Post-GWAS Analysis
Gene and Functional Annotation

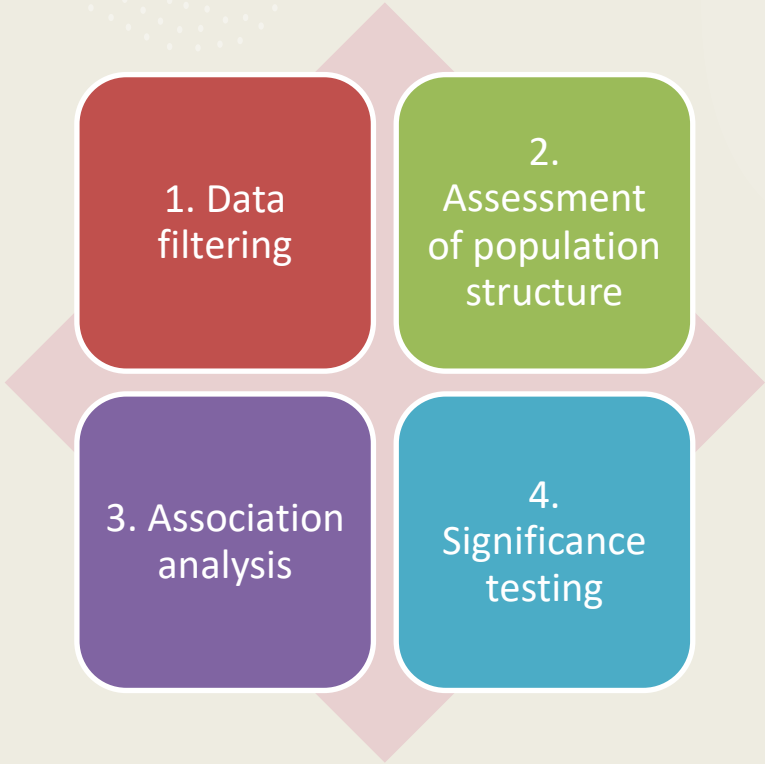
Training Material

Valentina Tsartsianidou /
Postdoctoral Researcher

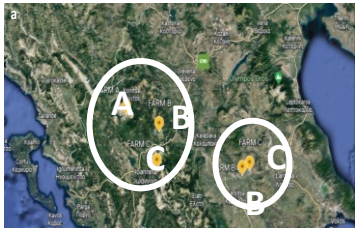
AUTH

14.05.2025

Summarize GWAS steps - 1st Lecture



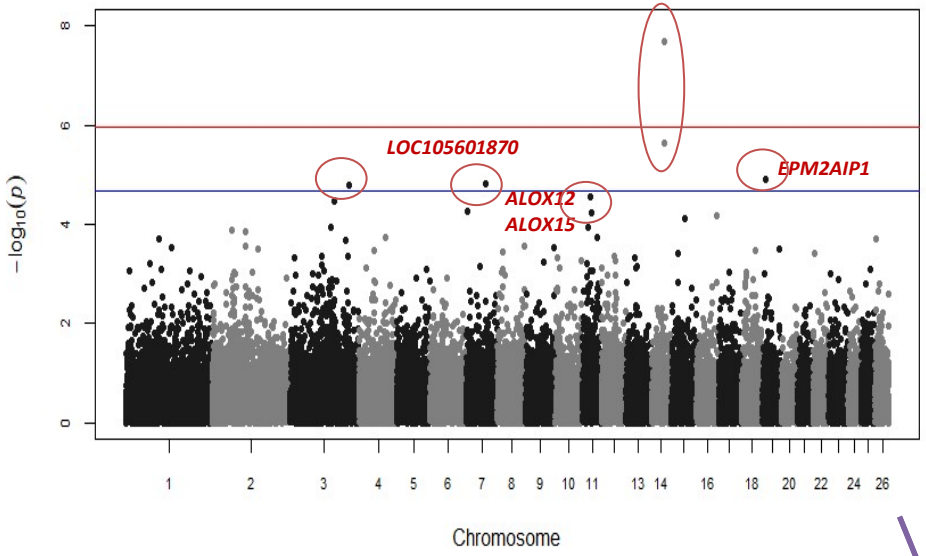
GWAS results – genome-wide Visualization in RStudio



