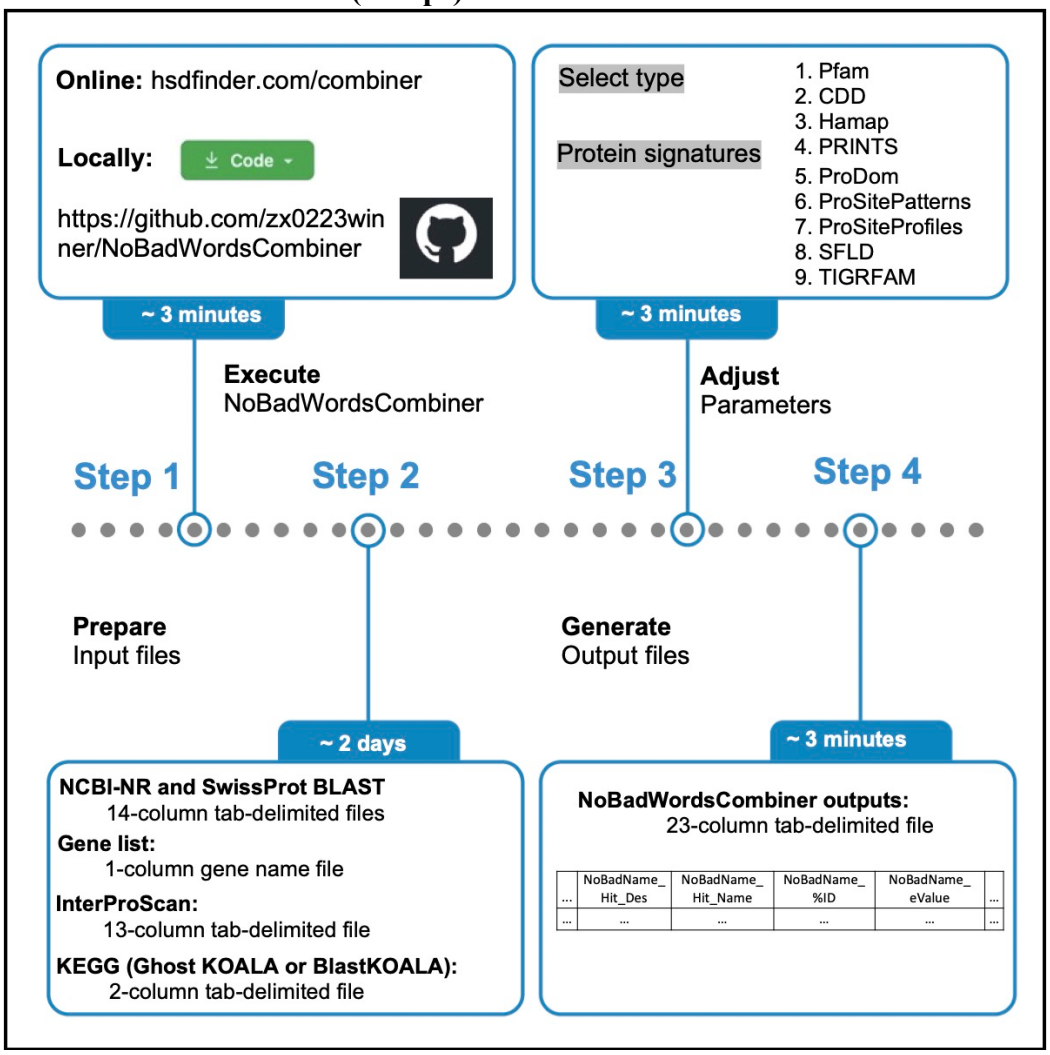
NoBadWordsCombiner Online Tutorial: Merge and minimize 'bad words' from BLAST hits against multiple eukaryotic gene annotation databases

1. Running the software online or locally.
2. Preparing the NCBI-NR and UniProtKB/Swiss-Prot protein BLAST-search result files.
3. Preparing the gene name list and a gene list with KO annotation from KEGG database.

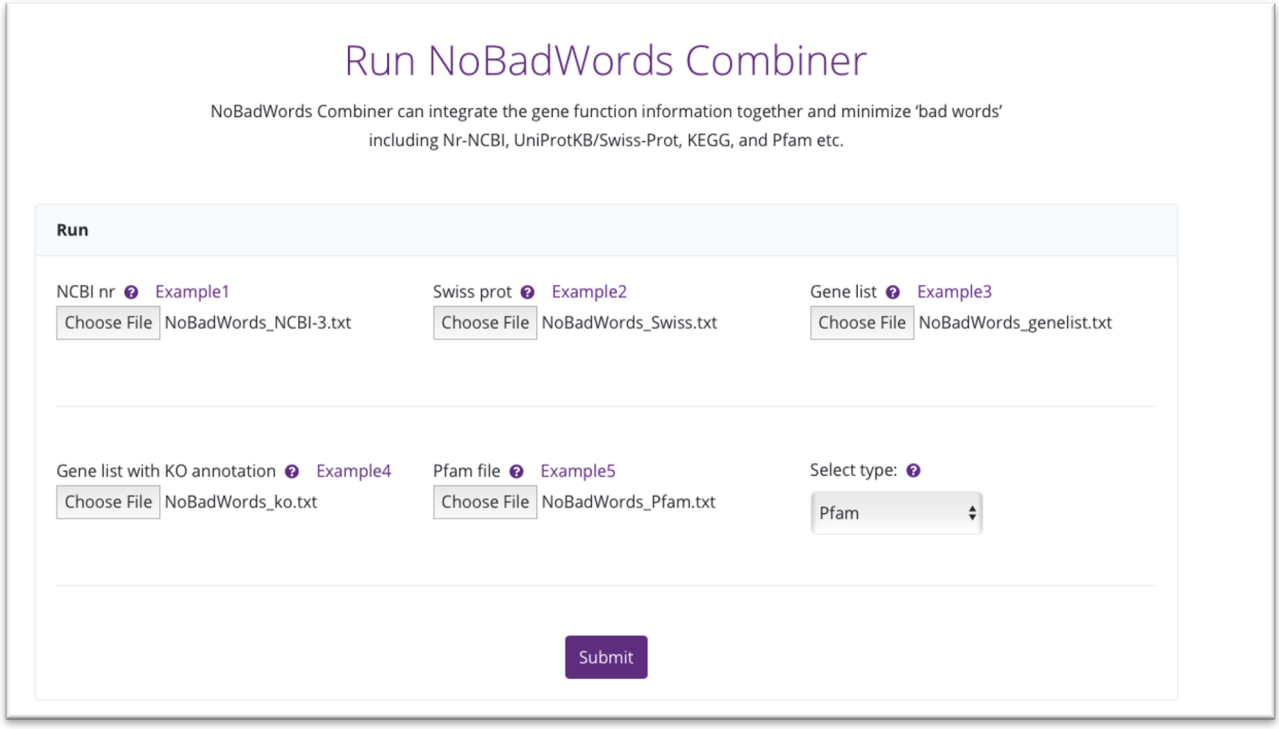
   Preparing the InterProScan search result file.
4. Output file of the NoBadWordsCombiner tool.

