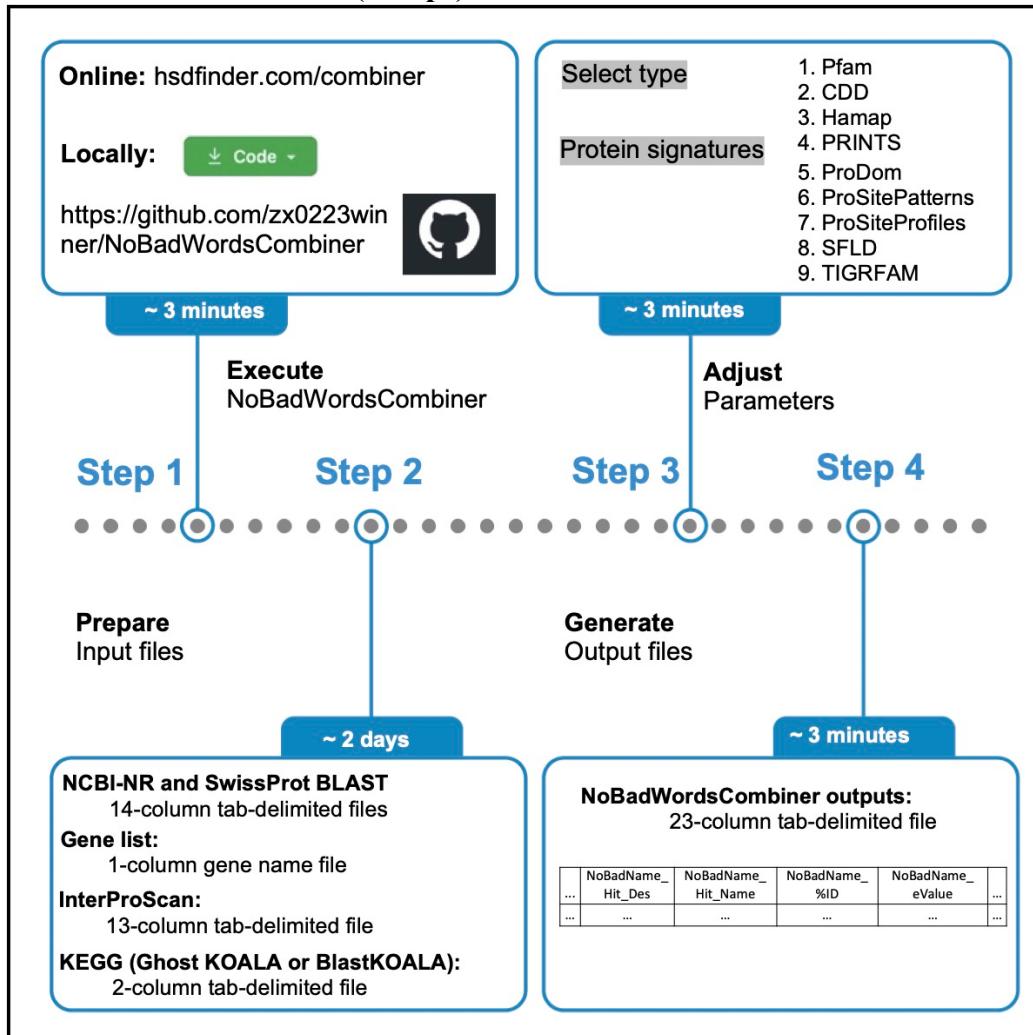


# NoBadWordsCombiner Online Tutorial: Merge and minimize ‘bad words’ from BLAST hits against multiple eukaryotic gene annotation databases

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.

### Run NoBadWords Combiner

NoBadWords Combiner can integrate the gene function information together and minimize 'bad words' including Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, and Pfam etc.