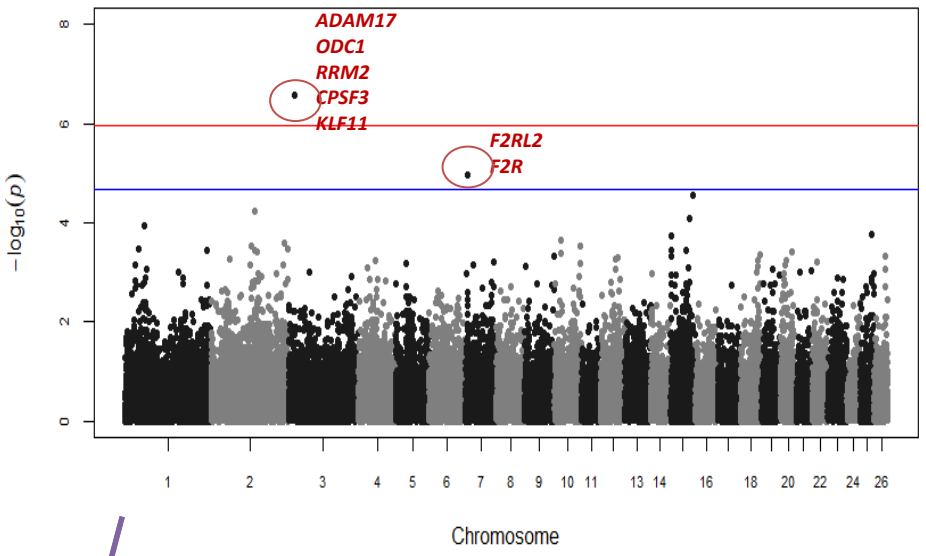
Manhattan plots

Association between SNPs and phenotype under study

Daily grazing duration



Altitude difference



Chromosomes of interest – 3, 7, 11, 14, 15, 19

Post-GWAS analysis
SNP and Gene Annotation

Search for candidate genes up/down-stream of the identified genomic markers **within a defined distance** (bp)

Why is this process needed?

Post-GWAS analysis
SNP and Gene Annotation

We link variants with gene function and structure to biologically interpret GWAS results and prioritize the genomic markers

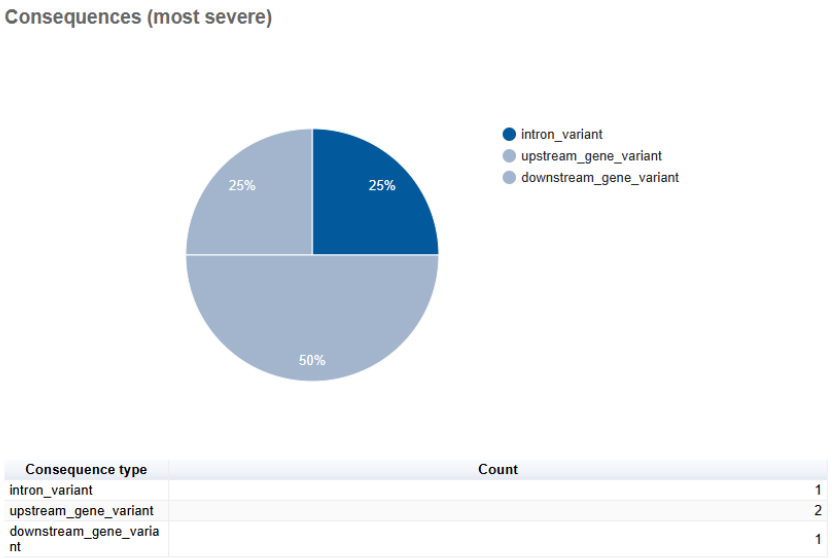
- Genomic Context

- Marker position: exonic, intronic intergenic, UTR, nearest gene or regulatory element

- Predicted Consequence of variant change -> Impact on gene product or regulatory element

- synonymous, missense, nonsense mutations

- Marker distance from the nearby genes



Consequence type	Count
intron_variant	1
upstream_gene_variant	2
downstream_gene_variant	1

Post-GWAS analysis
SNP and Gene Annotation

What genotype information is required to search for candidate genes closely located to our markers?