Workflow of HSDFinder (8 steps)

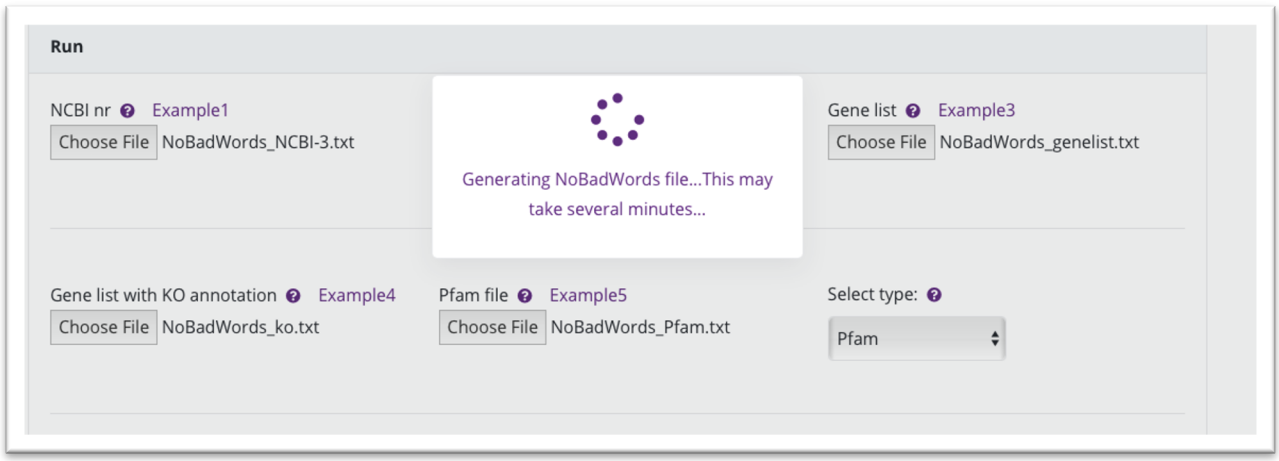


Step 1: Running the software online or locally
The NoBadWordsCombiner tool can either be operated online or locally within few minutes to process.

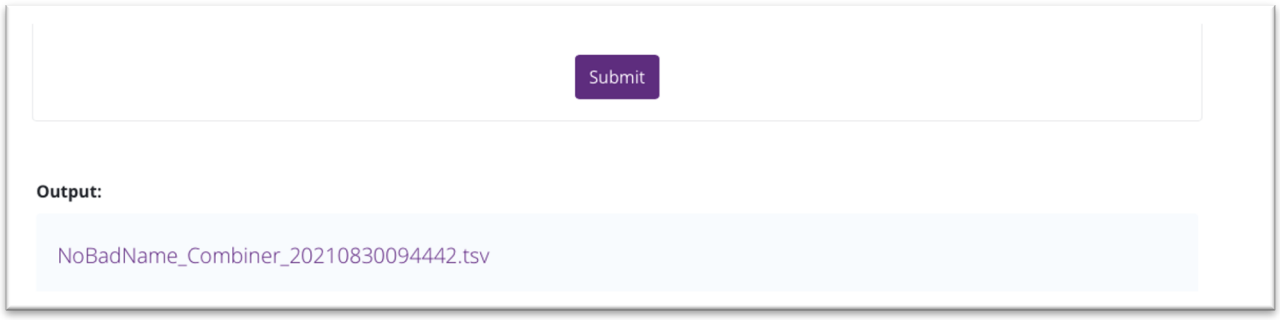
Online: User can simply click the submit button below after collecting and dragging in all the necessary input files. Within a few minutes, the result file will show up below the submission button. Then user can safely open and read the file referring to the protocol guideline.



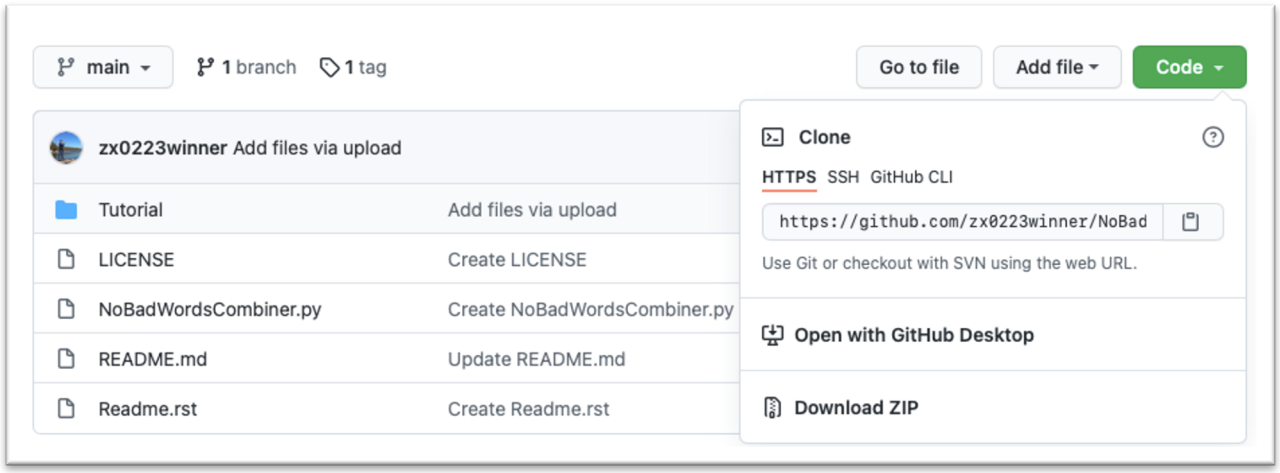
With the built-in example, it might take ~3mins to reach the analysis result.



The output file will turn out like this.



Or Locally: User can also download the code from GitHub to run it locally. Click the green Code button and select the Download Zip option.



Please refer to the Readme file to proceed the local environment running. Same to the online running, user can simply run the command on the dash shell with required documents.

Then user can run the code below:
> python NoBadWordsCombiner.py -n Input_1_NoBadWords_NCBI.txt -s Input_2_NoBadWords_Swiss.txt -g Input_3_NoBadWords_genelist.txt -k Input_4_NoBadWords_ko.txt -p Input_5_NoBadWords_Pfam.txt -t pfam -o Output-self-define-name.tsv

Or

> python NoBadWordsCombiner.py --ncbi_file Input_1_NoBadWords_NCBI.txt --swiss_file Input_2_NoBadWords_Swiss.txt --gene_file Input_3_NoBadWords_genelist.txt --ko_file Input_4_NoBadWords_ko.txt --pfam_file Input_5_NoBadWords_Pfam.txt –type pfam -output_file Output-self-define-name.tsv

If error occurs with requiring Python Pandas package, user can type the code below.

#Environmental Requirement: Pandas

#To collect pandas packages :
>sudo pip install pandas

Note: 'KO database and its category.keg' file shall be put in the same directory with NoBadWordsCombiner.py script.

Step 2: Preparing the NCBI-NR and UniProtKB/Swiss-Prot protein BLAST-search result files.

Note: This step might need up to 2 days with the test data. This is due to the queuing system in different databases such as InterProScan and KEGG. There is a possible steep learning curve for users with limited knowledge of bioinformatics, especially for those who are not familiar with the basic BLAST package and dash shell in a Linux/Unix environment.