**Run**

NCBI nr <a href="#">?</a> Example1 <input type="button" value="Choose File"/> NoBadWords_NCBI-3.txt	Swiss prot <a href="#">?</a> Example2 <input type="button" value="Choose File"/> NoBadWords_Swiss.txt	Gene list <a href="#">?</a> Example3 <input type="button" value="Choose File"/> NoBadWords_genelist.txt
Gene list with KO annotation <a href="#">?</a> Example4 <input type="button" value="Choose File"/> NoBadWords_ko.txt	Pfam file <a href="#">?</a> Example5 <input type="button" value="Choose File"/> NoBadWords_Pfam.txt	Select type: <a href="#">?</a> <input type="button" value="Pfam"/>
<input type="button" value="Submit"/>		

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

**Run**

NCBI nr <a href="#">?</a> Example1 <input type="button" value="Choose File"/> NoBadWords_NCBI-3.txt	 Generating NoBadWords file...This may take several minutes...	
Gene list with KO annotation <a href="#">?</a> Example4 <input type="button" value="Choose File"/> NoBadWords_ko.txt	Pfam file <a href="#">?</a> Example5 <input type="button" value="Choose File"/> NoBadWords_Pfam.txt	Select type: <a href="#">?</a> <input type="button" value="Pfam"/>

The output file will turn out like this.

Output:  
NoBadName\_Combiner\_20210830094442.tsv

**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.

main ▾ 1 branch 1 tag

zx0223winner Add files via upload

Tutorial	Add files via upload
LICENSE	Create LICENSE
NoBadWordsCombiner.py	Create NoBadWordsCombiner.py
README.md	Update README.md
Readme.rst	Create Readme.rst

Code ▾

Clone

HTTPS SSH GitHub CLI

<https://github.com/zx0223winner/NoBad>

Use Git or checkout with SVN using the web URL.

Open with GitHub Desktop

Download ZIP

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
<a href="#">Parent Directory</a>		-
<a href="#">ChangeLog</a>	2021-06-28 11:23	85
<a href="#">ncbi-blast-2.12.0+-1.src.rpm</a>	2021-06-28 11:23	51M
<a href="#">ncbi-blast-2.12.0+-1.src.rpm.md5</a>	2021-06-28 11:23	63
<a href="#">ncbi-blast-2.12.0+-1.x86_64.rpm</a>	2021-06-28 11:23	187M
<a href="#">ncbi-blast-2.12.0+-1.x86_64.rpm.md5</a>	2021-06-28 11:23	66
<a href="#">ncbi-blast-2.12.0+-src.tar.gz</a>	2021-06-28 11:23	56M
<a href="#">ncbi-blast-2.12.0+-src.tar.gz.md5</a>	2021-06-28 11:23	64
<a href="#">ncbi-blast-2.12.0+-src.zip</a>	2021-06-28 11:23	60M
<a href="#">ncbi-blast-2.12.0+-src.zip.md5</a>	2021-06-28 11:23	61
<a href="#">ncbi-blast-2.12.0+-win64.exe</a>	2021-06-28 11:23	94M
<a href="#">ncbi-blast-2.12.0+-win64.exe.md5</a>	2021-06-28 11:23	63
<a href="#">ncbi-blast-2.12.0+-x64-linux.tar.gz</a>	2021-06-28 11:24	236M
<a href="#">ncbi-blast-2.12.0+-x64-linux.tar.gz.md5</a>	2021-06-28 11:24	70
<a href="#">ncbi-blast-2.12.0+-x64-macosx.tar.gz</a>	2021-06-28 11:24	144M
<a href="#">ncbi-blast-2.12.0+-x64-macosx.tar.gz.md5</a>	2021-06-28 11:24	71
<a href="#">ncbi-blast-2.12.0+-x64-win64.tar.gz</a>	2021-06-28 11:24	93M
<a href="#">ncbi-blast-2.12.0+-x64-win64.tar.gz.md5</a>	2021-06-28 11:24	70
<a href="#">ncbi-blast-2.12.0+.dmg</a>	2021-06-28 11:24	147M
<a href="#">ncbi-blast-2.12.0+.dmg.md5</a>	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.

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#uniprotkb/blink>

Downloads release 2021\_03

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](ftp://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)<sup>i</sup> / [FAQ](#) | [xml](#) | [fasta](#) | [text](#)  
Unreviewed (TrEMBL)<sup>j</sup> / [FAQ](#) | [xml](#) | [fasta](#) | [text](#)  
Isoform sequences<sup>k</sup> / [FAQ](#) | [fasta](#)

[Taxonomic divisions](#) / [README](#)  
[Reference proteomes](#) / [README](#)  
[Pan proteomes](#) / [README](#)  
[ID mapping](#) / [README](#)  
[Proteomics mapping](#) / [README](#)  
[Variants](#) / [README](#)  
[Genome annotation tracks](#) / [README](#)  
[Documents](#)  
[YML schema](#)

