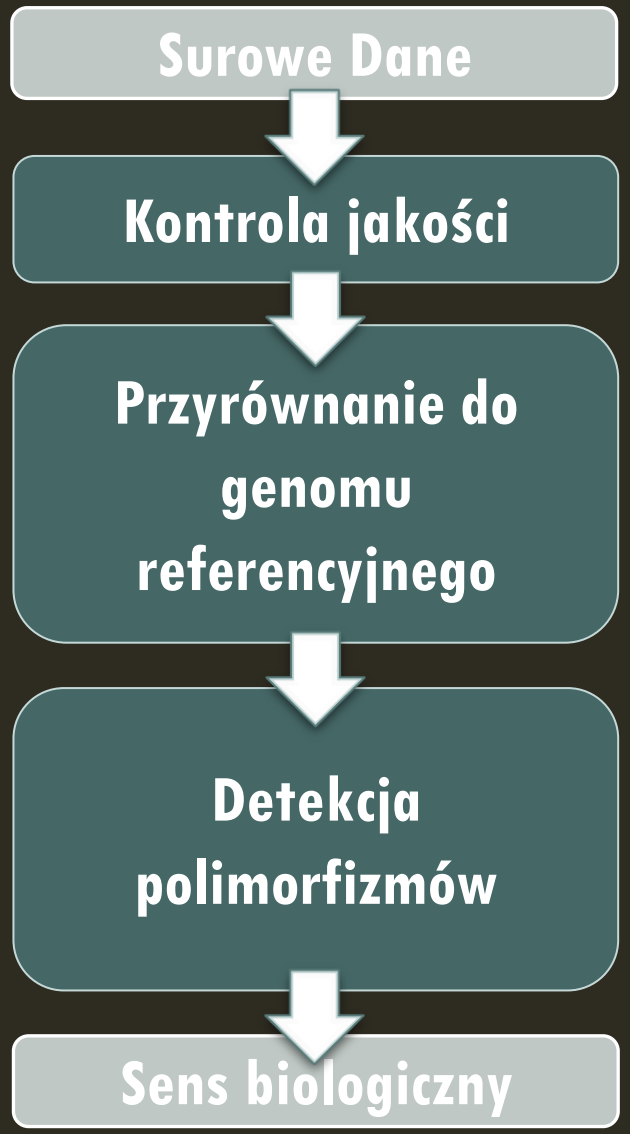
ADNOTACJA POLIMORFIZMÓW WYBRANE NARZĘDZIA

Analiza danych NGS
Wykład 7

SENS BIOLOGICZNY

Funkcjonalna adnotacja polimorfizmów

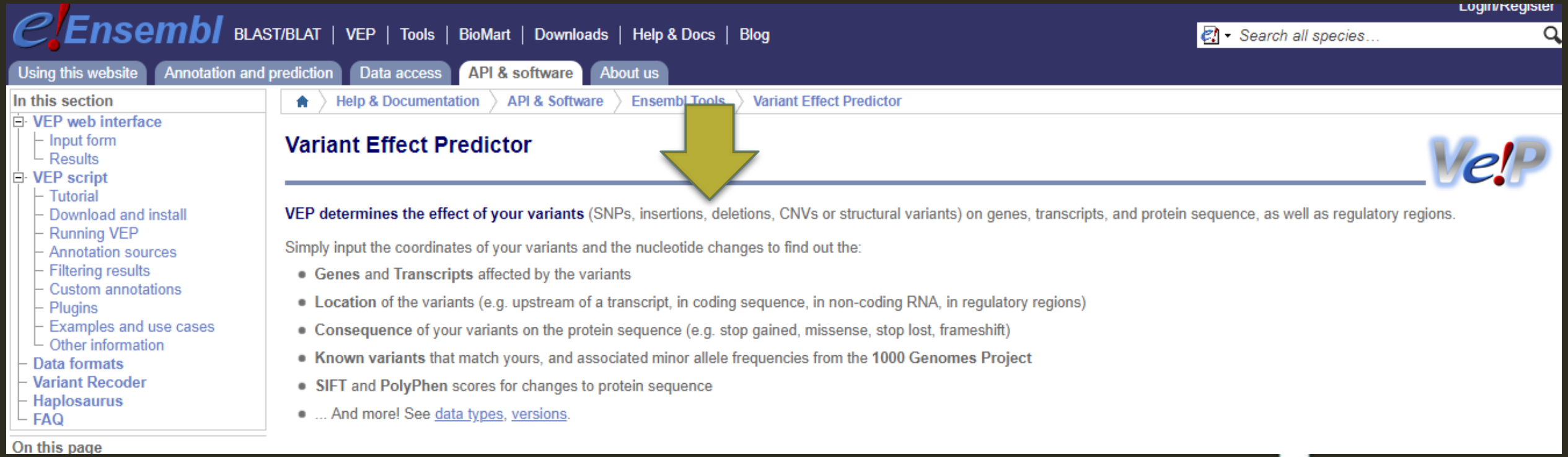
→ Sens biologiczny



VARIANT EFFECT PREDICTOR

WEB INTERFACE

VARIANT EFFECT PREDICTOR



Sens biologiczny

http://www.ensembl.org/info/docs/tools/vep/vep_formats.html

VARIANT EFFECT PREDICTOR

Variant Effect Predictor Data formats

Input

Both the web and script version of VEP can use the same input formats. Formats can be auto-detected by the VEP script, but must be manually selected when using the web interface. VEP can use VCF, variant identifiers and HGVS notations in addition to the default format

Default

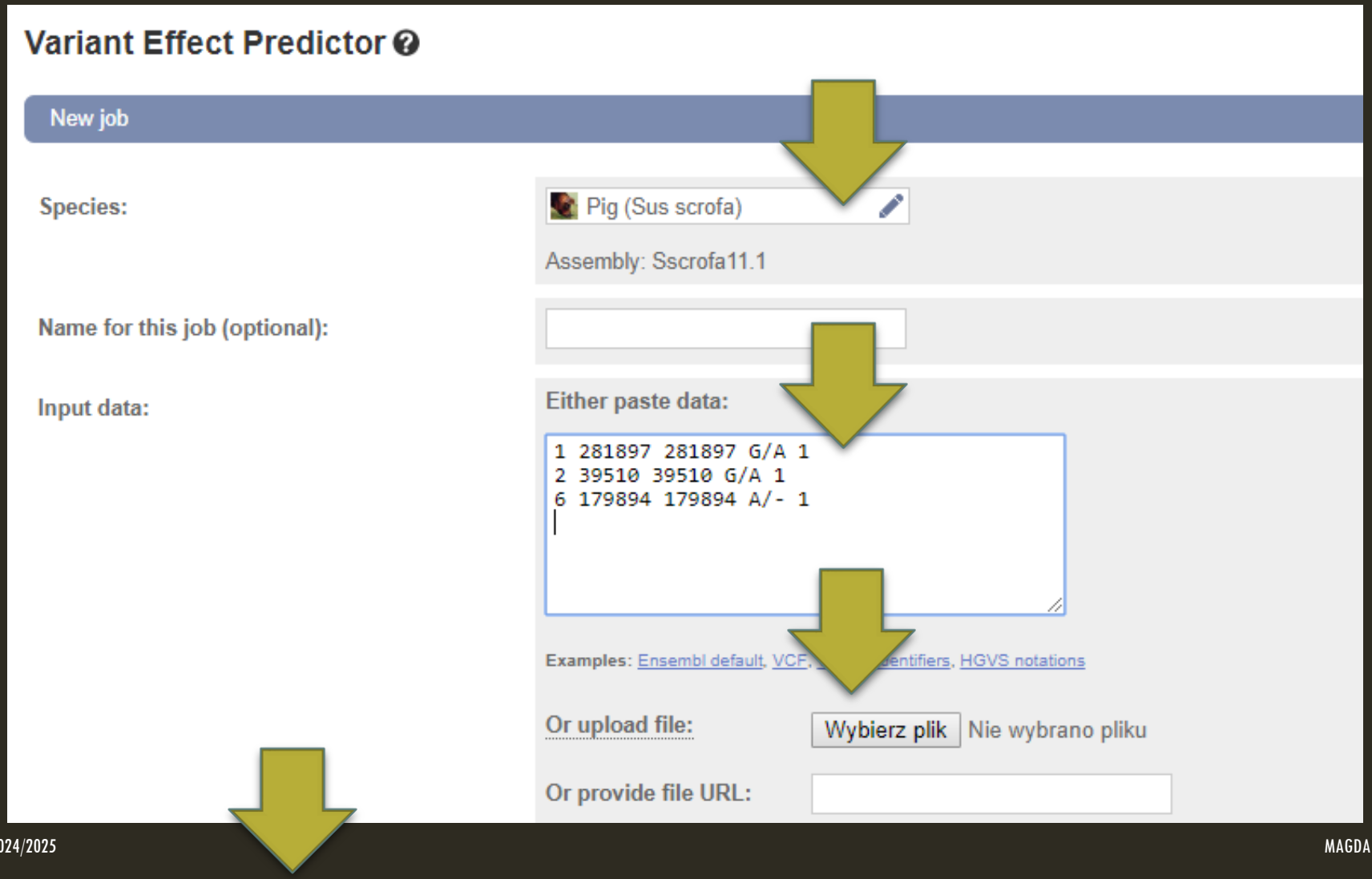
The default format is a simple **whitespace-separated** format (columns may be separated by space or tab characters), containing five required columns plus an optional identifier column:

1. **chromosome** - just the name or number, with no 'chr' prefix
2. **start**
3. **end**
4. **allele** - pair of alleles separated by a '/', with the reference allele first
5. **strand** - defined as + (forward) or - (reverse).
6. **identifier** - this identifier will be used in VEP's output. If not provided, VEP will construct an identifier from the given coordinates and alleles.

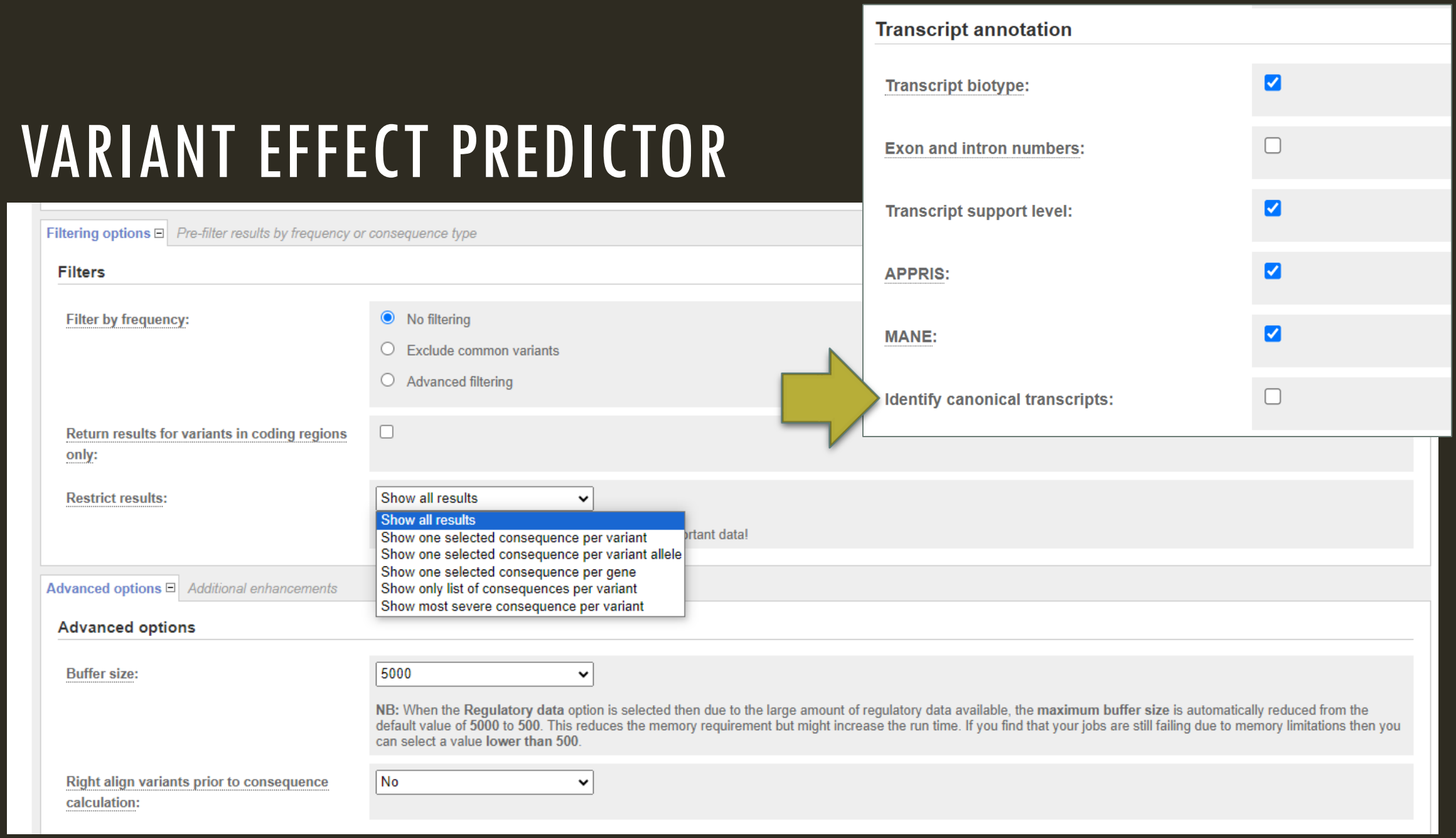
```
1   881907    881906    -/C   +
5   140532    140532    T/C   +
12  1017956   1017956   T/A   +
2   946507    946507    G/C   +
14  19584687  19584687  C/T   -
19  66520     66520     G/A   +    var1
8   150029    150029    A/T   +    var2
```

WEB INTERFACE

VARIANT EFFECT PREDICTOR

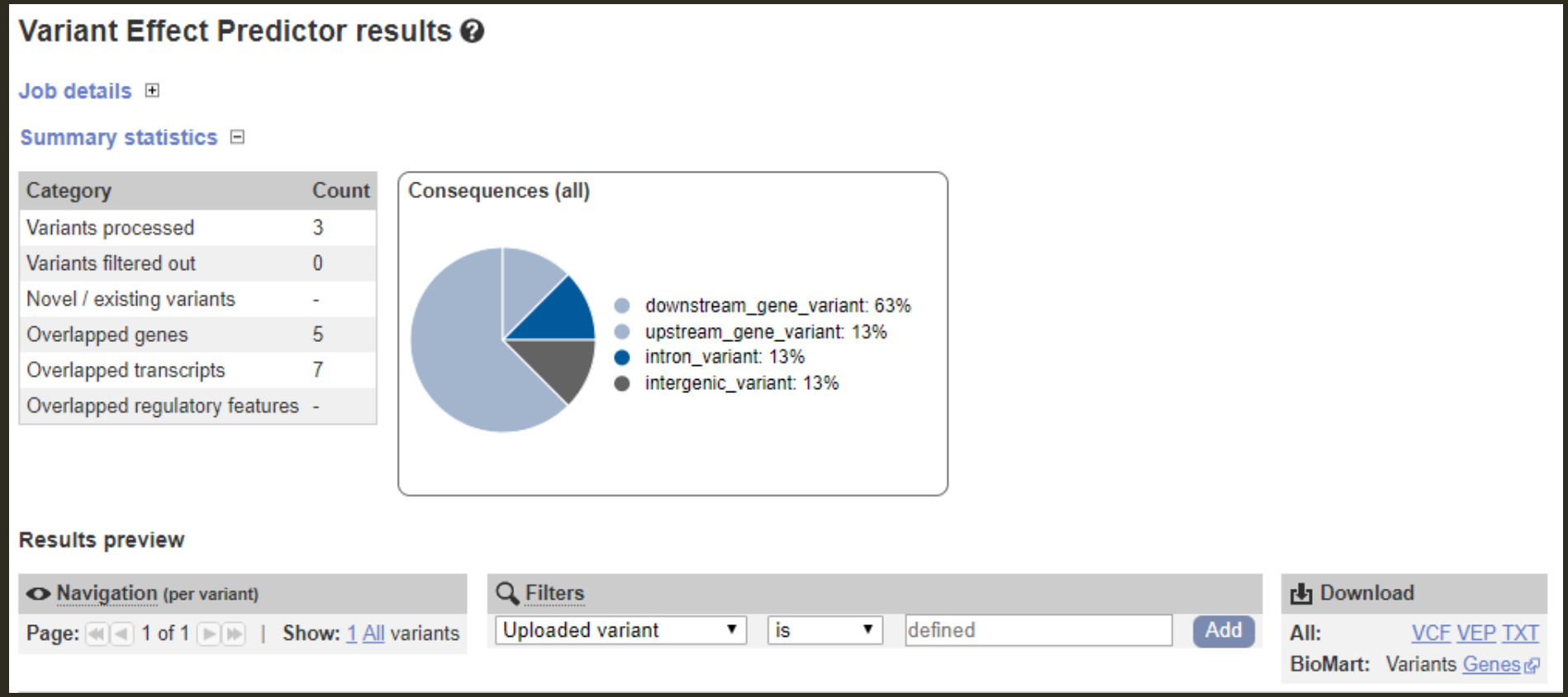


VARIANT EFFECT PREDICTOR



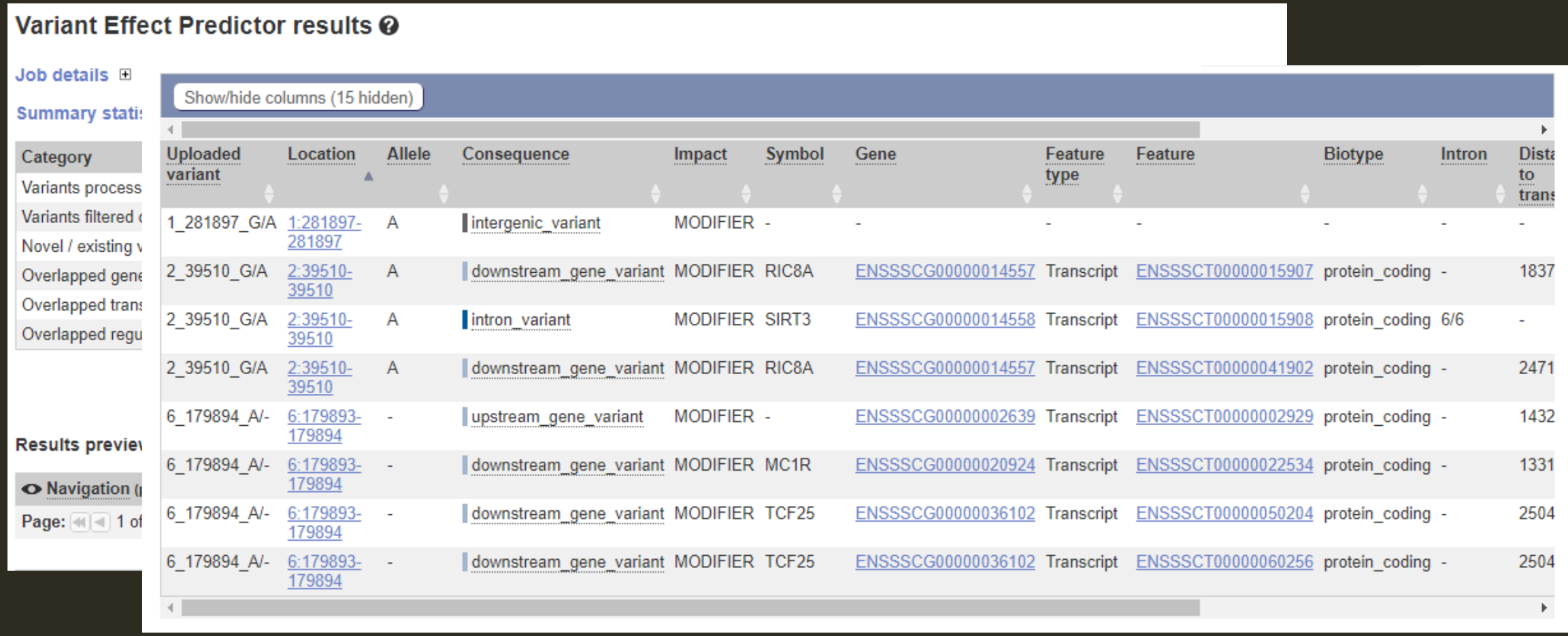
WEB INTERFACE

VARIANT EFFECT PREDICTOR



WEB INTERFACE

VARIANT EFFECT PREDICTOR



Variant Effect Predictor results

Job details

Summary statis

Category
Variants process
Variants filtered
Novel / existing v
Overlapped gene
Overlapped trans
Overlapped regu

