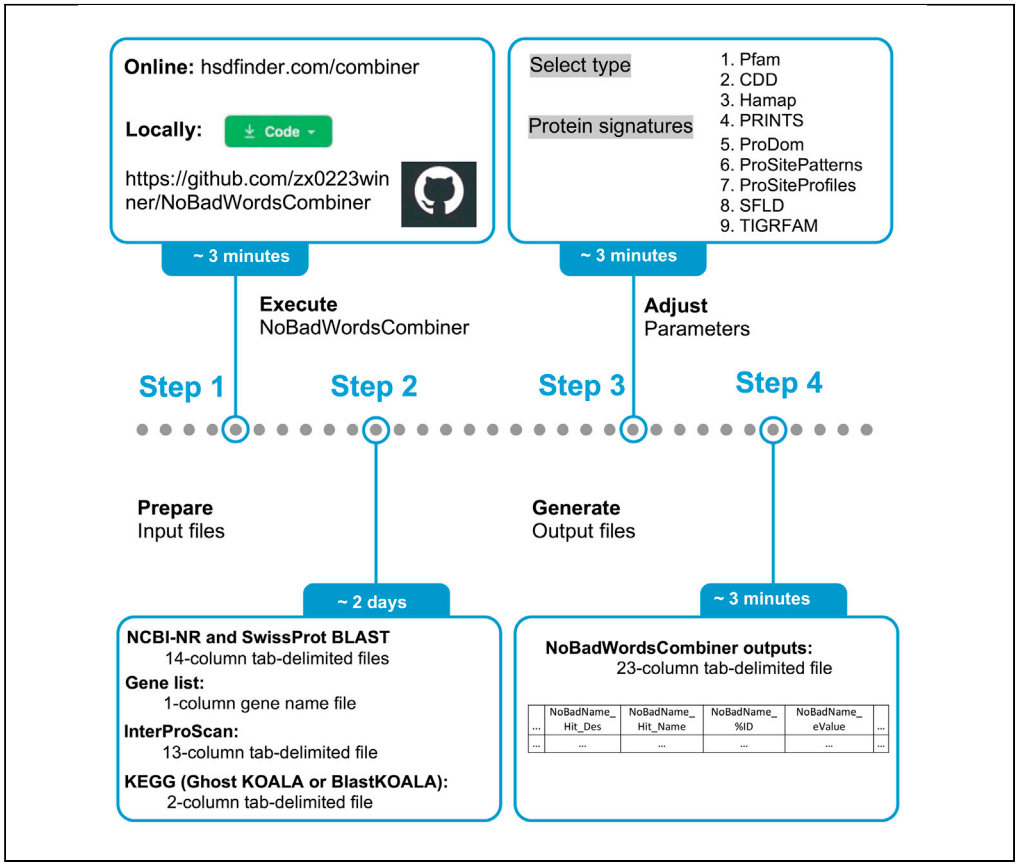
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for using NoBadWordsCombiner to merge and minimize "bad words" from BLAST hits against multiple eukaryotic gene annotation databases



Annotating protein-coding genes can be challenging, especially when searching for the best hits against multiple functional databases. This is partly because of "bad words" appearing as top hits, such as hypothetical or uncharacterized proteins. To help alleviate some of these issues, we designed a bioinformatics tool called NoBadWordsCombiner, which efficiently merges the hits from various databases, strengthening gene definitions by minimizing functional descriptions containing "bad words." Unlike other available tools, NoBadWordsCombiner is user friendly, but it does require users to have some general bioinformatics skills, including a basic understanding of the BLAST package and dash shell in Linux/Unix environments.

Xi Zhang, Yining Hu, David Roy Smith

xzha25@uwo.ca (X.Z.)
dsmit242@uwo.ca (D.R.S.)

Highlights

NoBadWordsCombiner tool can merge and minimize "bad words" during gene annotation

Bad words include hypothetical and uncharacterized protein descriptions

Gene definitions are strengthened by Pfam domains and KEGG pathways

An overview of the gene annotations is summarized in a 23-column table

Zhang et al., STAR Protocols 2, 100888
December 17, 2021
https://doi.org/10.1016/j.xpro.2021.100888

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for using NoBadWordsCombiner to merge and minimize "bad words" from BLAST hits against multiple eukaryotic gene annotation databases

Xi Zhang,[1,2,5,*] Yining Hu,[3] and David Roy Smith[4,6,*]

[1]Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, NS B3H 4R2, Canada

[2]Institute for Comparative Genomics, Dalhousie University, Halifax, NS B3H 4R2, Canada

[3]Department of Computer Science, Western University, London, ON N6A 5B7, Canada

[4]Department of Biology, Western University, London, ON N6A 5B7, Canada

[5]Technical contact

[6]Lead contact

*Correspondence: xzha25@uwo.ca (X.Z.), dsmit242@uwo.ca (D.R.S.)
https://doi.org/10.1016/j.xpro.2021.100888

SUMMARY

Annotating protein-coding genes can be challenging, especially when searching for the best hits against multiple functional databases. This is partly because of "bad words" appearing as top hits, such as hypothetical or uncharacterized proteins. To help alleviate some of these issues, we designed a bioinformatics tool called NoBadWordsCombiner, which efficiently merges the hits from various databases, strengthening gene definitions by minimizing functional descriptions containing "bad words." Unlike other available tools, NoBadWordsCombiner is user friendly, but it does require users to have some general bioinformatics skills, including a basic understanding of the BLAST package and dash shell in Linux/Unix environments.

For complete details on the use and execution of this protocol, please refer to Zhang et al. (2021a).

