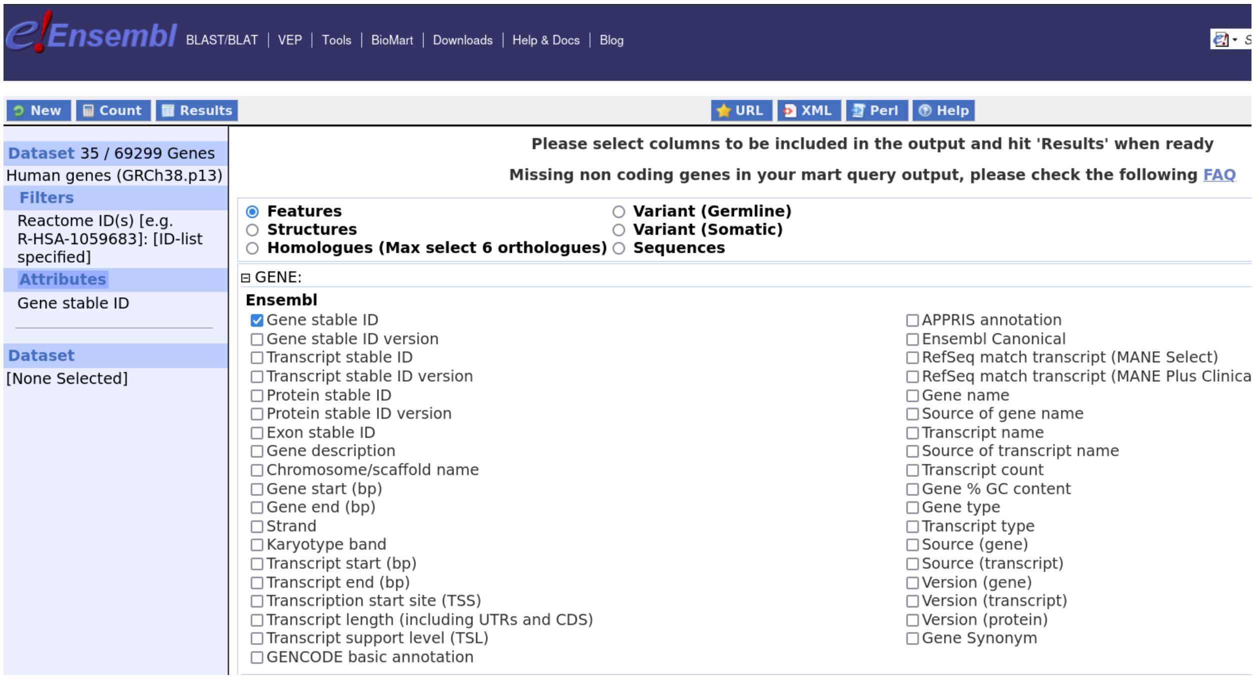
The pathways of Gluconeogenesis (the ID in REACTOME is R-HSA-70263) and the pathways of Glycolysis (the ID in REACTOME is R-HSA-70171) are both related to the metabolism of sugars.

You have to complete the following tasks in order to understand the regulation of gene expressions in these two pathways.

1. Download the gene names of these two pathways from the ENSEMBL (www.ensembl.org) database (use as Filter the ID of each pathway and as Attribute just the Gene Stable ID (you need just the Features here) as shown in the figure


e!Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results
URL
XML
Perl
Help

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following FAQ

Dataset 35 / 69299 Genes
Human genes (GRCh38.p13)
Filters
Reactome ID(s) [e.g. R-HSA-1059683]: [ID-list specified]
Attributes
Gene stable ID

Dataset
[None Selected]

● **Features**
○ **Structures**
○ **Homologues (Max select 6 orthologues)**
○ **Variant (Germline)**
○ **Variant (Somatic)**
○ **Sequences**

⊟ GENE:

Ensembl

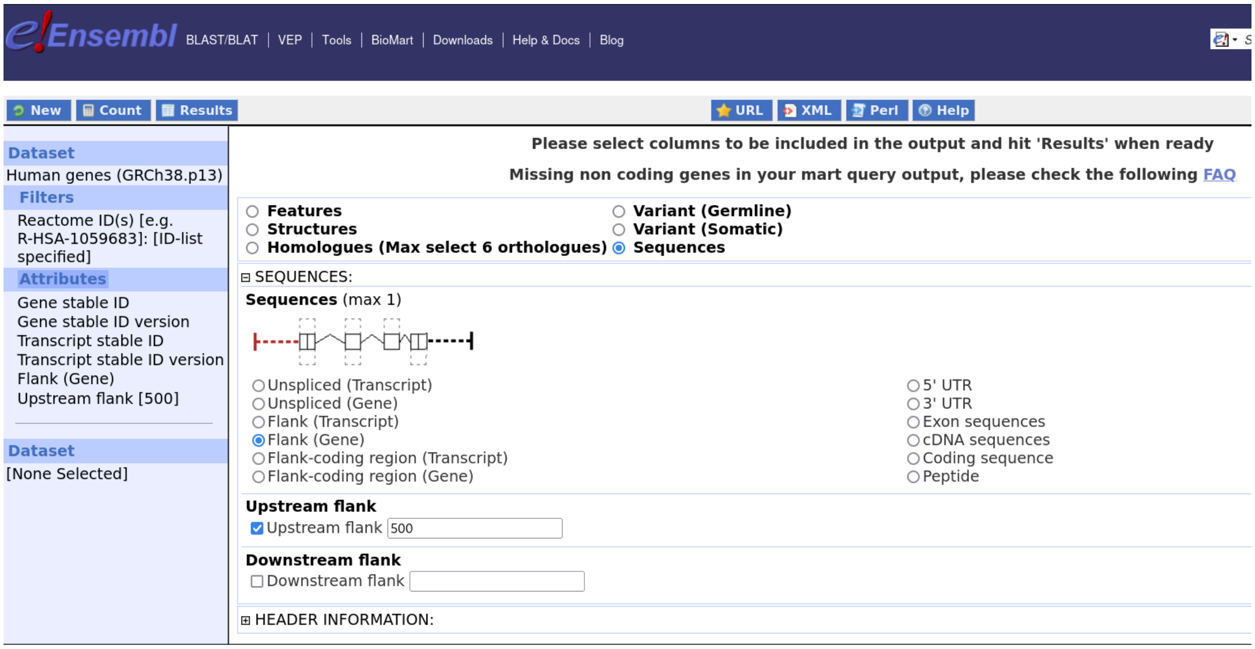
☑ Gene stable ID
☐ Gene stable ID version
☐ Transcript stable ID
☐ Transcript stable ID version
☐ Protein stable ID
☐ Protein stable ID version
☐ Exon stable ID
☐ Gene description
☐ Chromosome/scaffold name
☐ Gene start (bp)
☐ Gene end (bp)
☐ Strand
☐ Karyotype band
☐ Transcript start (bp)
☐ Transcript end (bp)
☐ Transcription start site (TSS)
☐ Transcript length (including UTRs and CDS)
☐ Transcript support level (TSL)
☐ GENCODE basic annotation
☐ APPRIS annotation
☐ Ensembl Canonical
☐ RefSeq match transcript (MANE Select)
☐ RefSeq match transcript (MANE Plus Clinica
☐ Gene name
☐ Source of gene name
☐ Transcript name
☐ Source of transcript name
☐ Transcript count
☐ Gene % GC content
☐ Gene type
☐ Transcript type
☐ Source (gene)
☐ Source (transcript)
☐ Version (gene)
☐ Version (transcript)
☐ Version (protein)
☐ Gene Synonym

Be careful to choose "Unique results only" when you will get the results.

Question 1
Assuming that in total we have 69299 genes:
Do these two lists of gene names have more common genes than expected by chance?

Hint: Use hypergeometric. The null hypothesis is that the two lists of genes are just random selections from the total of 69299 genes. The number of "successes" in the hypergeometric is the number of common genes.

2. Go again to ENSEMBL and now get the **Upstream 500 bp for each gene (hint: In the Sequence, go to SEQUENCES, then flank, and Upstream 500, as shown in the figure**


Ensembl
BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog
New
Count
Results
URL
XML
Perl
Help

Dataset
Human genes (GRCh38.p13)
Filters
Reactome ID(s) [e.g. R-HSA-1059683]: [ID-list specified]
Attributes
Gene stable ID
Gene stable ID version
Transcript stable ID
Transcript stable ID version
Flank (Gene)
Upstream flank [500]

Dataset
[None Selected]

Please select columns to be included in the output and hit 'Results' when ready

Missing non coding genes in your mart query output, please check the following FAQ

○ **Features** ○ **Variant (Germline)**
○ **Structures** ○ **Variant (Somatic)**
○ **Homologues (Max select 6 orthologues)** ● **Sequences**

⊟ SEQUENCES:

Sequences (max 1)

○ Unspliced (Transcript)
○ Unspliced (Gene)
○ Flank (Transcript)
● Flank (Gene)
○ Flank-coding region (Transcript)
○ Flank-coding region (Gene)

○ 5' UTR
○ 3' UTR
○ Exon sequences
○ cDNA sequences
○ Coding sequence
○ Peptide

Upstream flank
☑ Upstream flank 500

Downstream flank
☐ Downstream flank

⊞ HEADER INFORMATION:

- Download the sequences for the upstream regions of each gene in these two pathways and create two separate files one of the glycolysis (e.g. glycolysis.fasta) and one for the gluconeogenesis (e.g. gluconeogenesis.fa) [hint the glycolysis contains 67 genes and the gluconeogenesis 35 genes].

- Download the sequences for a random set of 1000 genes [hint: either download all the gene names and just select 1000 randomly and then get their upstream sequences for these 1000 genes,
  or
  get the upstream sequences for all genes and then choose 1000 by chance; The first solution is actually better)

3. Find all substrings of length **8** that are overrepresented in each pathway compared to the set of random genes and estimate the p-values (using the hypergeometric distribution as we did in the class)

Question 2:
How many substrings with p-value < 0.001 there are in the glycolysis genes and how many in the gluconeogenesis genes?

Are there common substrings that are overrepresented in the two pathways? i.e. how many substrings with p-value < 0.001 are common in both pathways?

Question 3:

1. Build the PWM for each pathway for the substring with the smallest p-value (in the case of ties - i.e. the same p-value just get the first that appears in the sorted (p-value sorted from smaller to larger) vector of substrings).

2. If the PWM for the glycolysis is the matrix M1, then use M1 to scan the glycolysis genes and get the maximum score for each sequence. Report the 67 maximum scores (i.e. 1 score for each of the 67 sequences).

3. If the PWM for the gluconeogenesis is the matrix M2, then use M2 to scan the gluconeogenesis genes and get the maximum score for each sequence. Report the 35 maximum scores (i.e. 1 score for each of the 35 sequences).

4. Use the M1 to scan the genes for the gluconeogenesis. What is the maximum score for each sequence? How does it compare to the M2 scores (from the task 3)

5. Use the M2 to scan the genes for the glycolysis. What is the maximum score for each sequence? How does it compare to the M1 scores (from the task 2).