Software requirements:

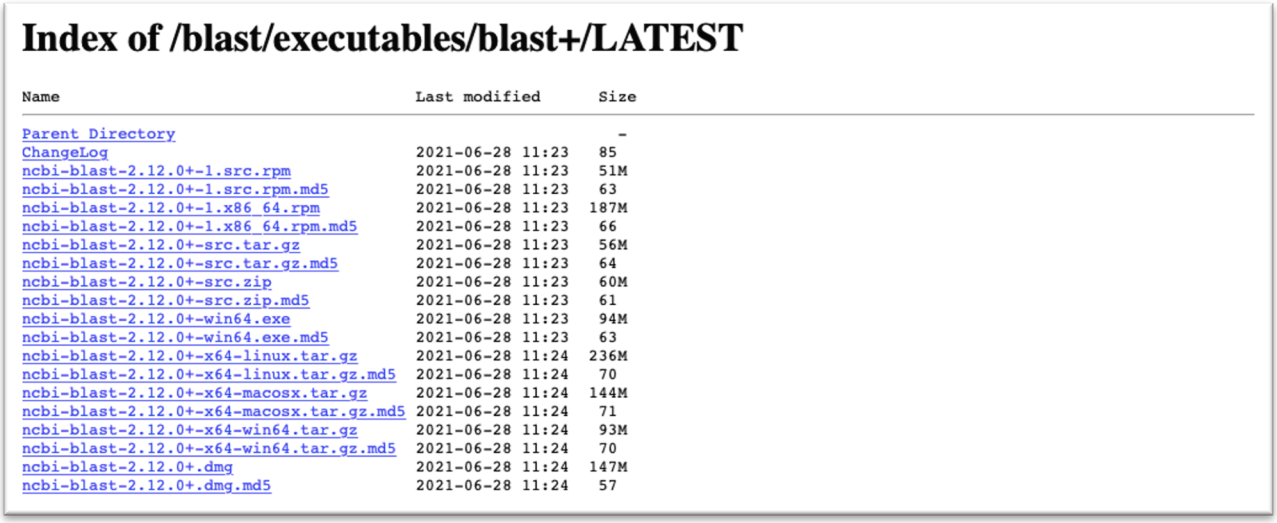
2.1.64-bit Linux Perl 5 (default on most Linux distributions) e.g., Ubuntu 16.04.7 LTS
>perl -v
If not the latest Perl 5, using the command below to update to the latest.
> sudo apt-get update
> sudo apt-get install perl

2.2.Python 3 (e.g., Python 3.8.5)
>python3 –version
If not the latest python3, using the command below to update to the latest.
> sudo apt-get update
> sudo apt-get install python3.8

2.3.Java JDK/JRE version 11
> java -version
If not the latest java v11, using the command below to update to the latest.
> sudo apt install default-jre

Preparing the NCBI-NR and UniProtKB/Swiss-Prot protein BLAST-search result files.

2.1.Download the BLAST Package via https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ . Please select the appropriate distribution based on your computer operating systems (Windows, MacOS or Linux) e.g., Linux: ncbi-blast-2.12.0+-x64-linux.tar.gz

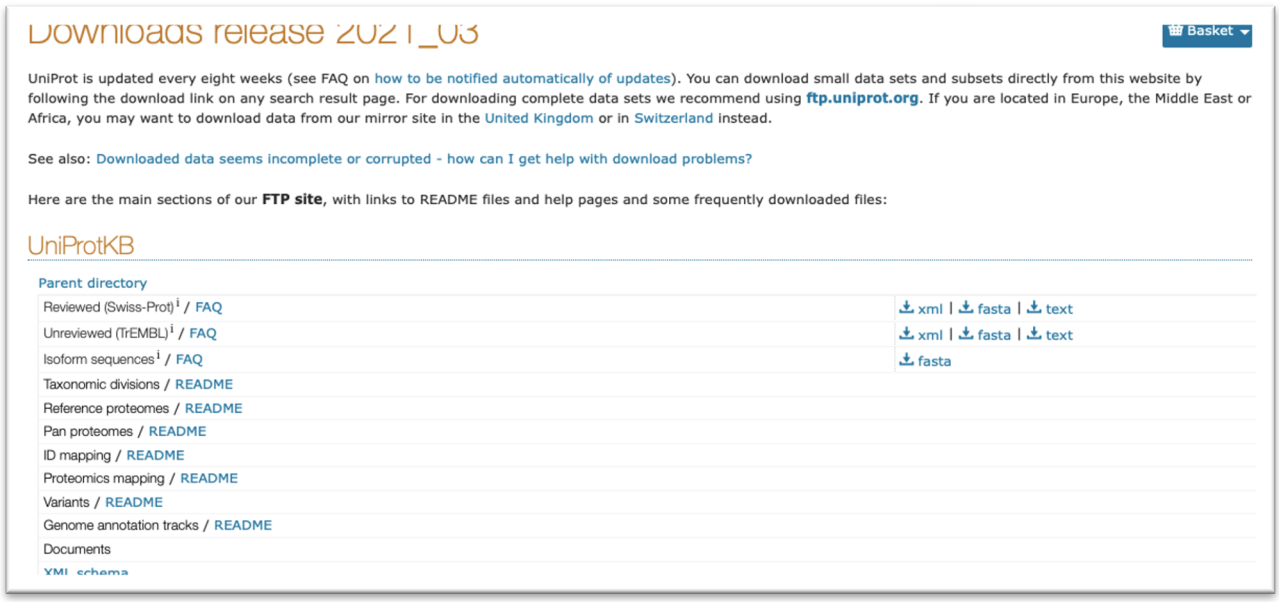


Index of /blast/executables/blast+/LATEST

Name	Last modified	Size
Parent Directory		-
ChangeLog	2021-06-28 11:23	85
ncbi-blast-2.12.0+-1.src.rpm	2021-06-28 11:23	51M
ncbi-blast-2.12.0+-1.src.rpm.md5	2021-06-28 11:23	63
ncbi-blast-2.12.0+-1.x86_64.rpm	2021-06-28 11:23	187M
ncbi-blast-2.12.0+-1.x86_64.rpm.md5	2021-06-28 11:23	66
ncbi-blast-2.12.0+-src.tar.gz	2021-06-28 11:23	56M
ncbi-blast-2.12.0+-src.tar.gz.md5	2021-06-28 11:23	64
ncbi-blast-2.12.0+-src.zip	2021-06-28 11:23	60M
ncbi-blast-2.12.0+-src.zip.md5	2021-06-28 11:23	61
ncbi-blast-2.12.0+-win64.exe	2021-06-28 11:23	94M
ncbi-blast-2.12.0+-win64.exe.md5	2021-06-28 11:23	63
ncbi-blast-2.12.0+-x64-linux.tar.gz	2021-06-28 11:24	236M
ncbi-blast-2.12.0+-x64-linux.tar.gz.md5	2021-06-28 11:24	70
ncbi-blast-2.12.0+-x64-macosx.tar.gz	2021-06-28 11:24	144M
ncbi-blast-2.12.0+-x64-macosx.tar.gz.md5	2021-06-28 11:24	71
ncbi-blast-2.12.0+-x64-win64.tar.gz	2021-06-28 11:24	93M
ncbi-blast-2.12.0+-x64-win64.tar.gz.md5	2021-06-28 11:24	70
ncbi-blast-2.12.0+.dmg	2021-06-28 11:24	147M
ncbi-blast-2.12.0+.dmg.md5	2021-06-28 11:24	57

2.2. Unzip the “NoBadWordsCombiner_file_examples.zip” file from https://github.com/zx0223winner/NoBadWordsCombiner/tree/main/Tutorial, the file named “Chlamydomonas_UWO241_protein.fasta” is the example FASTA file.
# Display the first ten rows of the FASTA file.
$ head Chlamydomonas_UWO241_protein.fasta
>g1.t1
MAATVENVVERVKSFSSVVRGVKSGKPDGATTQLVQETIEILATYCDFEEVVPVCLKFLDEVLTA
APQTSTLIRLEGGAK
IFPSIIRNFMGVDASILALCAKVMCKCASGSPAMQHHLVKEKGLPTLLLSCCSAHAGEPAVVGPLL
EVLVALARYSKGAT
ALSNANLVHACKELLVGLMGHWHAFGMVLKLIKSVMKHEGPCLAALKAGEVVRLLLGVARLVS
RMPDQRKLLKRASRTLW
VLSQRSLHPLPEMELNWPHTHTHTHTHTHTHT

2.3. Set up the manually curated UniProtKB/Swiss-Prot database and computationally calculated NCBI-NR database. The Reviewed “uniprot_sprot.fasta.gz” file can be downloaded directly from https://www.uniprot.org/downloads#uniprotkblink



Downloads release 2021_03

Basket

UniProt is updated every eight weeks (see FAQ on how to be notified automatically of updates). You can download small data sets and subsets directly from this website by following the download link on any search result page. For downloading complete data sets we recommend using **ftp.uniprot.org**. If you are located in Europe, the Middle East or Africa, you may want to download data from our mirror site in the United Kingdom or in Switzerland instead.