Results previev

Navigation

Page: 1 of

Show/hide columns (15 hidden)

Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Intron	Dista to trans
1_281897_G/A	1:281897-281897	A	intergenic_variant	MODIFIER	-	-	-	-	-	-	-
2_39510_G/A	2:39510-39510	A	downstream_gene_variant	MODIFIER	RIC8A	ENSSSCG00000014557	Transcript	ENSSSCT00000015907	protein_coding	-	1837
2_39510_G/A	2:39510-39510	A	intron_variant	MODIFIER	SIRT3	ENSSSCG00000014558	Transcript	ENSSSCT00000015908	protein_coding	6/6	-
2_39510_G/A	2:39510-39510	A	downstream_gene_variant	MODIFIER	RIC8A	ENSSSCG00000014557	Transcript	ENSSSCT00000041902	protein_coding	-	2471
6_179894_A/-	6:179893-179894	-	upstream_gene_variant	MODIFIER	-	ENSSSCG00000002639	Transcript	ENSSSCT00000002929	protein_coding	-	1432
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	MC1R	ENSSSCG00000020924	Transcript	ENSSSCT00000022534	protein_coding	-	1331
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	TCF25	ENSSSCG00000036102	Transcript	ENSSSCT00000050204	protein_coding	-	2504
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	TCF25	ENSSSCG00000036102	Transcript	ENSSSCT00000060256	protein_coding	-	2504

WEB INTERFACE

Variant Effect Predictor results

Job details

Summary statistics

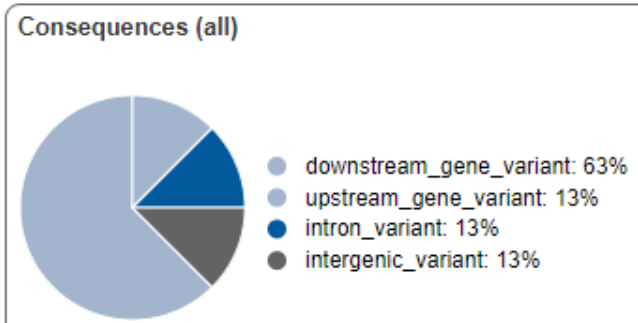
Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	4
Overlapped transcripts	14
Overlapped regulatory features	2

Variant Effect Predictor results

Job details

Summary statistics

Category	Count
Variants processed	3
Variants filtered out	0
Novel / existing variants	-
Overlapped genes	5
Overlapped transcripts	7
Overlapped regulatory features	-



http://www.sequenceontology.org/

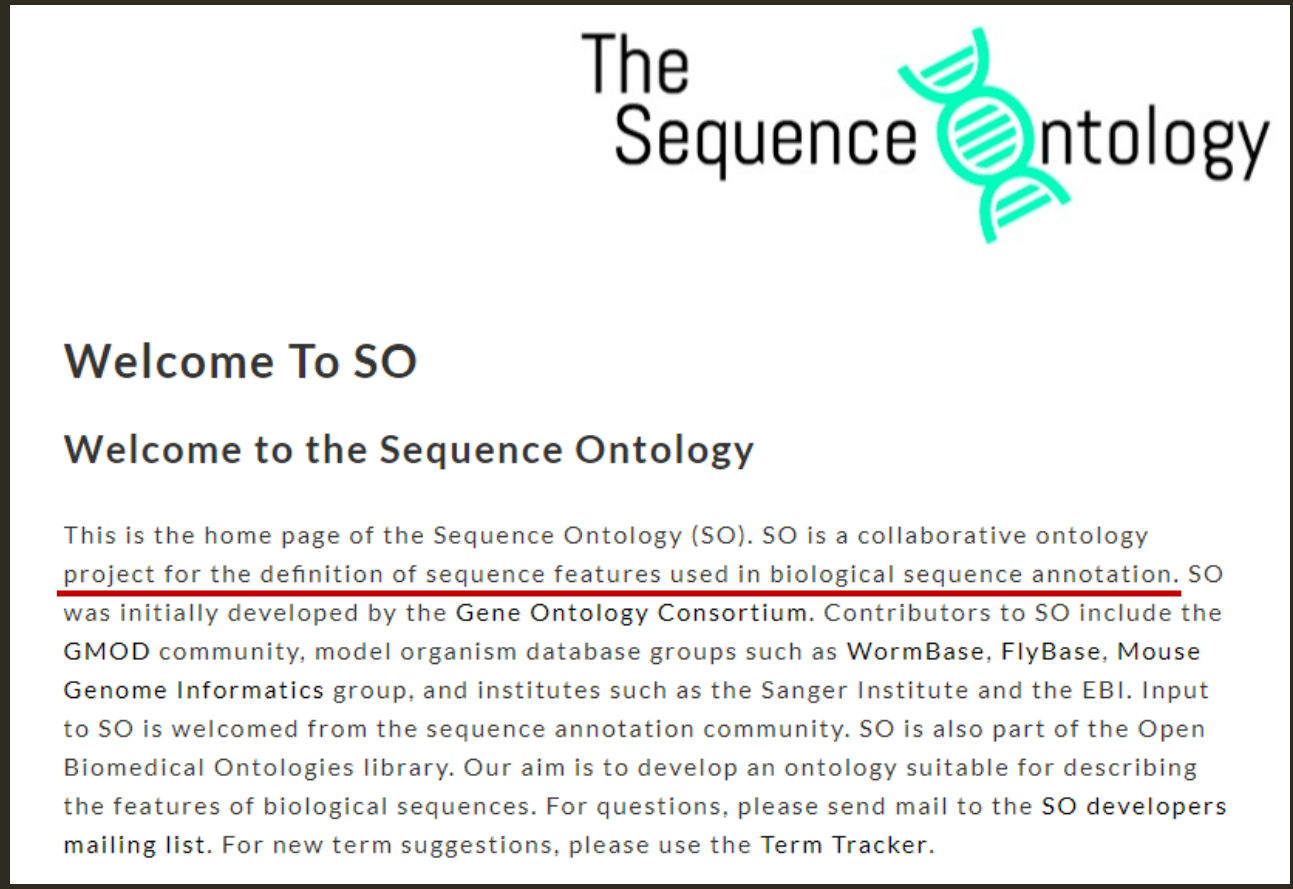
SO TERMS

Zbór terminów i relacji używanych do opisu cech sekwencji biologicznej.

Cechy biologiczne to takie, które są określone przez ich potencjał udziału w procesie biologicznym.

Opis funkcji, które mogą wykazywać określone sekwencje/fragmenty genomu.

Cel: ujednolicone i kontrolowane nazewnictwo do opisu podstawowych adnotacji sekwencji



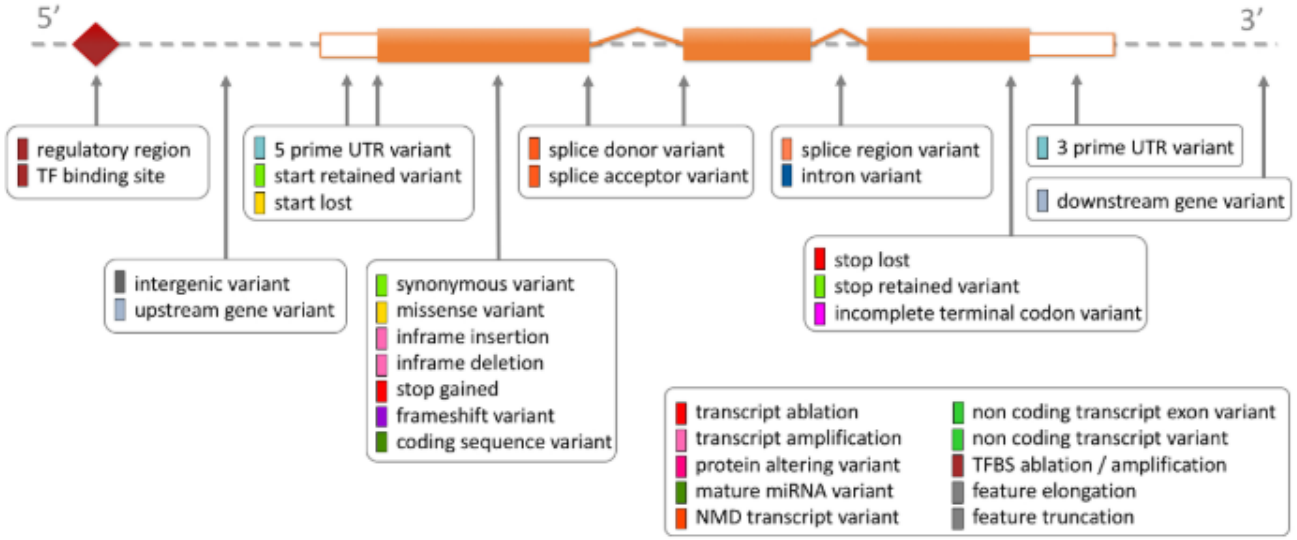
SO CLASSIFICATION

Ensembl Variation - Calculated variant consequences

For each variant that is mapped to the reference genome, we identify all overlapping Ensembl transcripts. We then use a rule-based approach to predict the effects that each allele of the variant may have on each transcript. The set of consequence terms, defined by the Sequence Ontology (SO), that can be currently assigned to each combination of an allele and a transcript is shown in the table below. Note that each allele of each variant may have a different effect in different transcripts.

This approach is applied to all germline variants and somatic mutations stored in the Ensembl databases. The resulting consequence type calls, along with information determined as part of the process, such as the cDNA and CDS coordinates, and the affected codons and amino acids in coding transcripts, are stored in the Ensembl Variation database and displayed on our website. For human and mouse variants any overlap with regulatory features is also displayed. For structural variants consequence terms are calculated on the fly for display on our website or API access. You can use this pipeline to annotate your own data via VEP.

See below a diagram showing the location of each display term relative to the transcript structure:



SO CLASSIFICATION

*	SO term	SO description	SO accession	Display term	IMPACT
	transcript_ablation	A feature ablation whereby the deleted region includes a transcript feature	SO:0001893	Transcript ablation	HIGH
	splice_acceptor_variant	A splice variant that changes the 2 base region at the 3' end of an intron	SO:0001574	Splice acceptor variant	HIGH
	splice_donor_variant	A splice variant that changes the 2 base region at the 5' end of an intron	SO:0001575	Splice donor variant	HIGH
	stop_gained	A sequence variant whereby at least one base of a codon is changed, resulting in a premature stop codon, leading to a shortened transcript	SO:0001587	Stop gained	HIGH
	frameshift_variant	A sequence variant which causes a disruption of the translational reading frame, because the number of nucleotides inserted or deleted is not a multiple of three	SO:0001589	Frameshift variant	HIGH
	stop_lost	A sequence variant where at least one base of the terminator codon (stop) is changed, resulting in an elongated transcript	SO:0001578	Stop lost	HIGH
	start_lost	A codon variant that changes at least one base of the canonical start codon	SO:0002012	Start lost	HIGH
	transcript_amplification	A feature amplification of a region containing a transcript	SO:0001889	Transcript amplification	HIGH
	inframe_insertion	An inframe non synonymous variant that inserts bases into in the coding sequence	SO:0001821	Inframe insertion	MODERATE
	inframe_deletion	An inframe non synonymous variant that deletes bases from the coding sequence	SO:0001822	Inframe deletion	MODERATE
	missense_variant	A sequence variant, that changes one or more bases, resulting in a different amino acid sequence but where the length is preserved	SO:0001583	Missense variant	MODERATE
	protein_altering_variant	A sequence_variant which is predicted to change the protein encoded in the coding sequence	SO:0001818	Protein altering variant	MODERATE
	splice_region_variant	A sequence variant in which a change has occurred within the region of the splice site, either within 1-3 bases of the exon or 3-8 bases of the intron	SO:0001630	Splice region variant	LOW
	incomplete_terminal_codon_variant	A sequence variant where at least one base of the final codon of an incompletely annotated transcript is changed	SO:0001626	Incomplete terminal codon variant	LOW
	start_retained_variant	A sequence variant where at least one base in the start codon is changed, but the start remains	SO:0002019	Start retained variant	LOW
	stop_retained_variant	A sequence variant where at least one base in the terminator codon is changed, but the terminator remains	SO:0001567	Stop retained variant	LOW

SO CLASSIFICATION

CONSEQUENCES:

High - the variant with high impact is assumed to have disruptive impact in the protein, probably causing protein truncation and loss of function

Moderate - non-disruptive but it potentially change protein effectiveness

Low is usually harmless or unlikely to change protein effectiveness

Modifier - incorporates non-coding variants or variants affecting non-coding genes. It is provided when there is no evidence of impact or when the impact is difficult to predict.

WEB INTERFACE

VARIANT EFFECT PREDICTOR

Variant Effect Predictor results

Job details

Summary statis

Category
Variants process
Variants filtered
Novel / existing v
Overlapped gene
Overlapped trans
Overlapped regu

Results previev

Navigation

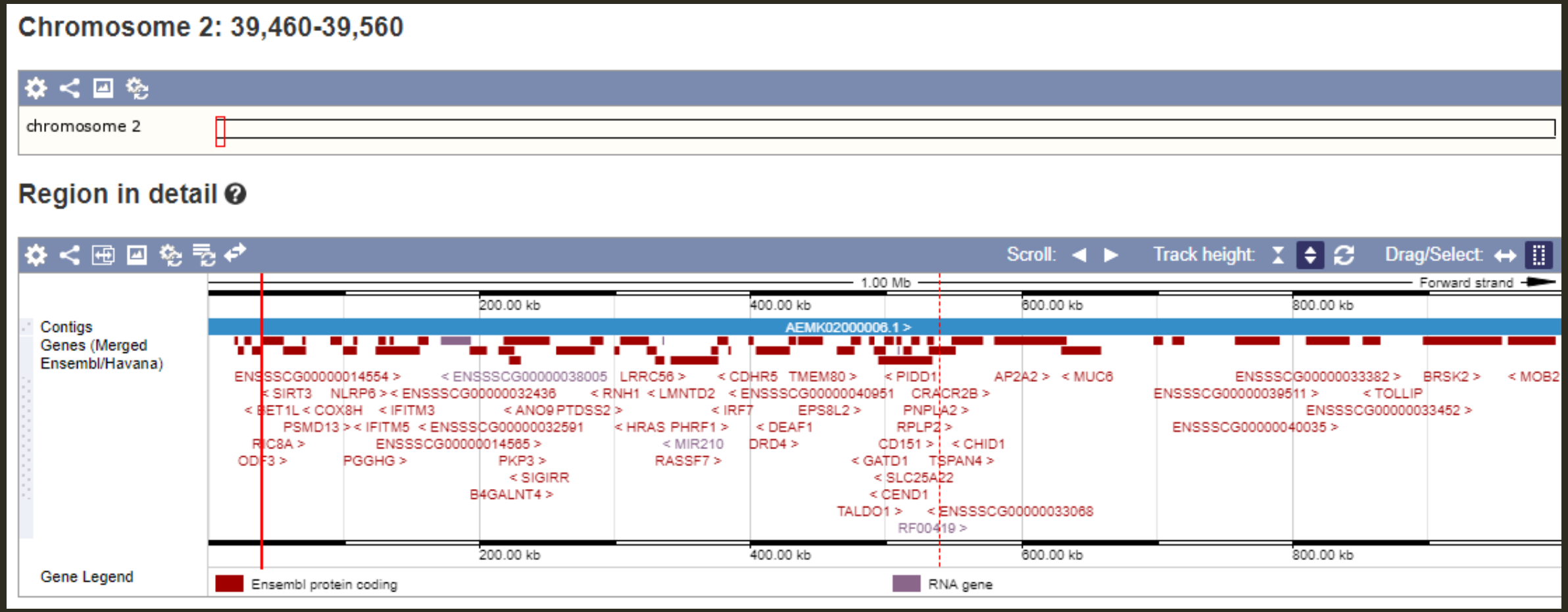
Page: 1 of

Show/hide columns (15 hidden)

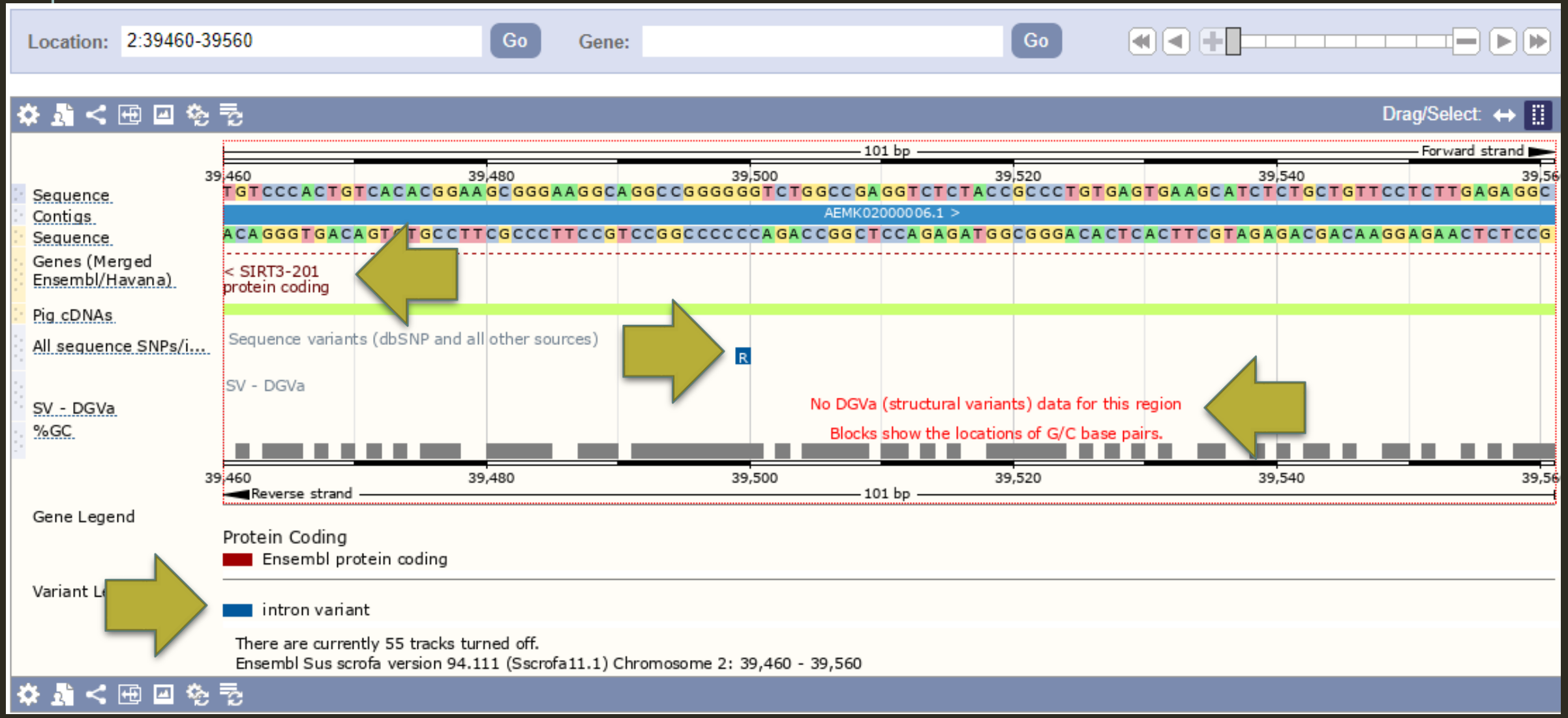
Uploaded variant	Location	Allele	Consequence	Impact	Symbol	Gene	Feature type	Feature	Biotype	Intron	Dista to trans
1_281897_G/A	1:281897-281897	A	intergenic_variant	MODIFIER	-	-	-	-	-	-	-
2_39510_G/A	2:39510-39510	A	downstream_gene_variant	MODIFIER	RIC8A	ENSSSCG00000014557	Transcript	ENSSSCT00000015907	protein_coding	-	1837
2_39510_G/A	2:39510-39510		intron_variant	MODIFIER	SIRT3	ENSSSCG00000014558	Transcript	ENSSSCT00000015908	protein_coding	6/6	-
2_39510_G/A	2:39510-39510		downstream_gene_variant	MODIFIER	RIC8A	ENSSSCG00000014557	Transcript	ENSSSCT00000041902	protein_coding	-	2471
6_179894_A/-	6:179893-179894	-	upstream_gene_variant	MODIFIER	-	ENSSSCG00000002639	Transcript	ENSSSCT00000002929	protein_coding	-	1432
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	MC1R	ENSSSCG00000020924	Transcript	ENSSSCT00000022534	protein_coding	-	1331
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	TCF25	ENSSSCG00000036102	Transcript	ENSSSCT00000050204	protein_coding	-	2504
6_179894_A/-	6:179893-179894	-	downstream_gene_variant	MODIFIER	TCF25	ENSSSCG00000036102	Transcript	ENSSSCT00000060256	protein_coding	-	2504