BEFORE YOU BEGIN

Next-generation sequencing (NGS) technologies can generate huge amounts of molecular sequence data (Yandell and Ence, 2012). Functional annotations of protein-coding genes from NGS data can be easily acquired via database searches, including NCBI-NR (Pruitt et al., 2005), UniProtKB/Swiss-Prot (Boutet et al., 2007), and TrEMBL (Boeckmann et al., 2003). But the results of these searches often include 'bad words', such as best hits to hypothetical proteins or uncharacterized proteins, which can confuse the interpretation of gene annotation results. Indeed, it was reported that 20–30% of the annotations from assembled chlamydomonadalean nuclear genomes are represented by hypothetical proteins, including those from the *Chlamydomonas reinhardtii* genome (Zhang et al., 2021a). For various other recently sequenced genomes, the percentage of hypothetical proteins can be even higher (Galperin, 2001). It can be time-consuming to manually curate the functional hits from Basic Local Alignment Search Tool (BLAST) searches. This can be especially true if trying to minimize hits containing 'bad words' (e.g., hypothetical proteins) when the redundant hits have meaningful functional annotations (i.e., without 'bad words').

Currently, there are very few user-friendly bioinformatics tools for merging and minimizing 'bad words' during functional gene annotation, and those that are available typically involve custom programing scripts with a steep learning curve (De Wit et al., 2012). Here, we present NoBadWordsCombiner, an open-source, user-friendly bioinformatics web tool for efficiently merging and minimizing 'bad words' scanned from various functional annotation databases. This tool can plugin to

external databases, such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) and InterProScan (Quevillon et al., 2005), to strengthen the definition of gene annotations. NoBadWordsCombiner does require users to have some basic familiarity with bioinformatics. They must be comfortable with the BLAST package (Altschul et al., 1997), the dash shell in Linux/Unix environments, and inputting files from third-party tools, such as InterProScan (Quevillon et al., 2005) and KEGG (BlastKOALA and GhostKOALA) (Kanehisa et al., 2016).

Recently, we sequenced, assembled, and annotated the nuclear genome of the Antarctic green alga *Chlamydomonas* sp. UWO241 (Zhang et al., 2021a), hereafter referred to as UWO241. During our analysis of this genome, we designed and applied the NoBadWordsCombiner tool during the functional annotation stage, which greatly minimized descriptions containing 'bad words'. The protocol presented here describes how to use NoBadWordsCombiner for merging and minimizing 'bad words' from eukaryotic gene annotation databases. The model psychrophilic green alga UWO241 is used as a case-study for this goal.

Overview

NoBadWordsCombiner merges the functional gene annotation BLAST hits from NCBI nr, UniProtKB/Swiss-Prot, and TrEMBL database searches. Specifically, it removes redundancy from descriptions containing hits to hypothetical or uncharacterized proteins (not including instances when all hits are hypothetical/uncharacterized). Then, the definition of the combined hits is strengthened via protein functional domains and pathway information based on data from the InterPro and KEGG databases. Finally, the overview of the gene annotations with the minimized 'bad words' is summarized in a mega table.

Downloading the software and prerequisites

NoBadWordsCombiner can be operated on the web (http://hsdfinder.com/combiner/) or the local environment (Linux and Python 3) after downloading the software package from GitHub (https://github.com/zx0223winner/ NoBadWordsCombiner). To run locally, pre-installed Python (preferably Python 3) and Linux (e.g., Ubuntu 20.04 LTS) environments are required. The BLAST and InterProScan software packages as well as the online KEGG pathways tools BlastKOALA and GhostKOALA (Kanehisa et al., 2016) can be accessed via the links in the key resources table.

> ***Note:*** A minimum specification requirement is a computer with 2 cores, 4 GB of RAM, and 256 GB storage, which should allow the 'bad words' to be merged and minimized within a few minutes.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Chlamydomonas sp. UWO241 (renamed *Chlamydomonas priscuii*)	(Zhang et al., 2021a)	Genbank: GCA_016618255
Software and algorithms		
BLAST v2.2.26	(Altschul et al., 1997)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
UniProtKB/Swiss-Prot	(Boutet et al., 2007)	https://www.uniprot.org/uniprot/?query=reviewed:yes
UniProtKB/TrEMBL	(Boeckmann et al., 2003)	https://www.uniprot.org/uniprot/?query=reviewed:no
NCBI-NR	(Pruitt et al., 2005)	https://www.ncbi.nlm.nih.gov/refseq/
InterProScan v4.7	(Quevillon et al., 2005)	http://www.ebi.ac.uk/interpro/download/
BlastKOALA or GhostKOALA	(Kanehisa and Goto, 2000; Kanehisa et al., 2016)	https://www.kegg.jp
NoBadWordsCombiner	This article	http://hsdfinder.com/combiner

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Python 3	N/A	https://www.python.org/downloads/
Django v3.1.5	N/A	https://www.djangoproject.com/download/
pandas v1.2.2	N/A	https://pandas.pydata.org
blastxml_to_tabular.py	(Cock et al., 2015)	https://github.com/peterjc/galaxy_blast/blob/master/tools/ncbi_blast_plus/blastxml_to_tabular.py

MATERIALS AND EQUIPMENT

The software implementation was written in Python 3 using the following custom scripts and platforms: *NoBadWordsCombiner.py*, which enables the 'bad words' to be merged and minimized from BLAST hits against multiple eukaryotic gene annotation databases and protein signature databases (e.g., Pfam); Django (3.1.5), a Python-based web platform, which maintains the web server; and pandas (1.2.2), the software library used for manipulating the data. *Blastxml_to_tabular.py* (Cock et al., 2015) is a custom Python script that can convert a BLAST XML file to the desired tabular output. The NCBI-NR and UniProtKB/TrEMBL databases, including the gene annotations, are computationally analyzed, whereas the UniProtKB/Swiss-Prot database is manually curated and, thus, contains fewer annotations of hypothetical proteins. A full list of the utilized packages and database, including links, can be found in the key resources table. The full NoBadWordsCombiner source code can be found in the GitHub repository. A useful hands-on tutorial (Online_NoBadWordsCombiner Tutorial.pdf) can also be accessed under the tutorial directory of GitHub.

The test input data consist of BLAST and protein signature results from InterProScan (Quevillon et al., 2005).

