

# FlaGs2 User Manual

## April 2022

<https://github.com/GCA-VH-lab/FlaGs2>

Chayan Kumar Saha

Jose A. Nakamoto

Gemma C. Atkinson

### Contents

|  |    |
|--|----|
| 1. Description                             | 2  |
| 2. Prerequisites and installation          | 5  |
| 3. Usage                                   | 6  |
| 4. Running the example files               | 7  |
| 5. Arguments                               | 8  |
| 6. Output files                            | 12 |
| 7. Running online with webFlaGs2 (Pending) | 15 |
| 8. Recommendations and tips                | 17 |
| 9. References                              | 18 |

### Appendices

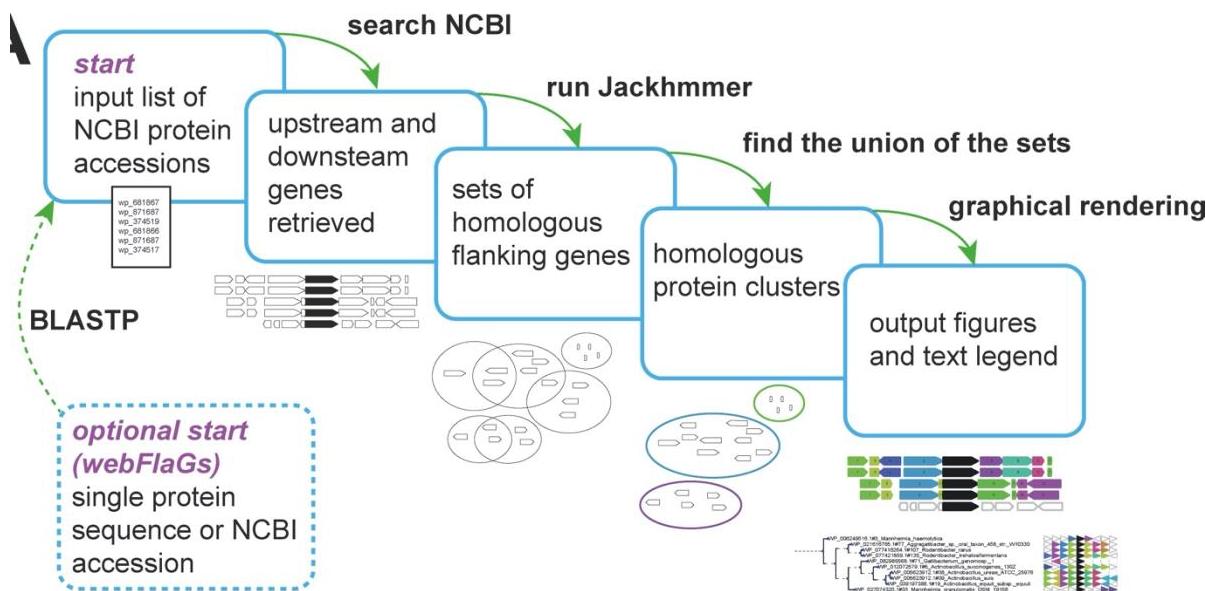
Appendix 1. Creating an input file for FlaGs2 from an online NCBI BlastP or PSI-Blast search.

Appendix 2. Creating an input file for FlaGs2 using BlaspP and our reduce protein database.

## 1. Description

FlaGs2 (for Flanking Genes) analyses genomic context around genes encoding a protein of interest using a Python3 script combined with Jackhmmer (Eddy 2011) and, optionally, ETE-toolkit (Huerta-Cepas, et al. 2016). FlaGs2 finds the flanking genes upstream and downstream of the genes of interest, clusters them based on homology and represents them visually with distinct identifiers (colour and number).

At initiation the tool verifies the input file which is either a list of NCBI protein accessions (-p option), or a tab delimited list of both protein accessions and corresponding genome assembly identifiers (-a). The latter option is to inform FlaGs2 which specific genome it should be searching with the protein accession; since NCBI protein accessions are non-redundant, identical protein sequences can be present in multiple genomes with the same accession number. If FlaGs2 is given protein accessions alone using the -p option, it takes each accession and finds the corresponding list of assembly identifiers (eg. *GCF\_000001635.23* or *GCA\_000001635.5*). Then it can either report all results for all identifiers (-r option), or for one representative identifier (default). The -r option should be used with caution as identical proteins can sometimes be found in thousands of genomes from closely related species, strains or isolates. See *Arguments* below for guidance.