PRZEGLĄDARKA GENOMOWA

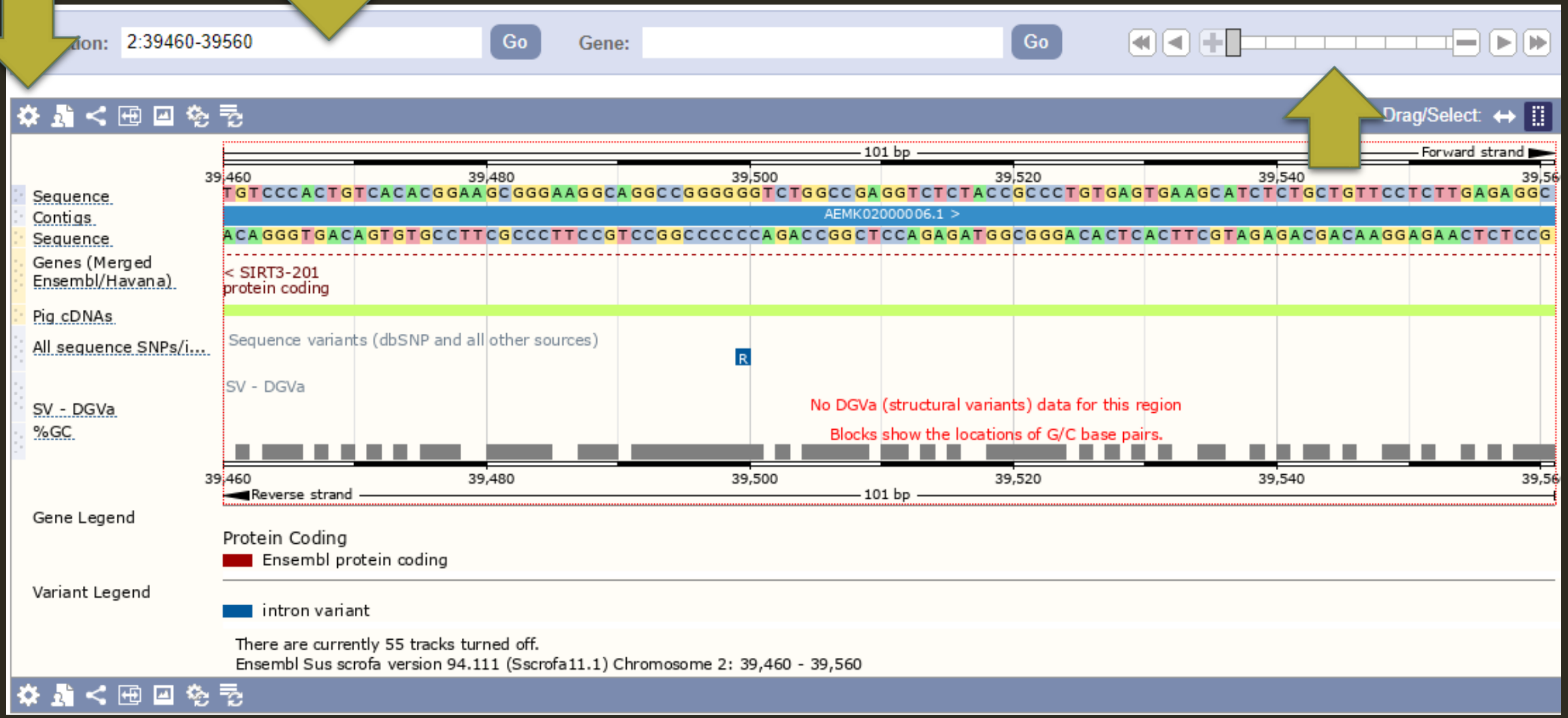
VARIANT EFFECT PREDICTOR



PRZEGLĄDARKA GENOMOWA

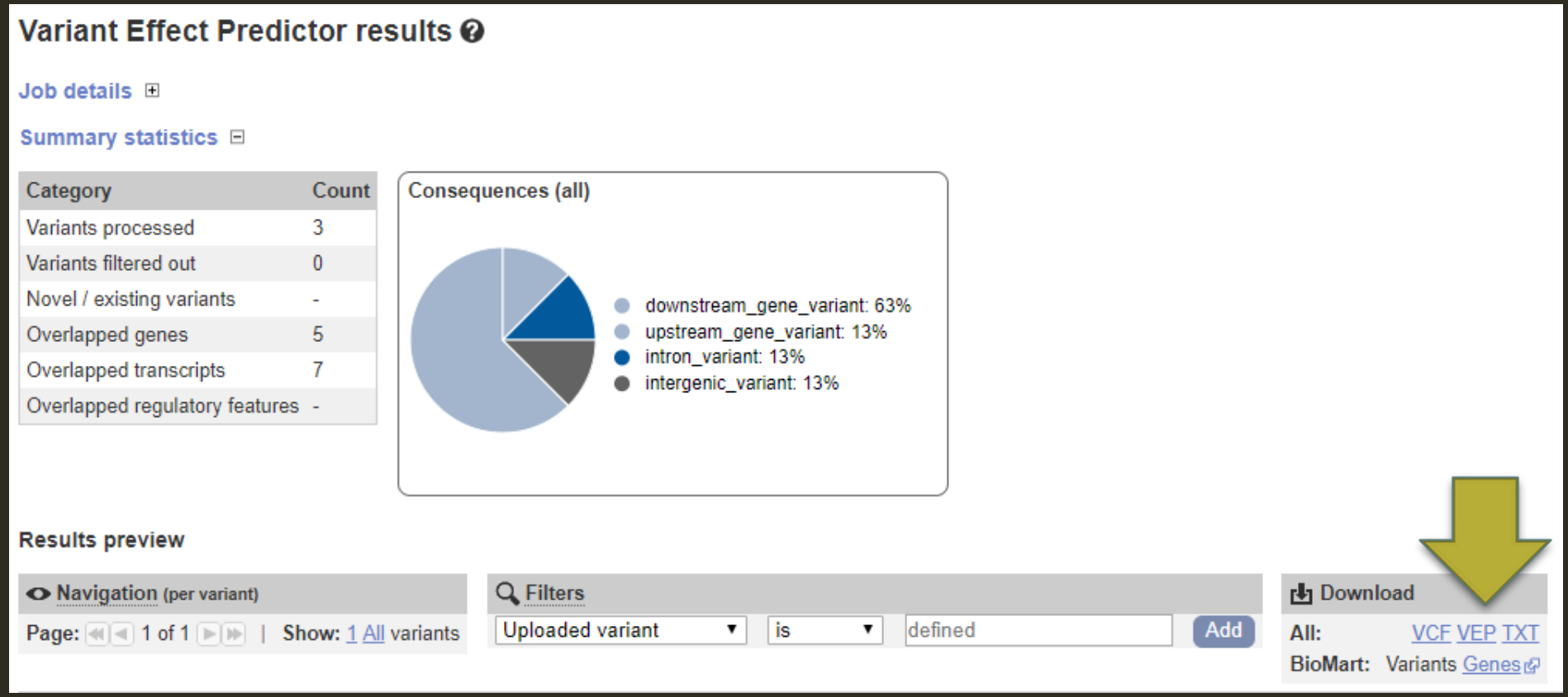


PRZEGLĄDARKA GENOMOWA



WEB INTERFACE

VARIANT EFFECT PREDICTOR



FORMAT VEP

```
#Uploaded_variation Location    Allele  Gene    Feature Feature_type    Consequence
cDNA_position   CDS_position    Protein_position    Amino_acids Codons  Existing_va
Extra
1_281897_G/A    1:281897-281897 A   -   -   -   intergenic_variant  -   -   -   -   -   -
IMPACT=MODIFIER
2_39510_G/A 2:39510-39510   A   ENSSSCG00000014557  ENSSSCT00000015907  Transcript
downstream_gene_variant -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=RIC8A;BIOTYPE=protein_coding;DISTANCE=1837;STRAND=1;SYMBOL_SOURCE=HGNC;H
GNC_ID=HGNC:29550;APPRIS=P1
2_39510_G/A 2:39510-39510   A   ENSSSCG00000014558  ENSSSCT00000015908  Transcript
intron_variant  -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=SIRT3;BIOTYPE=protein_coding;INTRON=6/6;STRAND=-1;SYMBOL_SOURCE=HGNC;HGN
C_ID=HGNC:14931;APPRIS=P1
2_39510_G/A 2:39510-39510   A   ENSSSCG00000014557  ENSSSCT00000041902  Transcript
downstream_gene_variant -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=RIC8A;BIOTYPE=protein_coding;DISTANCE=2471;STRAND=1;SYMBOL_SOURCE=HGNC;H
GNC_ID=HGNC:29550
6_179894_A/-    6:179893-179894 -   ENSSSCG00000002639  ENSSSCT00000002929  Transcript
upstream_gene_variant   -   -   -   -   -   -
IMPACT=MODIFIER;BIOTYPE=protein_coding;DISTANCE=1432;STRAND=-1;APPRIS=P1
6_179894_A/-    6:179893-179894 -   ENSSSCG00000020924  ENSSSCT00000022534  Transcript
downstream_gene_variant -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=MC1R;BIOTYPE=protein_coding;DISTANCE=1331;STRAND=-1;SYMBOL_SOURCE=Entrez
Gene;APPRIS=P1
6_179894_A/-    6:179893-179894 -   ENSSSCG00000036102  ENSSSCT00000050204  Transcript
downstream_gene_variant -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=TCF25;BIOTYPE=protein_coding;DISTANCE=2504;STRAND=-1;SYMBOL_SOURCE=HGNC;
HGNC_ID=HGNC:29181;APPRIS=A2
6_179894_A/-    6:179893-179894 -   ENSSSCG00000036102  ENSSSCT00000060256  Transcript
downstream_gene_variant -   -   -   -   -   -
IMPACT=MODIFIER;SYMBOL=TCF25;BIOTYPE=protein_coding;DISTANCE=2504;STRAND=-1;SYMBOL_SOURCE=HGNC;
HGNC_ID=HGNC:29181;APPRIS=P5
```

COMMAND LINE

VARIANT EFFECT PREDICTOR

- To run VEP using your own species and assembly, please use a --fasta file and --gff or --gtf annotation.

To run VEP with default options, use the following command:

```
./vep --cache -i input.txt -o output.txt
```

where `input.txt` contains data in one of the compatible input formats and `output.txt` is the output file to be created.

Options can be passed as the full string (e.g. --format), or as the shortest unique string among the options (e.g. --form for --format, since there is another option --force_overwrite).

You may use one or two hypen ("-") characters before each option name; --**cache** or -**cache**.

VEP options can also be read from:

- **Configuration files** using --config. Options set in configuration files are overriden if specified on the command line.
- **Environment variables** that start with prefix `VEP_`. For instance, you can set the cache flag with `export VEP_CACHE=1` and the input flag with `export VEP_INPUT="/path/to/input.txt"` before running `./vep`. configuration files or on the command line.

Running VEP on non-vertebrate species

To use VEP on non-vertebrate species, you need to use extra options:

- **Database** access: add the option --genomes to point to the correct database server.

```
./vep -i input.txt -o output.txt --species triticum_aestivum --database --genomes
```

Some of the non-vetebrate species databases (mainly bacteria and protists) are composed of a collection of species. For these databases the flag "--is_multispecies 1" is required.

```
./vep -i input.txt -o output.txt --species protists_euglenozoa1 --database --genomes --is_multispecies 1
```

- **Cache** access: use the option "--cache_version *eg_version*" where *eg_version* is the Ensembl genomes version number, which differs from the Ensembl vertebrates/VEP version numbers (111).

```
./vep -i input.txt -o output.txt --species triticum_aestivum --cache --cache_version 42
```

More information about VEP cache is available here.

SNPEFF

SNPEFF & SNPSIFT

Home

Home

- SnpEff
- SnpSift
- Citing SnpEff & SnpSift
  - Citing SnpEff
  - Citing SnpSift
- Microsoft Genomics
- Galaxy & GATK
- In memory of Dr. Xiangyi Lu

Usage examples

License

SnpEff & SnpSift

Genomic variant annotations, and functional effect prediction toolbox.

Download | **SnpEff Documentation** | **SnpSift Documentation**

Latest version 5.2c (2024-04-09)

SnpEff

Genetic variant annotation, and functional effect prediction toolbox. It annotates and predicts the effects of genetic variants on genes and proteins (such as amino acid changes).

Features:

- Supports over **38,000 genomes**.

SNPEFF & SNPSIFT

```
java -Xmx8g -jar snpEff.jar GRCh37.75 examples/test.chr22.vcf > test.chr22.ann.vcf

# Here is how the output looks like
$ head examples/test.chr22.ann.vcf
##SnpEffVersion="4.1 (build 2015-01-07), by Pablo Cingolani"
##SnpEffCmd="SnpEff  GRCh37.75 examples/test.chr22.vcf "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Ge
##INFO=<ID=NMD,Number=.,Type=String,Description="Predicted nonsense mediated decay effects for this variant. Form
#CHROM  POS ID  REF ALT QUAL    FILTER  INFO
22  17071756    .   T   C   .   .   ANN=C|3_prime_UTR_variant|MODIFIER|CCT8L2|ENSG00000198445|transcript|ENST0000
22  17072035    .   C   T   .   .   ANN=T|missense_variant|MODERATE|CCT8L2|ENSG00000198445|transcript|ENST0000035
22  17072258    .   C   A   .   .   ANN=A|missense_variant|MODERATE|CCT8L2|ENSG00000198445|transcript|ENST0000035
22  17072674    .   G   A   .   .   ANN=A|missense_variant|MODERATE|CCT8L2|ENSG00000198445|transcript|ENST0000035
```

SNPEFF INPUT

```
##fileformat=VCFv4.1
##fileDate=20191102
##source=ensembl;version=99;url=https://e99.ensembl.org/bos_taurus
##reference=ftp://ftp.ensembl.org/pub/release-99/fasta/bos_taurus/dna/
##INFO=<ID=dbSNP_150,Number=0,Type=Flag,Description="Variants (including SNPs and indels) imported from dbSNP [rem
##INFO=<ID=TSA,Number=1,Type=String,Description="Type of sequence alteration. Child of term sequence_alteration as
##INFO=<ID=E_Cited,Number=0,Type=Flag,Description="Cited.https://www.ensembl.org/info/genome/variation/prediction/
##INFO=<ID=E_Multiple_observations,Number=0,Type=Flag,Description="Multiple_observations.https://www.ensembl.org/i
ity.html#evidence_status">
##INFO=<ID=E_Freq,Number=0,Type=Flag,Description="Frequency.https://www.ensembl.org/info/genome/variation/predicti
##INFO=<ID=E_TOPMed,Number=0,Type=Flag,Description="TOPMed.https://www.ensembl.org/info/genome/variation/predictio
##INFO=<ID=E_Hapmap,Number=0,Type=Flag,Description="HapMap.https://www.ensembl.org/info/genome/variation/predictio
##INFO=<ID=E_Phenotype_or_Disease,Number=0,Type=Flag,Description="Phenotype_or_Disease.https://www.ensembl.org/inf
y.html#evidence_status">
##INFO=<ID=E_ESP,Number=0,Type=Flag,Description="ESP.https://www.ensembl.org/info/genome/variation/prediction/vari
##INFO=<ID=E_gnomAD,Number=0,Type=Flag,Description="gnomAD.https://www.ensembl.org/info/genome/variation/predictio
##INFO=<ID=E_1000G,Number=0,Type=Flag,Description="1000Genomes.https://www.ensembl.org/info/genome/variation/predi
##INFO=<ID=E_ExAC,Number=0,Type=Flag,Description="ExAC.https://www.ensembl.org/info/genome/variation/prediction/va
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO
1       207381  rs719849645     C       T       .       .       dbSNP_150;TSA=SNV
1       207419  rs517357290     G       A       .       .       dbSNP_150;TSA=SNV
1       207431  rs525766619     AG      A       .       .       dbSNP_150;TSA=deletion;E_Multiple_observations
1       207453  rs449980494     A       T       .       .       dbSNP_150;TSA=SNV
1       207567  rs438394308     G       C       .       .       dbSNP_150;TSA=SNV;E_Multiple_observations
1       207568  rs459198567     C       A       .       .       dbSNP_150;TSA=SNV
```

SNPEFF COMMAND LINE

Databases

```
(base) mielczarek@VERSUS:~/snpEff_annot> java -jar /home/mielczarek/anaconda2/share/snpeff-4.3.1t-3/snpEff.jar databases |
 grep Bos_taurus
UMD3.1.75                                                       Bos_taurus
                                                       http://downloads.sourceforge.net/project/snpeff/databases/v4_3/snp
Eff_v4_3_UMD3.1.75.zip
UMD3.1.86                                                       Bos_taurus
        OK                                             http://downloads.sourceforge.net/project/snpeff/databases/v4_3/snp
Eff_v4_3_UMD3.1.86.zip
```

SNPEFF COMMAND LINE

```
(base) mielc
 grep Bos_ta
UMD3.1.75

Eff_v4_3_UMD
UMD3.1.86
        OK
Eff_v4_3_UMD
```

Building databases

SnpEff needs a database to perform genomic annotations. There are pre-built databases for thousands of genomes, so chances are that your organism of choice already has a SnpEff database available. In the (unlikely?) event that you need to build one yourself, here we describe how to it.

Info

You can know which genomes are supported by running the following command:

```
java -jar snpEff.jar databases
```

Warning

Most people do NOT need to build a database, and can safely use a pre-built one. So unless you are working with a rare genome you most likely don't need to do it either.

SNPEFF OUTPUT

```
##fileformat=VCFv4.1
##fileDate=20191102
##source=ensembl;version=99;url=https://e99.ensembl.org/bos_taurus
##reference=ftp://ftp.ensembl.org/pub/release-99/fasta/bos_taurus/dna/
##INFO=<ID=dbSNP_150,Number=0,Type=Flag,Description="Variants (including SNPs and indels) imported from dbSNP [remapped to ARS-UCD1.2]">
##INFO=<ID=TSA,Number=1,Type=String,Description="Type of sequence alteration. Child of term sequence_alteration as defined by the sequence ontology project.">
##INFO=<ID=E_Cited,Number=0,Type=Flag,Description="Cited.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_Multiple_observations,Number=0,Type=Flag,Description="Multiple_observations.https://www.ensembl.org/info/genome/variation/prediction/variant_qual
ity.html#evidence_status">
##INFO=<ID=E_Freq,Number=0,Type=Flag,Description="Frequency.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_TOPMed,Number=0,Type=Flag,Description="TOPMed.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_Hapmap,Number=0,Type=Flag,Description="HapMap.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_Phenotype_or_Disease,Number=0,Type=Flag,Description="Phenotype_or_Disease.https://www.ensembl.org/info/genome/variation/prediction/variant_qualit
y.html#evidence_status">
##INFO=<ID=E_ESP,Number=0,Type=Flag,Description="ESP.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_gnomAD,Number=0,Type=Flag,Description="gnomAD.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_1000G,Number=0,Type=Flag,Description="1000Genomes.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##INFO=<ID=E_ExAC,Number=0,Type=Flag,Description="ExAC.https://www.ensembl.org/info/genome/variation/prediction/variant_quality.html#evidence_status">
##SnpEffVersion="4.3t (build 2017-11-24 10:18), by Pablo Cingolani"
##SnpEffCmd="SnpEff  UMD3.1.86 bos_taurus.vcf.gz "
##INFO=<ID=ANN,Number=.,Type=String,Description="Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_ID | Feature_Type | Featu
re_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO
' ">
##INFO=<ID=LOF,Number=.,Type=String,Description="Predicted loss of function effects for this variant. Format: 'Gene_Name | Gene_ID | Number_of_transcripts_in_
gene | Percent_of_transcripts_affected'">
##INFO=<ID=NMD,Number=.,Type=String,Description="P                                                                                                    ri
pts_in_gene | Percent_of_transcripts_affected'">
#CHROM  POS     ID      REF     ALT     QUAL    FI
1       207381  rs719849645     C       T       .
ENSBTAG00000005540|intergenic_region|ENSBTAG000000                                                                                                    3-
1       207419  rs517357290     G       A       .
ENSBTAG00000005540|intergenic_region|ENSBTAG000000                                                                                                    3-
```

```
INFO
.       dbSNP_150;TSA=SNV;ANN=T|intergenic_region|MODIFIE
ENSBTAG00000005540|||n.207381C>T||||||
.       dbSNP_150;TSA=SNV;ANN=A|intergenic_region|MODIFIE
ENSBTAG00000005540|||n.207419G>A||||||
```

ENRICHMENT ANALYSIS

ENRICHMENT ANALYSIS

One of the main uses of the GO is to perform enrichment analysis on gene sets.

