Final Project Proposal: GO Simplify

Beverlie Poblete

Tool Background

With the increasing number of advancements in biotechnology, we are able to collect high volumes of genomic data at decreasing costs. This has shifted the need for collecting biological data to a need of analyzing and making sense of the biological data. One way to tackle this problem was through ontology, which has enabled the organization and analysis of large datasets. Ontologies are most particularly useful in creating genomic annotations. The Gene Ontology (GO) database was developed to record annotations of the functional properties of gene products across species, which further contributes to gene function prediction. GO annotations are created by associating a gene or gene product with a GO term. These GO annotations thus capture a "snapshot" about how a gene's molecular function, location in the cell, and what biological processes it is involved in. In the GO database, each ontology consists of a set of GO terms, which are organized in a directed acyclic graph (DAG) where each GO term is a vertex of the graph, and the edges encode the relationship between GO terms. GO annotations stores the currently known functional knowledge of gene products where a positive annotation indicates the gene product carries out the function described by the GO term. A negative annotation in turn indicates the gene product does not perform the function described by the term. The Gene Ontology (GO) knowledgebase (http://geneontology.org/) is the world's largest source of information on the functions of genes. While exploring the knowledgebase, I found I had to click through multiple links to find related genes and gene products annotated to the same GO class of a certain gene product or related annotations. With my final project tool, GO Simplify, I aim to create a user friendly application that will take in a simple inputs of a gene/gene product and organism. The GO simplify tool will then display the associations between GO terms and the gene/gene product for the organism input. There will also be a follow-up option for users to display other genes/gene products annotated to the same GO term.

Tool Functionality

This tool requires two basic inputs: a gene or gene product and organism. With the inputs, the tool will query a subset database (that will be a subset of what is offered in the GO knowledgebase) and return and display the associations between your input gene/gene product and GO terms for the select organism. Following this display, a dropdown will be displayed containing the resulting GO terms from the previous search. The end user can select one of the options from the dropdown to see what other genes/gene products were annotated to the selected GO term. With this tool, I aim to simplify the data presented in the GO knowledgebase.

Tool Description

The following software technologies will be used in implementing GO Simplify:

1. **SQL relational database/table**
   This database will hold information of gene/product symbol, gene/product name, GO class (direct), and organism. The database will also hold information on genes and gene products annotated to positive regulation of select GO classes. I will manually curate the database to

hold a subset of data for ~50 gene/gene products. I will be collecting the data to populate my database from the GO knowledgebase.

2. **Python-based Computer Gateway Interface (CGI)**
   The CGI script will ingest the user's input of gene/gene product and organism and query the SQL table for the relevant gene/product symbol, gene/product name, and GO class (direct) for the input gene/gene product and organism. The CGI will also take in the follow-up user input from the dropdown of relevant GO classes to query the SQL database for related genes/gene products with annotations to the selected GO class. All information from both queries will be transmitted back to the user via the GUI.
3. **HTML/JavaScript-based graphic user interface (GUI)**
   HTML, JavaScript, and possibly CSS will be used to create a user-friendly web interface for taking in the user's input via HTML form and later drop-down and displaying the results of the query performed in the CGI script.

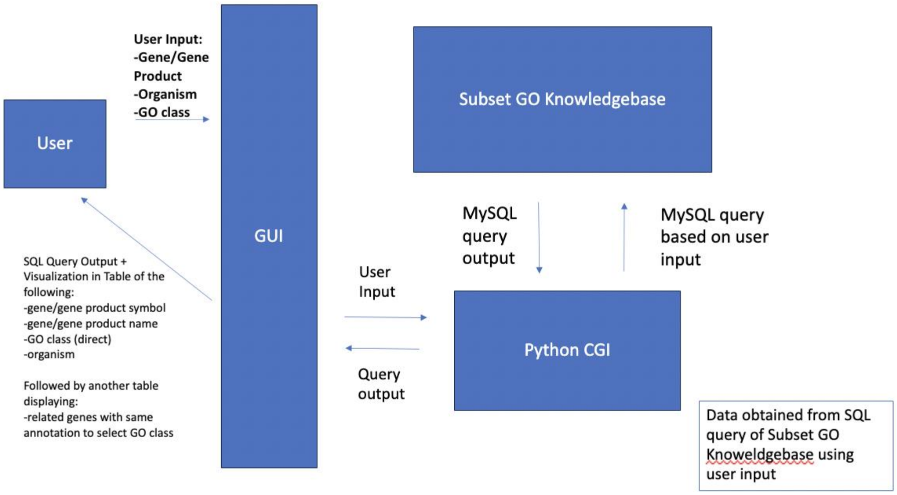


Figure 1. Software components and relative interaction and information exchange

<u>Tool Design and Development</u>

The following describe the proposed tool design and development:

