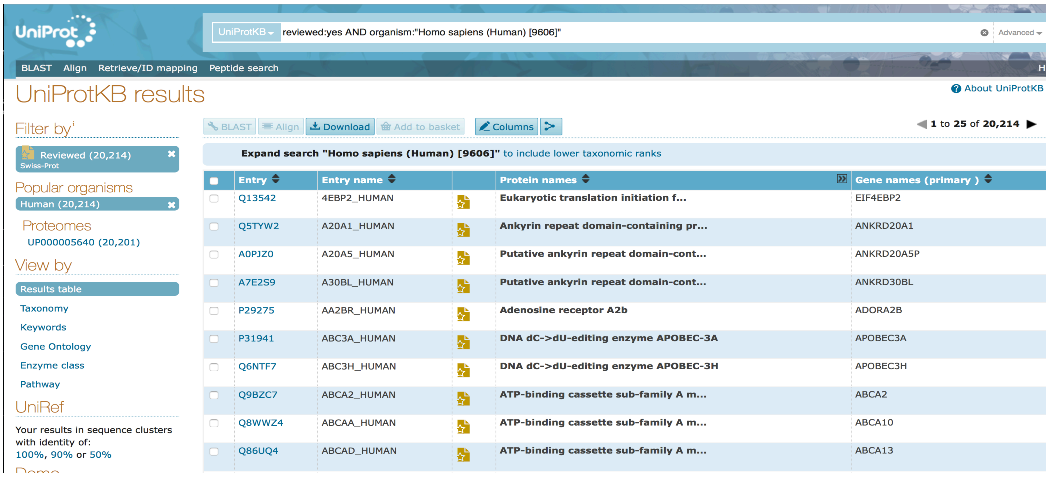
In order to use my program retrieving data for high confidence orthologue sequences of each gene, first, users need a gene name text file that contain only one gene name at each line. Additionally, if users only have csv files download from Uniprot human genome, make sure select column 'Gene names (primary)'. Users can run <python 00_process_csv.py> to process csv files into a text file that contain only primary gene names. If there are more than thousands of gene names in a single file, users can run <python 01_split_txt_check_dup.py> to split single gene name text file into multiple by users' preference of numbers in each split file. Once users have their gene text file, Users can retrieve orthologue data by run <python main.py gene_txt_dir_path gene_txt_name (y/n)>. The last argument is indication of whether this gene text file is generated from previous script <01_split_txt_check_dup.py>, if yes(y) then it will not check for duplicated gene names in this file.

Scripts are written in python 2.7.13 and tested on python 2.7.13. Make sure you are using the correct python version. Before running this program, make sure python library requests is installed on your computer. You can use pip command to install requests library.

Following are screenshots with better visualization of above description:



reviewed:yes AND organism:"Homo sapiens (Human) [9606]"