**Figure 1. The FlaGs2 workflow.** The user inputs a list of protein accession numbers – optionally with GCF assembly identifiers – and can specify the number of adjacent flanking genes to consider and the sensitivity of the Jackhmmer search through changing the E value cut-off and number of iterations. The web version of FlaGs2 (webFlaGs) can optionally use a single protein sequence or NCBI accession and begin by executing a BLASTP search against the RefSeq database (excluding eukaryotes) or a representative genome database to generate the input list of accessions. The output always includes a to-scale figure of flanking genes, a description of the flanking gene identities as a legend, and optionally, a phylogenetic tree annotated with colour- and number-coded pennant flags. Unconserved proteins are uncoloured.

After verification of the input query list, FlaGs2 uses the assembly identifier to retrieve the Genomic Feature Format (gff3) file for each protein query and uses this to identify annotated flanking genes from upstream and downstream regions. Then it retrieves the sequences for all

flanking genes and searches for homologues within the set using Jackhmmer from the Hmmer package ([hmmer.org](http://hmmer.org), (Eddy 2011)). Flanking genes are then clustered and each cluster is assigned a specific number. Sets of homologues found with the Jackhmmer searches are joined together to make one cluster (the union) if there is one or more common protein among the Jackhmmer search results (Figure 1). The numbering of clusters begins with 1. The lower the cluster number the more conserved the protein is, i.e the more frequently it is found among all the flanking genes. Finally, FlaGs2 generates a visual representation of the output using the matplotlib Python module, and this is saved as a pdf file (Figure 2). The black block arrow represents the query proteins and the rest represent the flanking genes. The figure is proportional to actual gene lengths and intergenic space. If the query protein is encoded on the negative strand, the entire context is then flipped to make it easily comparable. The number and colour represent the cluster, and the “...\_outdesc.txt” output file provides the legend for interpretation. Unconserved proteins are uncoloured and unnumbered: \_\_\_\_\_, RNA coding genes are outlined in green: \_\_\_\_\_, and pseudogenes are grey, outlined in dark blue: \_\_\_\_\_.

```

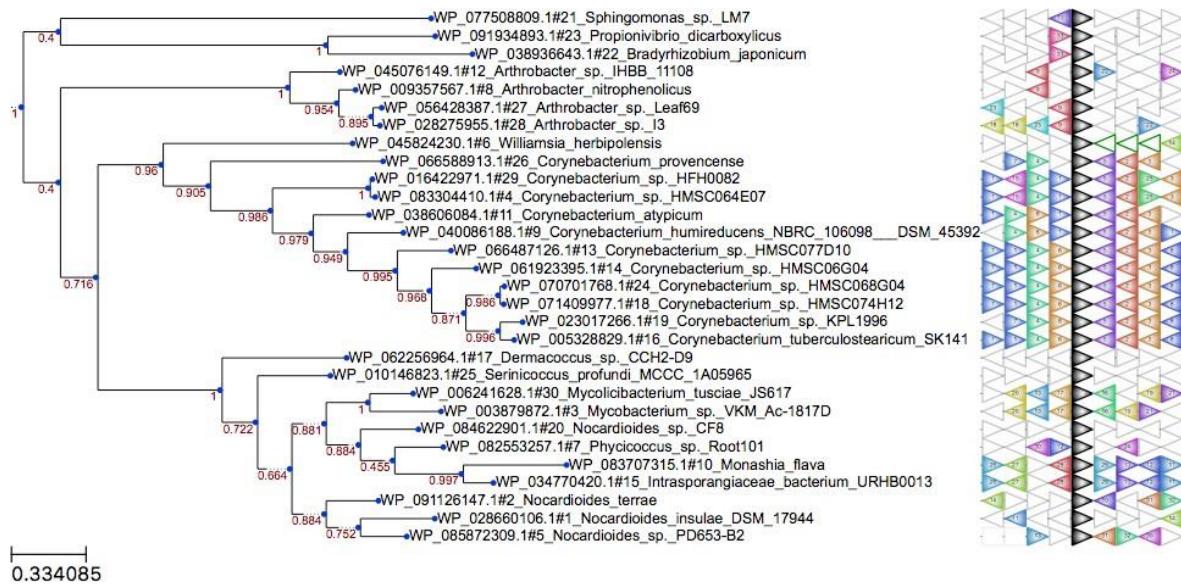
WP 077508809.1#21 Sphingomonas sp. LM7
WP 091934893.1#23 Propionivibrio dicarboxylicus
WP 038936643.1#22 Bradyrhizobium japonicum
WP 045076149.1#12 Arthrobacter sp. IHBB 11108
WP 009357567.1#9 Arthrobacter nitrophoenicus
WP 056428387.1#27 Arthrobacter sp. Leaf69
WP 028275955.1#28 Arthrobacter sp. I3
WP 045824230.1#6 Williamsia herbipolensis
WP 066588913.1#26 Corynebacterium provencense
WP 083304410.1#4 Corynebacterium sp. HMSC064E07
WP 016422971.1#29 Corynebacterium sp. HFH0082
WP 038606084.1#11 Corynebacterium atypicum
WP 066487126.1#13 Corynebacterium sp. HMSC077D10
WP 061923395.1#14 Corynebacterium sp. HMSC06G04
WP 071409977.1#18 Corynebacterium sp. HMSC074H12
WP 070701768.1#24 Corynebacterium sp. HMSC068G04
WP 005328829.1#16 Corynebacterium tuberculostearicum SK141
WP 023017266.1#19 Corynebacterium sp. KPL1996
WP 040086188.1#9 Corynebacterium humireducens NBRC 106098 = DSM 45392
WP 062256964.1#17 Dermacoccus sp. CCH2-D9
WP 010146823.1#25 Serinicoccus profundi MCCC 1A05965
WP 003879872.1#3 Mycobacterium sp. VKM Ac-1817D
WP 006241628.1#30 Mycolicibacterium tusciae JS617
WP 084622901.1#20 Nocardioides sp. CF8
WP 082553257.1#7 Phycococcus sp. Root101
WP 083707315.1#10 Monashia flava
WP 034770420.1#15 Intrasporangiaceae bacterium URHB0013
WP 091126147.1#2 Nocardioides terrae
WP 028660106.1#1 Nocardioides insulare DSM 17944
WP 085872309.1#5 Nocardioides sp. PD653-B2