1. I am planning on implementing a schema and relational database containing two tables. The first table will hold gene/gene product symbol, gene/product name, organism, GO class (name), and GO class key. The second table will contain a row per gene/product containing columns holding data for gene/product name, GO class name, and GO class key (foreign key for first table). This proposed database will be manually curated through queries in the GO database for a select number of gene/gene products for ~2 or 3 organisms to be able to complete this final project by the deadline. I will gather the information in an excel sheet prior to creating the database on the bfx3 server.

Small examples of what information the proposed tables will hold are shown in Table 1 and Table 2 below. Please note I limited the rows to 2-3 rows in the examples below to be concise.

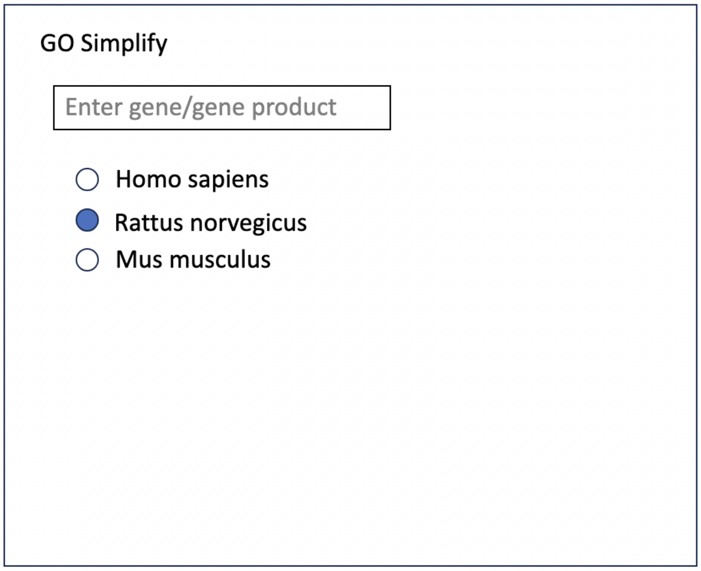
Gene/gene product symbol	**Gene/product name**	**organism**	**GO class (direct)**	**GO class key**
Tcf7l2	Transcription factor 7 like 2	Rattus norvegicus	Positive regulation of insulin secretion	1
Cckbr	Cholecystokinin A receptor	Rattus norvegicus	Insulin secretion	2
Cat	catalase	Rattus norvegicus	Response to insulin	3

Table 1*. Proposed Table 1*

Gene/product	**Gene/product name**	**GO class (direct)**	**Go class key**
Tcf7l2	Transcription factor 7 like 2	Positive regulation of insulin secretion	1
Casr	Calcium-sensing receptor	Positive regulation of insulin secretion	1

Table 2. *Proposed Table 2*

2. One CGI script will ingest the user's input of gene/gene product and organism and query the SQL table for the relevant gene/product symbol, gene/product name, and GO class (direct) for the input gene/gene product and organism. Another CGI script will also take in the follow-up user input from the dropdown of relevant GO classes to query the SQL database for related genes/gene products with annotations to the selected GO class. All information from both queries will be transmitted back to the user via the GUI.
3. A GUI created using HTML/JavaScript/CSS will be prepared to take user input via HTML form. The first input will be for the gene/gene product string and then organism (most likely radio button to choose amongst 2-3 organisms). Using this user input, the GUI will display the SQL query output returned by the CGI script as a table on the web page. After this initial table is displayed, a dropdown will be displayed with a dropdown of the GO classes returned from the first query. The user can select an option from the dropdown. The GUI will take this input and possibly call another python CGI script that will perform another query to find related gene/gene products with annotations to the same GO class. The output from this second query will be displayed to the user in the GUI as a second table of results.