> ***Note:*** Five mandatory tab-delimited tables are needed to run the tool. The first and second input documents of the NCBI-NR and SwissProt database BLAST results have 14 columns (Tables 1 and 2). These two tables are parsed from the local BLAST results via a custom Python script (*Blastxml_to_tabular.py*). The third and fourth input files were designed as a 1-column gene name list file and a 2-column KEGG annotation file, respectively (Tables 3 and 4). The fifth input document of the InterProScan results has 13 columns (Table 5). The KO accession with each gene model identifier was retrieved from the KEGG database (Kanehisa and Goto, 2000). In the following step-by-step protocol, we use the deduced protein sequences from the UWO241 genome annotation (Zhang et al., 2021a) to show how to generate these tables.

STEP-BY-STEP METHOD DETAILS

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

> **Timing: ~2 days (depending on the amount of the data, computing power, and Internet speed.)**

Upload protein BLAST-search result files from your genome of interest in tab-separated values (tsv) format as the input files (Tables 1 and 2) of NoBadWordsCombiner. This protocol will go over how to acquire local BLAST-search results via an example FASTA file. The example file as well as the hands-on tutorial (Online_NoBadWordsCombiner Tutorial.pdf) can be acquired from GitHub under the tutorial directory (Figure 1A).

> ***Note:*** You can ignore this step and proceed with your own protein data set if you know how to acquire the appropriate BLASTP search results.

Table 1. Input file example of NCBI nr database BLAST result

Query Acc	Query_ Length	HitDescription	HitName	Hit Length	Hit Bits	HSP_ rank	%ID	eValue	Query_ Start	Query_ End	Hit_ start	Hit_ end	HSP_ length
g1.t1	817	hypothetical protein CEUSTIGMA_g3421.t1 [Chlamydomonas eustigma]	gi|1238995578|dbj| GAX75978.1|	1443	260.766	1	54.2635659	1.41E-75	10	774	11	268	258
g2.t1	399	ankyrin, partial [Anaeromyces robustus]	gi|1183350135|gb| ORX78377.1|	235	65.4698	1	40.2298851	3.61E-10	19	279	18	96	87
g3.t1	3567	hypothetical protein CEUSTIGMA_g3419.t1 [Chlamydomonas eustigma]	gi|1238995576|dbj| GAX75976.1|	1103	172.17	1	38.4615385	1.15E-39	805	1674	330	597	299
g4.t1	963	hypothetical protein CEUSTIGMA_g3418.t1 [Chlamydomonas eustigma]	gi|1238995575|dbj| GAX75975.1|	623	310.457	1	89.5061728	1.17E-97	469	954	172	333	162
g6.t1	291	hypothetical protein CHLRE_10g421079v5 [Chlamydomonas reinhardtii]	gi|1335042461|gb| PNW77074.1|	103	82.8037	1	58.3333333	1.66E-18	103	282	34	93	60
g7.t1	7908	hypothetical protein CEUSTIGMA_g3945.t1 [Chlamydomonas eustigma]	gi|1238994727|dbj| GAX76500.1|	2934	156.377	1	32.6785714	7.48E-34	6334	7761	2313	2860	560
g9.t1	471	hypothetical protein CEUSTIGMA_g3416.t1 [Chlamydomonas eustigma]	gi|1238995573|dbj| GAX75973.1|	139	164.466	1	62.1212121	3.00E-49	76	468	11	139	132
g10.t1	1827	hypothetical protein GPECTOR_108g190 [Gonium pectorale]	gi|1004134917|gb| KXZ42995.1|	463	331.257	1	78.8288288	1.18E-103	580	1245	48	269	222
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 2. Input file example of SwissProt database BLAST result

Query Acc	Query_ Length	HitDescription	HitName	Hit Length	Hit Bits	HSP_ rank	%ID	eValue	Query_ Start	Query_ End	Hit_ start	Hit_ end	HSP_ length
g2.t1	399	2-5A-dependent ribonuclease OS=Mus musculus OX=10090 GN=Rnasel PE=1 SV=2	sp|Q05921| RN5A_ MOUSE	735	48.1358	1	34.8837209	4.14E-06	25	267	125	206	86
g3.t1	3567	DNA mismatch repair protein MSH6 OS=Arabidopsis thaliana OX=3702 GN=MSH6 PE=1 SV=2	sp|O04716| MSH6_ ARATH	1324	53.5286	1	41.8181818	1.61E-05	379	543	121	175	55
g4.t1	963	Eukaryotic peptide chain release factor GTP-binding subunit ERF3A OS=Homo sapiens OX=9606 GN=GSPT1 PE=1 SV=1	sp|P15170| ERF3A_ HUMAN	499	234.958	1	72.2972973	2.94E-72	511	954	69	216	148
g9.t1	471	Thylakoid-associated protein slr0729 OS=Synechocystis sp. (strain PCC 6803 / Kazusa) OX=1111708 GN=slr0729 PE=4 SV=1	sp|P72673| Y729_ SYNY3	101	47.7506	1	29.4736842	1.49E-06	187	468	11	99	95
g10.t1	1827	Threonylcarbamoyl-AMP synthase OS=Schizosaccharomyces pombe (strain 972 / ATCC 24843) OX=284812 GN=sua5 PE=3 SV=1	sp|O94530| SUA5_ SCHPO	408	217.238	1	50.8264463	3.67E-63	580	1242	58	299	242
g15.t1	270	Protein transport protein Sec61 subunit beta OS=Chlamydo monas reinhardtii OX=3055 GN=SEC61B PE=1 SV=1	sp|A8I6P9| SC61B_ CHLRE	89	65.4698	1	53.7037037	2.15E-14	106	261	36	89	54
g16.t1	897	Probable prolyl 4-hydroxylase 4 OS=Arabidopsis thaliana OX=3702 GN=P4H4 PE=2 SV=1	sp|Q8LAN3| P4H4_ ARATH	298	157.147	1	41.0788382	9.49E-45	1	708	69	289	241
g17.t1	1104	GATA transcription factor 3 OS=Arabidopsis thaliana OX=3702 GN=GATA3 PE=2 SV=2	sp|Q8L4M6| GATA3_ ARATH	269	62.003	1	56.097561	1.17E-09	79	201	171	211	41
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Table 3. Input file example of gene name list