```



**Figure 2. The to-scale pdf output.** The black gene is the query, and otherwise colours and numbers represent the clusters. The “...\_outdesc.txt” output file provides the legend for interpretation. Genes and intergenic spaces are to scale.

An optional feature of FlaGs2 is the generation of a phylogenetic tree of the query sequences, on which the flanking genes are annotated (Figure 3). This is achieved using the ETE-toolkit ([etetoolkit.org](http://etetoolkit.org), (Huerta-Cepas, et al. 2016)). The ETE “mafft\_default-trimal01-none-fasttree\_full” workflow is used to generate the tree (see ETE user manual, <http://etetoolkit.org/cookbook/>). The flanking genes and the queries are represented as triangular pennant flag-like shapes, with colours and numbers representing clusters. The “...\_outdesc.txt” output file again provides the legend for interpretation.



**Figure 3. The annotated phylogenetic tree output.** The tree is made with ETE3 using the mafft\_default-trimal01- none-fasttree\_full method and is saved as an SVG vector image file.

## 2. Prerequisites and use of the easy install script

FlaGs2.py is available for download here: <https://github.com/GCA-VH-lab/FlaGs2>

FlaGs2 has been primarily tested in Mac and Linux environments. Windows users are directed to use the server version at [webflags.se](#) instead. To locally run FlaGs2 on Windows, we recommend using the Windows Subsystem for Linux 2 (wsl2), Ubuntu 20.04 (found in the Microsoft Store) and Xserver to simulate a Linux environment.