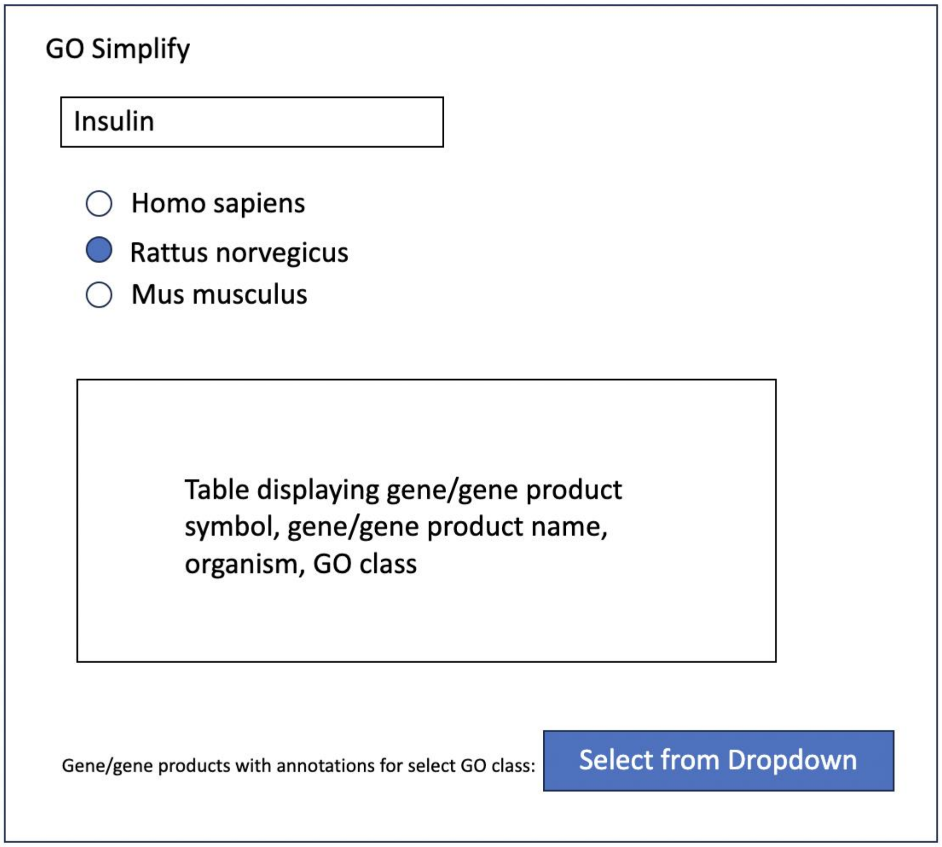


GO Simplify

Enter gene/gene product

○ Homo sapiens
● Rattus norvegicus
○ Mus musculus

Figure 2. First user input screen



GO Simplify

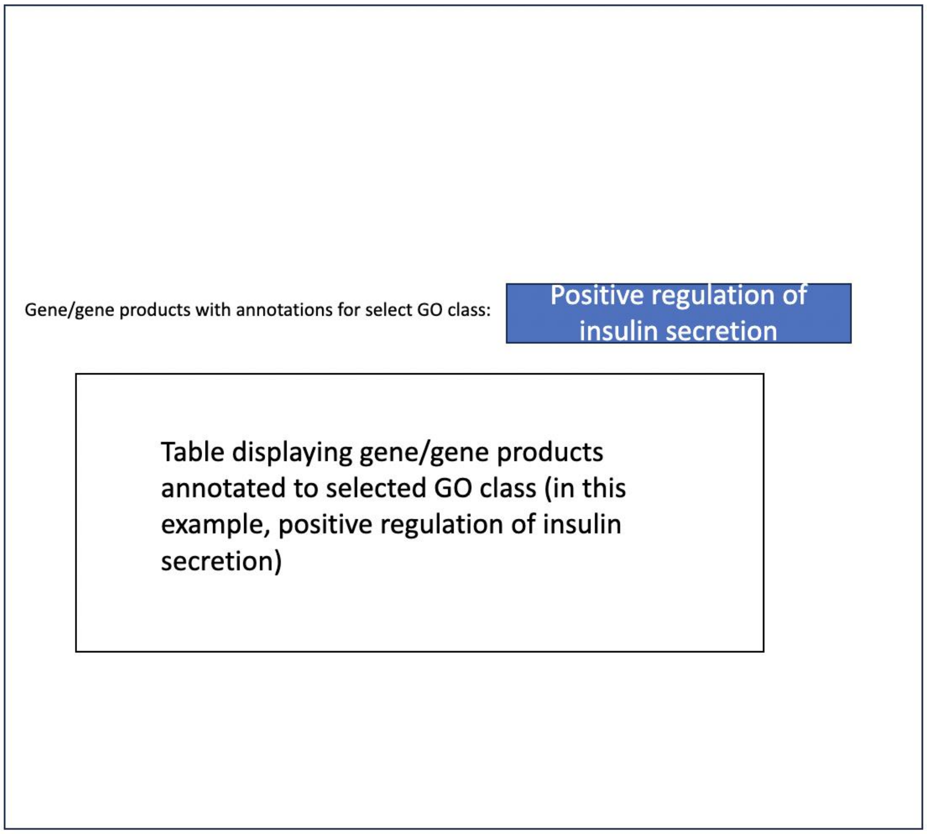
Insulin

○ Homo sapiens
● Rattus norvegicus
○ Mus musculus

Table displaying gene/gene product symbol, gene/gene product name, organism, GO class

Gene/gene products with annotations for select GO class:

Figure 3. Output/Display after first user input submitted



Gene/gene products with annotations for select GO class:

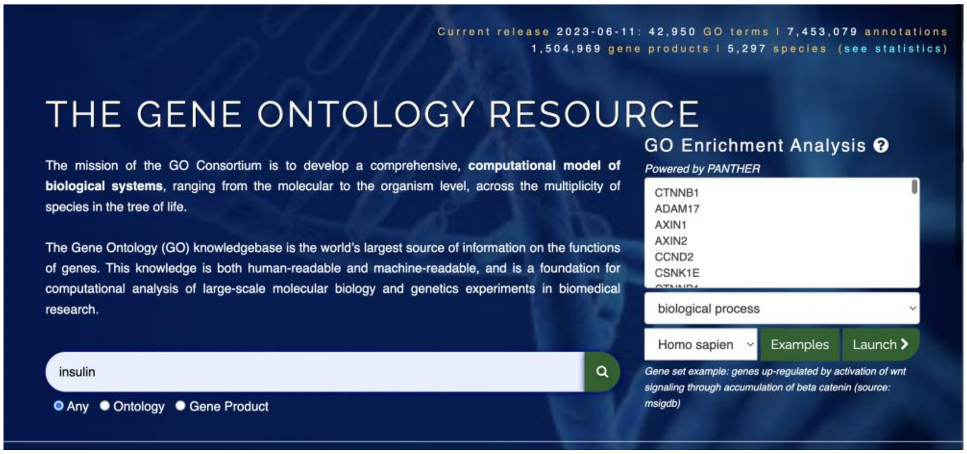
Table displaying gene/gene products annotated to selected GO class (in this example, positive regulation of insulin secretion)

Figure 4. Output/Display after second user input (dropdown)

Anticipated Challenges/Obstacles:

I anticipate the data mining section of the final project will be the most difficult and time-consuming. While the GO database contains a large number of entries, I plan on sub-setting my database to 50 gene/gene products for 2-3 organisms. I need to select a limit of how many rows to display in my results to not overload the page with too much data. I also may need to limit the number of genes per GO class key in the proposed second table to keep my database small.

Appendix – How I plan on mining data from GO database:



Current release 2023-06-11: 42,950 GO terms | 7,453,079 annotations
1,504,969 gene products | 5,297 species (see statistics)

THE GENE ONTOLOGY RESOURCE

The mission of the GO Consortium is to develop a comprehensive, **computational model of biological systems**, ranging from the molecular to the organism level, across the multiplicity of species in the tree of life.

The Gene Ontology (GO) knowledgebase is the world's largest source of information on the functions of genes. This knowledge is both human-readable and machine-readable, and is a foundation for computational analysis of large-scale molecular biology and genetics experiments in biomedical research.