See also: Downloaded data seems incomplete or corrupted - how can I get help with download problems?

Here are the main sections of our **FTP site**, with links to README files and help pages and some frequently downloaded files:

UniProtKB

Parent directory	
Reviewed (Swiss-Prot) / FAQ	xml | fasta | text
Unreviewed (TrEMBL) / FAQ	xml | fasta | text
Isoform sequences / FAQ	fasta
Taxonomic divisions / README	
Reference proteomes / README	
Pan proteomes / README	
ID mapping / README	
Proteomics mapping / README	
Variants / README	
Genome annotation tracks / README	
Documents	
XML schema	

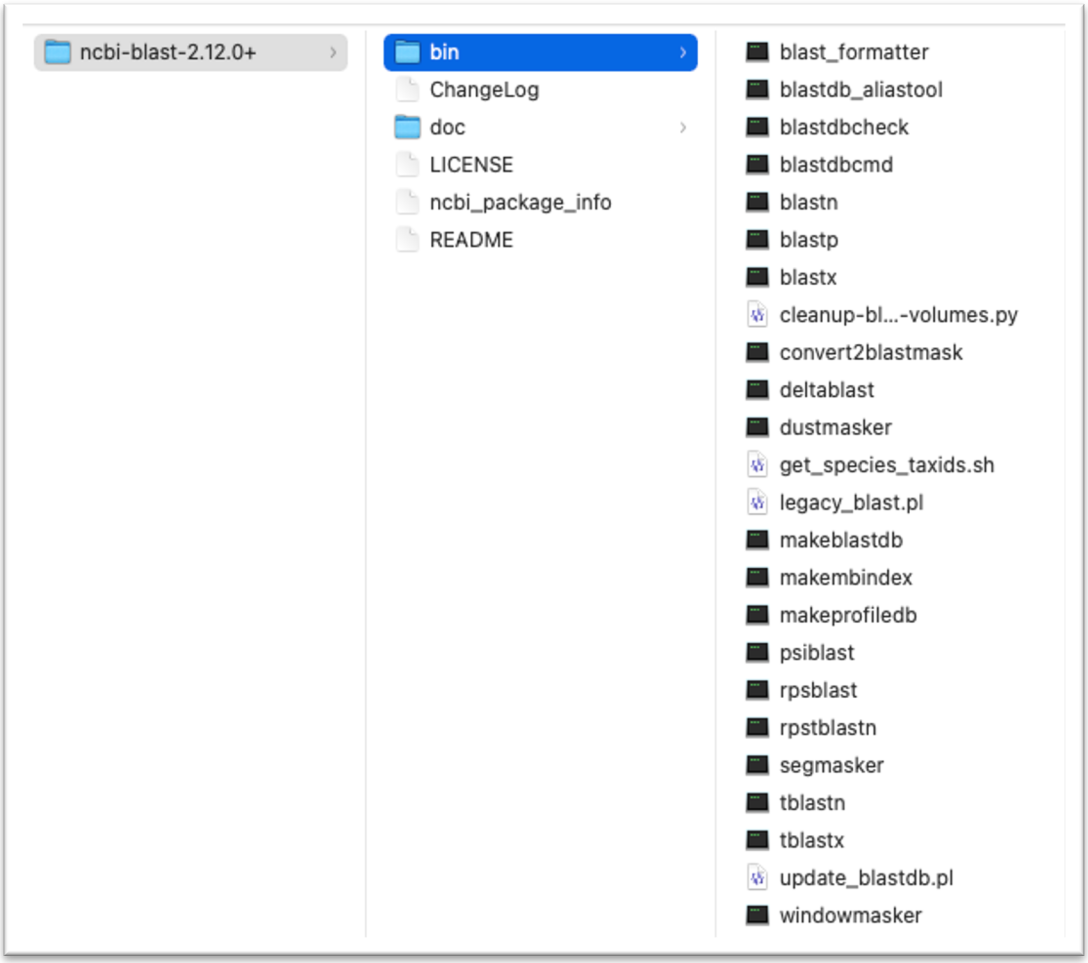
2.4. The makeblastdb command is from the bin of Blast package. The makeblastdb command is to construct a protein database by taking in the FASTA file with the parameter (-in), setting up the database type (e.g., protein) with the parameter (-dbtype protein), and titling the name of database (e.g., uniprot_prot_database) with parameters (-title database_name). -out option is the database ouput name (e.g., uniprot_db)

>./makeblastdb -in uniprot_sprot.fasta -dbtype prot -title uniprot_prot_database -out uniprot_db

```
/home/xizhang@perun9> makeblastdb -in uniprot_prot.fasta -dbtype prot -title uniprot_prot_database -out uniprot_db


Building a new DB, current time: 08/31/2021 12:20:22
New DB name:   /misc/home/xizhang/uniprot_db
New DB title:  uniprot_prot_database
Sequence type: Protein
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 16325 sequences in 0.721572 seconds.
/home/xizhang@perun9>
```

It will yield several database format files with name of uniport_db.phr, uniport_db.pin, uniport_db.psp.



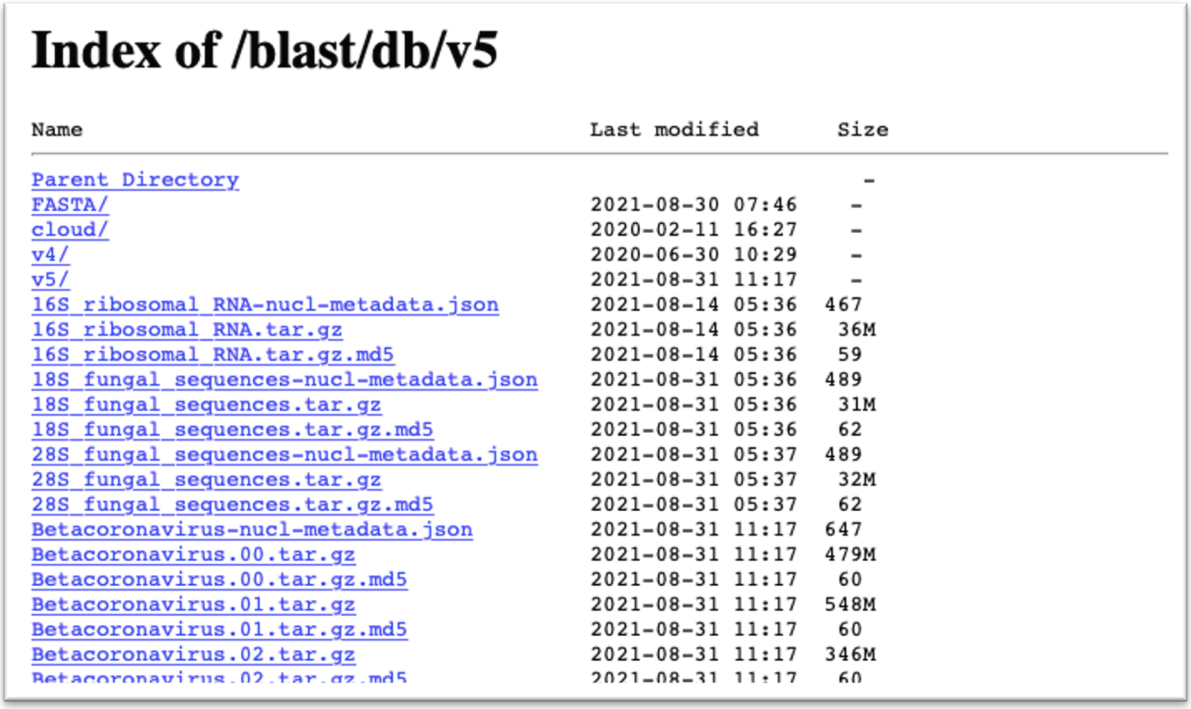
2.5. Using BLASTP search option to blast the amino acid sequences against uniport_db database.
> ./blastp -query Chlamydomonas_UWO241_protein.fasta -db uniprot_db -out BLASTP_UWO241_uniprot.xml -evalue 1e-5 -outfmt 5