Statistically significant SNP markers identified through GWAS

Daily grazing duration

3 190036206 190036206 A/G
3 145252109 145252109 A/G
3 131380348 131380348 G/A
3 181785624 181785624 G/A
7 62298200 62298200 A/C
7 4217464 4217464 A/G
11 26623188 26623188 A/C
11 29347981 29347981 A/G
11 18248852 18248852 A/G
11 49829927 49829927 C/A
14 39858010 39858010 A/G
14 39795104 39795104 A/G
19 10373068 10373068 G/A

Altitude difference

15 71858906 71858906 G/A
15 63232034 63232034 G/A
15 3499482 3499482 G/A
7 7334440 7334440 C/A
3 18569142 18569142 A/C

Chromosome number
Marker position
bi-allelic genotype

Post-GWAS analysis
SNP and Gene Annotation

How is the annotation region defined?

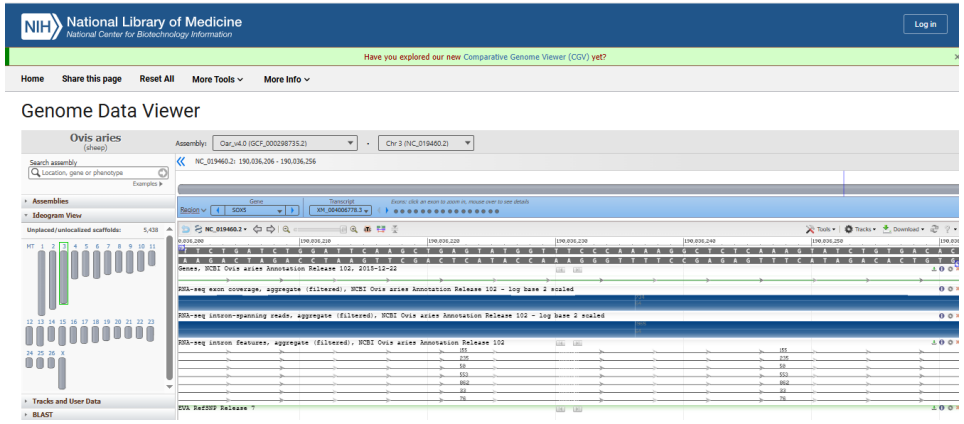
It can be decided based on pairwise Linkage Disequilibrium estimates

OR

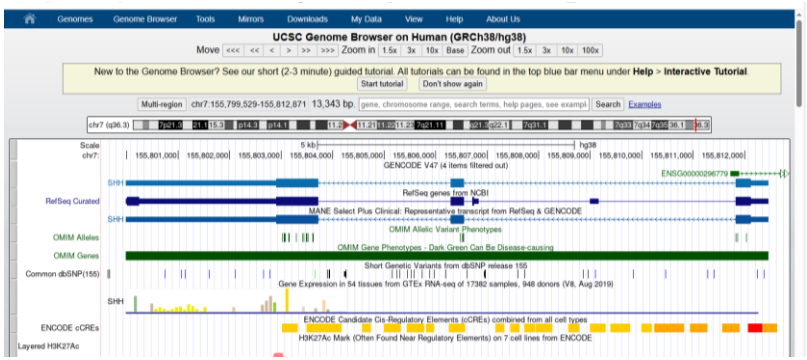
Use an annotation framework based on relevant literature appropriate for the specific analysis conducted

Post-GWAS analysis
SNP and Gene Annotation

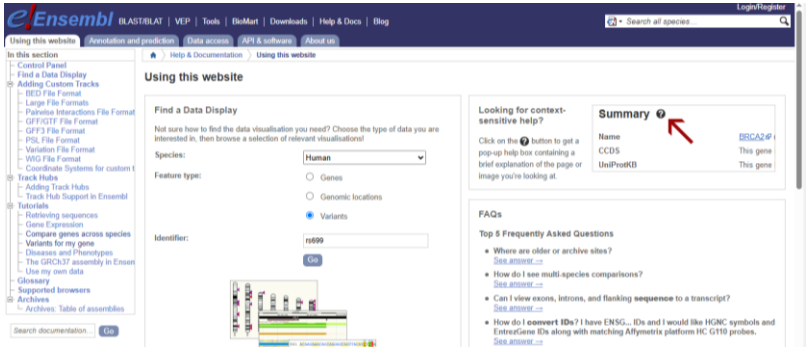
Extensively used Genome Browsers – manual process



https://www.ncbi.nlm.nih.gov/gdv/browser/genome/?id=GCF_000298735.2



https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chr7%3A155799529%2D155812871&hgsid=2545600844_HvqkPU09kLltfDUA1MZZ9hmPcGyd



https://www.ensembl.org/info/website/index.html

1. Convert Genome Coordinates between assemblies

Our data correspond to the *Ovis aries* genome assembly Oar_v4.0 → Texel breed - 2015