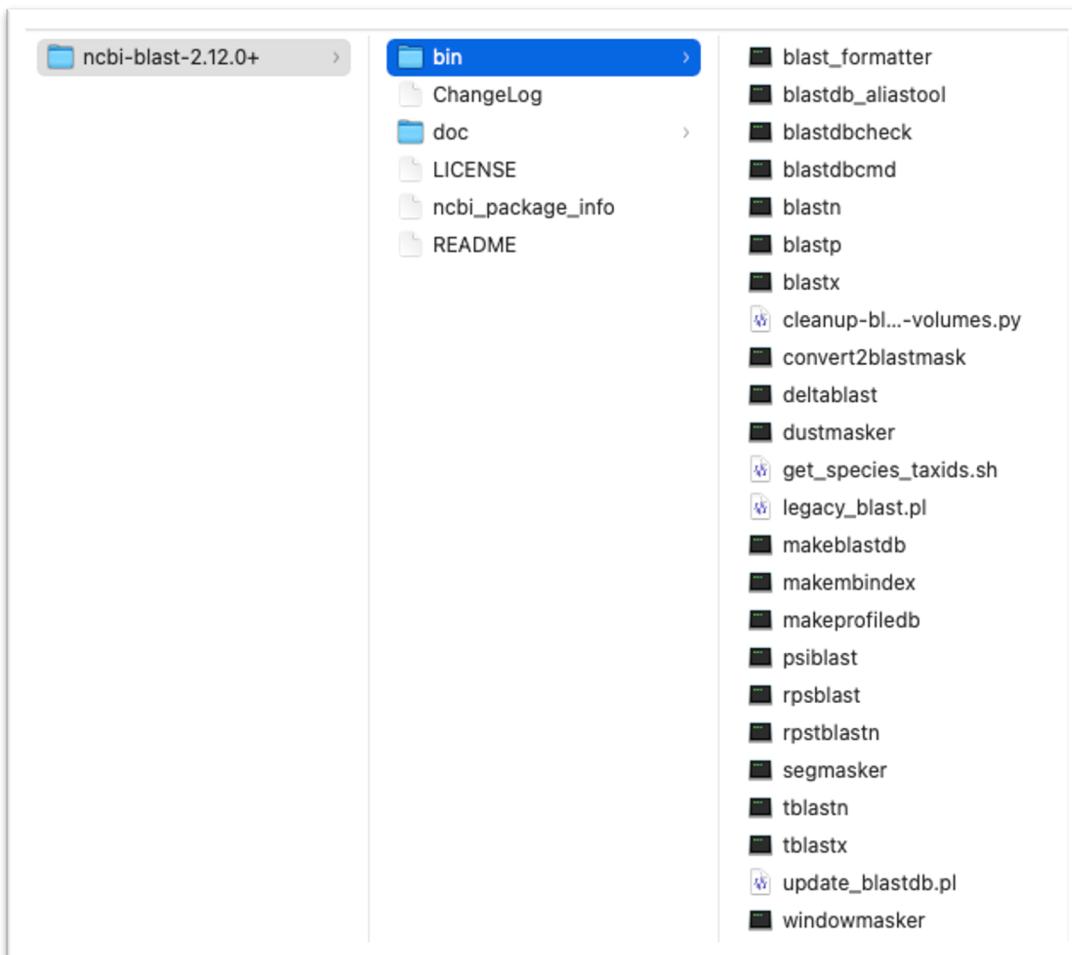
2.4.The makeblastdb command is from the bin of Blast package. The makeblastdb command is to construct a protein database by taking in the FASTA file with the parameter (-in), setting up the database type (e.g., protein) with the parameter (-dbtype protein), and titling the name of database (e.g., uniprot\_prot\_database) with parameters (-title database\_name). -out option is the database output 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 uniprot\_db.phr, uniprot\_db.pin, uniprot\_db.psp.