Gene name
g1.t1
g2.t1
g3.t1
g4.t1
g5.t1
g6.t1
g7.t1
…

Table 4. Input file example of KO accession with each gene model identifier retrieved from the KEGG database

Gene identifier	KO accession
g59.t1	K10849
g60.t2	K17087
g61.t2	N/A
g62.t1	N/A
g63.t2	N/A
g64.t1	N/A
g65.t1	K15172
g66.t1	K02519
…	…

Table 5. Input file example of InterProScan database result

Protein accession	Unique code	Sequence length	Protein signature	Signature accession	Signature description	Start location	Stop location	E-value	Status	Date	InterPro accession	InterPro description
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	SUPER FAMILY	SSF82153	N/A	129	260	9.42E-10	T	31-03-2019	IPR036378	FAS1 domain superfamily
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	ProSite Profiles	PS50213	FAS1/ BIgH3 domain profile.	111	257	9.579	T	31-03-2019	IPR000782	FAS1 domain
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	Pfam	PF02469	Fasciclin domain	123	259	5.80E-09	T	31-03-2019	IPR000782	FAS1 domain
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	ProSite Patterns	PS00183	Ubiquitin-conjugating enzymes active site.	69	84	-	T	31-03-2019	IPR023313	Ubiquitin-conjugating enzyme, active site
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	SMART	SM00212	N/A	2	148	1.10E-36	T	31-03-2019	N/A	N/A
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	PANTHER	PTHR44511	N/A	2	119	1.50E-59	T	31-03-2019	N/A	N/A
g5250.t1	f246997202ceeb0ebfd5ea2f454be9a2	262	SMART	SM00554	N/A	145	260	7.20E-07	T	31-03-2019	IPR000782	FAS1 domain
g15700.t1	e9641f2405b85bc4c48a85029514acf0	799	Mobi DBLite	mobidb-lite	consensus disorder prediction	347	361	-	T	31-03-2019	N/A	N/A
…	…	…	…	…	…	…	…	…	…	…	…	…

1. Download the BLAST package and FASTA file. A BLAST-search result example file is found in the ZIP file in the GitHub 'tutorial' directory under the name NoBadWordsCombiner_file_examples.zip

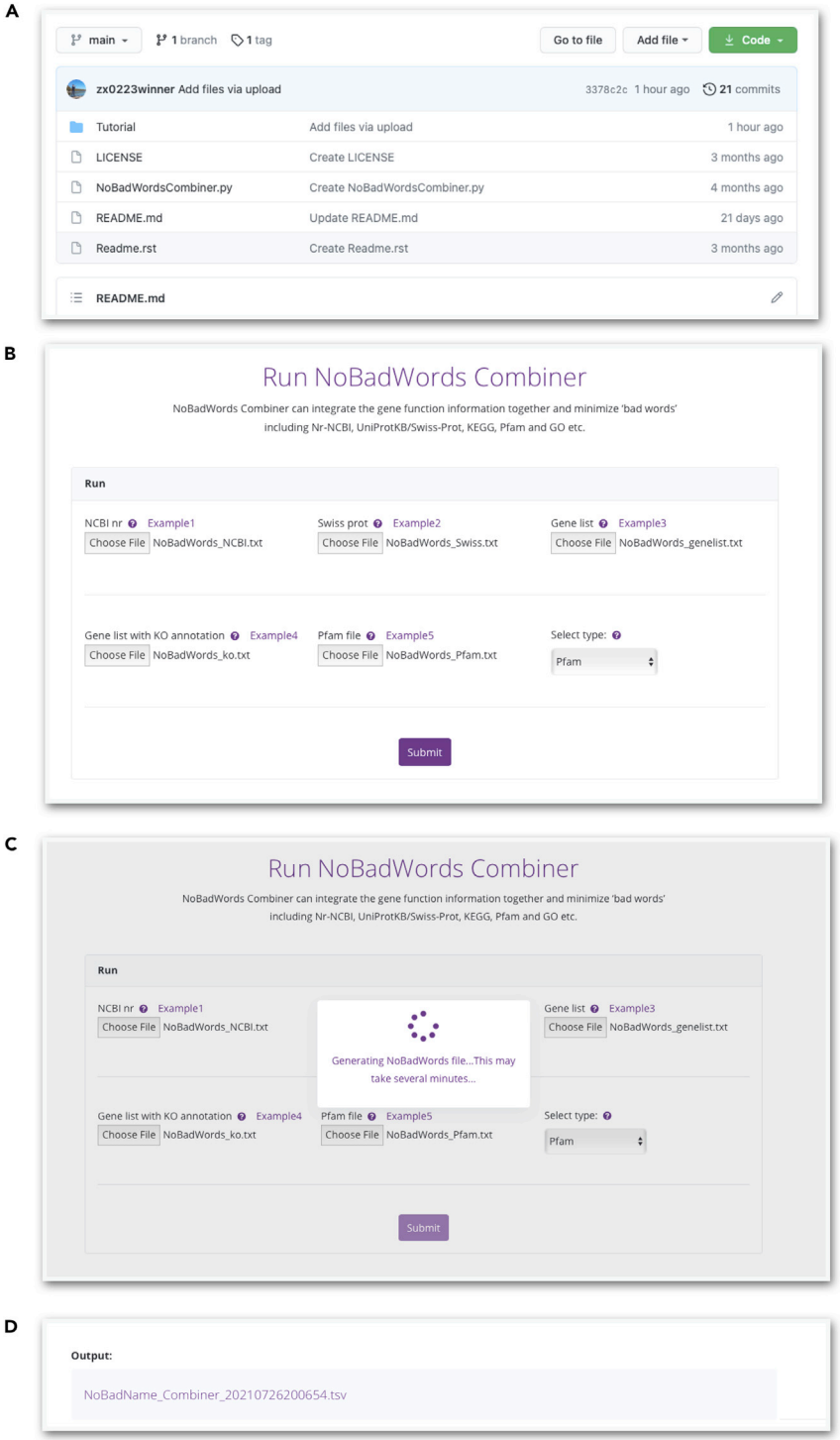


Figure 1. The NoBadWordsCombiner home page

(A) The GitHub web interface of this tool.