Conversion of genome coordinates using ARS-UI_Ramb_v2.0 → Rambouillet breed - 2022

in-depth functional annotation of the sheep genome
(https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giab096/6521876)

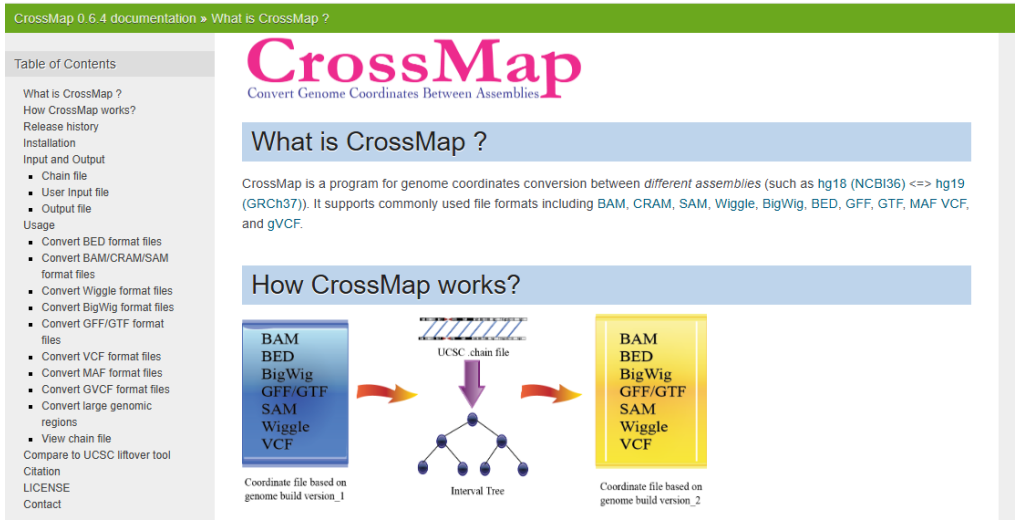
Coordinate Conversion - CrossMap

Genomic tool to convert genomic coordinates between assemblies

- Input: BED, VCF, GFF files

```
chr3    189803626    189803626    OAR3_204624914.1
chr3    145065882    145065882    OAR3_155103999.1
chr3    131220126    131220126    OAR3_140271376.1
chr3    181785624    181785624    OAR3_195793605_X.1
```

- Requires: Chain file
  - text file encoding alignments between different assemblies in segment blocks
    - contiguous aligned segments defining also the gaps in both target and query genomes



More info: https://crossmap.sourceforge.net/

2. Ensembl VEP (Variant Effect Predictor) – Variant Annotation

- Input:
  - -List of variants

```
3 190834971 190834971 A/G +
3 145489075 145489075 A/G +
3 131527909 131527909 G/A +
3 182812896 182812896 G/A +
```