UniProtKB results

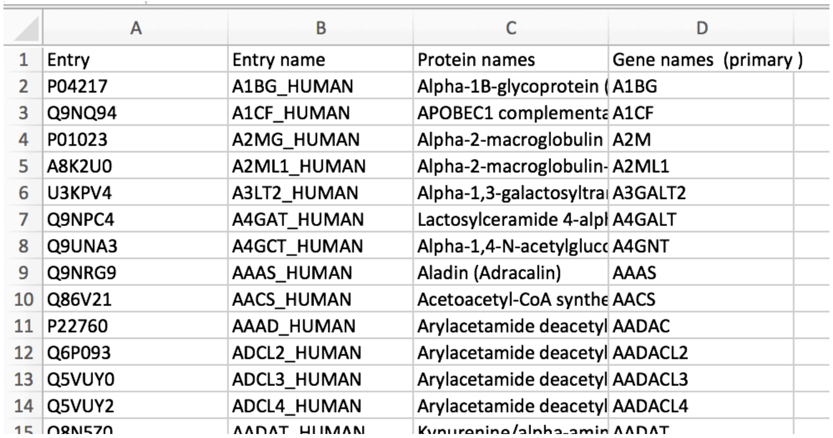
1 to 25 of 20,214

Expand search "Homo sapiens (Human) [9606]" to include lower taxonomic ranks

Your results in sequence clusters with identity of: 100%, 90% or 50%

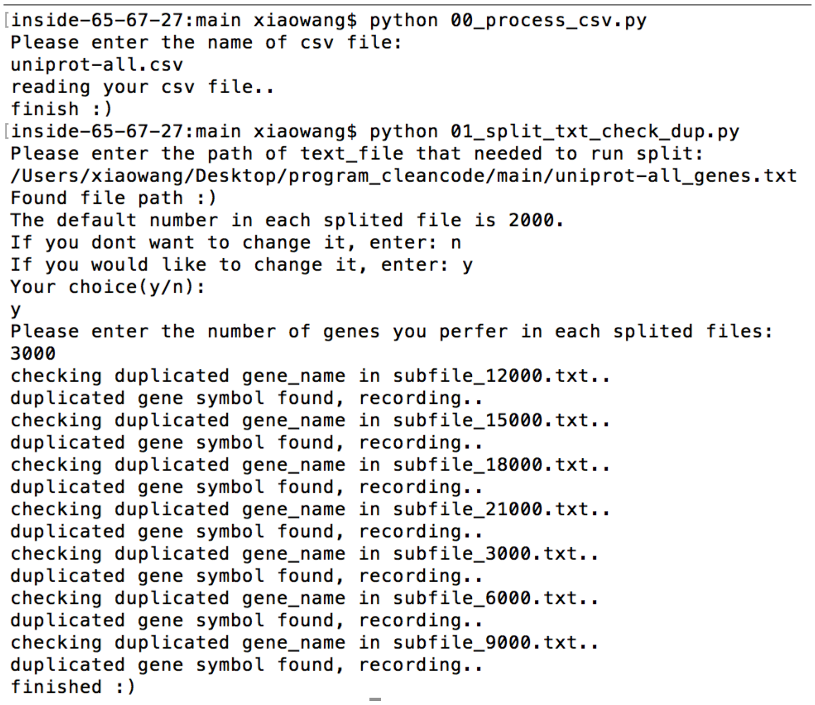
Entry	Entry name	Protein names	Gene names (primary)
Q13542	4EBP2_HUMAN	Eukaryotic translation initiation f...	EIF4EBP2
Q5TYW2	A20A1_HUMAN	Ankyrin repeat domain-containing pr...	ANKRD20A1
A0PJZ0	A20A5_HUMAN	Putative ankyrin repeat domain-cont...	ANKRD20A5P
A7E2S9	A30BL_HUMAN	Putative ankyrin repeat domain-cont...	ANKRD30BL
P29275	AA2BR_HUMAN	Adenosine receptor A2b	ADORA2B
P31941	ABC3A_HUMAN	DNA dC->dU-editing enzyme APOBEC-3A	APOBEC3A
Q6NTF7	ABC3H_HUMAN	DNA dC->dU-editing enzyme APOBEC-3H	APOBEC3H
Q9BZC7	ABCA2_HUMAN	ATP-binding cassette sub-family A m...	ABCA2
Q8WWZ4	ABCAA_HUMAN	ATP-binding cassette sub-family A m...	ABCA10
Q86UQ4	ABCAD_HUMAN	ATP-binding cassette sub-family A m...	ABCA13

Screenshot of Uniprot human proteome project with Gene name (primary) column.