(B) Uploading the necessary input files.

(C) The interface of running the tool.

(D) The output example of the tool.

(Figure 1A). Also, the 'blastdbv5-user-guide.pdf' document in the GitHub 'tutorial' directory contains complementary vignettes to help guide users.

a. Download the BLAST Package via https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/. Please select the appropriate version based on your computer operating system (Windows, MacOS, or Linux)
b. Unzip the 'NoBadWordsCombiner_file_examples.zip' file; 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

MAATVENVVERVKSFSSVVRGVKSGKPDGATTQLVQETIEILATYCDFEEVVPV

CLKFLDEVLTAAPQTSTLIRLEGGAK

IFPSIIRNFMGVDASILALCAKVMCKCASGSPAMQHHLVKEKGLPTLLLSCCSA

HAGEPAVVGPLLEVLVALARYSKGAT

ALSNANLVHACKELLVGLMGHWHAFGMVLKLIKSVMKHEGPCLAALKAGEVVRL

LLGVARLVSRMPDQRKLLKRASRTLW

VLSQRSLHPLPEMELNWPHTHTHTHTHTHTHT

>g2.t1

MMMLAYRFGFTTLMYATVKGHADAMRLLLKHPSADTAAMMMLTDIRGCTALMFA

AQDGHVNAIRMLLDHPSADVAARIAV

RSTVGISALTSAAGFAAGQPTLSRRASPARSCTPLLFLLRRVAVEPQLCDTQ

>g3.t1

MVPTDGARHGWTATSLPAILGAASHAKITVQQLVVGGPPPSCPYGPEIVGRSLS

LFSKSAKTWDRAPGGVVSAFCAATGE
```

2. Set up the manually curated UniProtKB/Swiss-Prot database and computationally calculated NCBI-NR database.
   a. The 'uniprot_sprot.fasta.gz' file can be downloaded directly from https://www.uniprot.org/downloads. When downloading, choose Reviewed (Swiss-Prot) in FASTA format under the parent directory. The NCBI-NR BLAST v5 databases can be accessed via https://ftp.ncbi.nlm.nih.gov/blast/db. Some necessary files (e.g., nr.00.tar.gz, nr.00.tar.gz.md5, and nr.01.tar.gz) can be automatically downloaded via a custom Perl script at step 2c.
   b. The 'makeblastdb' command will 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., uniport_prot_database) with parameters (-title database_name). The '-out' option will yield the database output name (e.g., uniport_db).

```
# Note: if your FASTA data are represented by nucleotides, you can

change the database type with the parameter (-dbtype

nucl)

> ./makeblastdb -in uniprot_sprot.fasta -dbtype prot -

title uniprot_prot_database -out uniprot_db
```

c. 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/).
This command will download the NCBI-NR database (https://ftp.ncbi.nlm.nih.gov/blast/db/v5) with the name 'nr' without using the makeblastdb command to redo it. It could take minutes to hours for processing, depending on the speed of the Internet.
i. Users can first check all available databases 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
# ...
```

ii. Users can then run the command below to download the nr database, which includes 55 volumes of data (>100 Gb). Alternatively, users 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
#...
```

d. Use the BLASTP search option to blast the amino acid sequences against the uniport_db and nr_v5 databases. The BLASTP command can do the protein similarity search by searching the query file (Chlamydomonas_UWO241_protein.fasta) against the protein database created from the former step with default parameters, such as '-evalue' (indicating the significance of the BLAST hits), '-outfmt' (meaning the format of the BLAST result), and '-out' (telling the file name of the output file; e.g., BLASTP_UWO241_uniprot.xml).

```
>./blastp -query Chlamydomonas_UWO241_protein.fasta -db
uniprot_db -out BLASTP_UWO241_uniprot.xml -evalue 1e-5 -
outfmt 5
```

e. The BLAST XML file (-outfmt 5) can include useful information compared to the BLAST Tabular file (-outfmt 6), such as the aligned sequence, the sequence of the hit, and the description of

hits in the database. However, the XML format is not human-readable. Users will need to employ a commonly used parser (*Blastxml_to_tabular.py*) (Cock et al., 2015), which is a custom python script, to convert a BLAST XML file to a desired tabular output (tab-delimited file).

i. Users can first download the script *Blastxml_to_tabular.py* via the link from the key resources table. Then run 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_uniprot.xml >
BLASTP_UWO241_uniprot.tsv
```

ii. The parameters behind the option '-c' in the *Blastxml_to_tabular.py* script will yield desired columns in the output tab-delimited file. For example, 'qseqid' refers to query sequence ID and 'qlen' refers to query length. These desired parameters will create a 14-column table (e.g., Tables 1 and 2).

f. To speed up the NCBI BLAST search, users can specify one or more (comma-delimited) taxids, or a file containing multiple taxids on the command-line. For example, to search against the *Chlamydomonas reinhardtii* (Taxonomy ID: 147) and *Chlamydomonas eustigma* (Taxonomy ID: 57939) nuclear genomes, use the command shown below. Also, we recommend that users browse the 'blastdbv5-user-guide.pdf' document in the GitHub 'tutorial' directory to familiarize themselves with the creation of either taxids or taxidlist.

```
>./blastp -query Chlamydomonas_UWO241_protein.fasta -db
nr -taxids 147,57939 -out BLASTP_UWO241_NCBI-NR.xml -
evalue 1e-5 -outfmt 5
# Multiple taxonomy IDs are delimited by ','.
#Similar to Step 2e, 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,pide
nt,evalue,qstart,qend,sstart,send,length
BLASTP_UWO241_NCBI-NR.xml > BLASTP_UWO241_NCBI-NR.tsv
# The option '-c' refers to the desired output columns
which can be set in comma-delimited format (e.g.,
qseqid,qlen,salltitles)
```

⚠ CRITICAL: Make sure to use the BLASTP option, which allows for greater sensitivity. The BLAST output parameter must be in format 5. Users can adjust the parameter of

the E-value, but we recommend that it be no greater than 1e-5 (to ensure accurate predictions). Troubleshooting 1.