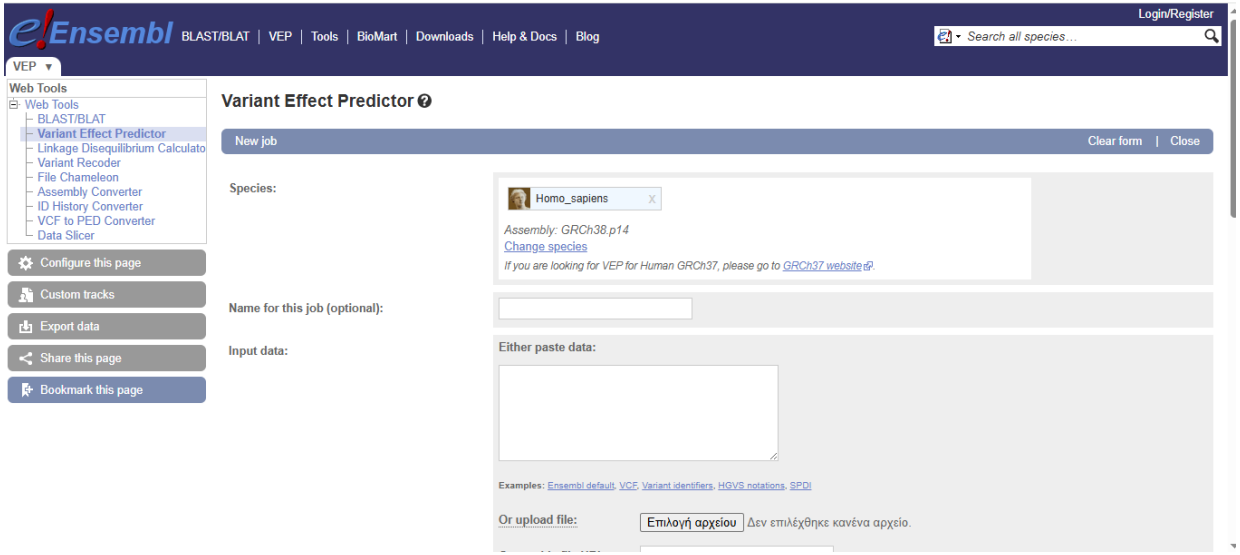
VEP default input

-Cache file = Ensembl annotation source: refers to a specific genome built

- File with all transcripts, regulatory features and variant data for a species
- It should be compatible to the vep version installed (https://www.ensembl.org/info/docs/tools/vep/script/vep_cache.html#cache)

- Output:
  - -Gene name
  - -transcript ID
  - -consequence
  - -distance

Web interface



https://www.ensembl.org/Multi/Tools/VEP?db=core

Example Output – VEP

```
## Codons : Reference and variant codon sequence
## Existing_variation : Identifier(s) of co-located known variants
## Extra column keys:
## IMPACT : Subjective impact classification of consequence type
## DISTANCE : Shortest distance from variant to transcript
## STRAND : Strand of the feature (1/-1)
## FLAGS : Transcript quality flags
## VEP command-line: vep --cache --database 0 --distance 500000 --input_file input_vep.Ensembl --output_file output_vep.Ensembl --species ovis_aries
#Uploaded_variation     Location        Allele  Gene    Feature Feature_type    Consequence     cDNA_position   CDS_position    Protein_position        Amino_acids     Codons  Existing_variation      Extra
3_190834971_A/G 3:190834971     G       ENSOARG00020039732      ENSOART00020005659      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=409525;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020006508      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=319939;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020006599      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=320099;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020006708      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=320098;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020006726      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=319936;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020006747      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=319939;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020006859      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=265242;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020007341      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=43693;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020007365      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=43693;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020007387      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=43693;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020007433      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=55413;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020007461      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=43687;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020007507      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=319936;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020003767      ENSOART00020043009      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=320096;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020044492      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=42871;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020045070      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=42871;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020030797      ENSOART00020049895      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=291786;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020034889      ENSOART00020052086      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=226257;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020053005      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=278348;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020059129      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=216518;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020060262      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=278348;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020060954      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=249387;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020040754      ENSOART00020062327      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=417103;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020040754      ENSOART00020063707      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=417111;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020004542      ENSOART00020066162      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=42871;STRAND=1
3_190834971_A/G 3:190834971     G       ENSOARG00020039732      ENSOART00020067972      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=409676;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020068667      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=227631;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020036459      ENSOART00020070106      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=31970;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020070693      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=278348;STRAND=-1
3_190834971_A/G 3:190834971     G       ENSOARG00020004466      ENSOART00020071286      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=216518;STRAND=-1
3_145489075_A/G 3:145489075     G       ENSOARG00020000431      ENSOART00020000557      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=28833;STRAND=-1
3_145489075_A/G 3:145489075     G       ENSOARG00020000431      ENSOART00020000571      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=29663;STRAND=-1
3_145489075_A/G 3:145489075     G       ENSOARG00020000431      ENSOART00020059714      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=28833;STRAND=-1
3_145489075_A/G 3:145489075     G       ENSOARG00020000431      ENSOART00020076609      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=28833;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014046      ENSOART00020011577      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=288466;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020013894      ENSOART00020021288      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=305238;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020013894      ENSOART00020021355      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=305169;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014046      ENSOART00020021482      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=258124;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014046      ENSOART00020021576      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=266310;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014046      ENSOART00020021657      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=258124;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014249      ENSOART00020021831      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=145132;STRAND=-1
3_131527909_G/A 3:131527909     A       ENSOARG00020014442      ENSOART00020022076      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=65519;STRAND=1
3_131527909_G/A 3:131527909     A       ENSOARG00020014571      ENSOART00020022295      Transcript      downstream_gene_variant -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=61279;STRAND=1
3_131527909_G/A 3:131527909     A       ENSOARG00020014635      ENSOART00020023247      Transcript      upstream_gene_variant   -       -       -       -       -       -       IMPACT=MODIFIER;DISTANCE=88214;STRAND=1
3_131527909_G/A 3:131527909     A       ENSOARG00020014635      ENSOART00020023352      Transcript      upstream_gene_variant                                                   IMPACT=MODIFIER;DISTANCE=88214;STRAND=1
```