insulin

● Any ○ Ontology ○ Gene Product

GO Enrichment Analysis

Powered by PANTHER

CTNNB1
ADAM17
AXIN1
AXIN2
CCND2
CSNK1E

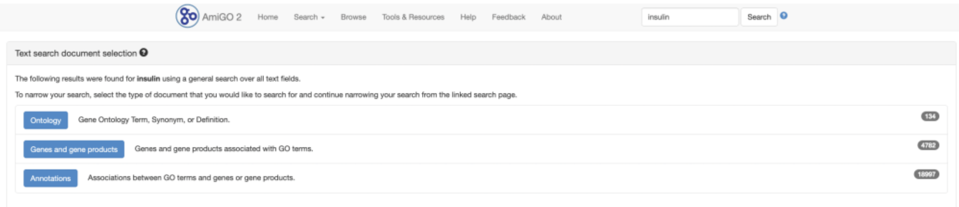
biological process

Homo sapien

Examples Launch

Gene set example: genes up-regulated by activation of wnt signaling through accumulation of beta catenin (source: msigdb)

1. On GO database, enter gene product



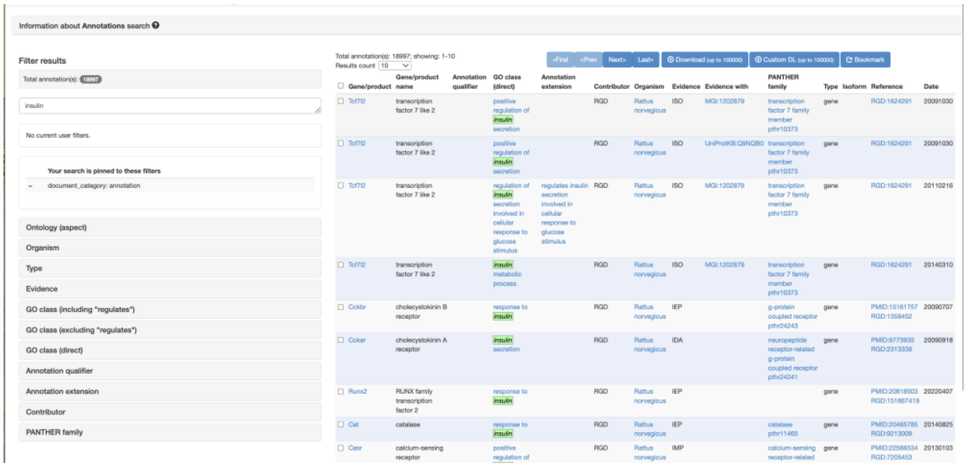
Text search document selection

The following results were found for **insulin** using a general search over all text fields.

To narrow your search, select the type of document that you would like to search for and continue narrowing your search from the linked search page.

Ontology	Gene Ontology Term, Synonym, or Definition.	134
Genes and gene products	Genes and gene products associated with GO terms.	4782
Annotations	Associations between GO terms and genes or gene products.	18997

2. Select "Annotations"



Information about Annotations search

Filter results

Total annotation(s): 18997

insulin

No current user filters.