3. This will give two BLAST result files formed by 14-column spreadsheets, including key information, such as query name and percentage identity.
4. The 14-column explanation of parsed BLAST search result files (Tables 1 and 2).
   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)
5. If users want to upload different BLAST files or mistakenly submitted an incorrect file, they can reload the browser page or simply overlap with another file. Troubleshooting 2

Preparing the gene name list and a gene list with KO annotations from the KEGG database

Timing: ~3 h (depending on the queuing time of GhostKOALA)

Users can retrieve the third and fourth files from the genome FASTA file and the KEGG database, which include the correlation of the KO accession with each gene model identifier (Figure 1B). The gene name file is the baseline to merge all the different functional annotations.

6. Users can acquire the gene name list file by the following command lines. We use the FASTA file from the UWO241 genome as an example (Table 3).
   # 'grep' is the command used in the dash shell to grasp each line containing the word pattern of '>'. 'sed' is used to substitute all the '>' into none, which generates a clean name list file.
   a. 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
```

   b. 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
```

As for a gene list with KO annotation, users have the option to use the GhostKOALA (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:

7. BlastKOALA accepts a smaller dataset and is suitable for annotating high-quality genomes.
   a. Upload the query amino acid sequences in FASTA format.
   b. Enter the taxonomy group of your genome.
   c. Enter the KEGG GENES database file to be searched.
   d. Enter your email address. An email will be sent to you for confirmation of your input data. Click the link in the email to initiate your job; you will receive another email once it is finished.
8. GhostKOALA accepts larger datasets (e.g., 300 Mb) and is suitable for annotating metagenomes.
   a. Upload the query amino acid sequences in FASTA format.
   b. Enter the KEGG GENES database file to be searched.
   c. Enter your email address. Same as above (7d).
9. From the email link of KEGG, users can download the gene list with the associated KO annotations. The format of the output file is referred to in Table 4. Explanation of the 2-column input file for KO accession (Table 4):
   a. Gene identifier (e.g., g59.t1)
   b. KO accession (e.g., K10849)
10. Use the GhostKOALA or BlastKOALA analysis tool of KEGG to acquire the KO annotation file of your genome (https://www.kegg.jp/ghostkoala/). We provide an example of a KO annotation file under the GitHub directory of the tutorial with the name NoBadWordsCombiner_file_examples.zip. Troubleshooting 3

Preparing the InterProScan search result file

Timing: ~3 h

Upload an InterProScan search result file of your genome in tab-delimited format as the fifth input file (Table 5). Users must individually download and install InterProScan to acquire the input file for the NoBadWordsCombiner tool. The latest InterProScan documentation can be found via the link https://interproscan-docs.readthedocs.io/en/latest/index.html. Here, we provide the necessary steps for using InterProScan:

11. Installation requirements:
    a. InterProScan is developed to run on Linux and no versions are planned for Windows or Apple (MAC OS X) operating systems.
    b. Software requirements: 64-bit Linux; Perl 5; Python 3; Java JDK/JRE version 11.
    c. Obtaining the core InterProScan software (Direct link: ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.51-85.0/interproscan-5.51-85.0-64-bit.tar.gz).

Note: this is a large file (around 8 Gb).

12. Running InterProScan:
    a. 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
```

b. 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
```

13. Output files:
    a. InterProScan should run without any warning, and it will create a tsv output file (i.e., Chlamydomonas_UWO241_protein.fasta.tsv) containing several member database matches, including Pfam. For your convenience, an example of an InterProScan search result is found in the ZIP file under the GitHub directory of tutorial with the name NoBadWordsCombiner_file_examples.zip. Troubleshooting 4

14. 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)

    Note: Before clicking the submission button, users 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 the default in order to collect larger database entries and because it has been widely used in many sequence analysis and genome annotation projects. Users can select other protein signatures, such as CDD, which can utilize 3D structures to decipher sequence structure and functional relationships. The descriptions of the various protein signatures are shown below:
        i. Pfam: A large collection of protein families, each represented by multiple sequence alignments and hidden Markov models (HMMs).
        ii. CDD: Prediction of Conserved Domains Database.
        iii. Hamap is a system for the classification and annotation of protein sequences.
        iv. PRINTS: A fingerprint is a group of conserved motifs used to characterize a protein family.
        v. ProDom is a comprehensive set of protein domain families automatically generated from the UniProt Knowledge Database.
        vi. ProSitePatterns and ProSiteProfiles: PROSITE consists of documentation entries describing protein domains, families, and functional sites as well as associated patterns and profiles to identify them.

vii. SFLDs are protein families based on Hidden Markov Models or HMMs.
viii. TIGRFAMs are protein families based on Hidden Markov Models or HMMs.

Output file of the NoBadWordsCombiner tool

Timing: ~3 min

15. Tap the submit button and a pending image will jump out (Figure 1C). It usually takes less than three minutes to run with a 200 Mb genome-sized file (Figure 1D). Troubleshooting 5
16. The output of 23-column tab-delimited mega table (Table 6)
    a. ID (e.g., 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)
17. The reason we created two columns for the header (e.g., NoBadName_Hit_Des or HitDescription) is to reduce ambiguity. A uniform header is needed when merging different databases. We also included the percentage identity and E-value for each type of BLAST search result, which can be easily compared. If two BLAST database hit descriptions are both without 'bad words', the one with lower E-value will be chosen.