3. Functional Annotation

Process of determining gene functions, pathways associated with specific sequences/genes

Gene Ontology (GO) = the widest annotation source of gene products

standardized system to describe gene functions and their products across species in a consistent way with structured vocabulary

GO Has Three Main Categories-Sources:

- **Biological Process (BP)**

  What the gene product does in a wide biological context.

  e.g., cell division, immune response, signal transduction.

- **Molecular Function (MF)**

  The basic activity of the gene product at the molecular level.

  e.g., ATP binding, enzyme activity, receptor binding.

- **Cellular Component (CC)**

  Where in the cell the gene product is active.

  e.g., nucleus, mitochondrion, plasma membrane.

3. Functional Annotation

Types of functional annotation

Code	Meaning
EXP	Inferred from Experiment
IDA	Inferred from Direct Assay
IEA	Inferred from Electronic Annotation
ISS	Inferred from Sequence or Structural Similarity
TAS	Traceable Author Statement
IMP	Inferred from Mutant Phenotype

3. Functional Enrichment

A statistical test to determine whether certain functions/pathways are over-represented in a defined set of genes

- Gene list as an input
- Select a reference genome including all annotated genes
- Test each annotation term (GO term) for enrichment
- Correct for multiple testing (e.g., Bonferroni, FDR)

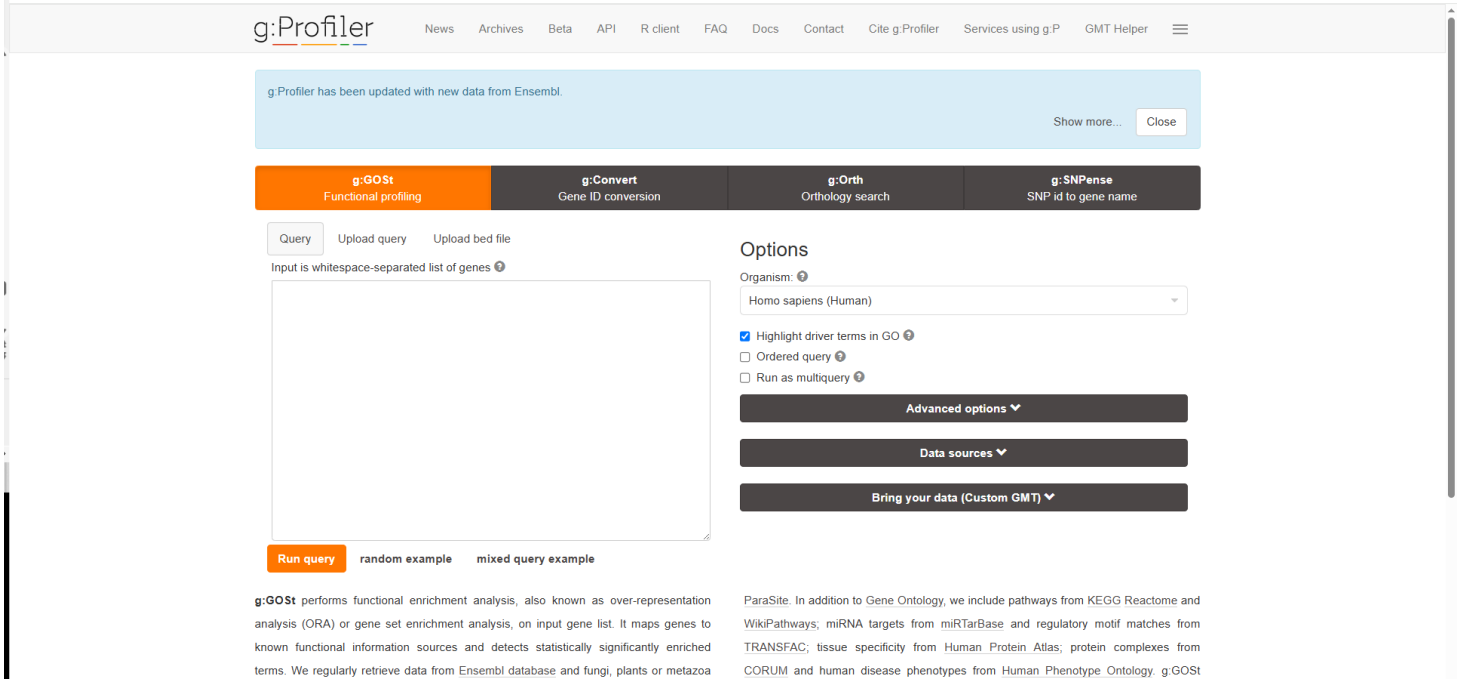
Functional Enrichment with g:Profiler2 in RStudio