Note: The BLAST XML file (-outfmt 5) can include useful information comparing to the BLAST Tabular file (-outfmt 6), such as the aligned sequence, the sequence of the hit, and the description of hits into the database. However, the XML format is not human-readable.

2.6. Users will need to employ a commonly used parser (*Blastxml_to_tabular.py*) from the link(https://github.com/zx0223winner/NoBadWordsCombiner/tree/main/Tutorial), which is a custom python script, to convert a BLAST XML file to a desired tabular output (tab-delimited file).
> python blastxml_to_tabular.py -c qseqid,qlen,salltitles,sseqid,slen,bitscore,qframe,pident,evalue,qstart,qend,sstart,send,length BLASTP_UWO241_uniprot.xml > BLASTP_UWO241_uniprot.tsv

After these many steps, we can acquire our one of the input files i.e, BLASTP_UWO241_uniprot.tsv.

2.7. Similar to above, users are going to prepare the tedious, but informative NCBI-NR database BLAST serach. To download the NCBI-NR v5 databases, use the Perl script update_blastdb.pl which is in the downloaded BLAST+ package (https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/). Check the step 4.
This command will download the NCBI-NR database (https://ftp.ncbi.nlm.nih.gov/blast/db/v5) .



Index of /blast/db/v5

Name	Last modified	Size
Parent Directory		-
FASTA/	2021-08-30 07:46	-
cloud/	2020-02-11 16:27	-
v4/	2020-06-30 10:29	-
v5/	2021-08-31 11:17	-
16S_ribosomal_RNA-nucl-metadata.json	2021-08-14 05:36	467
16S_ribosomal_RNA.tar.gz	2021-08-14 05:36	36M
16S_ribosomal_RNA.tar.gz.md5	2021-08-14 05:36	59
18S_fungal_sequences-nucl-metadata.json	2021-08-31 05:36	489
18S_fungal_sequences.tar.gz	2021-08-31 05:36	31M
18S_fungal_sequences.tar.gz.md5	2021-08-31 05:36	62
28S_fungal_sequences-nucl-metadata.json	2021-08-31 05:37	489
28S_fungal_sequences.tar.gz	2021-08-31 05:37	32M
28S_fungal_sequences.tar.gz.md5	2021-08-31 05:37	62
Betacoronavirus-nucl-metadata.json	2021-08-31 11:17	647
Betacoronavirus.00.tar.gz	2021-08-31 11:17	479M
Betacoronavirus.00.tar.gz.md5	2021-08-31 11:17	60
Betacoronavirus.01.tar.gz	2021-08-31 11:17	548M
Betacoronavirus.01.tar.gz.md5	2021-08-31 11:17	60
Betacoronavirus.02.tar.gz	2021-08-31 11:17	346M
Betacoronavirus.02.tar.gz.md5	2021-08-31 11:17	60

2.8. User can first check all available database via the command below.

> perl update_blastdb.pl --blastdb_version 5 –showall

This will give the results like this:

Connected to NCBI; downloading BLASTDBv5

human_genome

landmark

…

2.9. User can then run the command below to automatedly download the nr database which includes 55 volumes of data (>100 Gb). Or user can manually download these 110 files (i.e., nr.00.tar.gz, nr.00.tar.gz.md5, etc.) from the link: https://ftp.ncbi.nlm.nih.gov/blast/db/v5.

> perl update_blastdb.pl --blastdb_version 5 nr --decompress

This will bring the results like this:

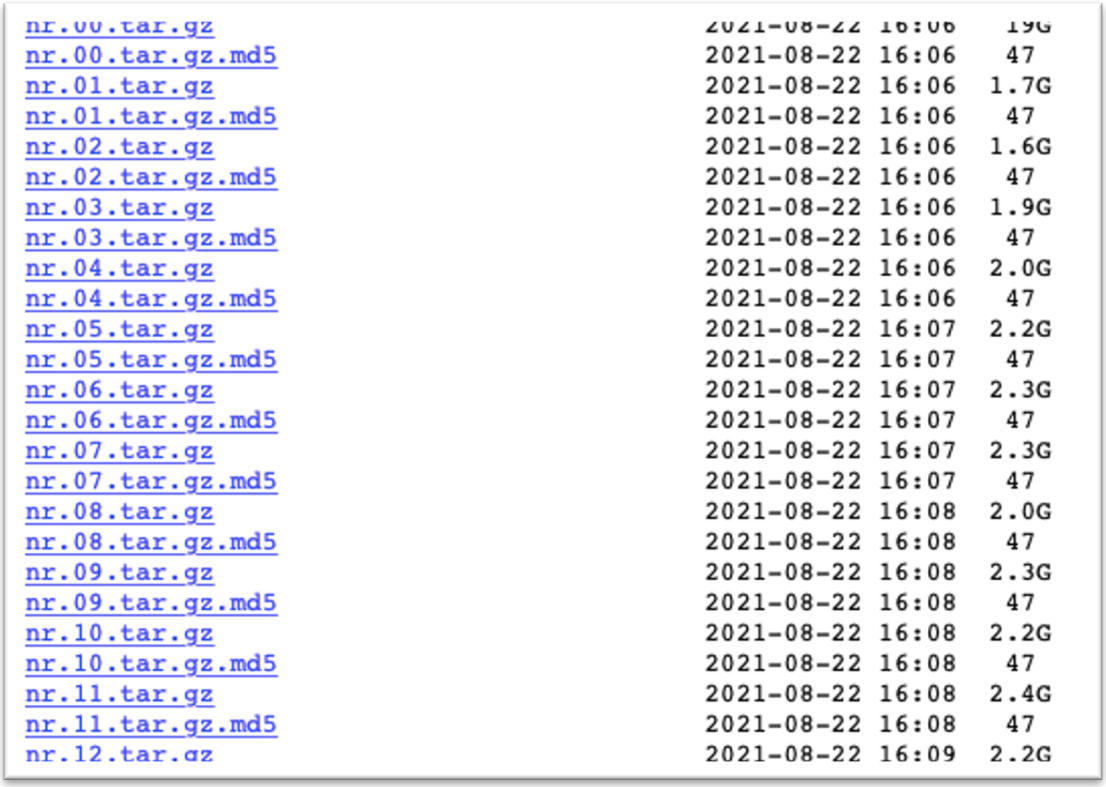
Connected to NCBI; downloading BLASTDBv5

Downloading nr (55 volumes) ...