For Mac and Linux users anaconda is required. We recommend the use of Anaconda Python to simplify installation of other required and optional dependencies. It can be downloaded from <https://www.anaconda.com/download/>.

A bash script **build.sh** is bundled with the download, which creates a conda environment named **eFlaGs2** that contains a collection of conda packages that is required specifically for running FlaGs. The user needs to type the following command to create the environment:

```
bash ./build.sh
```

Once the environment is created, user can type the following command to activate the environment and run FlaGs.

```
conda activate eFlaGs2
```

To deactivate the environment, user needs to use the following command.

```
conda deactivate
```

### 3. Usage:

Mac OS or Linux OS:

```
./FlaGs2.py <options>
or python3 FlaGs2.py <options>
or (if Python 3 is the default python) python FlaGs2.py <options>
```

For options and arguments, see below.

## 4. Running the example files:

N.B. a valid email address is required by NCBI to monitor the number of requests to their server per second per user. You will not receive emails from the local version of FlaGs2.

Without tree (only Biopython neccessary):

```
python3 FlaGs2.py -a GCF_accession_input.txt -o  
GCF_accession_output -u example@gmail.com -vb
```

or

```
python3 FlaGs2.py -p accession_input.txt -o accession_output -  
u
```

With tree (Biopython and ETE necessary), and tree order of output:

```
python3 FlaGs2.py -a GCF_accession_input.txt -t -to -o  
GCF_accession_output -u example@gmail.com -vb
```

or

```
python3 FlaGs2.py -p accession_input.txt -t -to -o  
accession_output -u example@gmail.com -vb
```

With tree, and input order of output:

```
python3 FlaGs2.py -a GCF_accession_input.txt -t -o  
GCF_accession_output -u example@gmail.com -vb
```

or

```
python3 FlaGs2.py -p accession_input.txt -t -o  
accession_output -u example@gmail.com -vb
```

## 5. Arguments:

### 1. "-a", "--assemblyList"

Protein accession with Genome Assembly Identifier eg. GCF\_000001765.3 in a text input file separated by a tab.

Input File Example:

```
GCF_000001765.3 WP_047256880.1 #tab separated GCF_000002753.1 WP_012725678.1
.....
```

### 2. "-p", "--proteinList"

Protein Accession WP\_047256880.1 in a text Input file separated by newline. If these are not RefSeq accessions (WP\_), the program will attempt to convert to a RefSeq accession.

Input File Example:

```
WP_047256880.1
WP_012725678.1
.....
```

### 3. "-l", "--localGenomeList"

FlaGs2 v1.0.5 can use local genomes too. For that, the user needs to provide the following files:

- i. Compressed (gzip) FASTA format of the predicted protein products annotated on the genome assembly (eg. Assembly\_1.faa.gz, file should contain “.faa.gz” suffix)

An example file named as ‘GCF\_000428725.1\_ASM42872v1\_protein.faa.gz’ of the recommended format can be downloaded from this link [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GCF\\_000428725.1\\_ASM42872v1/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GCF_000428725.1_ASM42872v1/). Or use the following command from the terminal:

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GCF_000428725.1_ASM42872v1/GCF_000428725.1_ASM42872v1_protein.faa.gz
```

- ii. Compressed (gzip) Annotation of the genomic sequence(s) in Generic Feature Format Version 3 (GFF3) (eg. Assembly\_1.gff.gz, file should contain the “.gff.gz” suffix)

The example file named as ‘GCF\_000428725.1\_ASM42872v1\_genomic.gff.gz’ of recommended format can be downloaded from this link [ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GCF\\_000428725.1\\_ASM42872v1/](ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GCF_000428725.1_ASM42872v1/).

F\_ 000428725.1\_ASM42872v1/. Or use the following command from the terminal:

```
wget
ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/428/725/GC
F_000428725.1_ASM42872v1/GCF_000428725.1_ASM42872v1_genomi
c.g ff.gz
```

With the “-l” flag set, the user specifies the input text file with assembly ID and the query protein accession.

Input file example:

```
GCF_000001765.3 WP_047256880.1 #tab separated
GCF_000002753.1 WP_012725678.1
```

In this case, both of the Fasta and gff file should contain the Assembly Identifier (the GCF number) as the prefix. For example, for assembly identifier GCF\_000001765.3, the compressed Fasta file should be named GCF\_000001765.3.faa.gz and the GFF file should be named GCF\_000001765.3.gff.gz

Or input file example with user-defined assembly names:

```
Assembly_1 WP_047256880.1 #tab separated
Assembly_2 WP_012725678.1
```

For assembly identifier Assembly\_1, the compressed Fasta file should be named as Assembly\_1.faa.gz and the GFF file should be named as Assembly\_1.gff.gz

#### 4. “-ld”, “--localGenomeDirectory”

This flag allows the user to specify the path or directory where .faa.gz and .gff.gz are stored. By default, FlaGs2 script will look for the .faa.gz and .gff.gz files in the same directory where the script is located or running from.

The user can only use this flag when the ‘-l’ flag is being used. For example: If the files are stored in ~/FlaGs/local directory user can specify by following:

```
-ld ~/FlaGs/local
```

#### 5. “-r”, “--redundant”

This will initiate a search for flanking genes in all available GCFs that encode this identical protein sequence (using -r A or -r a) for each query, or a specific number of GCFs (for example to search a maximum of 5 genomes encoding identical protein sequences, the user would write -r 5).

Warning: the -r A option should be used with caution as identical proteins can sometimes be found in thousands of genomes in RefSeq. Using -r with a number to limit is highly recommended.

## 6. "-e", "--ethreshold"

This e-value is used by Jackhmmer as a cutoff parameter to detect homology among all flanking genes. By default it is 1e-10.

## 7. "-n", "--number"

This number represents number of Jackhmmer iterations that allows to find more distant homolog, by default it is set as 3.

## 8. "-g", "--gene"

Using this parameter, the user can define the number of upstream and downstream genes to look for each query in the input list. By default, it is 4, which means for each query it will try to find 4 upstream and 4 downstream flanking genes and then process further.

## 9. "-t", "--tree"

This requires an ETE3 installation. The option enables showing flanking genes along with a phylogenetic tree.

## 10. "-ts", "--tshape"

This requires an ETE3 installation and thus this option only works when -t is used. This parameter can increase or decrease the size of triangle shapes that represent flanking genes, by default it is 12.

## 11. "-tf", "--tfontsize"

This also requires an ETE3 installation and thus this option only works when -t is used. This parameter can increase or decrease the size of font inside triangles that represent flanking genes, by default it is 4.

## 12. "-to", "--tree\_order"

In combination with -t, it will first generate the tree output, and then use the tree order to generate the other (without tree) output file.

This is very useful option for stitching the two output files together in (for example) Illustrator to make a figure with both the tree, and the to-scale neighbourhood output.

## 13. "-u", "--user\_email"

A valid email address is required by NCBI to monitor the number of requests to their server per second per user. (there is a limit of 3 queries per second, which may be exceeded if you run multiple instances of FlaGs2 in parallel; but see the API-key advice below). You will not receive emails from the local version of FlaGs.

## 14. "-api", "--api\_key"

Valid API-key allows 10 queries per seconds, which makes the tool run faster. This may sometimes allow the user to run multiple instances of FlaGs2 in parallel, but in general this is not well tested or advised.

For details on the API key, please check:

<https://ncbiinsights.ncbi.nlm.nih.gov/2017/11/02/new-api-keys-for-the-e-utilities/>

15. "-o", "--out\_prefix"

Any Keyword to define your output eg. MyQuery. This will appear in the name of your output files.

16. "-k", "--keep"

If the user wants to keep the intermediate files eg. gff3, this option is used (warning: these are large files). By default, intermediate files will be removed.

17. "-v", "--version"

Retrieve the version number of the program.

18. "-vb", "--verbose"

This option will show the work progress for each query as STDOUT. This is a very useful option, which we recommend using.

## 6. Output files

In addition to the figure output files described in *1. Description* above, FlaGs2 generates the following output files:

1. Information about the flanking genes for each of the queries retrieved from databases is stored in a tab delimited file named with the “\_operon.tsv” suffix. For example, information about two example queries WP\_028660106.1 and WP\_091126147.1 is shown in Figure 4.

| Query  | Length<br>(aa) | Query<br>Strand | Direction<br>relative<br>to query | Cluster<br>Number | Start<br>position<br>relative<br>to query | End<br>position<br>relative<br>to query | Start<br>position<br>in<br>genome | End<br>position<br>in<br>genome | Protein<br>accessions<br>and order<br>for<br>output | Genome<br>accession | Genome<br>Assembly<br>Identifier |
|--|----------------|-----------------|-----------------------------------|-------------------|---|---|-----------------------------------|---------------------------------|---|---------------------|----------------------------------|