For example, given a set of genes that are up-regulated under certain conditions, an enrichment analysis will find which GO terms are over-represented (or underrepresented) using annotations for that gene set.

GO ENRICHMENT ANALYSIS

Gene Ontology

GENE ONTOLOGY

- GO → Gene Ontology
- 1998
- Hierarchia ontologii
- Różne gatunki

http://geneontology.org/docs/ontology-documentation/

GENE ONTOLOGY

- Usystematyzowane ontologie kategoryzują produkty genów w zakresie
  - procesów biologicznych
  - komponentów komórki
  - funkcji molekularnych

Molecular Function	Molecular-level activities performed by gene products. Molecular function terms describe activities that occur at the molecular level, such as “catalysis” or “transport”. GO molecular function terms represent activities rather than the entities (molecules or complexes) that perform the actions, and do not specify where, when, or in what context the action takes place. Molecular functions generally correspond to activities that can be performed by individual gene products (*i.e.* a protein or RNA), but some activities are performed by molecular complexes composed of multiple gene products. Examples of broad functional terms are *catalytic activity* and *transporter activity*; examples of narrower functional terms are *adenylate cyclase activity* or *Toll-like receptor binding*. To avoid confusion between gene product names and their molecular functions, GO molecular functions are often appended with the word “activity” (a *protein kinase* would have the GO molecular function *protein kinase activity*).
Cellular Component	The locations relative to cellular structures in which a gene product performs a function, either cellular compartments (*e.g.*, *mitochondrion*), or stable macromolecular complexes of which they are parts (*e.g.*, the *ribosome*). Unlike the other aspects of GO, cellular component classes refer not to processes but rather a cellular anatomy.
Biological Process	The larger processes, or ‘biological programs’ accomplished by multiple molecular activities. Examples of broad biological process terms are *DNA repair* or *signal transduction*. Examples of more specific terms are *pyrimidine nucleobase biosynthetic process* or *glucose transmembrane transport*. Note that a biological process is not equivalent to a pathway. At present, the GO does not try to represent the dynamics or dependencies that would be required to fully describe a pathway.

GO ONTOLOGY

Funkcja molekularna	Komponent komórki	Proces biologiczny
seria zdarzeń realizowanych przez jeden lub więcej połączonych funkcji molekularnych	komponent komórki	czynności występujące na poziomie molekularnym
np. *signal transduction, pyrimidine metabolic process*	np. *nucleus, ribosome*	np. *catalytic activity, binding activity*

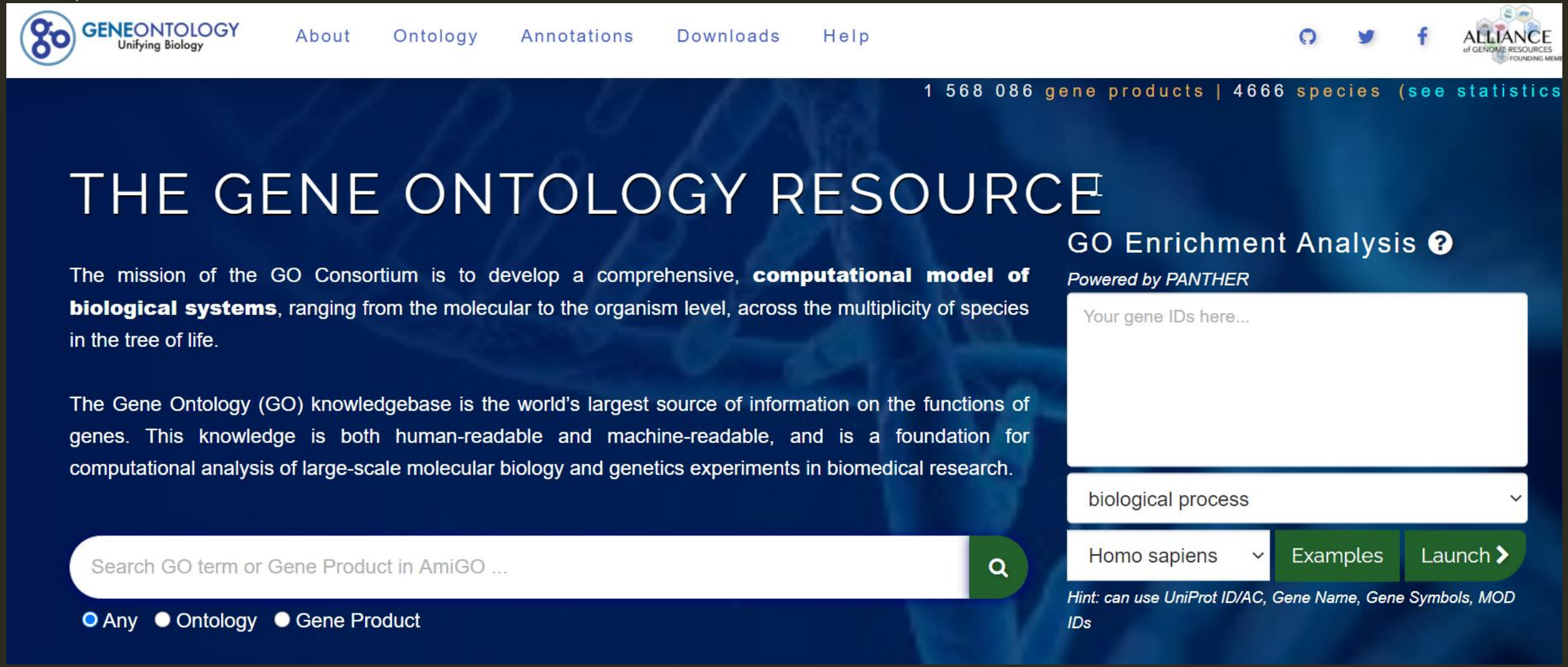
Funkcjonalność genu/białka jest definiowana w sposób standardowy i jednoznaczny.

GO ONTOLOGY

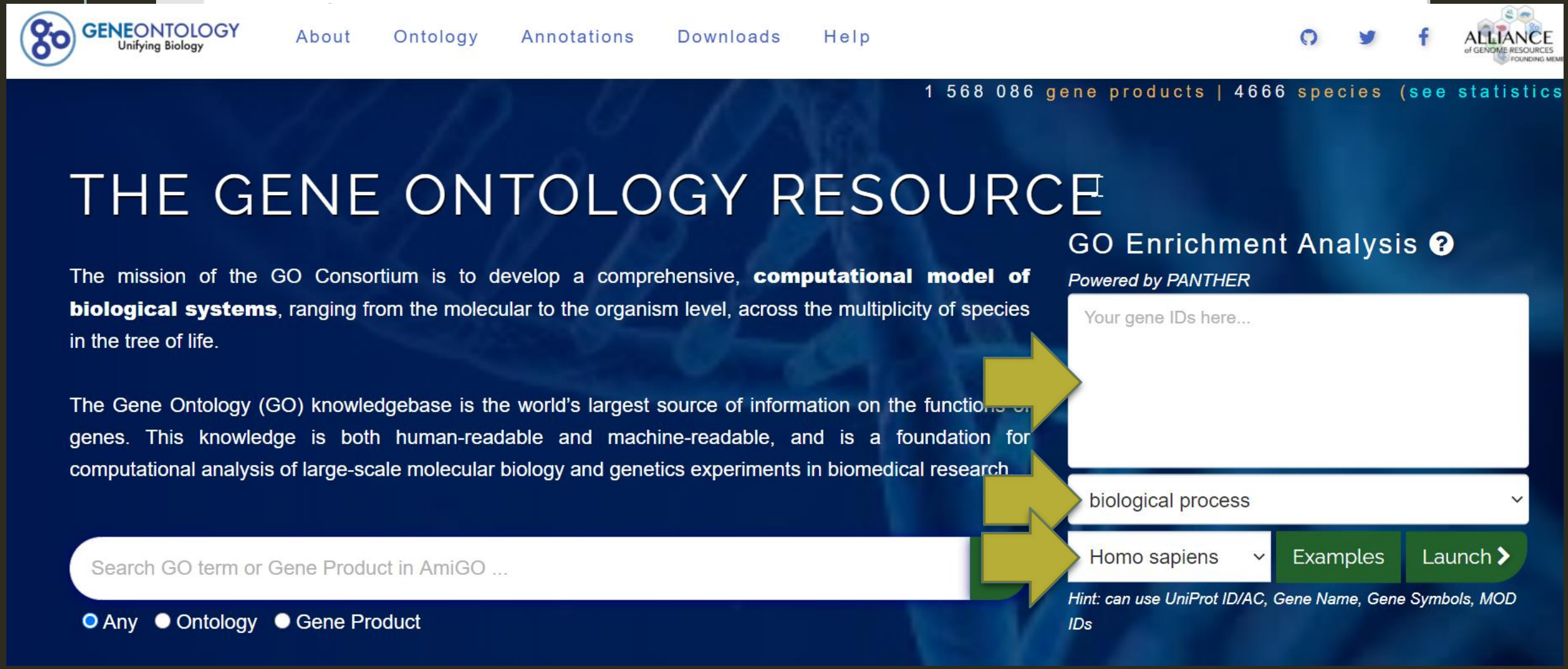
Funkcja molekularna	Komponent komórki	Funkcja molekularna
• Growth • Bone growth • Endochondral bone growth • Growth plate cartilage developement	• Chromosomal part • Chromatin • Euchromatin • Nuclear euchromatin	• Binding • ATP binding • MutS complex

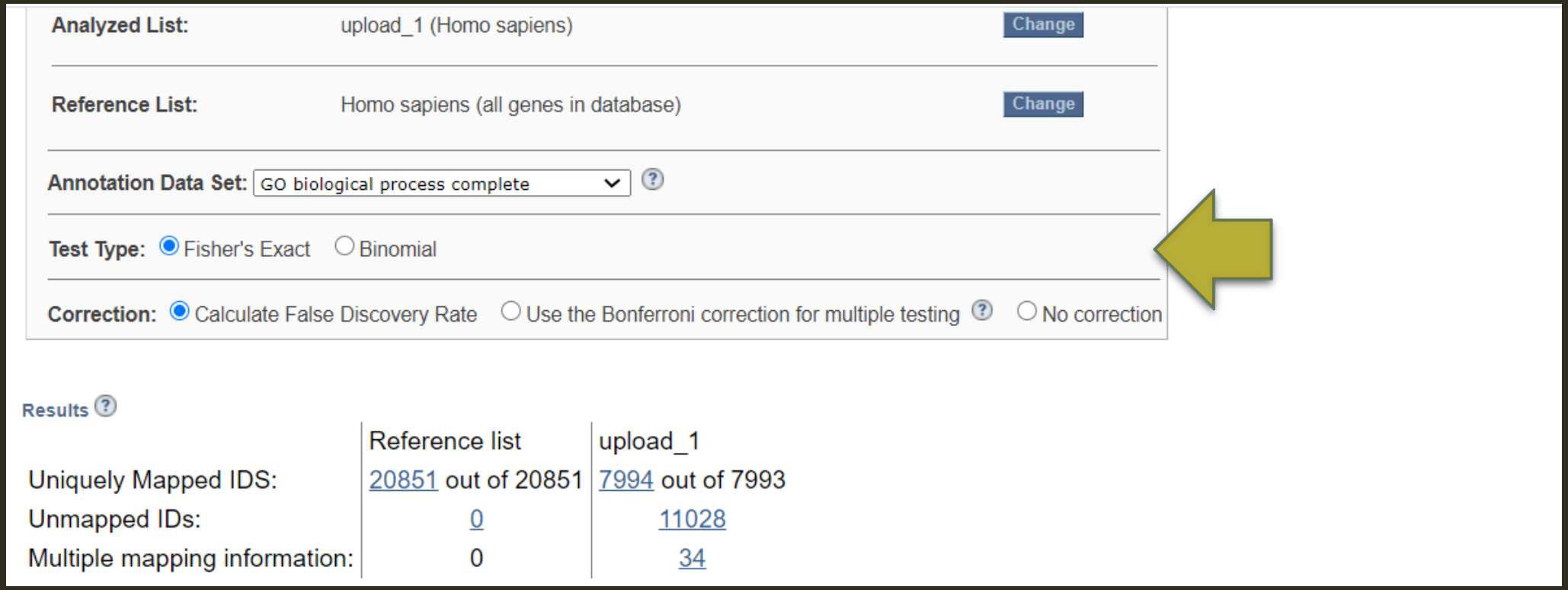
Hierarchia ontologii – hierarchia opisów o coraz większej dokładności. Szczyt hierarchii i dostarcza ogólnego obrazu klasy funkcjonalnej, natomiast niższe poziomy określają funkcję dokładniej.

GENE ONTOLOGY



```
1  #Uploaded_variation Location    Allele  Gene    Feature Feature_type    Consequence
   cDNA_position   CDS_position    Protein_position    Amino_acids Codons  Existing_variation
   Extra
2  1_281897_G/A    1:281897-281897 A   -   -   -   intergenic_variant  -   -   -   -   -   -
   IMPACT=MODIFIER
```



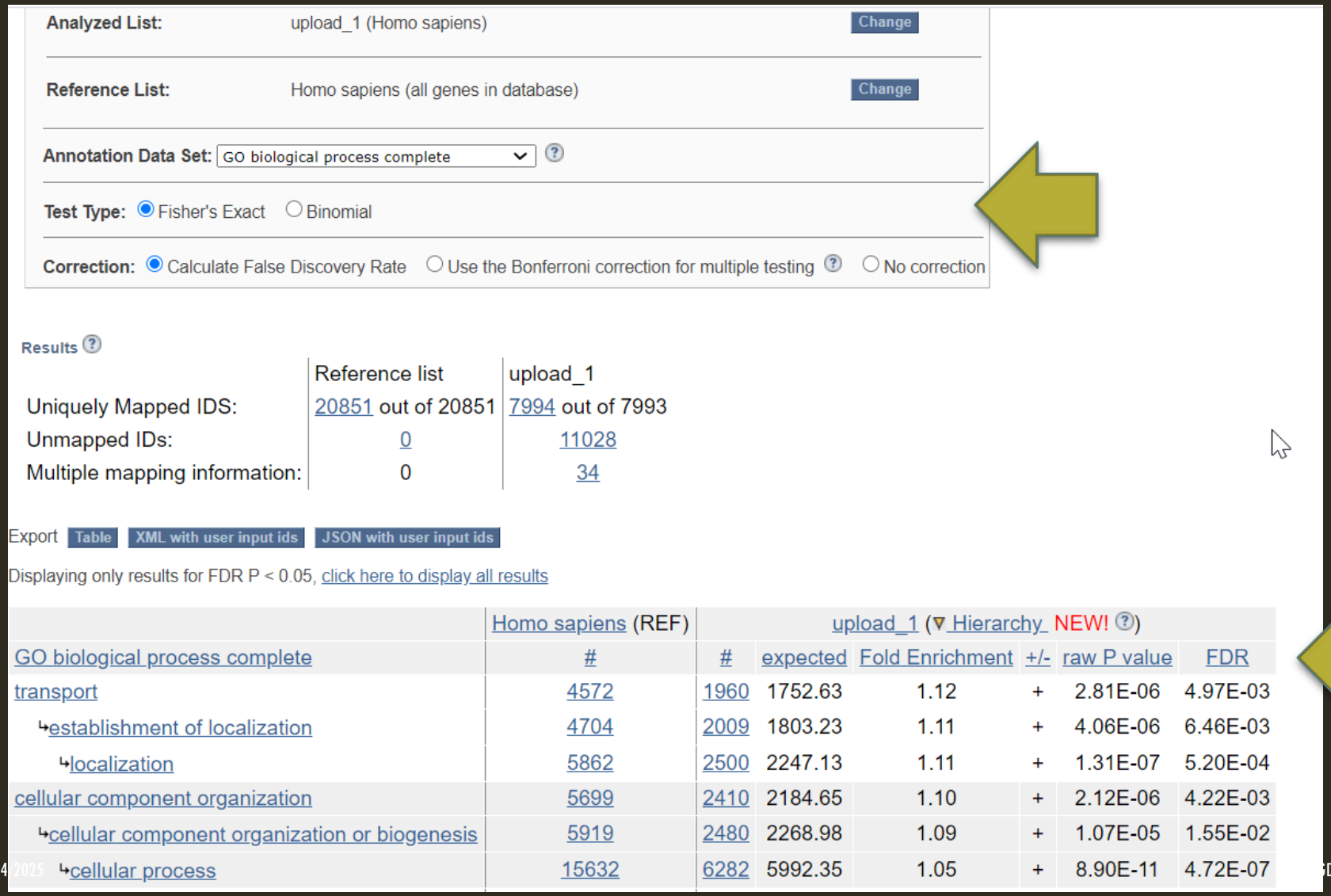


H_0: Ontologia genu nie jest częściej(rzadziej) reprezentowana wśród genów odpowiedzialnych za daną cechę

H_1: Ontologia genu jest częściej(rzadziej) reprezentowana wśród genów odpowiedzialnych za daną cechę

→ sprawdzenie, czy pewne kategorie (np. ścieżki biologiczne, funkcje genów) są nadreprezentowane w danym zbiorze danych w porównaniu do tego, czego byśmy się spodziewali przez przypadek.

GO ONTOLOGY: PROCESY BIOLOGICZNE



Analyzed List: upload_1 (Homo sapiens) Change

Reference List: Homo sapiens (all genes in database) Change

Annotation Data Set: GO biological process complete

Test Type: Fisher's Exact / Binomial

Correction: Calculate False Discovery Rate / Use the Bonferroni correction for multiple testing / No correction

Results

	Reference list	upload_1
Uniquely Mapped IDS:	20851 out of 20851	7994 out of 7993
Unmapped IDs:	0	11028
Multiple mapping information:	0	34

Export Table XML with user input ids JSON with user input ids

Displaying only results for FDR P < 0.05, click here to display all results

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW!)					
GO biological process complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
transport	4572	1960	1752.63	1.12	+	2.81E-06	4.97E-03
↳establishment of localization	4704	2009	1803.23	1.11	+	4.06E-06	6.46E-03
↳localization	5862	2500	2247.13	1.11	+	1.31E-07	5.20E-04
cellular component organization	5699	2410	2184.65	1.10	+	2.12E-06	4.22E-03
↳cellular component organization or biogenesis	5919	2480	2268.98	1.09	+	1.07E-05	1.55E-02
↳cellular process	15632	6282	5992.35	1.05	+	8.90E-11	4.72E-07

GO ONTOLOGY: FUNKCJA MOLEKULARNA

	Homo sapiens (REF)	upload_1 (▽ Hierarchy NEW! ?)					
GO molecular function complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
metal ion transmembrane transporter activity	433	224	165.99	1.35	+	2.93E-04	4.82E-02
↳inorganic molecular entity transmembrane transporter activity	834	399	319.70	1.25	+	2.36E-04	4.01E-02
protein kinase activity	585	294	224.25	1.31	+	1.49E-04	2.63E-02
↳catalytic activity, acting on a protein	2274	1011	871.71	1.16	+	3.75E-05	7.76E-03
↳catalytic activity	5866	2515	2248.67	1.12	+	2.86E-08	1.70E-05
↳kinase activity	772	381	295.94	1.29	+	4.77E-05	9.08E-03
↳transferase activity	2353	1037	902.00	1.15	+	8.03E-05	1.47E-02
↳phosphotransferase activity, alcohol group as acceptor	691	346	264.89	1.31	+	4.74E-05	9.39E-03
cytoskeletal protein binding	999	497	382.96	1.30	+	1.50E-06	5.97E-04
↳protein binding	14135	5703	5418.50	1.05	+	4.53E-09	3.08E-06
↳binding	16492	6632	6322.03	1.05	+	8.51E-14	1.01E-10

GO ONTOLOGY: KOMPONENT KOMÓRKOWY

	Homo sapiens (REF)	upload_1 (▼ Hierarchy NEW! ?)					
GO cellular component complete	#	#	expected	Fold Enrichment	+/-	raw P value	FDR
axoneme	127	79	48.68	1.62	+	9.79E-04	4.78E-02
↳cytoskeleton	2309	1053	885.13	1.19	+	8.57E-07	1.14E-04
↳intracellular organelle	12896	5201	4943.54	1.05	+	3.87E-07	5.97E-05
↳intracellular	14726	5944	5645.05	1.05	+	2.34E-10	7.82E-08
↳organelle	13901	5638	5328.79	1.06	+	2.87E-10	8.20E-08
↳cellular anatomical entity	18790	7456	7202.94	1.04	+	7.04E-18	1.41E-14
↳ciliary plasm	129	81	49.45	1.64	+	6.30E-04	3.71E-02
↳plasma membrane bounded cell projection cytoplasm	216	125	82.80	1.51	+	3.15E-04	2.25E-02
↳cytoplasmic region	258	152	98.90	1.54	+	3.66E-05	3.66E-03
↳cytoplasm	11740	4861	4500.40	1.08	+	3.69E-12	1.85E-09
↳plasma membrane bounded cell projection	2269	1053	869.80	1.21	+	7.37E-08	1.23E-05
↳cell projection	2367	1096	907.36	1.21	+	5.10E-08	9.29E-06
adherens junction	174	101	66.70	1.51	+	1.12E-03	4.86E-02
↳anchoring junction	839	397	321.62	1.23	+	5.05E-04	3.26E-02
↳cell junction	2066	986	791.98	1.24	+	3.47E-09	7.71E-07

http://geneontology.org/docs/go-enrichment-analysis/

GO ONTOLOGY: INTERPRETACJA WYNIKÓW

Interpreting the results table

The results page displays a table that lists **significant shared GO terms (or parents of GO terms)** used to describe the set of genes that users entered on the previous page, the **background frequency**, the **sample frequency**, **expected p-value**, an indication of **over/underrepresentation** for each term, and **p-value**. In addition, the results page displays all the criteria used in the analysis. Any unresolved gene names will be listed on top of the table.