article

https://pmc.ncbi.nlm.nih.gov/articles/PMC7859841/pdf/f1000research-9-29655.pdf

- Utilizes data from multiple reliable databases GO, KEGG, Reactome etc. for functional annotation across various species
- Input: gene list obtained from VEP tool
- Format: Ensembl gene IDs (preferred), Gene symbols

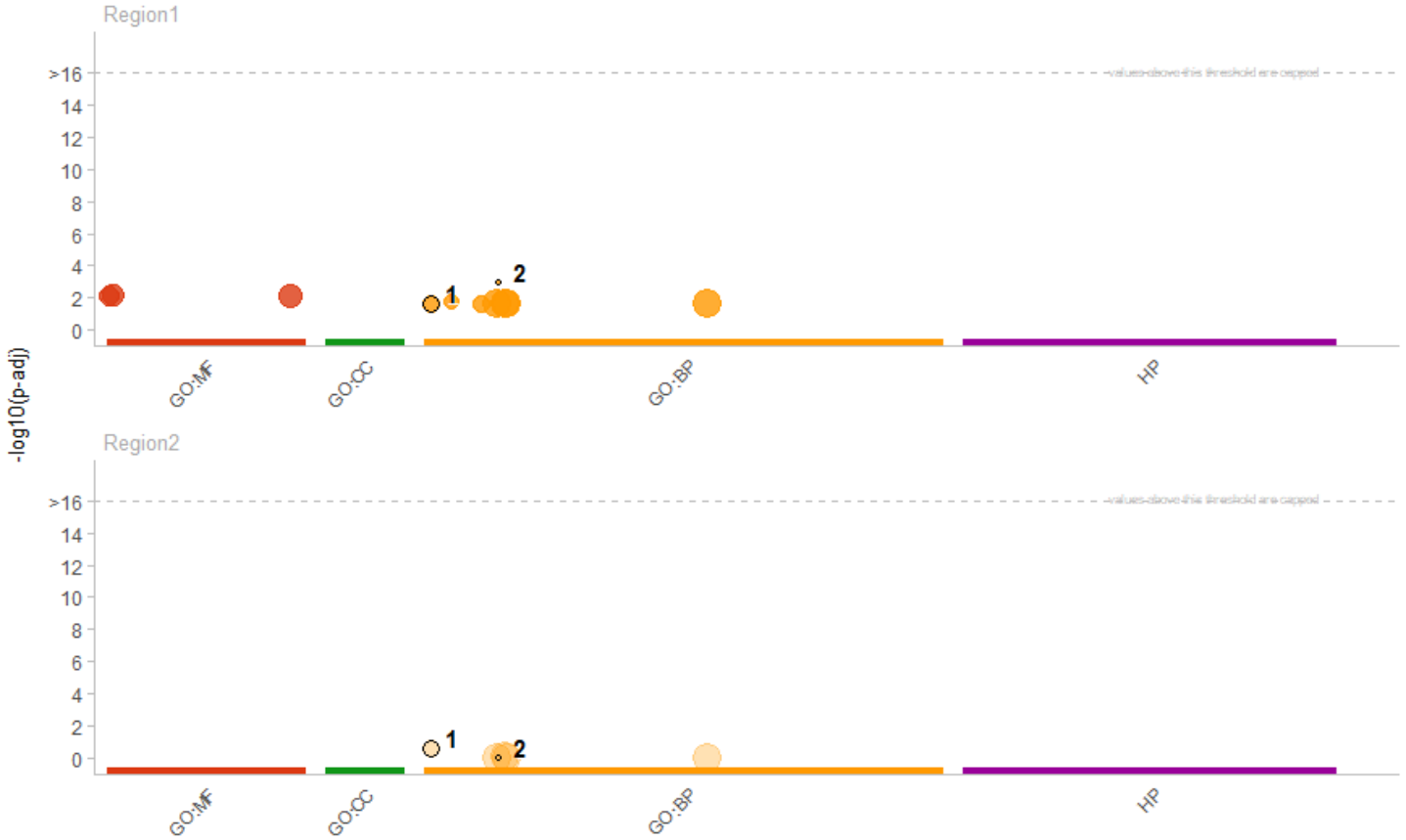
Web interface



https://biit.cs.ut.ee/gprofiler/gost

g:Profiler2 output

Manhattan plot of enriched GO terms using 2 gene lists (chr3) as input