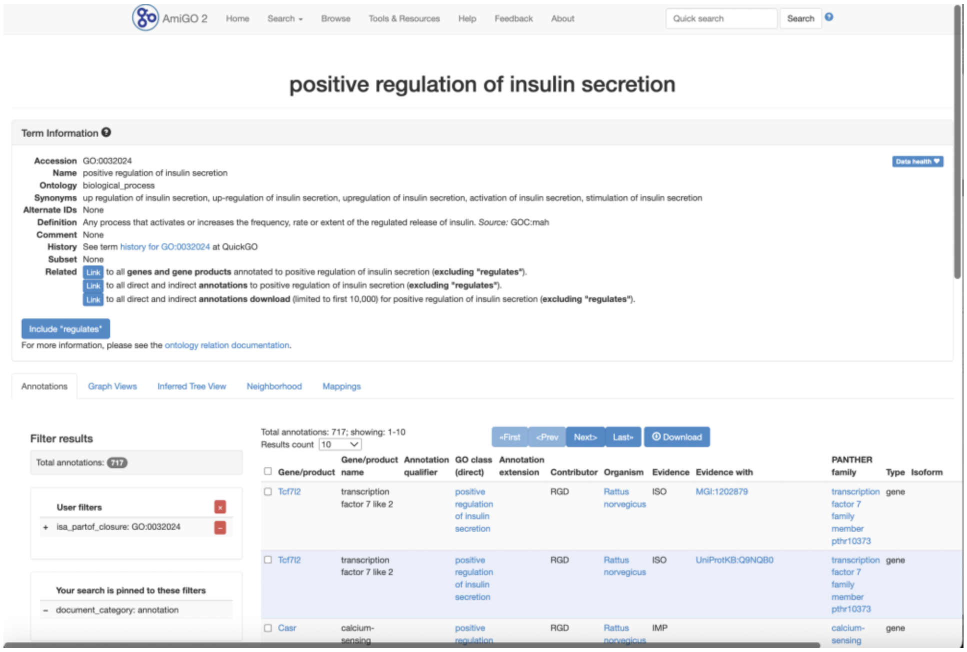
Your search is pinned to these filters

– document_category: annotation

Total annotation(s): 18997; showing: 1-10
Results count 10

☐	Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform	Reference	Date
☐	Tcf7l2	transcription factor 7 like 2		positive regulation of insulin secretion		RGD	Rattus norvegicus	ISO	MGI:1202879	transcription factor 7 family member pthr10373	gene		RGD:1624291	20091030
☐	Tcf7l2	transcription factor 7 like 2		positive regulation of insulin secretion		RGD	Rattus norvegicus	ISO	UniProtKB:Q9NQB0	transcription factor 7 family member pthr10373	gene		RGD:1624291	20091030
☐	Tcf7l2	transcription factor 7 like 2		regulation of insulin secretion involved in cellular response to glucose stimulus	regulates insulin secretion involved in cellular response to glucose stimulus	RGD	Rattus norvegicus	ISO	MGI:1202879	transcription factor 7 family member pthr10373	gene		RGD:1624291	20110216
☐	Tcf7l2	transcription factor 7 like 2		insulin metabolic process		RGD	Rattus norvegicus	ISO	MGI:1202879	transcription factor 7 family member pthr10373	gene		RGD:1624291	20140310
☐	Cckbr	cholecystokinin B receptor		response to insulin		RGD	Rattus norvegicus	IEP		g-protein coupled receptor pthr24243	gene		PMID:15161757 RGD:1358452	20090707
☐	Cckar	cholecystokinin A receptor		insulin secretion		RGD	Rattus norvegicus	IDA		neuropeptide receptor-related g-protein coupled receptor pthr24241	gene		PMID:9773935 RGD:2313338	20090918
☐	Runx2	RUNX family transcription factor 2		response to insulin		RGD	Rattus norvegicus	IEP			gene		PMID:20818503 RGD:151667419	20220407
☐	Cat	catalase		response to insulin		RGD	Rattus norvegicus	IEP		catalase pthr11465	gene		PMID:20465785 RGD:9213009	20140825
☐	Casr	calcium-sensing receptor		positive regulation of		RGD	Rattus norvegicus	IMP		calcium-sensing receptor-related	gene		PMID:22566534 RGD:7205453	20130103

3. Use results here to populate table 1 in database



positive regulation of insulin secretion

Term Information

Accession	GO:0032024
Name	positive regulation of insulin secretion
Ontology	biological_process
Synonyms	up regulation of insulin secretion, up-regulation of insulin secretion, upregulation of insulin secretion, activation of insulin secretion, stimulation of insulin secretion
Alternate IDs	None
Definition	Any process that activates or increases the frequency, rate or extent of the regulated release of insulin. *Source:* GOC:mah
Comment	None
History	See term history for GO:0032024 at QuickGO
Subset	None
Related	Link to all **genes and gene products** annotated to positive regulation of insulin secretion (**excluding "regulates"**). Link to all direct and indirect **annotations** to positive regulation of insulin secretion (**excluding "regulates"**). Link to all direct and indirect **annotations download** (limited to first 10,000) for positive regulation of insulin secretion (**excluding "regulates"**).

For more information, please see the ontology relation documentation.