Background frequency and sample frequency

Background frequency is the number of genes annotated to a GO term in the entire background set, while *sample frequency* is the number of genes annotated to that GO term in the input list. For example, if the input list contains 10 genes and the enrichment is done for biological process in S. cerevisiae whose background set contains 6442 genes, then if 5 out of the 10 input genes are annotated to the GO term: DNA repair, then the sample frequency for DNA repair will be 5/10. Whereas if there are 100 genes annotated to DNA repair in all of the S. cerevisiae genome, then the background frequency will be 100/6442.

Overrepresented or underrepresented

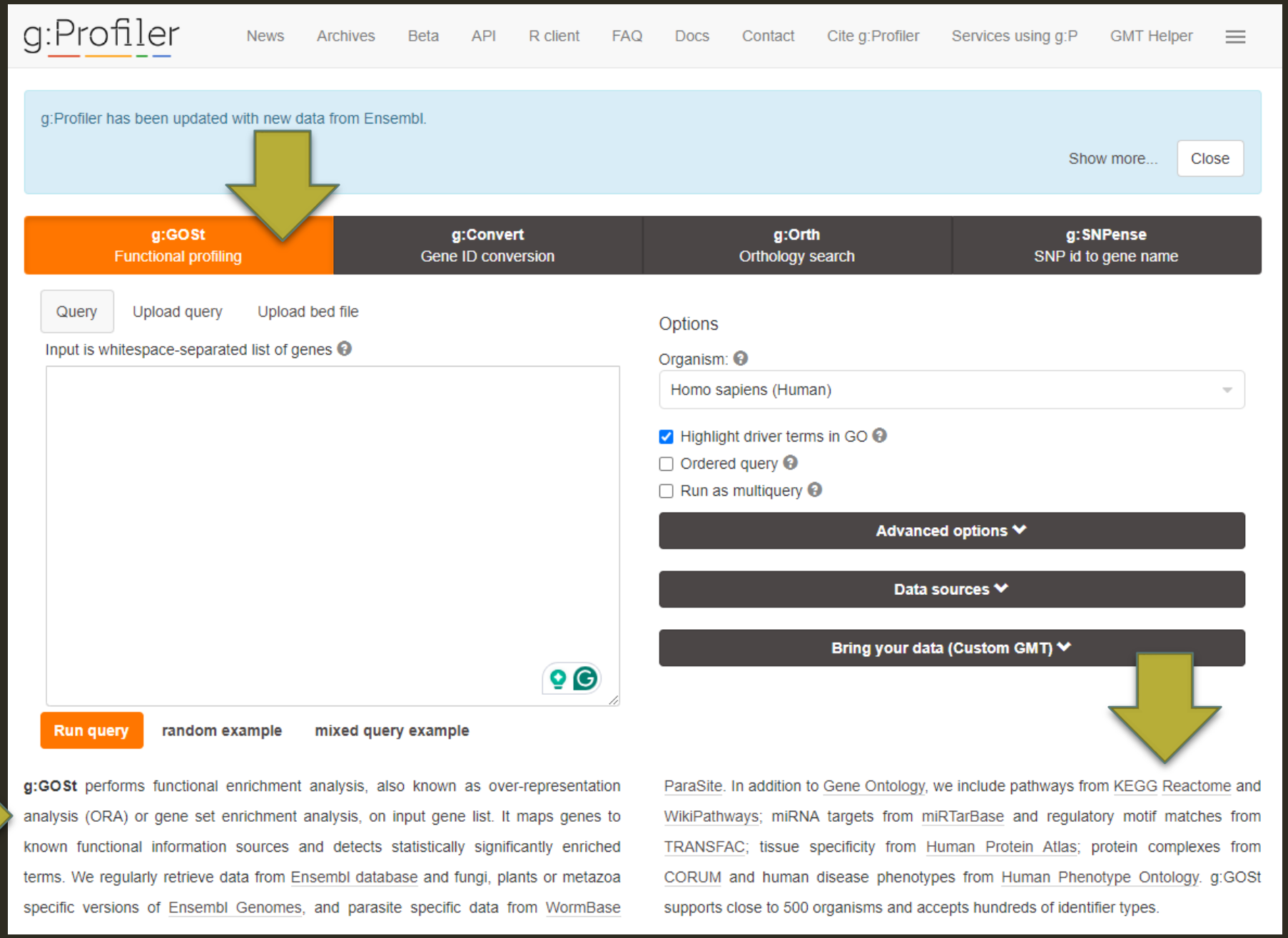
The symbols + and - indicate over or underrepresentation of a term.

