

Genes

Match input gene ID's with corresponding Entrez ID's.

Inputs - Data: Data set.

Outputs - Data: Instances with meta data that the user has manually selected in the widget. - Genes: All genes from the input with included gene info summary and matcher result.

To work with widgets in the bioinformatics add-on data sets must be properly annotated. We need to specify: - Location of genes in a table (rows, columns) - ID from the [NCBI Gene database](#) (Entrez ID) - Organism (Taxonomy ID)

Genes is a useful widget that presents information on the genes from the [NCBI Gene database](#) and outputs annotated data table. You can also select a subset and feed it to other widgets. By clicking on the gene Entrez ID in the list, you will be taken to the NCBI site with the information on the gene.

Gene Name Matcher

Info

6370 genes in input data
5848 genes match Entrez database
522 genes with match conflicts

Organism

Saccharomyces cerevisiae

Gene IDs in the input data

Stored in data column

class

Stored as feature (column) names

Output

Exclude unmatched genes
 Replace feature IDs with gene names

Commit Automatically

Filter: 5

Input ID	Entrez ID	Name	Description	Synonyms	Other IDs
YGR270W	853186	YTA7	Yta7p		SGD: S000003502
YIL075C	854735	RPN2	proteasome ...	SEN3	SGD: S000001337
YDL007W	851557	RPT2	proteasome ...	YHS4, YTA5	SGD: S000002165
YER094C	856830	PUP3	proteasome ...	SCS32	SGD: S000000896
YFR004W	850554	RPN11	proteasome ...	MPR1	SGD: S000001900
YDR427W	852037	RPN9	proteasome ...	NAS7	SGD: S000002835
YKL145W	853712	RPT1	proteasome ...	CIM5, YTA3	SGD: S000001628
YGL048C	852834	RPT6	proteasome ...	CIM3, CRL3,...	SGD: S000003016
YFR050C	850611	PRE4	proteasome ...		SGD: S000001946
YDL097C	851461	RPN6	proteasome ...	NAS4	SGD: S000002255
YOR259C	854433	RPT4	proteasome ...	CRL13, PCS1...	SGD: S000005785
YPR108W	856223	RPN7	proteasome ...		SGD: S0000006312

IDs from the input data without corresponding Entrez ID

-- EMPTY, Q0010, Q0017, Q0032, Q0092, Q0142, Q0143, Q0144, Q0182, Q0297, YAL004W, YAL034CB, YAL035CA, YAL042CA, YAL043CA, YAL045C, YAL056CA, YAL058CA, YAL066W, YAL069W, YAR030C, YAR047C, YAR053W, YAR060C, YAR069C, YAR070C, YAR073W, YAR075W, YBL012C, YBL053W, YBL06...

Example

First we load *brown-selected.tab* (from *Browse documentation data sets*) with the **File** widget and feed our data to the Genes widget. Orange recognized the organism correctly, but we have to tell it where our gene labels are. To do this, we tick off *Stored as feature (column) name* and select *gene* attribute from the list. Then we can observe gene info provided from the NCBI Gene data database. In the **Data Table** we can see the Entrez ID column included as a meta attribute. The data is also properly annotated (see *Data Attributes* section in **Data Info** widget).

The screenshot displays the Orange Data Mining interface with the following components:

- Gene Name Matcher Widget:**
 - Info:** 186 genes in input data, 186 genes match Entrez database, 0 genes with match conflicts. Organism: *Saccharomyces cerevisiae*.
 - Filter:** A table listing input IDs, Entrez IDs, names, descriptions, and synonyms for various proteasome subunits.
 - Output:** Includes options for 'Stored as feature (column) names' (checked) and 'Replace feature IDs with gene names' (checked).
- Data Table Widget:**

	function	gene	Entrez ID	alpha 0	alpha 7	alpha 14	alpha 21
1	Proteas	YGR270W	853186	?	-0.023	0.057	0.007
2	Proteas	YIL075C	854735	-0.031	-0.031	-0.060	0.037
3	Proteas	YDL007W	851557	-0.013	?	0.067	-0.025
4	Proteas	YER094C	856830	0.003	0.025	0.067	0.083
5	Proteas	YFR004W	850554	-0.068	-0.003	-0.041	0.022
6	Proteas	YDR427W	852037	-0.012	-0.009	-0.009	?
7	Proteas	YKL145W	853712	0.012	0.008	-0.006	-0.025
8	Proteas	YGL048C	852834	0.067	-0.064	0.011	0.022
9	Proteas	YFR050C	850611	0.093	0.027	0.044	0.066
10	Proteas	YDL097C	8514			0.050	0.019
11	Proteas	YOR259C	8544			0.030	-0.007
12	Proteas	YPR108W	8562			0.073	0.064
13	Proteas	YFR071W	8567			0.043	-0.002
- Data Info Widget:**
 - Data Set Size:** Rows: 186, Columns: 82.
 - Features:** Discrete: (none), Numeric: 79.
 - Targets:** Discrete outcome with 3 values.
 - Meta Attributes:** Discrete: (none), Numeric: (none), Textual: 2.
 - Location:** Data is stored in memory.
 - Data Attributes:** taxonomy_id: 4932, gene_as_attribute_name: False, gene_id_column: Entrez ID.