2.5.Using BLASTP search option to blast the amino acid sequences against uniprot\_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,leng th 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
<a href="#">Parent Directory</a>		-
<a href="#">FASTA/</a>	2021-08-30 07:46	-
<a href="#">cloud/</a>	2020-02-11 16:27	-
<a href="#">v4/</a>	2020-06-30 10:29	-
<a href="#">v5/</a>	2021-08-31 11:17	-
<a href="#">16S ribosomal RNA-nucl-metadata.json</a>	2021-08-14 05:36	467
<a href="#">16S ribosomal RNA.tar.gz</a>	2021-08-14 05:36	36M
<a href="#">16S ribosomal RNA.tar.gz.md5</a>	2021-08-14 05:36	59
<a href="#">18S fungal sequences-nucl-metadata.json</a>	2021-08-31 05:36	489
<a href="#">18S fungal sequences.tar.gz</a>	2021-08-31 05:36	31M
<a href="#">18S fungal sequences.tar.gz.md5</a>	2021-08-31 05:36	62
<a href="#">28S fungal sequences-nucl-metadata.json</a>	2021-08-31 05:37	489
<a href="#">28S fungal sequences.tar.gz</a>	2021-08-31 05:37	32M
<a href="#">28S fungal sequences.tar.gz.md5</a>	2021-08-31 05:37	62
<a href="#">Betacoronavirus-nucl-metadata.json</a>	2021-08-31 11:17	647
<a href="#">Betacoronavirus.00.tar.gz</a>	2021-08-31 11:17	479M
<a href="#">Betacoronavirus.00.tar.gz.md5</a>	2021-08-31 11:17	60
<a href="#">Betacoronavirus.01.tar.gz</a>	2021-08-31 11:17	548M
<a href="#">Betacoronavirus.01.tar.gz.md5</a>	2021-08-31 11:17	60
<a href="#">Betacoronavirus.02.tar.gz</a>	2021-08-31 11:17	346M
<a href="#">Betacoronavirus.02.tar.gz.md5</a>	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  