ENRICHMENT ANALYSIS GPROFILER

Gene Ontology, KEGG, Reactome
… i inne

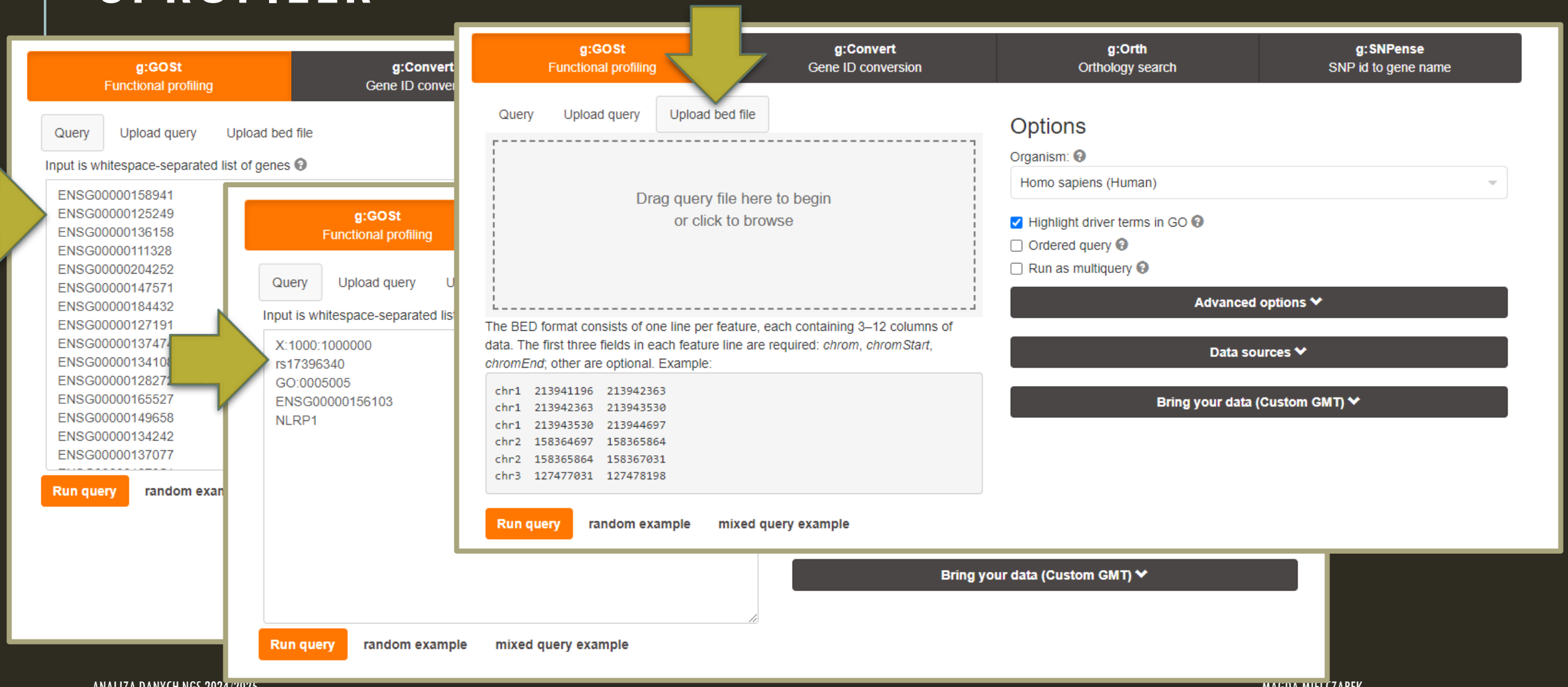
https://biit.cs.ut.ee/gprofiler/gost

GPROFILER G:GOST



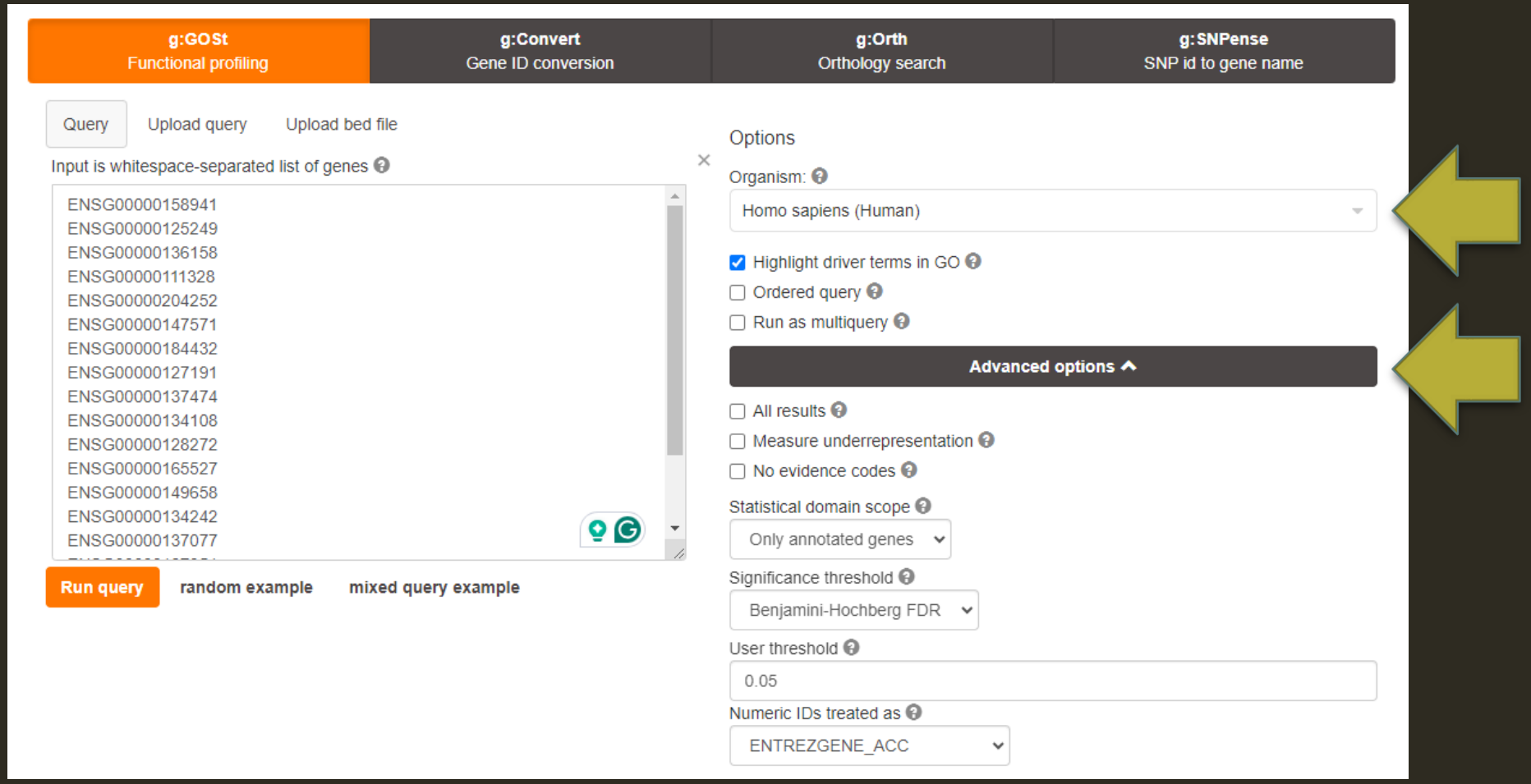
https://biit.cs.ut.ee/gprofiler/gost

GPROFILER



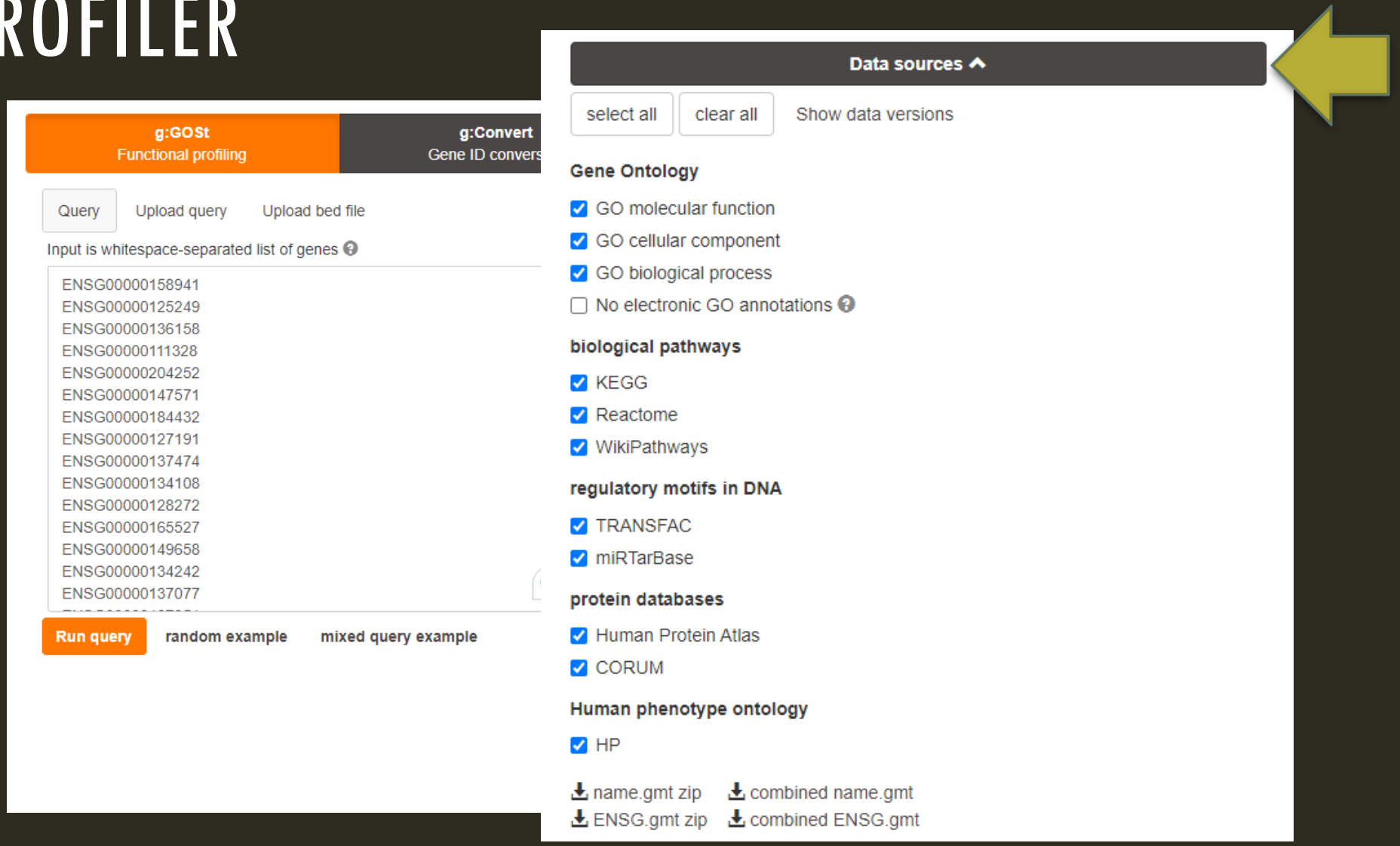
https://biit.cs.ut.ee/gprofiler/gost

GPROFILER

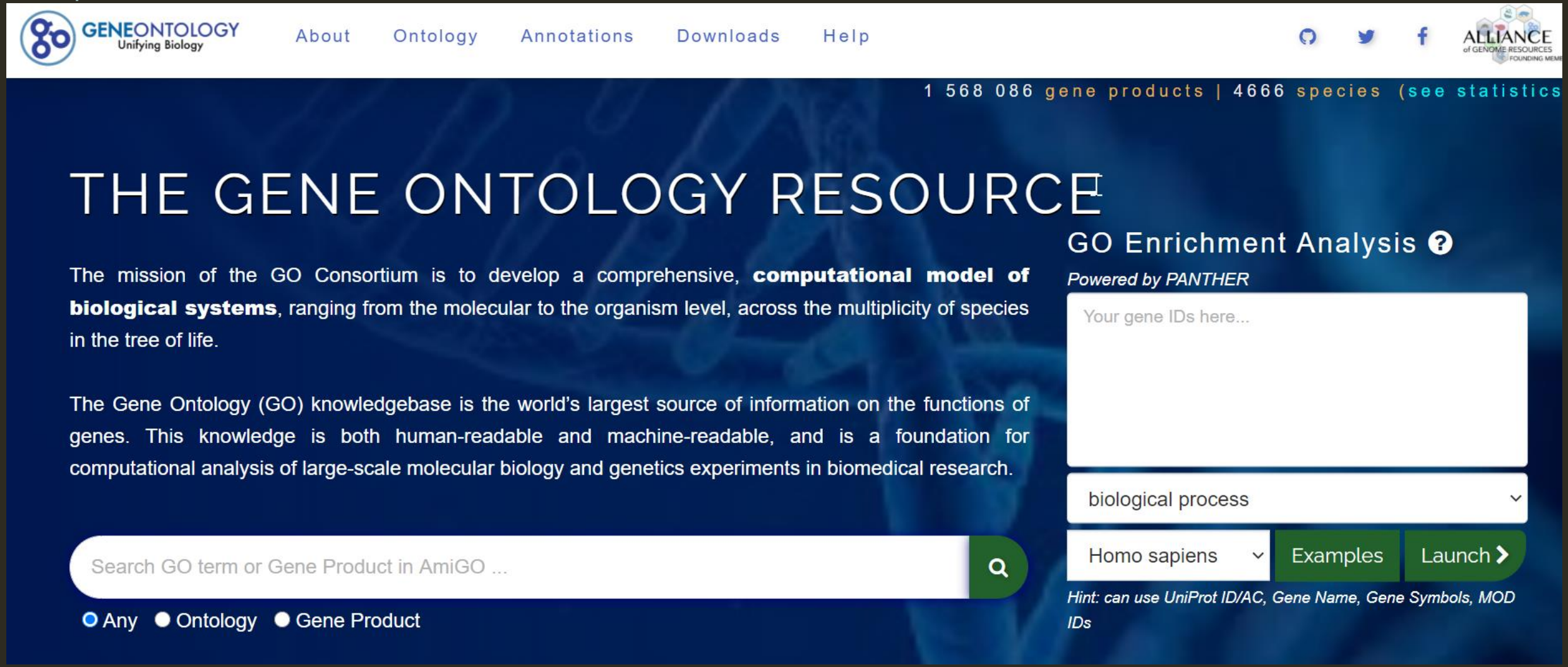


https://biit.cs.ut.ee/gprofiler/gost

GPROFILER



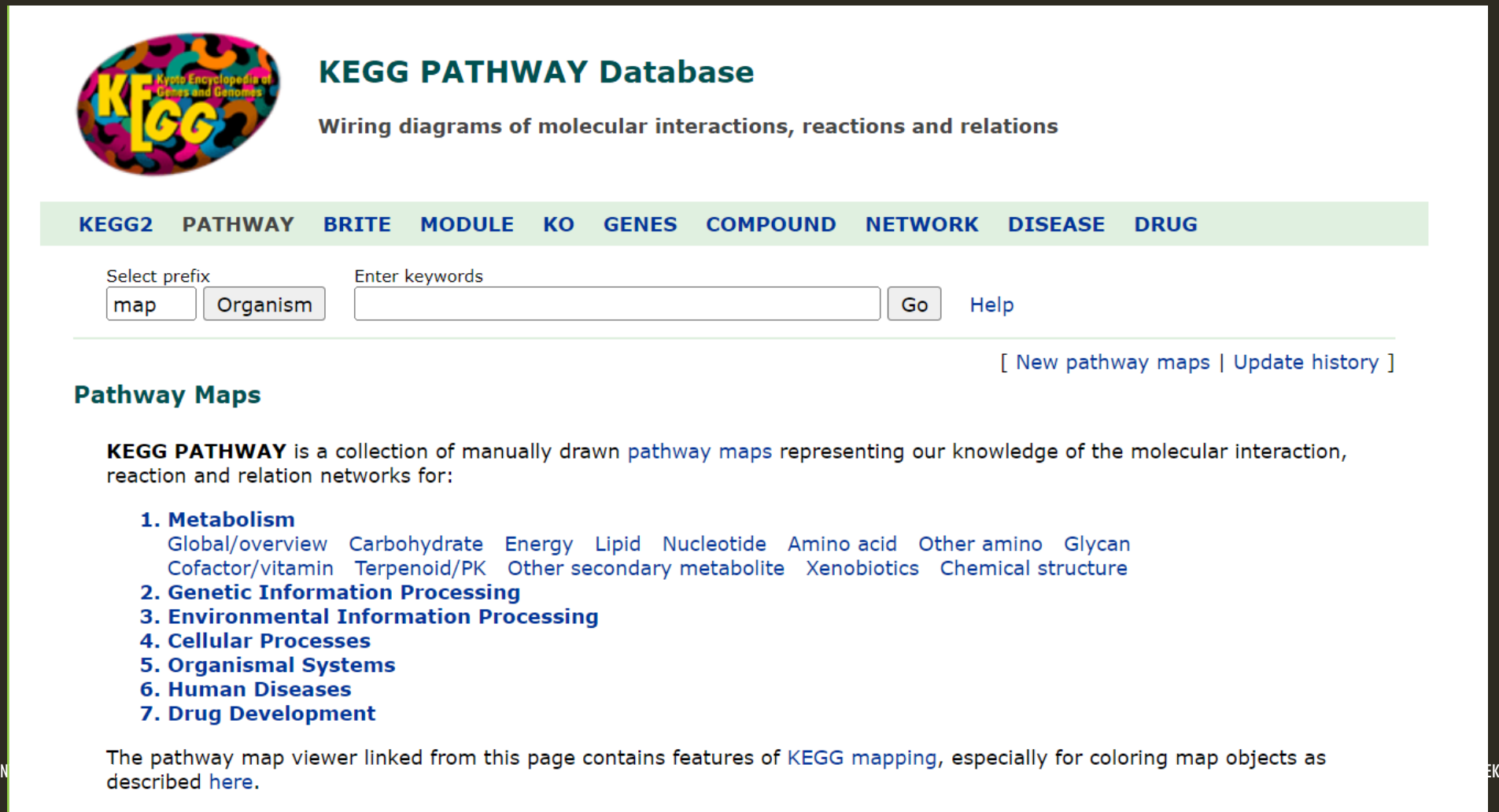
GENE ONTOLOGY



KEGG

- KEGG → Kyoto Encyclopedia of Genes and Genomes
- Cel
  - opis funkcji systemów biologicznych
  - biologiczna interpretacja danych molekularnych
- Połączenie informacji:
  - genomicznych (dla gatunków o znanej sekwencji całego genomu)
  - chemicznych
  - ogólnoustrojowych danych metabolicznych
- Kanehisa Laboratories → Japonia
- 1995
- 15 powiązanych ze sobą baz danych

KEGG


KEGG PATHWAY Database
Wiring diagrams of molecular interactions, reactions and relations
KEGG2 PATHWAY BRITE MODULE KO GENES COMPOUND NETWORK DISEASE DRUG
Select prefix
map
Organism
Enter keywords
Go
Help
[New pathway maps | Update history]
Pathway Maps
KEGG PATHWAY is a collection of manually drawn pathway maps representing our knowledge of the molecular interaction, reaction and relation networks for:
1. Metabolism
Global/overview Carbohydrate Energy Lipid Nucleotide Amino acid Other amino Glycan
Cofactor/vitamin Terpenoid/PK Other secondary metabolite Xenobiotics Chemical structure
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular Processes
5. Organismal Systems
6. Human Diseases
7. Drug Development
The pathway map viewer linked from this page contains features of KEGG mapping, especially for coloring map objects as described here.

KEGG

- Wizualizacja interakcji molekularnych i sieci reakcji
- Utworzone na podstawie informacji z publikacji naukowych

Rodzaje ścieżek:

- Metabolizm np. metabolizm lipidów
- Przetwarzanie informacji genetycznej np. transkrypcja
- Przetwarzanie informacji ze środowiska np. szlaki sygnałowe
- Procesy na poziomie komórki np. transport i katabolizm
- Procesy na poziomie organizmu np. układ immunologiczny
- Procesy chorobotwórcze np. choroby neurodegeneracyjne
- Projektowanie leków np. leki przeciwzapalne

KEGG

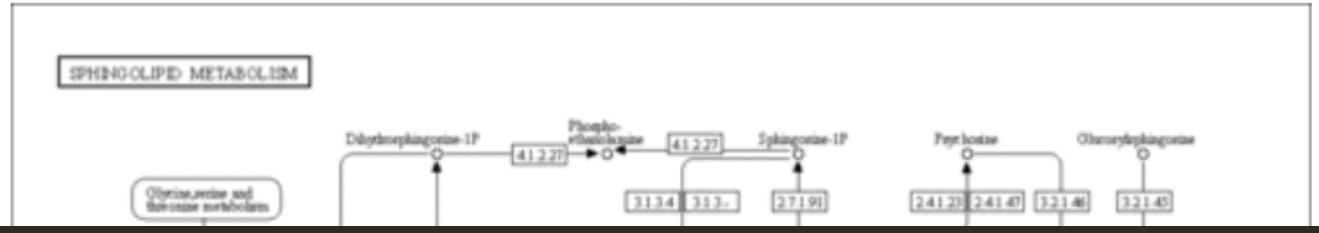
Sphingolipid metabolism - Homo sapiens (hsa00600) → pełna informacja
https://www.genome.jp/dbget-bin/www_bget?pathway:hsa00600

KEGG

PATHWAY: hsa00600

Help

Entry	hsa00600 Pathway
Name	Sphingolipid metabolism - Homo sapiens (human)
Class	Metabolism; Lipid metabolism BRITE hierarchy
Pathway map	hsa00600 Sphingolipid metabolism

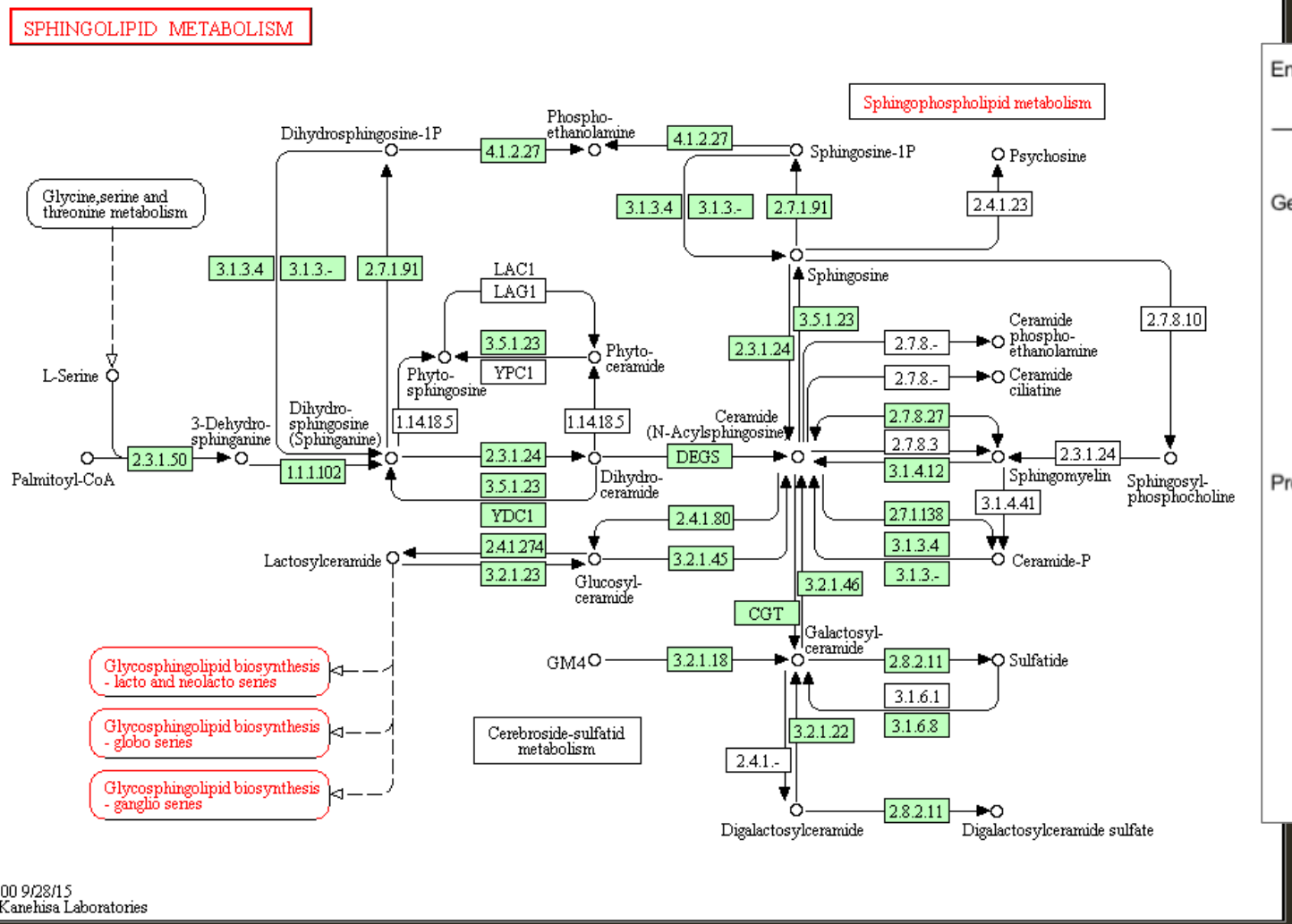


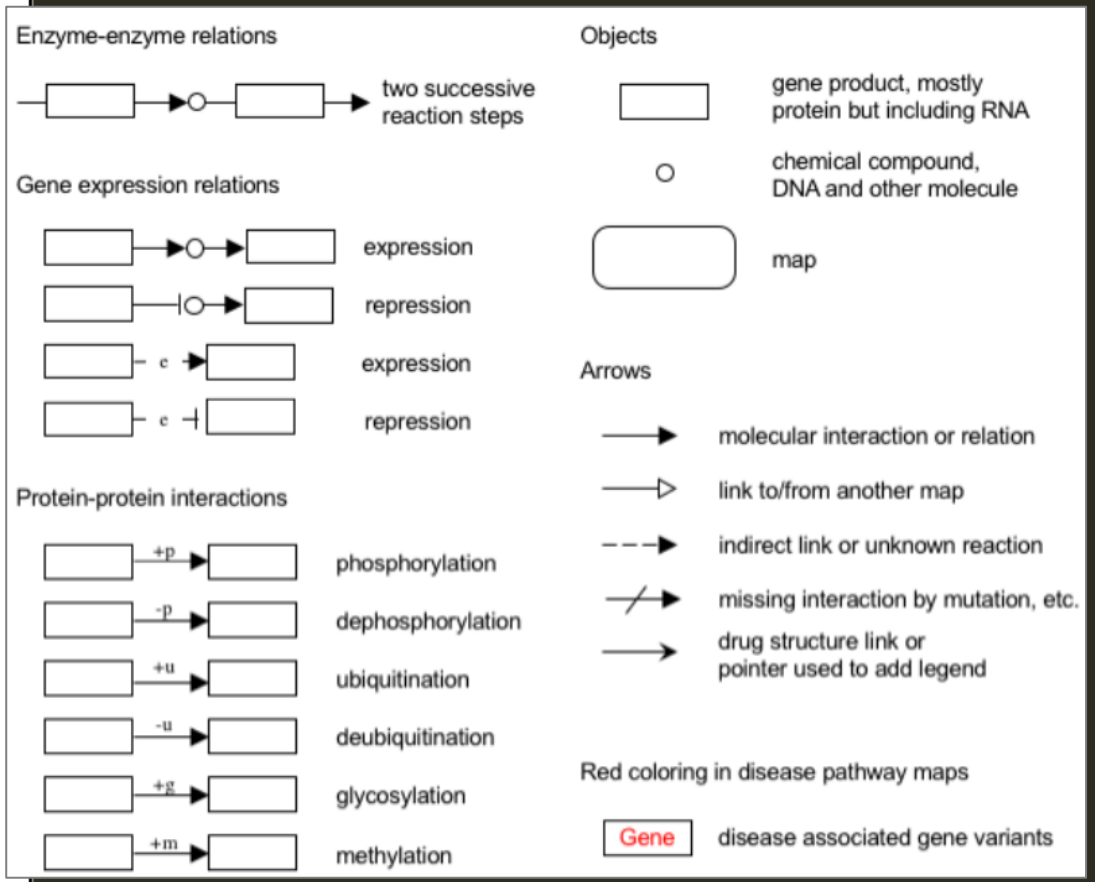
All links

Network (9)
KEGG NETWORK (9)
Genome (1)
KEGG GENOME (1)
Gene (54)
KEGG GENES (54)
All databases (64)

Download RDF

KEGG (HSA00600)

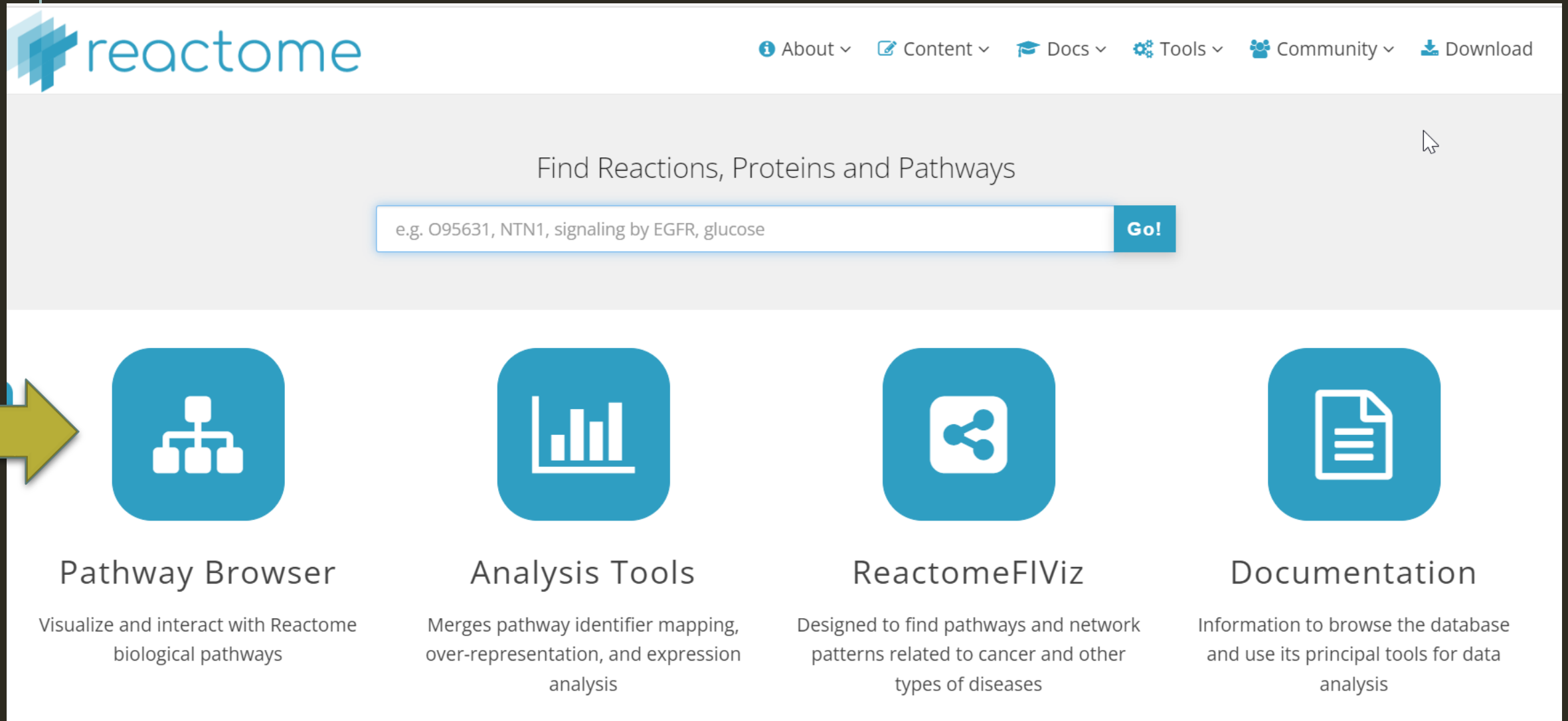




REACTOME

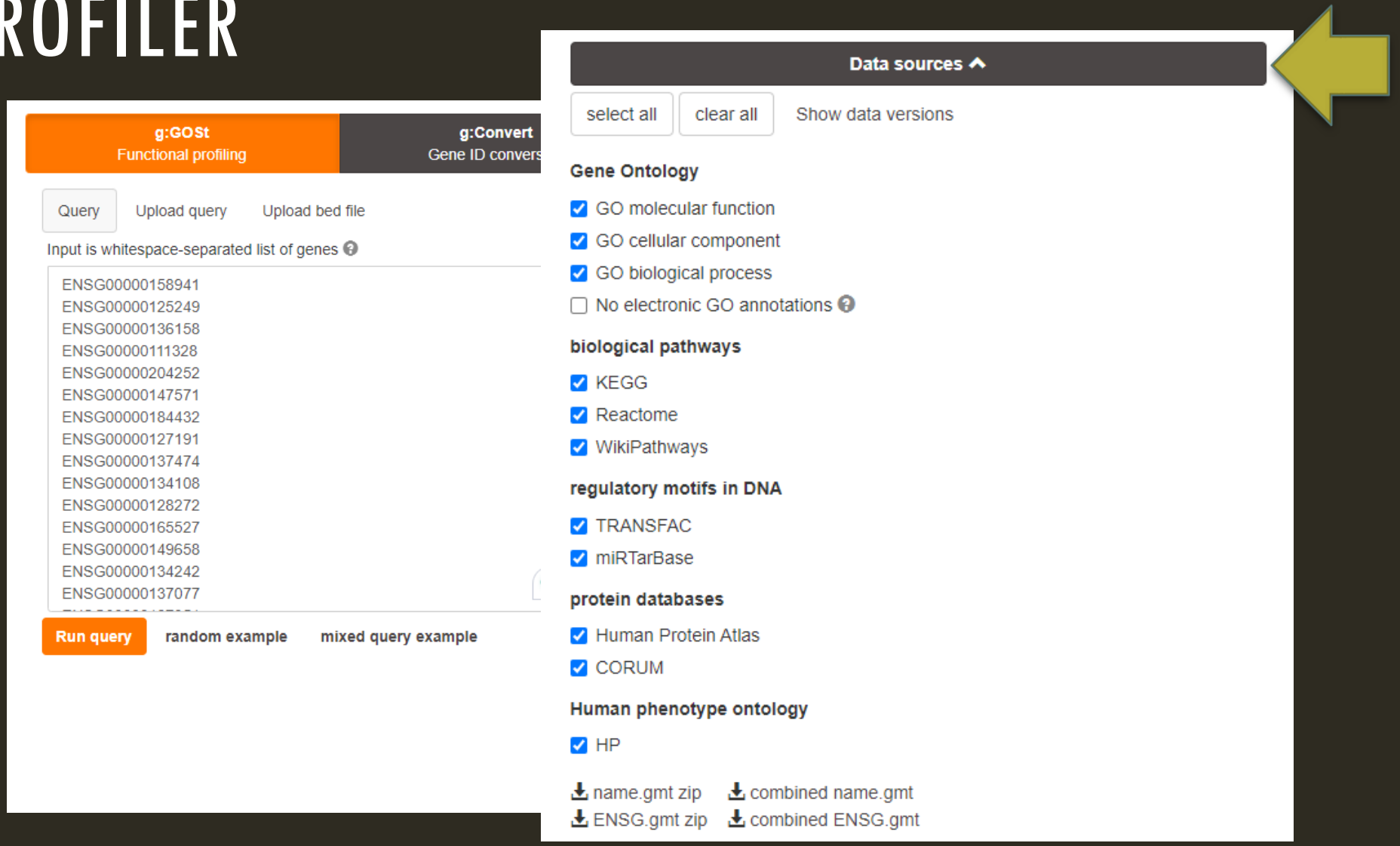
- Informacje o różnego rodzaju ścieżkach, głównie u człowieka
- Oferuje narzędzia do analizy danych (np. enrichment)
- Reactome dostarcza m. in. graficzne reprezentacje szlaków naprawy DNA oraz szlaków dojrzewania i degradacji RNA (nie wszystko jest w KEGG).