EXPECTED OUTCOMES

NoBadWordsCombiner is a free and straightforward online bioinformatics software tool for merging and minimizing hypothetical or uncharacterized proteins from various eukaryotic functional annotation databases. It provides a mega table combined with protein annotation database entries, such as InterPro, Pfam, and KEGG KO. To compare the data in the mega table, users can have an overview of the gene's annotation patterns, including information on pathways, gene family domain, and gene family.

The aim of this tool is to assist with high-quality gene model annotations of eukaryotic nuclear genomes. We provided a real example of this tool in action: the file contains the gene models and their functional descriptions for the UWO241 genome (Zhang et al., 2021a), which greatly

Table 6. Output file example of 23-column mega table via NoBadWordsCombiner

ID	Gene	Length	NoBad Name_ Hit_Des	NoBad Name_ Hit_Name	NoBad Name_ %ID	NoBad Name_ eValue	NCBI_Hit_Des	NCBI_ Hit_ Name	NCBI_ %ID	NCBI_ eValue	Swiss_ Hit_Des	Swiss_ Hit_ Name	Swiss_ %ID	Swiss_ eValue	KEGG_ KO	KEGG_ Des	Pfam	Pfam_ No	Pfam_ Des	Pfam_ evalue	Inter pro_ No	Inter pro_ domain
0	Query Acc	Query_ Length	Hit Description	Hit Name	%ID	e Value	Hit Description	Hit Name	%ID	e Value	Hit Description	Hit Name	%ID	e Value	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
1	g1.t1	817	hypothetical protein CEUSTIGMA_g3421.t1 [Chlamydomonas eustigma]	gi|1238995578|dbj|GAX75978.1|	54.26356595	1.41E-75	hypothetical protein CEUSTIGMA_g3421.t1 [Chlamydomonas eustigma]	gi|1238995578|dbj|GAX75978.1|	54.26356595	1.41E-75	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	g2.t1	399	2-5A-dependent ribonuclease OS=Mus musculus OX=10090 GN=Rnasel PE=1 SV=2	sp|Q05921|RN5A_MOUSE	34.88372093	4.14E-06	ankyrin, partial [Anaeromyces robustus]	gi|1183350135|gb|ORX78377.1|	40.2298851	3.61E-10	2-5A-dependent ribonuclease OS=Mus musculus OX=10090 GN=Rnasel PE=1 SV=2	sp|Q05921|RN5A_MOUSE	34.88372093	4.14E-06	N/A	N/A	Pfam	PF12796	Ankyrin repeats (3 copies)	1.80E-11	IPR020683	Ankyrin repeat-containing domain
3	g3.t1	3567	DNA mismatch repair protein MSH6 OS=Arabidopsis thaliana OX=3702 GN=MSH6 PE=1 SV=2	sp|O04716|MSH6_ARATH	41.8181818	1.61E-05	hypothetical protein CEUSTIGMA_g3419.t1 [Chlamydomonas eustigma]	gi|1238995576|dbj|GAX75976.1|	38.461538	1.15E-39	DNA mismatch repair protein MSH6 OS=Arabidopsis thaliana OX=3702 GN=MSH6 PE=1 SV=2	sp|O04716|MSH6_ARATH	41.8181818	1.61E-05	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	g4.t1	963	Eukaryotic peptide chain release factor GTP-binding subunit ERF3A OS=Homo sapiens OX=9606 GN=GSPT1 PE=1 SV=1	sp|P15170|ERF3A_HUMAN	72.2972973	2.94E-72	hypothetical protein CEUSTIGMA_g3418.t1 [Chlamydomonas eustigma]	gi|1238995575|dbj|GAX75975.1|	89.5061728	1.17E-97	Eukaryotic peptide chain release factor GTP-binding subunit ERF3A OS=Homo sapiens OX=9606 GN=GSPT1 PE=1 SV=1	sp|P15170|ERF3A_HUMAN	72.2972973	2.94E-72	K03267	ERF3, GSPT; peptide chain release factor subunit 3	Pfam	PF00009	Elongation factor Tu GTP binding domain	1.70E-34	IPR000795	Transcription factor, GTP-binding domain
5	g6.t1	291	hypothetical protein CHLRE_10g421079v5 [Chlamydomonas reinhardtii]	gi|1335042461|gb|PNW77074.1|	58.33333333	1.66E-18	hypothetical protein CHLRE_10g421079v5 [Chlamydomonas reinhardtii]	gi|1335042461|gb|PNW77074.1|	58.33333333	1.66E-18	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

(*Continued on next page*)

Table 6. ***Continued***

ID	Gene	Length	NoBad Name_ Hit_Des	NoBad Name_ Hit_Name	NoBad Name_ %ID	NoBad Name_ eValue	NCBI_Hit_Des	NCBI_ Hit_ Name	NCBI_ %ID	NCBI_ eValue	Swiss_ Hit_Des	Swiss_ Hit_ Name	Swiss_ %ID	Swiss_ eValue	KEGG_ KO	KEGG_ Des	Pfam	Pfam_ No	Pfam_ Des	Pfam_ evalue	Inter pro_ No	Inter pro_ domain
6	g7.t1	7908	hypothetical protein CEUSTIGMA_g3945.t1 [Chlamydomonas eustigma]	gi|1238994727|dbj|GAX76500.1|	32.6785714	7.48E-34	hypothetical protein CEUSTIGMA_g3945.t1 [Chlamydomonas eustigma]	gi|1238994727|dbj|GAX76500.1|	32.6785714	7.48E-34	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	g9.t1	471	Thylakoid-associated protein slr0729 OS=Synechocystis sp. (strain PCC 6803 / Kazusa) OX=1111708 GN=slr0729 PE=4 SV=1	sp|P72673|Y729_SYNY3	29.4736842	1.49E-06	hypothetical protein CEUSTIGMA_g3416.t1 [Chlamydomonas eustigma]	gi|1238995573|dbj|GAX75973.1|	62.1212121	3.00E-49	Thylakoid-associated protein slr0729 OS=Synechocystis sp. (strain PCC 6803 / Kazusa) OX=1111708 GN=slr0729 PE=4 SV=1	sp|P72673|Y729_SYNY3	29.4736842	1.49E-06	N/A	N/A	Pfam	PF11378	Protein of unknown function (DUF3181)	6.90E-26	IPR021518	Protein of unknown function DUF3181
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

aided downstream analyses of this genome, such as detecting highly similar duplicated genes (HSDs) as well as horizontally transferred genes and gene family expansions (Zhang et al., 2021b).