#...
```

	2021-08-22 16:06	19G
<a href="#">nr.v0.tar.gz</a>	2021-08-22 16:06	47
<a href="#">nr.00.tar.gz.md5</a>	2021-08-22 16:06	1.7G
<a href="#">nr.01.tar.gz</a>	2021-08-22 16:06	47
<a href="#">nr.01.tar.gz.md5</a>	2021-08-22 16:06	1.6G
<a href="#">nr.02.tar.gz</a>	2021-08-22 16:06	47
<a href="#">nr.02.tar.gz.md5</a>	2021-08-22 16:06	1.9G
<a href="#">nr.03.tar.gz</a>	2021-08-22 16:06	47
<a href="#">nr.03.tar.gz.md5</a>	2021-08-22 16:06	2.0G
<a href="#">nr.04.tar.gz</a>	2021-08-22 16:06	47
<a href="#">nr.04.tar.gz.md5</a>	2021-08-22 16:07	2.2G
<a href="#">nr.05.tar.gz</a>	2021-08-22 16:07	47
<a href="#">nr.05.tar.gz.md5</a>	2021-08-22 16:07	2.3G
<a href="#">nr.06.tar.gz</a>	2021-08-22 16:07	47
<a href="#">nr.06.tar.gz.md5</a>	2021-08-22 16:07	2.3G
<a href="#">nr.07.tar.gz</a>	2021-08-22 16:07	47
<a href="#">nr.07.tar.gz.md5</a>	2021-08-22 16:08	2.0G
<a href="#">nr.08.tar.gz</a>	2021-08-22 16:08	47
<a href="#">nr.08.tar.gz.md5</a>	2021-08-22 16:08	2.3G
<a href="#">nr.09.tar.gz</a>	2021-08-22 16:08	47
<a href="#">nr.09.tar.gz.md5</a>	2021-08-22 16:08	2.2G
<a href="#">nr.10.tar.gz</a>	2021-08-22 16:08	47
<a href="#">nr.10.tar.gz.md5</a>	2021-08-22 16:08	2.4G
<a href="#">nr.11.tar.gz</a>	2021-08-22 16:08	47
<a href="#">nr.11.tar.gz.md5</a>	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.**

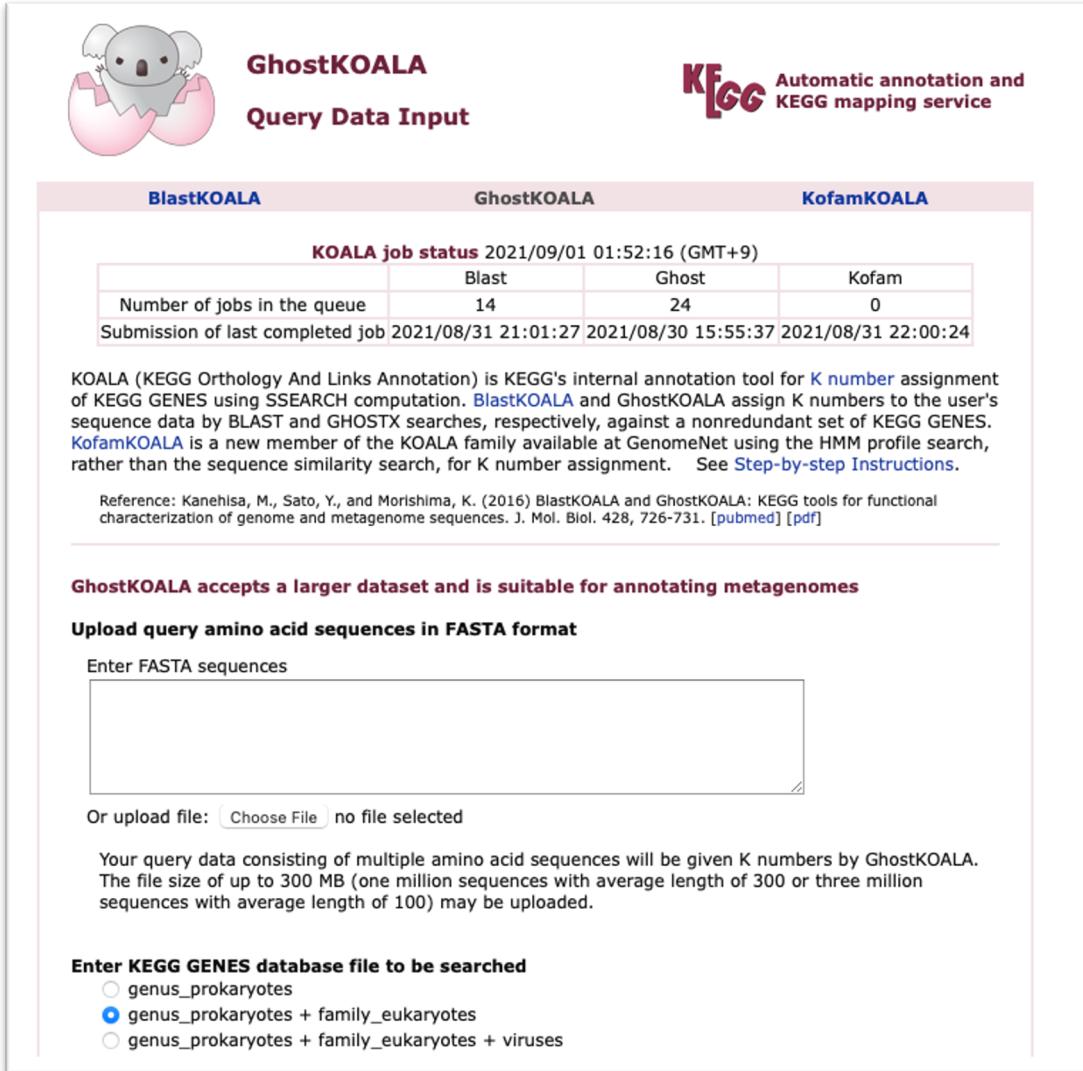
3.1. Users can first test the function of grep via the command below:

```
> 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/>/\n/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  $\geq$  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:



The screenshot shows the KEGG GhostKOALA interface. At the top left is a koala icon with the text "GhostKOALA" and "Query Data Input". At the top right is the KEGG logo with the text "Automatic annotation and KEGG mapping service". Below this is a navigation bar with tabs for "BlastKOALA", "GhostKOALA", and "KofamKOALA". A table titled "KOALA job status" shows the following data:

	Blast	Ghost	Kofam
Number of jobs in the queue	14	24	0
Submission of last completed job	2021/08/31 21:01:27	2021/08/30 15:55:37	2021/08/31 22:00:24

The main content area contains the following text:

KOALA (KEGG Orthology And Links Annotation) is KEGG's internal annotation tool for [K number](#) assignment of KEGG GENES using SSEARCH computation. [BlastKOALA](#) and GhostKOALA assign K numbers to the user's sequence data by BLAST and GHOSTX searches, respectively, against a nonredundant set of KEGG GENES. [KofamKOALA](#) is a new member of the KOALA family available at GenomeNet using the HMM profile search, rather than the sequence similarity search, for K number assignment. See [Step-by-step Instructions](#).

Reference: Kanehisa, M., Sato, Y., and Morishima, K. (2016) BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. 428, 726-731. [[pubmed](#)] [[pdf](#)]

**GhostKOALA accepts a larger dataset and is suitable for annotating metagenomes**

**Upload query amino acid sequences in FASTA format**

Enter FASTA sequences

Or upload file:  no file selected

Your query data consisting of multiple amino acid sequences will be given K numbers by GhostKOALA. The file size of up to 300 MB (one million sequences with average length of 300 or three million sequences with average length of 100) may be uploaded.

**Enter KEGG GENES database file to be searched**

genus\_prokaryotes  
 genus\_prokaryotes + family\_eukaryotes  
 genus\_prokaryotes + family\_eukaryotes + viruses

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

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

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

**Enter KEGG GENES database file to be searched**

- genus\_prokaryotes
- genus\_prokaryotes + family\_eukaryotes
- genus\_prokaryotes + family\_eukaryotes + viruses

The database files for GhostKOALA are somewhat different from those for BlastKOALA. For each group of KEGG organisms at the genus or family level, a nonredundant dataset is generated by taking all protein-coding genes from the representative genome and additional genes from the other genomes with two criteria. One is the same as in BlastKOALA, different K numbers, and the other is unique to GhostKOALA, different CD-HIT clusters, which are computed with 50% identity cutoff. In addition, the database file for viruses is created by CD-HIT with 90% identity cutoff from the viruses category of KEGG GENES. These additions are meant for analyzing taxonomic compositions of metagenomes.

**Enter your email address**

An email will be sent to you for confirmation of your input data. You will have to click on the link in the email to initiate your job. When the job is finished, you will receive another email for browsing the result and performing KEGG Mapper analysis. You cannot request another job until the current one is finished or canceled. *Notice: Your email address will not be used for any other purpose.*

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

InterPro

Classification of protein families

Home Search Browse Results Release notes **Download** Help About

[/ Download](#)

## Download i

Name	Description	Data	File name	Format	Links
InterProScan 5.52-86.0	Download and install the latest version of InterProScan (64-bit Linux)	v86.0	interproscan-5.52-86.0-64-bit.tar.gz	gzipped	<a href="#"></a>

**InterProScan**

To ensure you have the latest data and software enhancements we always recommend you download the latest version of InterProScan. However all previous releases are archived on the [FTP Site](#). You can find, clone, and download the full InterProScan source code on the [Github repository](#).

**InterPro**

Name	Description	File name	Format	Links
InterPro entry list	TSV file listing basic InterPro entry information - the accessions, types and names.	entry.list	TSV	<a href="#"></a>

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

interproscan-docs  
latest

Search docs

**CONTENTS:**

- Introduction
- Release notes: InterProScan 5.52-86.0
- Installation requirements
- Obtaining a copy of InterProScan
- Running InterProScan
- Input formats
- Output formats
- Nucleic acid sequences scan
- The InterProScan Lookup Match Service
- Running InterProScan 5 in Cluster Mode
- Running InterProScan 5 in CONVERT mode
- Improving performance
- Activating Phobius/SignalP/TMHMM analyses
- Providing your feedback
- Known issues
- FAQ
- Installing and compiling binaries used in Interproscan
- Configuration Options
- Cluster mode benchmark run

Docs » InterProScan documentation

## InterProScan documentation

### Contents:

- [Introduction](#)
  - [What is InterProScan?](#)
  - [Supported platforms](#)
  - [To install and run InterProScan](#)
    - [LSF cluster users](#)
- [Release notes: InterProScan 5.52-86.0](#)
  - [What's new](#)
    - [Data update](#)
    - [Software updates](#)
    - [Other updates](#)
    - [Known issues](#)
    - [Reporting issues](#)
- [Installation requirements](#)
  - [How to check these on a system?](#)
    - [Which version of Linux am I running?](#)
    - [Testing your Perl installation](#)
    - [Testing your Python installation](#)
    - [Testing the Java environment](#)
- [Obtaining a copy of InterProScan](#)
  - [Obtaining the core InterProScan software](#)
  - [Index hmm models](#)
  - [Panther models](#)
  - [Using the Local Pre-calculated Match Lookup Service \(optional\)](#)
- [Running InterProScan](#)

### 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.*