REACTOME


reactome
About
Content
Docs
Tools
Community
Download
Find Reactions, Proteins and Pathways
e.g. O95631, NTN1, signaling by EGFR, glucose
Go!
Pathway Browser
Visualize and interact with Reactome biological pathways
Analysis Tools
Merges pathway identifier mapping, over-representation, and expression analysis
ReactomeFIViz
Designed to find pathways and network patterns related to cancer and other types of diseases
Documentation
Information to browse the database and use its principal tools for data analysis

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

GPROFILER



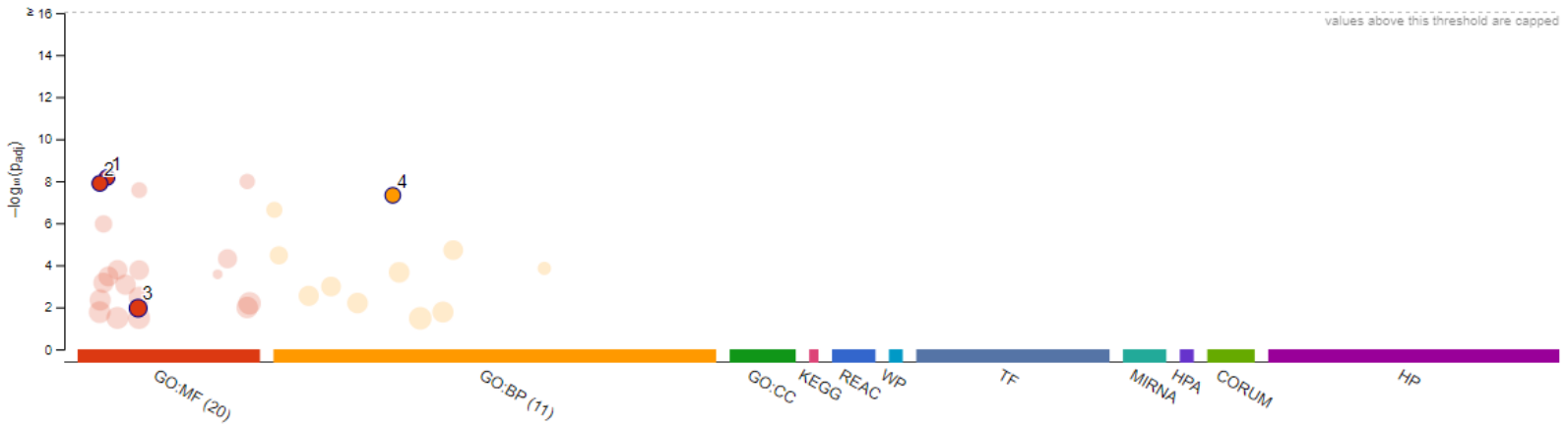
https://biit.cs.ut.ee/gprofiler/gost

GPROFILER

Overview | Detailed Results | GO Context | Query Info

Export to PNG | Show query URL | Show short link

☐ For print
☑ Capped
☑ Make unselected terms transparent



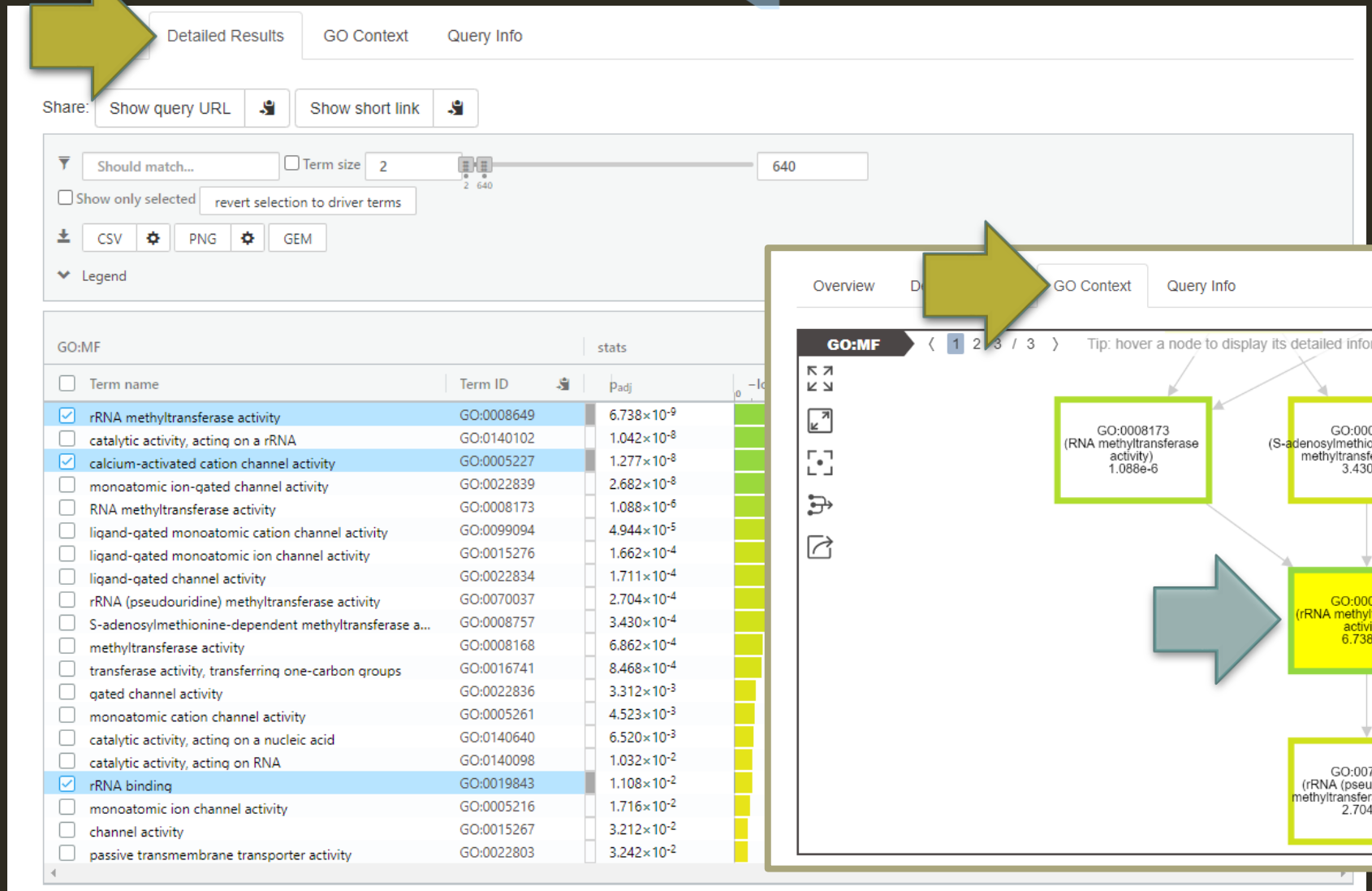
ID	Source	Term ID	Term Name	p_{adj} (query_1)
1	GO:MF	GO:0008649	rRNA methyltransferase activity	6.738×10^{-9}
2	GO:MF	GO:0005227	calcium-activated cation channel activity	1.277×10^{-8}
3	GO:MF	GO:0019843	rRNA binding	1.108×10^{-2}
4	GO:BP	GO:0031167	rRNA methylation	4.740×10^{-8}

version e111_eg58_p18_30541362
date 5.05.2024, 21:15:18
organism hsapiens

g:Profiler

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

GPROFILER

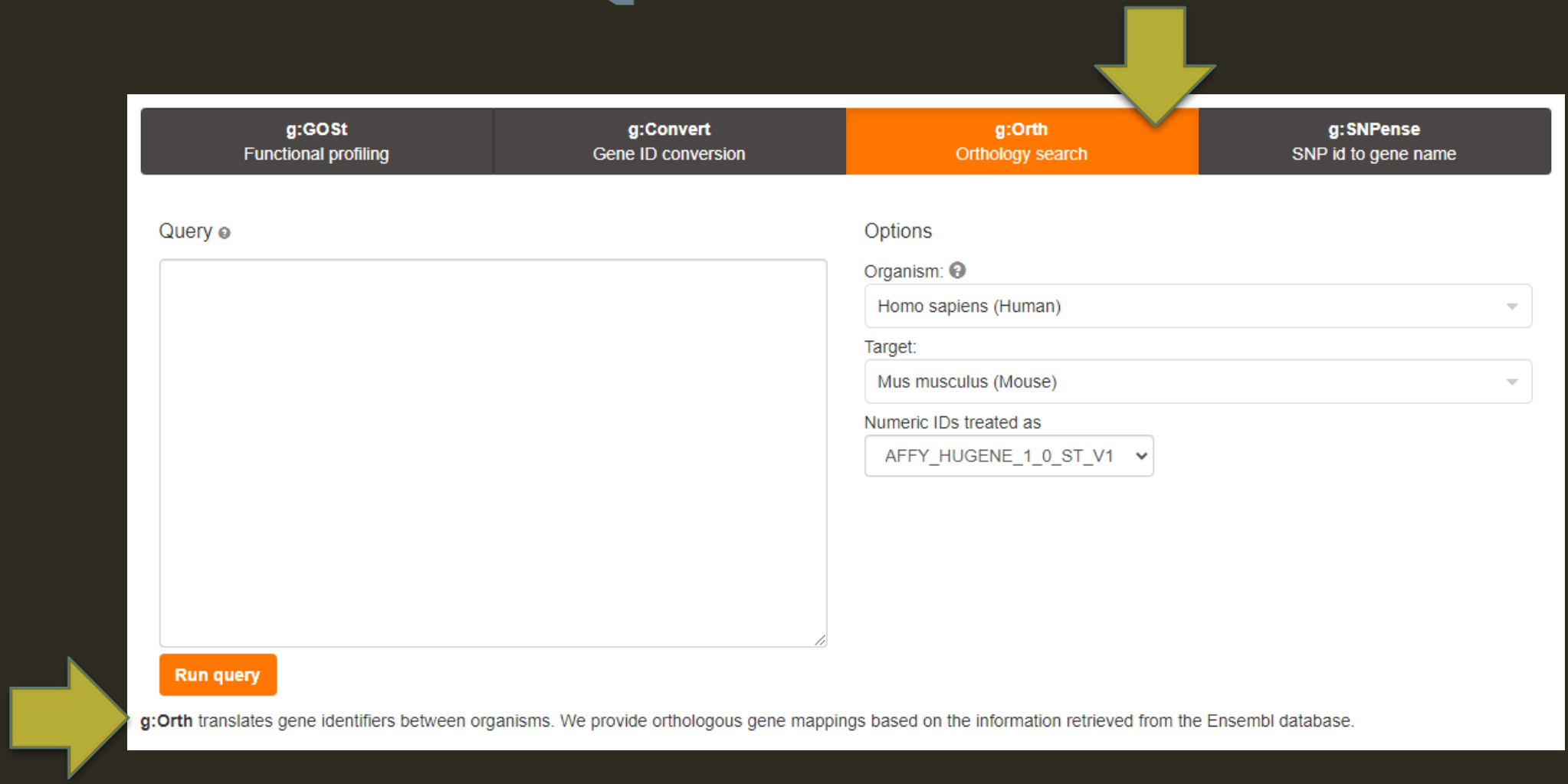


https://biit.cs.ut.ee/gprofiler/gost

GPROFILER

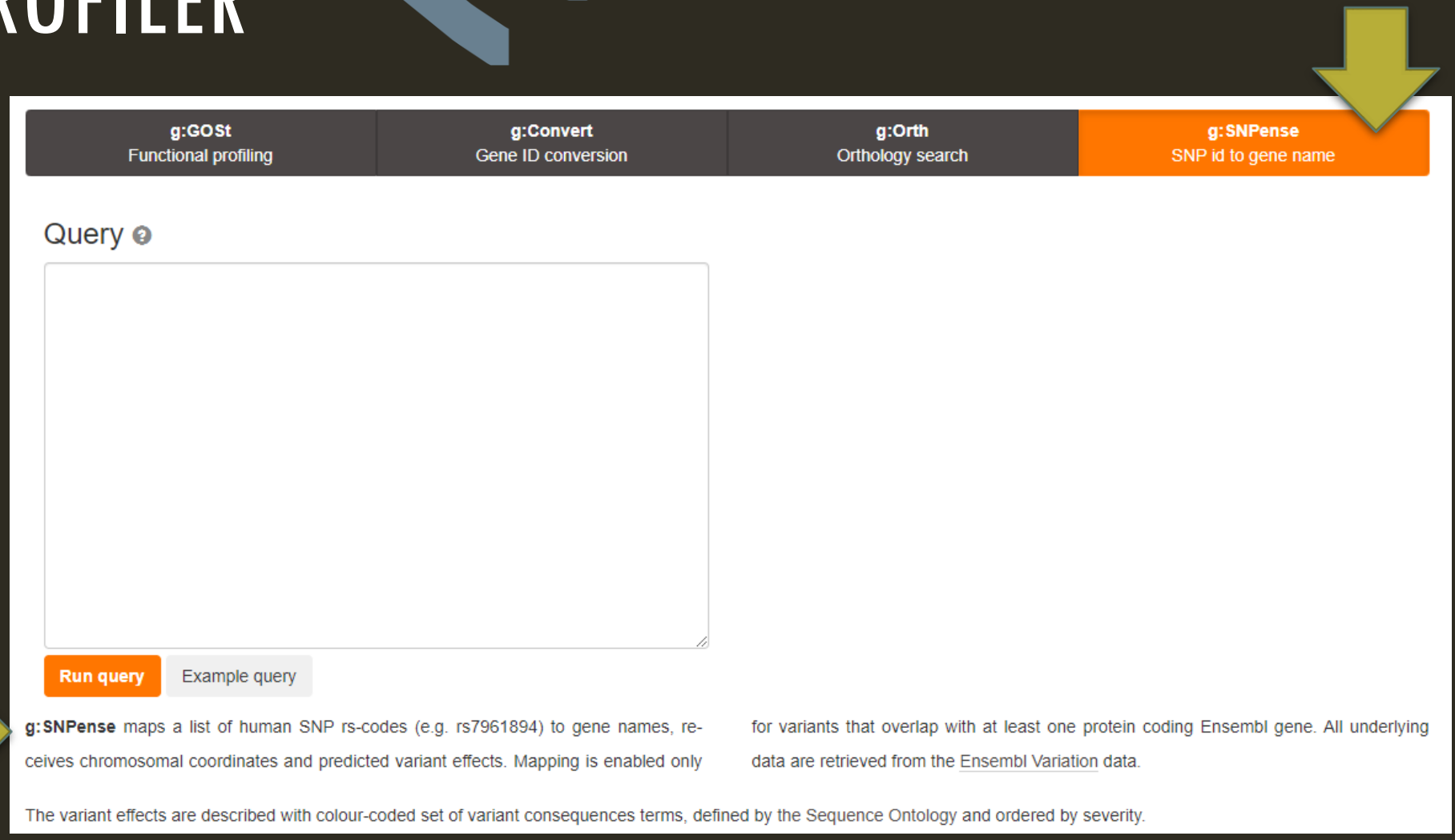
https://biit.cs.ut.ee/gprofiler/gost

GPROFILER



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

GPROFILER



KOBAS ENRICHMENT ANALYSIS

Gene Ontology, KEGG, Reactome
… i inne

KOBAS

KOBAS-intelligence | Home | Annotation | Enrichment | Download | Help | Citing | Mirror Sites

Result Retrieve

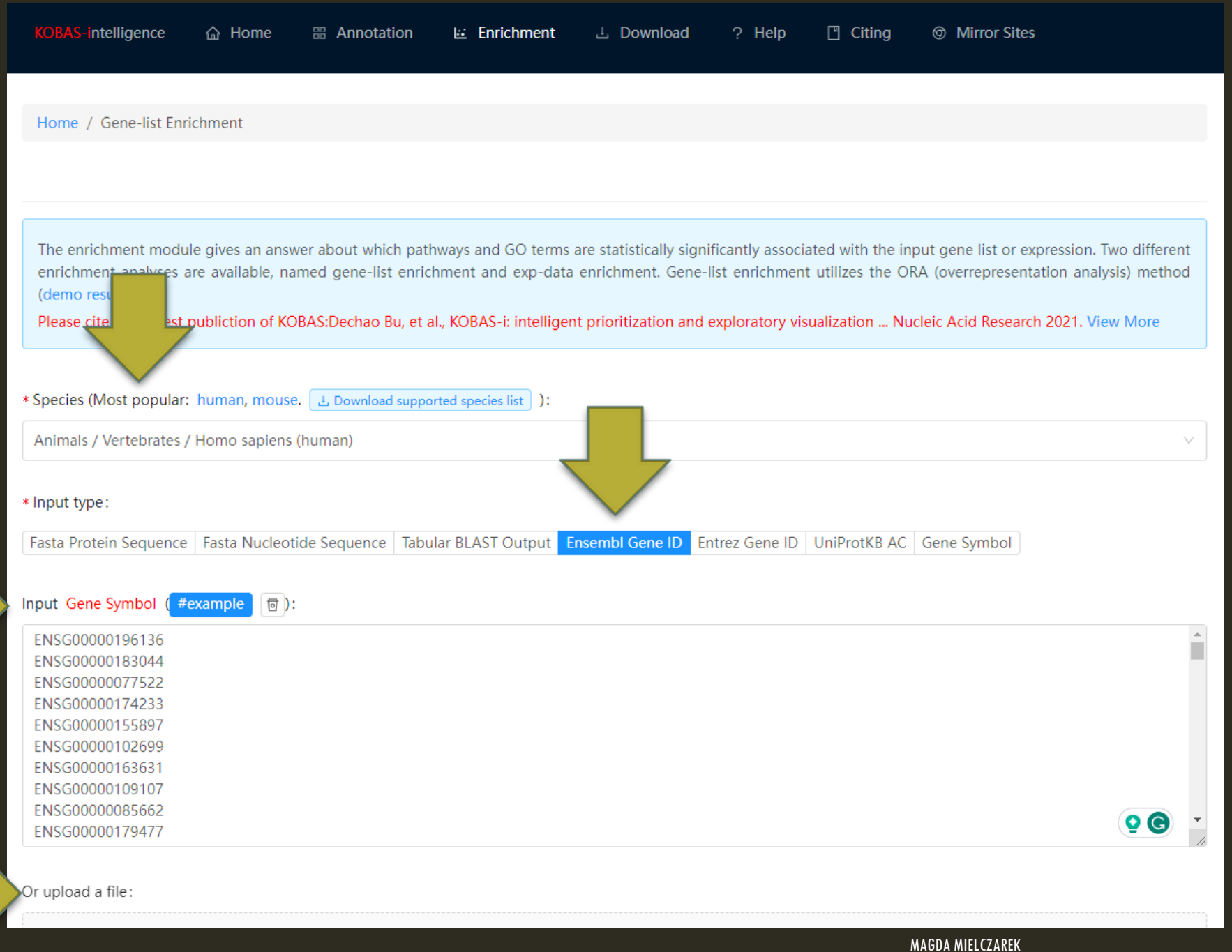
Input your task id here: Input your task id here search