Downloading nr.00.tar.gz...

Downloading nr.00.tar.gz.md5

#...



nr.00.tar.gz	2021-08-22 16:06	19G
nr.00.tar.gz.md5	2021-08-22 16:06	47
nr.01.tar.gz	2021-08-22 16:06	1.7G
nr.01.tar.gz.md5	2021-08-22 16:06	47
nr.02.tar.gz	2021-08-22 16:06	1.6G
nr.02.tar.gz.md5	2021-08-22 16:06	47
nr.03.tar.gz	2021-08-22 16:06	1.9G
nr.03.tar.gz.md5	2021-08-22 16:06	47
nr.04.tar.gz	2021-08-22 16:06	2.0G
nr.04.tar.gz.md5	2021-08-22 16:06	47
nr.05.tar.gz	2021-08-22 16:07	2.2G
nr.05.tar.gz.md5	2021-08-22 16:07	47
nr.06.tar.gz	2021-08-22 16:07	2.3G
nr.06.tar.gz.md5	2021-08-22 16:07	47
nr.07.tar.gz	2021-08-22 16:07	2.3G
nr.07.tar.gz.md5	2021-08-22 16:07	47
nr.08.tar.gz	2021-08-22 16:08	2.0G
nr.08.tar.gz.md5	2021-08-22 16:08	47
nr.09.tar.gz	2021-08-22 16:08	2.3G
nr.09.tar.gz.md5	2021-08-22 16:08	47
nr.10.tar.gz	2021-08-22 16:08	2.2G
nr.10.tar.gz.md5	2021-08-22 16:08	47
nr.11.tar.gz	2021-08-22 16:08	2.4G
nr.11.tar.gz.md5	2021-08-22 16:08	47
nr.12.tar.gz	2021-08-22 16:09	2.2G

Note: This step might take hours to download all the file depending on the Internet speed. It is better to leave the screen on while downloading, cause the process might be aborted due to the screen saver setting.

2.10. The downloaded NCBI-nr database can be BLAST directly without using makeblastdb command to redo it.

> ./blastp -query Chlamydomonas_UWO241_protein.fasta -db nr -out BLASTP_UWO241_NCBI-NR.xml -evalue 1e-5 -outfmt 5

#Similar to Step 6, the BLASTP_UWO241_NCBI-NR.xml file will be converted to BLASTP_UWO241_NCBI-NR.tsv via the command below.

> python blastxml_to_tabular.py -c qseqid,qlen,salltitles,sseqid,slen,bitscore,qframe,pident,evalue,qstart,qend,sstart,send,length BLASTP_UWO241_NCBI-NR.xml > BLASTP_UWO241_NCBI-NR.tsv

Note: user can also use '-taxids' or '-taxidlist'options to limit the BLAST search taxonomy if necessary. Please refer to the BLAST guide (blastdbv5_user_guide.pdf) from here: https://github.com/zx0223winner/NoBadWordsCombiner/tree/main/Tutorial

2.11. Outputs: This will give two BLAST result files formed by 14-column spreadsheets including the key information from query name to percentage identity etc.

The 14-column explanation of parsed BLAST search result files.

a. QueryAcc (e.g., g2.t1)
b. Query_Length (e.g., 399)
c. HitDescription (e.g., ankyrin, partial [Anaeromyces robustus])
d. HitName (e.g., gi|1183350135|gb|ORX78377.1|)
e. HitLength (e.g., 235)
f. HitBits (e.g., 65.4698)
g. HSP_rank (e.g., 1)
h. %ID (e.g., 40.2298851)
i. eValue (e.g., 3.61E-10)
j. Query_Start (e.g., 19)
k. Query_end (e.g., 279)
l. Hit_start (e.g., 18)
m. Hit_end (e.g., 96)
n. HSP_length (e.g., 87)

Step 3. Preparing the gene name list and a gene list with KO annotation from KEGG database. Preparing the InterProScan search result file.

3.1.Users can first test the function of grep via the command below:

> grep ">" Chlamydomonas_UWO241_protein.fasta| wc

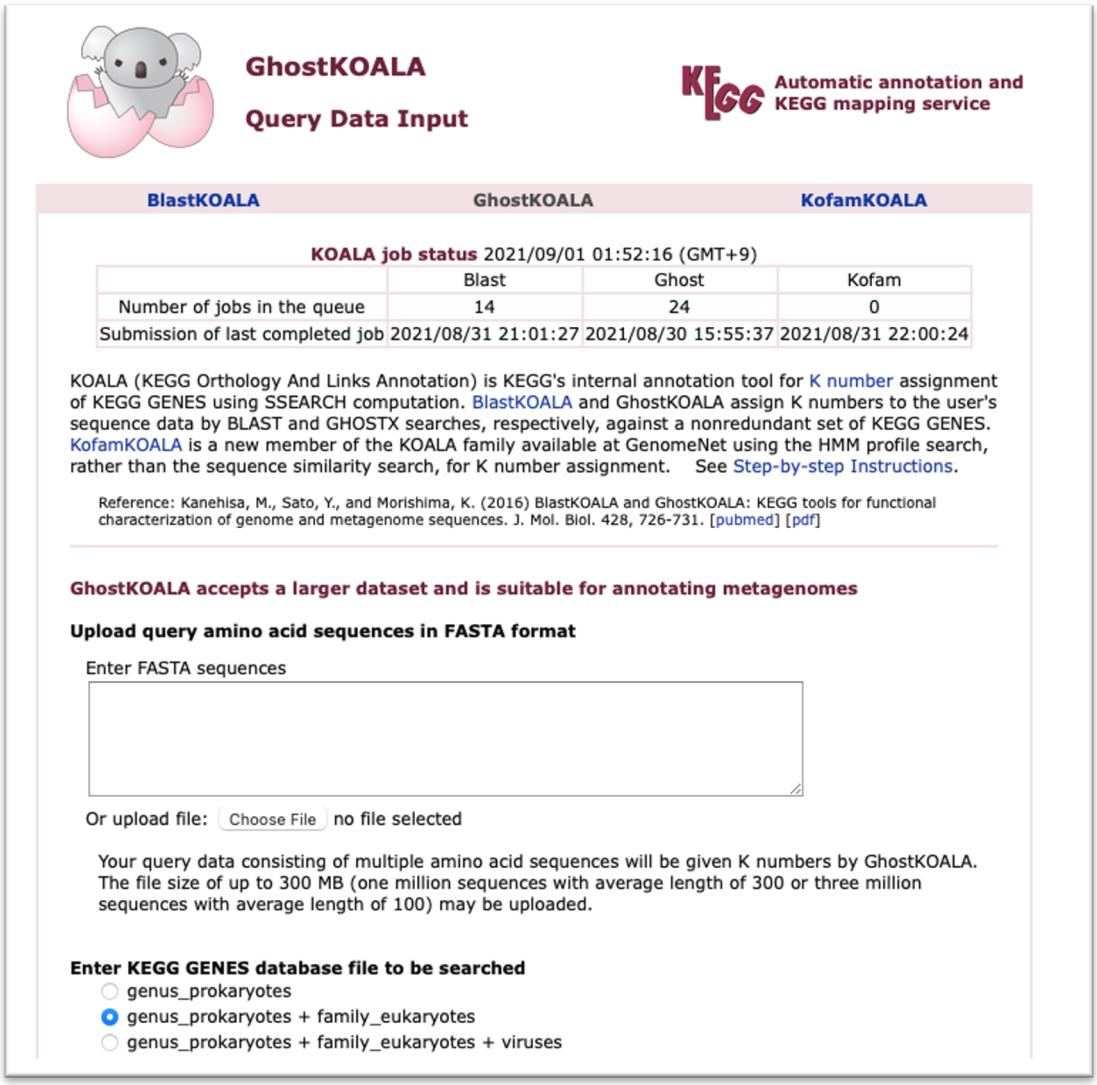
This should turn out the results as follows:

16325 16325 168617

3.2.Users can then carry out the following step to acquire a gene name file.