LIMITATIONS

NoBadWordsCombiner is limited to presenting gene annotations in a mega table without any plots or statistical interpretations, such as the total number of 'bad words', the frequency of genes containing 'bad words', or what types of genes have 'bad words'. E-values are only used to measure the better BLAST results when both hits contain no 'bad words'. If the genome is misassembled, using E-values alone might infer false positives. In the future, we hope to incorporate the threshold of aligned length, percentage of pairwise identity, and number of domains into the algorithm. For example, only when a certain criterion is satisfied (e.g., pairwise aligned length $\geq$ 50 amino acids, percentage identity $\geq$ 30%, and at least one domain), will the E-value be used to judge a better BLAST functional hit.

The web tool is reliant on third-party tools to generate the input files, such as InterProScan (Quevillon et al., 2005) and the KEGG tools Ghost KOALA or BlastKOALA (Kanehisa et al., 2016). Users need to be familiar with the basic BLAST package and dash shell in Linux/Unix environments. Notably, there is a steep learning curve for users without any bioinformatics or programming experience. It is our hope to further develop the tool, removing some of the middle steps. For now, we have provided the build-in reference files for each input file as well as example data to facilitate the usability of the tool.

In the future, NoBadWordsCombiner will be further improved, including continuous updating by considering more functional eukaryotic databases. It will also be expanded so it can work on other types of genomic data, such as prokaryotic and organelle genomes.

TROUBLESHOOTING

Problem 1

Why does BLASTP need to be chosen as an option? What E-value shall I choose? (Step 2)

Potential solution

Make sure to use the BLASTP option because amino acid sequences are generally more highly conserved than their corresponding nucleotide sequences. We recommend the E-value to be no larger than 1e-5 to ensure accurate prediction.

Problem 2

Can I resubmit the input files? (Step 5)

Potential solution

Yes. Simply re-fresh (reload) the browser page.

Problem 3

Why is the KEGG KO annotation file needed and what does it look like? (Step 10)

Potential solution

The example file has been provided with the name 'Input_4_NoBadWords_ko' from GitHub. The file documents the correlation of KO accession with each gene model identifier, which can be used to strengthen the gene functional category.

Problem 4

Is it difficult to run InterProScan? (Step 13)

Potential solution

No. It is straightforward to run the tool. A real example of a InterProScan result has been provided at GitHub in the NoBadWordsCombiner_file_examples.zip file named 'Input_5_NoBadWords_Pfam'. It is a tab-delimited file including the protein signatures, such as Pfam domain and InterPro annotations.

Problem 5

How does the tool proceed if the BLAST hits inferring hypothetical or uncharacterized proteins come from multiple databases? (step 17)

Potential solution

If BLAST database hit descriptions from multiple databases all contain 'bad words', the one with the lowest e-value will be chosen.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact David Roy Smith (dsmit242@uwo.ca) and technical contact Xi Zhang (xzha25@uwo.ca)

Materials availability

This study did not generate new unique reagents.

Data and code availability

The eukaryotic genomic datasets supporting the conclusions of this article are available from NCBI (https://www.ncbi.nlm.nih.gov). The NoBadWordsCombiner source code has been deposited at https://github.com/zx0223winner/NoBadWordsCombiner. The web server of NoBadWordsCombiner is freely available under http://hsdfinder.com/combiner.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC). We appreciate the constructive suggestions from all the anonymous reviewers.

AUTHOR CONTRIBUTIONS

The study was conceptualized by X.Z. and D.R.S. The data were analyzed by X.Z., and Y.H. implemented the NoBadWordsCombiner website. X.Z. and D.R.S. drafted the manuscript, and all authors commented to produce the manuscript for peer review.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., and Phan, I. (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). UniProtKB/Swiss-Prot: the manually annotated section of the UniProt KnowledgeBase. In Plant Bioinformatics. Methods in Molecular Biology, *406*, D. Edwards, ed (Humana Press). https://doi.org/10.1007/978-1-59745-535-0_4.

Cock, P.J., Chilton, J.M., Grüning, B., Johnson, J.E., and Soranzo, N. (2015). NCBI BLAST+ integrated into Galaxy. Gigascience *4*, 39.

De Wit, P., Pespeni, M.H., Ladner, J.T., Barshis, D.J., Seneca, F., Jaris, H., Therkildsen, N.O., Morikawa, M., and Palumbi, S.R. (2012). The simple fool's guide to population genomics via RNA-Seq: an introduction to high-throughput sequencing data analysis. Mol. Ecol. Resour. *12*, 1058–1067.

Galperin, M.Y. (2001). Conserved 'hypothetical'proteins: new hints and new puzzles. Comp. Funct. Genomics *2*, 14–18.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

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.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. *33*, D501–D504.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res. *33*, 116–120.

Yandell, M., and Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. Nat. Rev. Genet. *13*, 329–342.

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021a). Draft genome sequence of the Antarctic green alga Chlamydomonas sp. UWO241. iScience *24*, 102084.

Zhang, X., Hu, Y., and Smith, D.R. (2021b). Protocol for HSDFinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes. STAR Protoc. *2*, 100619.