(Go to demo results: Annotation, Gene list enrichment, Gene list enrichment visualization[New], Exp-data enrichment[New], Exp-data enrichment visualization[New])

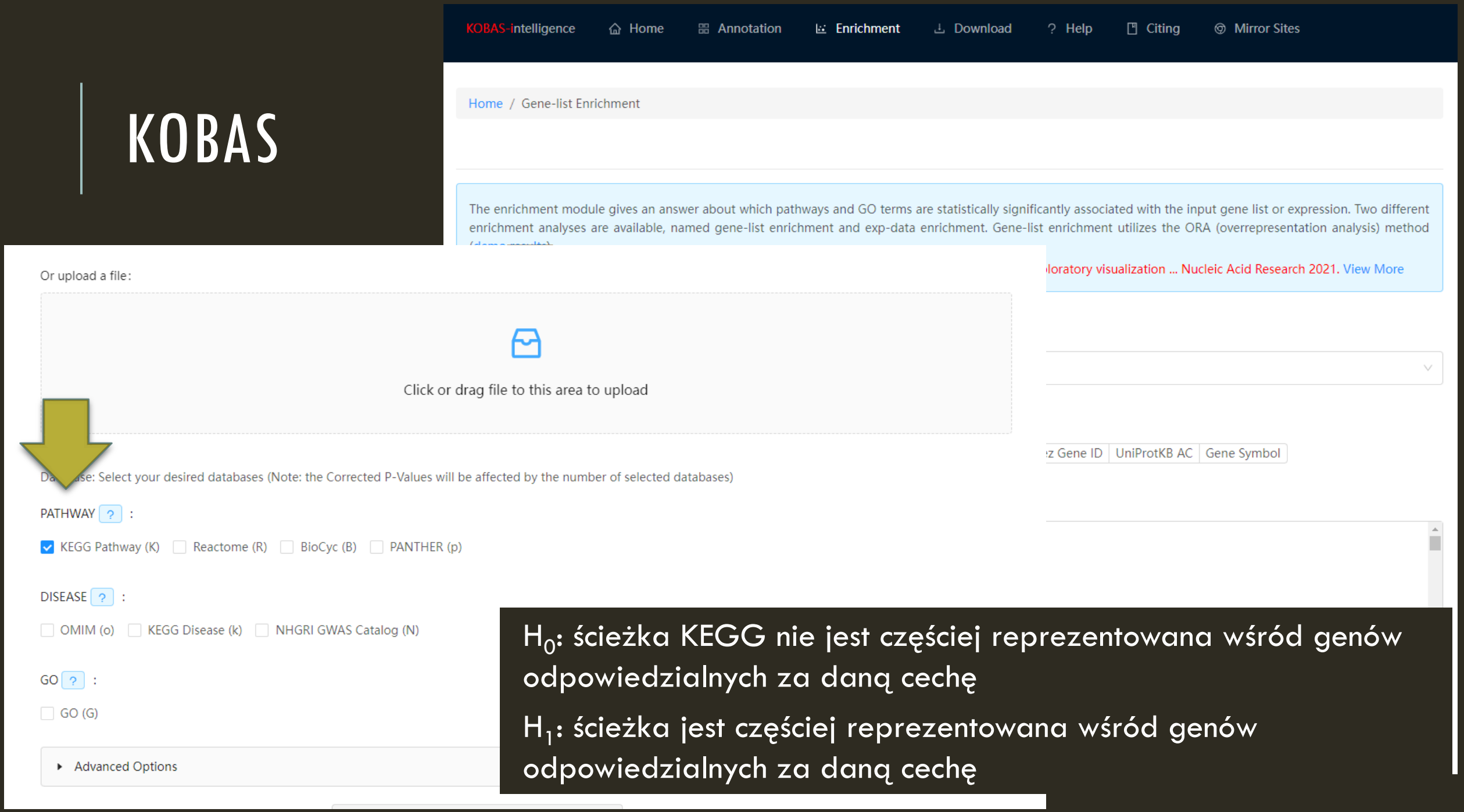
What is KOBAS?

KOBAS consists of two parts called the "annotation module" and the "enrichment module" (called "identify module" in the previous version) (view framework). The annotation module accepts the gene-list as input, including IDs or sequences, and generates annotations for each gene based on multiple databases of pathways, diseases, and GO information . The enrichment module gives an answer about which pathways and GO terms are statistically significantly associated with the input gene list or expression. Two different enrichment analyses are available, named gene-list enrichment and exp-data enrichment.

KOBAS


KOBAS-intelligence
Home
Annotation
Enrichment
Download
Help
Citing
Mirror Sites
Home / Gene-list Enrichment
The enrichment module gives an answer about which pathways and GO terms are statistically significantly associated with the input gene list or expression. Two different enrichment analyses are available, named gene-list enrichment and exp-data enrichment. Gene-list enrichment utilizes the ORA (overrepresentation analysis) method (demo resu
Please cite
publiction of KOBAS:Dechao Bu, et al., KOBAS-i: intelligent prioritization and exploratory visualization ... Nucleic Acid Research 2021. View More
* Species (Most popular: human, mouse. Download supported species list):
Animals / Vertebrates / Homo sapiens (human)
* Input type:
Fasta Protein Sequence
Fasta Nucleotide Sequence
Tabular BLAST Output
Ensembl Gene ID
Entrez Gene ID
UniProtKB AC
Gene Symbol
Input Gene Symbol (#example):
ENSG00000196136
ENSG00000183044
ENSG00000077522
ENSG00000174233
ENSG00000155897
ENSG00000102699
ENSG00000163631
ENSG00000109107
ENSG00000085662
ENSG00000179477
Or upload a file:

KOBAS



H_0: ścieżka KEGG nie jest częściej reprezentowana wśród genów odpowiedzialnych za daną cechę

H_1: ścieżka jest częściej reprezentowana wśród genów odpowiedzialnych za daną cechę

KOBAS

Enrichment Result

Go to visualization demo | Visualization | Download total terms

Database: KEGG PATHWAY × Input: min ~ max Total: min ~ max

P-Value: No filter <0.05 <0.01 <0.001 Corrected P-Value: No filter <0.05 <0.01 <0.001

#Term	Database	ID	Input	Total	P-Value	Corrected P-Value	Genes
Human cytomegalovirus infectio...	KEGG PATHWAY	hsa05163	32	223	3.78e-21	1.05e-18	Detail
Pathways in cancer	KEGG PATHWAY	hsa05200	44	530	5.00e-20	6.92e-18	Detail
Circadian entrainment	KEGG PATHWAY	hsa04713	23	97	9.06e-20	8.37e-18	Detail
Dopaminergic synapse	KEGG PATHWAY	hsa04728	25	131	2.15e-19	1.49e-17	Detail
Retrograde endocannabinoid sig...	KEGG PATHWAY	hsa04723	24	141	1.17e-17	6.47e-16	Detail
Amphetamine addiction	KEGG PATHWAY	hsa05031	18	68	1.81e-16	8.38e-15	Detail

KOBAS

Enrichment Result

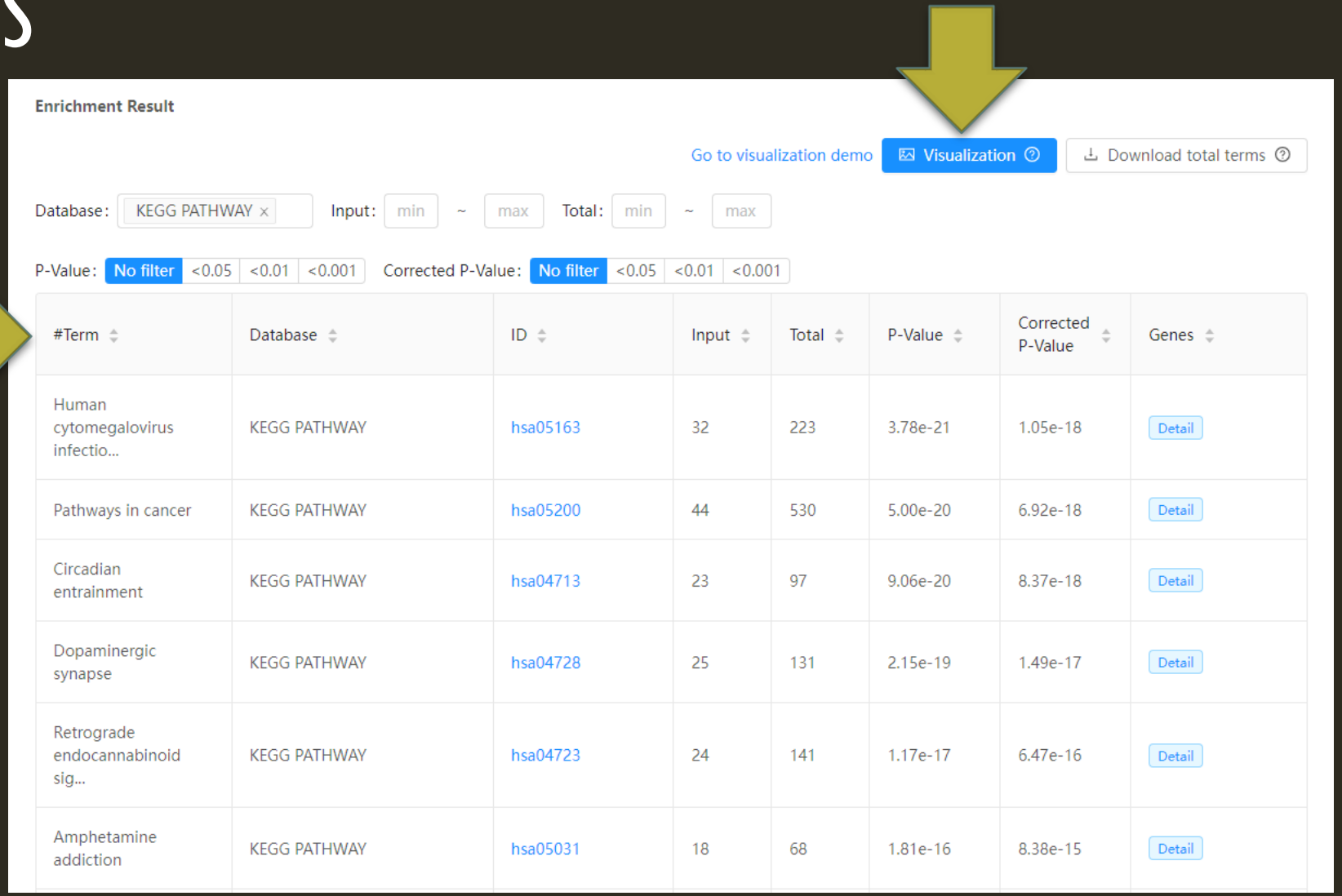
Go to visualization demo | Visualization | Download total terms

Database: KEGG PATHWAY × Input: min ~ max Total: min ~ max

P-Value: No filter <0.05 <0.01 <0.001 Corrected P-Value: No filter <0.05 <0.01 <0.001

#Term	Database	ID	Input	Total	P-Value	Corrected P-Value	Genes
Human cytomegalovirus infectio...	KEGG PATHWAY	hsa05163	32	223	3.78e-21	1.05e-18	Detail
Pathways in cancer	KEGG PATHWAY	hsa05200	44	530	5.00e-20	6.92e-18	Detail
Circadian entrainment	KEGG PATHWAY	hsa04713	23	97	9.06e-20	8.37e-18	Detail
Dopaminergic synapse	KEGG PATHWAY	hsa04728	25	131	2.15e-19	1.49e-17	Detail
Retrograde endocannabinoid sig...	KEGG PATHWAY	hsa04723	24	141	1.17e-17	6.47e-16	Detail
Amphetamine addiction	KEGG PATHWAY	hsa05031	18	68	1.81e-16	8.38e-15	Detail

KOBAS



Enrichment Result

Go to visualization demo | Visualization | Download total terms

Database: KEGG PATHWAY x Input: min ~ max Total: min ~ max

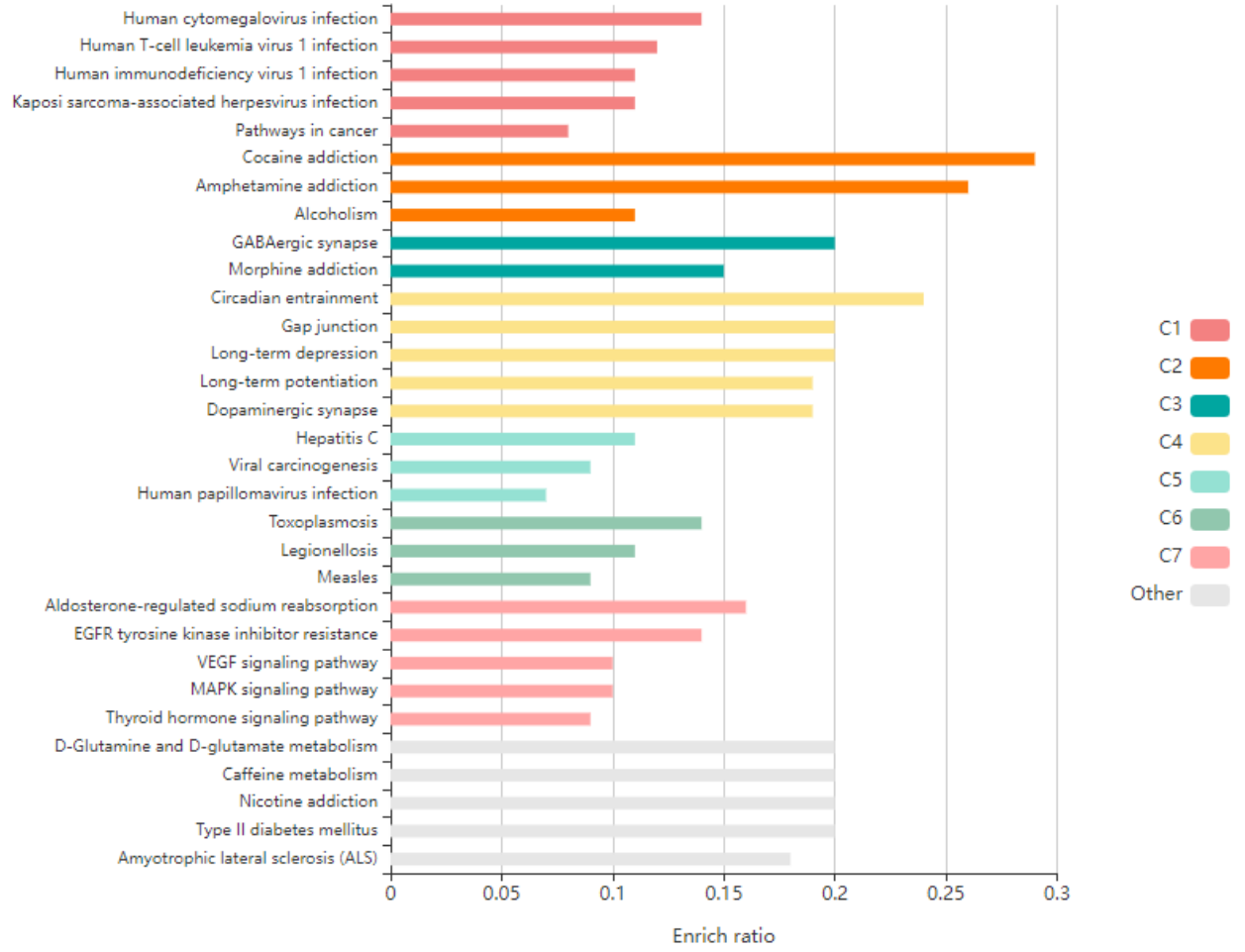
P-Value: No filter | <0.05 | <0.01 | <0.001 Corrected P-Value: No filter | <0.05 | <0.01 | <0.001

#Term	Database	ID	Input	Total	P-Value	Corrected P-Value	Genes
Human cytomegalovirus infectio...	KEGG PATHWAY	hsa05163	32	223	3.78e-21	1.05e-18	Detail
Pathways in cancer	KEGG PATHWAY	hsa05200	44	530	5.00e-20	6.92e-18	Detail
Circadian entrainment	KEGG PATHWAY	hsa04713	23	97	9.06e-20	8.37e-18	Detail
Dopaminergic synapse	KEGG PATHWAY	hsa04728	25	131	2.15e-19	1.49e-17	Detail
Retrograde endocannabinoid sig...	KEGG PATHWAY	hsa04723	24	141	1.17e-17	6.47e-16	Detail
Amphetamine addiction	KEGG PATHWAY	hsa05031	18	68	1.81e-16	8.38e-15	Detail

KOBAS

Enriched terms visualized in barplot

Each row represents an enriched function, and the length of the bar represents the enrich ratio，which is calculated as "input gene number"/ "backgound gene number". The color of the bar is the same as the color in the circular network in above, which represents different clusters. For each cluster, if there are more than 5 terms, top 5 with the highest enrich ratio will be displayed.



ENRICHMENT ANALYSIS SHINYGO

Gene Ontology, KEGG

ShinyGO 0.80

Change species | *Human*

Demo genes | Reset

Change the species if it is not human. Then just paste a list of genes and click Submit. Gene IDs can be NCBI, Ensembl, symbol, or other common types.

Background (recommended) | Submit

FDR cutoff: 0,05

pathways to show: 20

Pathway size: Min. 2 | Max. 5000

- [x] Remove redundancy
- [x] Abbreviate pathways
- [ ] Use pathway DB for gene counts
- [ ] Show pathway IDs

Gene IDs examples

Try iDEP for RNA-Seq data analysis

Enrichment | Chart | Tree | Network | KEGG | Genes | Groups | Plots | Genome | STRING | About

ShinyGO: a graphical gene-set enrichment tool for animals and plants

10/25/24: Migrated to new server. R upgraded to V 4.4.0.

4/21/2024: New barchart with GO terms on the bars.

4/12/2024: Max set size is increased to 5000 from 2000. Some meaningful GO terms (RNA biosynthetic proc.) contains 4000+ genes.

You can still use the old versions using links on the About tab. To support this effort, please cite our paper, like over 2000 users did. Just including URL is not enough. Email Jenny (gelabinfo@gmail.com) for questions, suggestions or data contributions. Follow Dr Ge on Twitter and LinkedIn for updates.

Feb. 11, 2022: Like ShinyGO but your genome is not covered? Customized ShinyGO is now available. Its database includes several custom genomes requested by users. To request to add a new species/genome, fill in this Form.

For-profit organizations: contact us for licensing, local installation, or customization services.

GO Enrichment analysis, plus a lot more!

Just paste your gene list to get enriched GO terms and othe pathways for over 14,000 species. based on annotation from Ensembl and STRING-db. Produce KEGG pathway diagrams with your genes highlighted, hierarchical clustering trees and networks summarizing overlapping terms/pathways, protein-protein interaction networks, gene characterristics plots, and enriched promoter motifs.

Enrichment FDR	Genes in list	Total genes	Functional Category
6.5E-220	86	101	DNA damage checkpoint
5.3E-215	86	108	DNA integrity checkpoint
3.2E-188	86	169	Cell cycle checkpoint
1.1E-131	87	659	Cellular response to DNA damage stimulus
1.2E-113	87	1039	Cell cycle process
3.6E-102	87	1395	Cell cycle
1.1E-100	45	57	Mitotic DNA integrity checkpoint
5.0E-99	87	1517	Cellular response to stress
1.9E-98	43	51	Mitotic DNA damage checkpoint