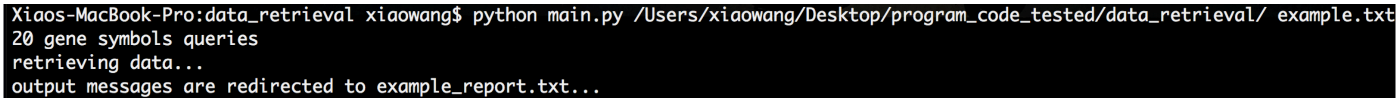
	A	B	C	D
1	Entry	Entry name	Protein names	Gene names (primary)
2	P04217	A1BG_HUMAN	Alpha-1B-glycoprotein (	A1BG
3	Q9NQ94	A1CF_HUMAN	APOBEC1 complementa	A1CF
4	P01023	A2MG_HUMAN	Alpha-2-macroglobulin	A2M
5	A8K2U0	A2ML1_HUMAN	Alpha-2-macroglobulin-	A2ML1
6	U3KPV4	A3LT2_HUMAN	Alpha-1,3-galactosyltra	A3GALT2
7	Q9NPC4	A4GAT_HUMAN	Lactosylceramide 4-alpl	A4GALT
8	Q9UNA3	A4GCT_HUMAN	Alpha-1,4-N-acetylgluc	A4GNT
9	Q9NRG9	AAAS_HUMAN	Aladin (Adracalin)	AAAS
10	Q86V21	AACS_HUMAN	Acetoacetyl-CoA synthe	AACS
11	P22760	AAAD_HUMAN	Arylacetamide deacetyl	AADAC
12	Q6P093	ADCL2_HUMAN	Arylacetamide deacetyl	AADACL2
13	Q5VUY0	ADCL3_HUMAN	Arylacetamide deacetyl	AADACL3
14	Q5VUY2	ADCL4_HUMAN	Arylacetamide deacetyl	AADACL4
15	Q8N5Z0	AADAT_HUMAN	Kynurenine/alpha-amin	AADAT

Screenshot of csv file download from Uniprot Web with selected columns.



```
[inside-65-67-27:main xiaowang$ python 00_process_csv.py
Please enter the name of csv file:
uniprot-all.csv
reading your csv file..
finish :)
[inside-65-67-27:main xiaowang$ python 01_split_txt_check_dup.py
Please enter the path of text_file that needed to run split:
/Users/xiaowang/Desktop/program_cleancode/main/uniprot-all_genes.txt
Found file path :)
The default number in each splited file is 2000.
If you dont want to change it, enter: n
If you would like to change it, enter: y
Your choice(y/n):
y
Please enter the number of genes you perfer in each splited files:
3000
checking duplicated gene_name in subfile_12000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_15000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_18000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_21000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_3000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_6000.txt..
duplicated gene symbol found, recording..
checking duplicated gene_name in subfile_9000.txt..
duplicated gene symbol found, recording..
finished :)
```

Screenshot of terminal output <00_process_csv.py> and <01_split_txt_check_dup.py>



```
Xiaos-MacBook-Pro:data_retrieval xiaowang$ python main.py /Users/xiaowang/Desktop/program_code_tested/data_retrieval/ example.txt
20 gene symbols queries
retrieving data...
output messages are redirected to example_report.txt...
```

Screen shot of data retrieving script <main.py> command line argument, and output messages.

Important notes:

High confidence orthologues sequence along with human reference sequence are written in <all_seqs_with_latin_names$gene_symbol.fasta>. Gene names with less than 11 orthologues does not have fasta file(filtered), and not enough orthologous gene names are recorded as txt file. Protein sequence IDs are also recorded in <all_seqs_with_protIDs$gene_symbol.fasta>. Duplicated species in each gene_names fasta files are checked and recorded as <duplicated_species_genename.txt>. Failed cases are already handled, and recorded as txt file. All files generated from program are manage into its own folder.

The speed of data retrieving is heavily depending on network speed and data size. Since I used REST (Representational State Transfer) API (application program interface) from Ensembl database, there is rate limiting on request times. Rate limiting is handled by waiting, however, it will reach maximum depth recursion if waited too many times. Be aware on the number of processes you are using, especially on static router server.

For further questions, send me an email:
xiaow.wang@mail.utoronto.ca