> grep ">" Chlamydomonas_UWO241_protein.fasta | sed 's/>//g' > UWO241-gene_name_list.txt

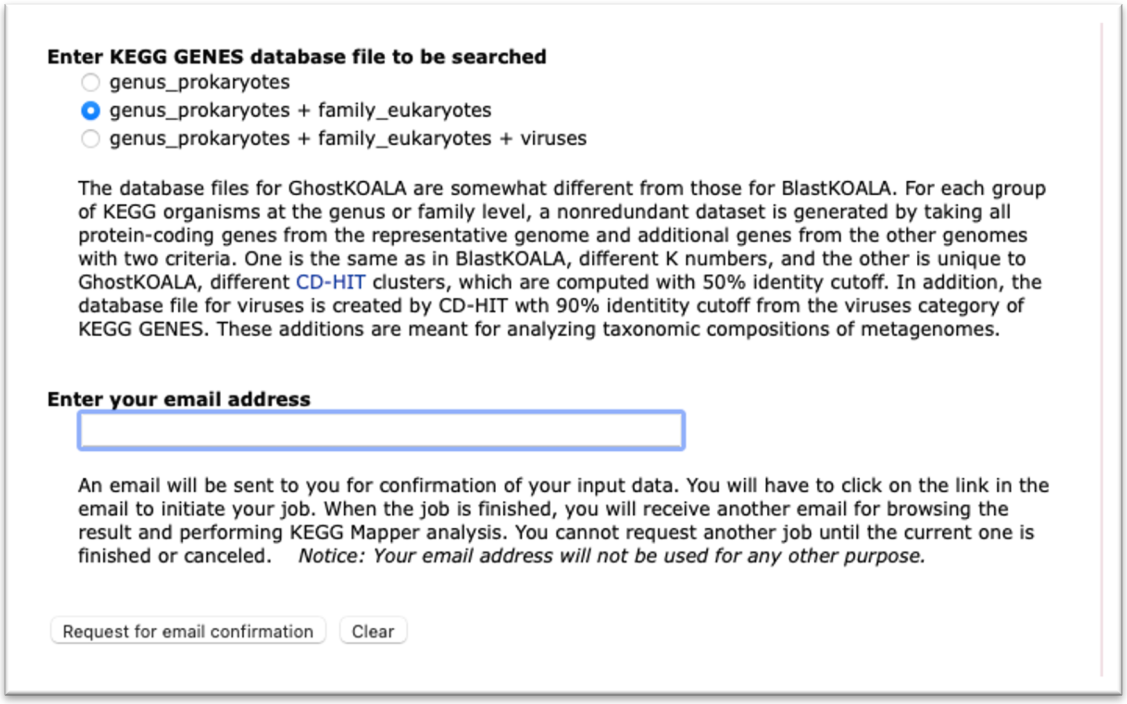
3.3. As for a gene list with KO annotation, users have the option to use the Ghost KOALA (Genome size ≥ 300MB) or BlastKOALA analysis tool of KEGG to acquire the KO annotation file of the genome (https://www.kegg.jp/ghostkoala/). Below, we provide the necessary steps for using the tools:



3.4. GhostKOALA accepts a larger dataset (e.g., 300 Mb) and is suitable for annotating metagenomes.

a. Upload query amino acid sequences in FASTA format.
b. Enter KEGG GENES database file to be searched.
c. Enter your email address.

Note: it might take hours depending on the queuing system of KEGG.

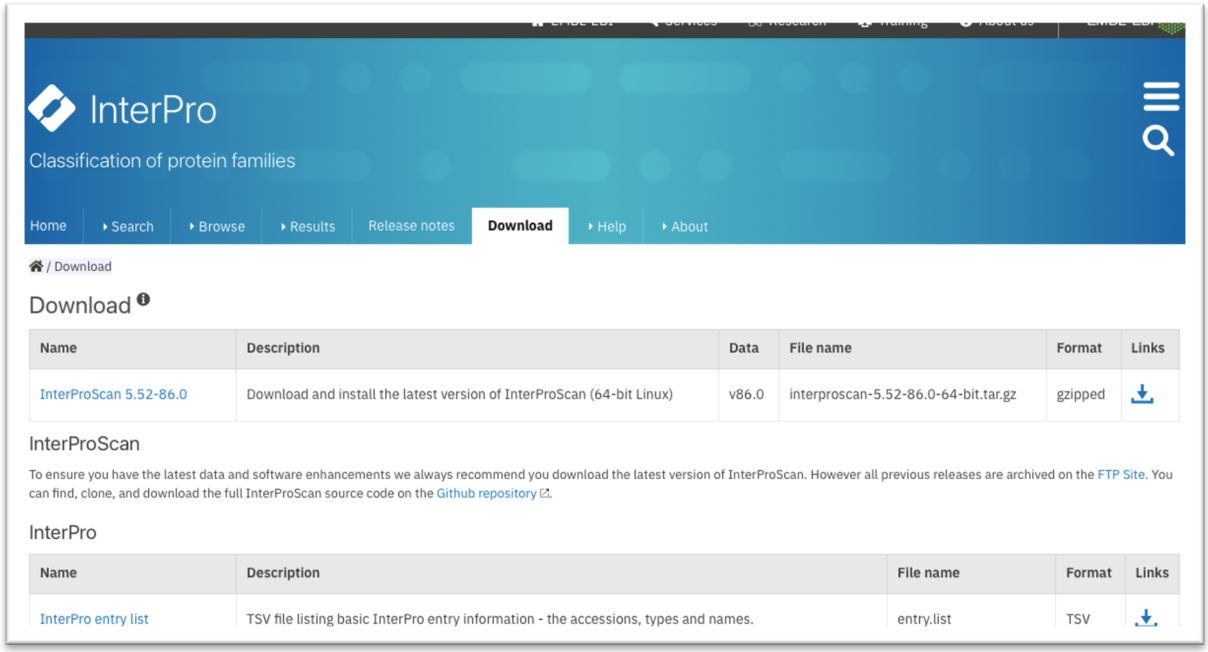


3.5. From the email link of KEGG, the user can download the gene list annotated with KO annotation. Explanation of the 2-column input file for KO accession.
   a. Gene identifier (e.g., g59.t1)
   b. KO accession (e.g., K10849)