Filter results

Total annotations: 717

User filters

+ isa_partof_closure: GO:0032024

Your search is pinned to these filters

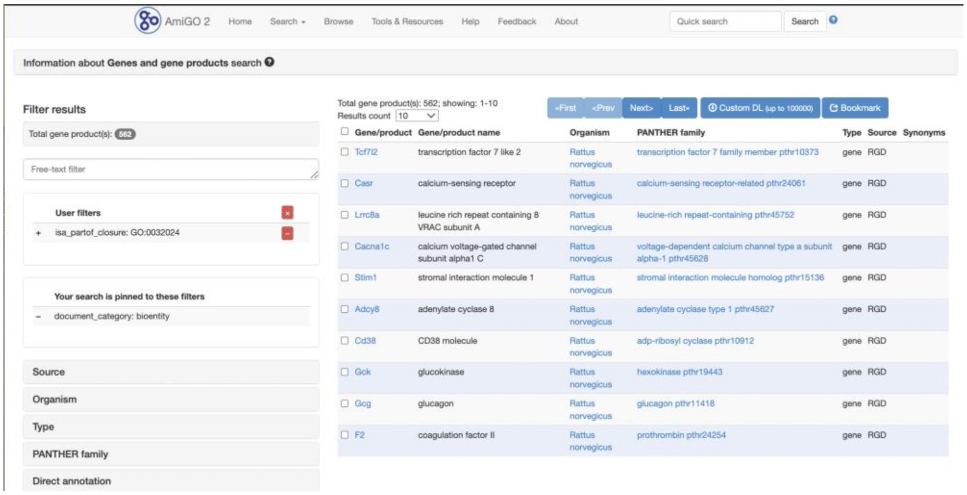
– document_category: annotation

Total annotations: 717; showing: 1-10

Results count 10

□ Gene/product	Gene/product name	Annotation qualifier	GO class (direct)	Annotation extension	Contributor	Organism	Evidence	Evidence with	PANTHER family	Type	Isoform
□ Tcf7l2	transcription factor 7 like 2		positive regulation of insulin secretion		RGD	Rattus norvegicus	ISO	MGI:1202879	transcription factor 7 family member pthr10373	gene	
□ Tcf7l2	transcription factor 7 like 2		positive regulation of insulin secretion		RGD	Rattus norvegicus	ISO	UniProtKB:Q9NQB0	transcription factor 7 family member pthr10373	gene	
□ Casr	calcium-sensing		positive regulation		RGD	Rattus norvegicus	IMP		calcium-sensing	gene	

4. Clicking “GO class” link in step 3 will take me to page for selected GO class



Information about Genes and gene products search

Filter results

Total gene product(s): 562

User filters

+ isa_partof_closure: GO:0032024

Your search is pinned to these filters

– document_category: bioentity

Total gene product(s): 562; showing: 1-10

Results count 10

□ Gene/product	Gene/product name	Organism	PANTHER family	Type	Source	Synonyms
□ Tcf7l2	transcription factor 7 like 2	Rattus norvegicus	transcription factor 7 family member pthr10373	gene	RGD	
□ Casr	calcium-sensing receptor	Rattus norvegicus	calcium-sensing receptor-related pthr24061	gene	RGD	
□ Lrrc8a	leucine rich repeat containing 8 VRAC subunit A	Rattus norvegicus	leucine-rich repeat-containing pthr45752	gene	RGD	
□ Cacna1c	calcium voltage-gated channel subunit alpha1 C	Rattus norvegicus	voltage-dependent calcium channel type a subunit alpha-1 pthr45628	gene	RGD	
□ Stim1	stromal interaction molecule 1	Rattus norvegicus	stromal interaction molecule homolog pthr15136	gene	RGD	
□ Adcy8	adenylate cyclase 8	Rattus norvegicus	adenylate cyclase type 1 pthr45627	gene	RGD	
□ Cd38	CD38 molecule	Rattus norvegicus	adp-ribosyl cyclase pthr10912	gene	RGD	
□ Gck	glucokinase	Rattus norvegicus	hexokinase pthr19443	gene	RGD	
□ Gcg	glucagon	Rattus norvegicus	glucagon pthr11418	gene	RGD	
□ F2	coagulation factor II	Rattus norvegicus	prothrombin pthr24254	gene	RGD	

5. Clicking on “Link” to all genes and gene products annotated to.... under “Related” in step 4 will take me to table showing genes and gene products with annotation to selected GO class. Use this data to populate proposed table 2 of database.

Sources:

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., & Lewis, S. (2008). Amigo: Online access to ontology and annotation data. *Bioinformatics*, *25*(2), 288–289. https://doi.org/10.1093/bioinformatics/btn615

Hill, D. P., Smith, B., McAndrews-Hill, M. S., & Blake, J. A. (2008). Gene ontology annotations: What they mean and where they come from. *BMC Bioinformatics*, *9*(S5). https://doi.org/10.1186/1471-2105-9-s5-s2

Zhao, Y., Wang, J., Chen, J., Zhang, X., Guo, M., & Yu, G. (2020). A literature review of gene function prediction by modeling gene ontology. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/fgene.2020.00400