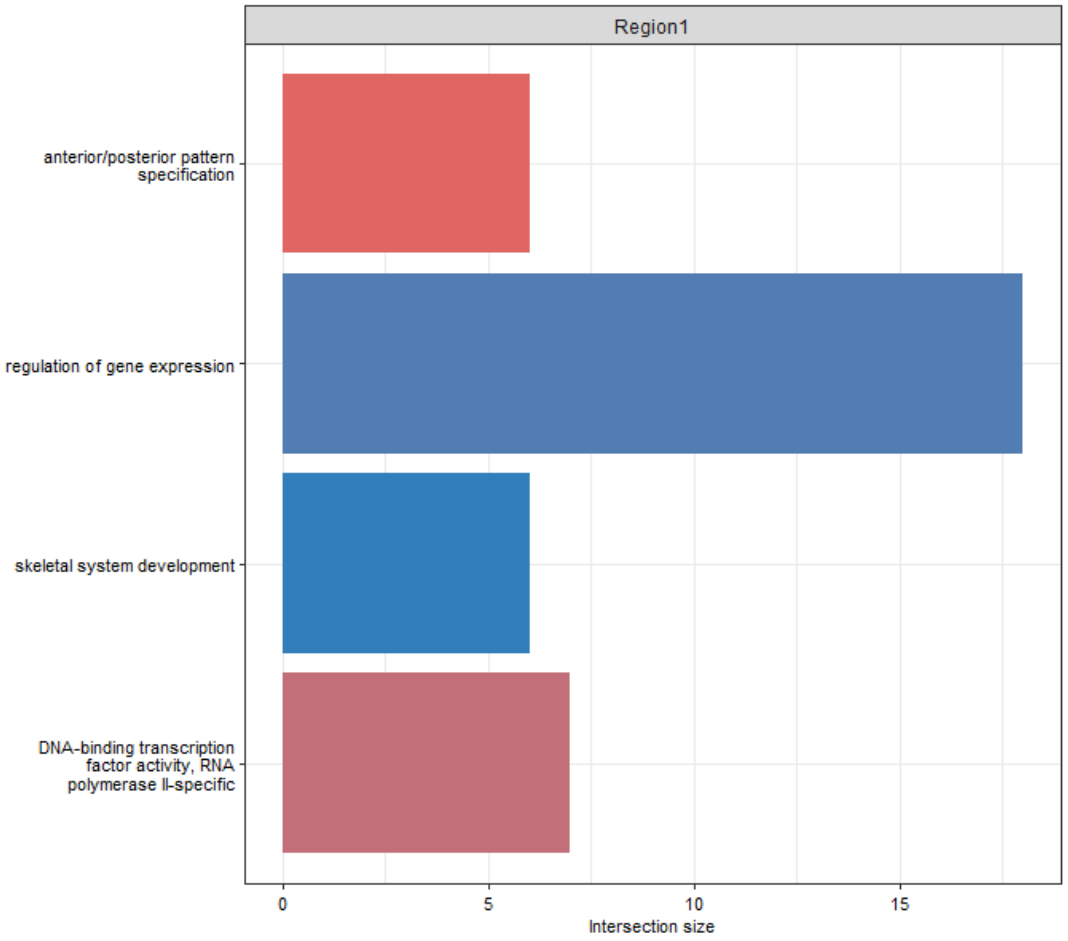


id	source	term_id	term_name	term_size	p_value Region1	p_value Region2
1	GO:BP	GO:0001501	skeletal system development	396	2.3e-02	2.4e-01
2	GO:BP	GO:0009952	anterior/posterior pattern specification	169	1.1e-03	1.0e+00

g:Profiler (biit.cs.ut.ee/gprofiler)

g:Profiler2 output

Bar plot of enriched GO terms showing number of candidate genes in each biological process detected



enrichplot()

Workflow Summary