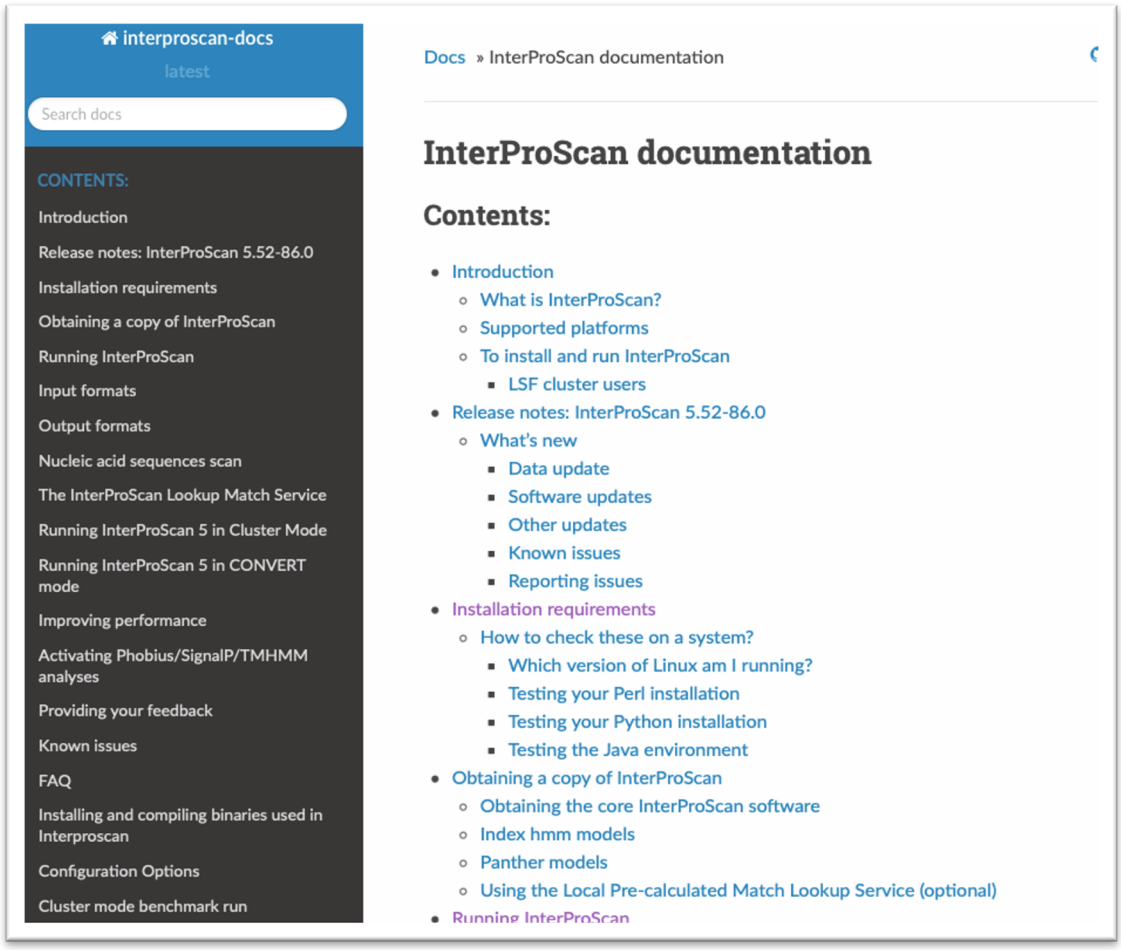
Preparing the InterProScan search result file.

3.6. Downloading the InterProScan from the link https://www.ebi.ac.uk/interpro/download/. Once the InterProScan package is uncompressed, it can be run directly from the command line.

#If run this script with no arguments, the usage instructions will be presented.
>./interproscan.sh
Run the shell script below:
interproscan.sh is the command taking in the input file with parameter (-i) and setting up the format of output file (e.g., tsv format). '-dp' is to ensure all the database matches proceeded in local environment.
>./interproscan.sh -i Chlamydomonas_UWO241_protein.fasta -f tsv -dp



The latest InterProScan documentation can be found via the link https://interproscan-docs.readthedocs.io/en/latest/index.html.



3.7. The 13-column explanation of InterProScan search result file (Table 5)
   a. Protein accession (e.g., g5250.t1)
   b. Sequence unique code (e.g., f246997202ceeb0ebfd5ea2f454be9a2)
   c. Sequence length (e.g., 262)
   d. Protein signature (e.g., Pfam)
   e. Signature accession (e.g., PF02469)
   f. Signature description (e.g., Fasciclin domain)
   g. Start location (e.g., 123)
   h. Stop location (e.g., 259)
   i. E-value (or score) (e.g., 5.80E-09)
   j. Status - is the status of the match (T: true)
   k. Date - is the date of the run (e.g., 31-03-2019)
   l. InterPro annotations - accession (e.g., IPR000782)
   m. InterPro annotations - description (e.g., FAS1 domain)

Step 4. Output file of the NoBadWordsCombiner tool.

Note: Before clicking the submission button, user can select one of nine protein signatures (i.e., Pfam, CDD, Hamap, PRINTS, ProDom, ProSitePattern, ProSiteProfiles, SFLD, or TIGRFAM). We set the Pfam domain parameter as default, in order to collect larger database entries and because they have been widely used in previous sequence analysis and genome annotation projects. Users can also select other protein signatures, such as CDD, which can utilize 3D structure to decipher sequence structure and functional relationships.

4.1. The output of 23-column tab-delimited mega table
   a. ID (eg., 2)
   b. Gene or QueryAcc (e.g., g2.t1)
   c. Length or Query_Length (e.g., 817)
   d. NoBadName_Hit_Des or HitDescription (e.g., 2-5A-dependent ribonuclease OS=Mus musculus OX=10090 GN=Rnasel PE=1 SV=2)
   e. NoBadName_Hit_Name or HitName (e.g., sp|Q05921|RN5A_MOUSE)
   f. NoBadName_%ID or %ID (e.g., 34.8837209)
   g. NoBadName_eValue or eValue (e.g., 4.14E-06)
   h. NCBI_Hit_Des or HitDescription (e.g., ankyrin, partial [Anaeromyces robustus])
   i. NCBI_Hit_Name or HitName (e.g., gi|1183350135|gb|ORX78377.1|)
   j. NCBI_%ID or %ID (e.g., 40.2298851)
   k. NCBI_eValue or eValue (e.g., 3.61E-10)
   l. Swiss_Hit_Des or HitDescription (e.g., 2-5A-dependent ribonuclease OS=Mus musculus OX=10090 GN=Rnasel PE=1 SV=2)
   m. Swiss_Hit_Name or HitName (e.g., sp|Q05921|RN5A_MOUSE)
   n. Swiss_%ID or %ID (e.g., 34.8837209)
   o. Swiss_eValue or eValue (e.g., 4.14E-06)
   p. KEGG_KO (e.g., K03267)
   q. KEGG_Des (e.g., ERF3, GSPT; peptide chain release factor subunit 3)
   r. Protein signatures (e.g., Pfam)

s. Pfam_No (e.g., PF12796)
t. Pfam_Des (e.g., Ankyrin repeats (3 copies))
u. Pfam_evalue (e.g., 1.80E-11)
v. Interpro_No (e.g., IPR020683)
w. Interpro_domain (e.g., Ankyrin repeat-containing domain)

Note: users can document any issues or debugging via the GitHub NoBadWordsCombiner repository, including code warnings or errors; this is the most common way to report potential bugs in the GitHub community.

Limitation: We do hope to further develop the tool, making it more user friendly, including trying to remove some of the middle steps. Unfortunately, we are not yet able to provide a "one-click solution" because of the incompatibility of the various databases employed by the tool. That said, we still believe that our tool is comparatively easier to use than some of the other options currently available to scientists. Indeed, presently there are very few tools that can search eukaryotic genome projects, efficiently merging hits from various databases and strengthening gene definitions by minimizing functional descriptions containing 'bad words'. Thus, we believe that our tool will provide a well-needed service to the bioinformatics and genomics community.