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_297  | =              | +               | D                                 | -3322             | -3026                                     | 78469                                   | 78765                             | WP_028660107.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_654  | =              | +               | D                                 | -3005             | -2352                                     | 77795                                   | 78448                             | WP_028660107.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_1254 | =              | +               | D                                 | -2310             | -1057                                     | 76500                                   | 77753                             | WP_051218224.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_783  | =              | +               | D                                 | -917              | -135                                      | 75578                                   | 76360                             | WP_051218222.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_1800 | =              | +               | 4                                 | 1                 | 1800                                      | 73643                                   | 75442                             | WP_028660106.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_456  | =              | +               | D                                 | 1977              | 2432                                      | 73011                                   | 73466                             | WP_028660105.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_2052 | =              | +               | D                                 | 2461              | 4512                                      | 70931                                   | 72982                             | WP_051218221.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_1461 | =              | +               | D                                 | 4598              | 6065                                      | 69388                                   | 70488                             | WP_051218217.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_028660106.1#1 Nocardioides_insulae_DSM_17944_1512 | =              | +               | 1                                 | 6029              | 7540                                      | 67903                                   | 69414                             | WP_036555377.1#1 NZ_KE383928.1  |   | GCF_000422825.1     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 1407           | +               | +                                 | 1                 | -3909                                     | -2503                                   | 72851                             | 74257                           | WP_091126140.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 564            | +               | +                                 | D                 | -2411                                     | -1848                                   | 74349                             | 74912                           | WP_091126192.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 924            | +               | +                                 | D                 | -1881                                     | -958                                    | 74879                             | 75802                           | WP_091126142.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 948            | +               | +                                 | D                 | -961                                      | -14                                     | 75799                             | 76746                           | WP_091126144.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 1773           | +               | +                                 | 4                 | 1   | 1773                                    | 76761                             | 78533                           | WP_091126147.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 882            | +               | +                                 | D                 | 1792                                      | 2673                                    | 78552                             | 79433                           | WP_091126149.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 498            | +               | +                                 | D                 | 2812                                      | 3309                                    | 79572                             | 80069                           | WP_091230136.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 990            | +               | +                                 | 2                 | 3374                                      | 4363                                    | 80134                             | 81123                           | WP_091126153.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |
| WP_091126147.1#2 Nocardioides_terrae                 | 516            | +               | +                                 | 3                 | 4394                                      | 4909                                    | 81154                             | 81669                           | WP_091126155.1#2 NZ_FOLB01000014.1 GCF_900112345.1  |                     |                                  |

**Figure 4. The flanking genes data file (example).** This tab delimited text file contains the information retrieved from NCBI for each query and their flanking genes. In this example, the query accession WP\_028660106.1 is renamed as WP\_028660106.1#1|Nocardioides\_insulae\_DSM\_17944, where #1 indicates the input order (can be different if tree order option is selected). Nocardioides\_insulae\_DSM\_17944 is the species name and strain information that is retrievable from the NCBI RefSeq Database.

2. A general summary report is also generated for each query, for example if the query accession is valid or invalid, if the query accession was converted or updated to another accession associated with an identical RefSeq sequence, if the query failed, or if flanking gene information is found or not. This is a tab delimited file with the “\_QueryStatus.txt” suffix
3. A general flanking gene summary report is also generated for each valid query if flanking gene information is found or not in a tab delimited file with “\_flankgene\_Report.log” suffix
4. A file with the suffix “\_jackhits.tsv” contains the *Jackhmmer* output
5. After the clustering steps a file named with the suffix “\_clusters.tsv” is generated with detailed information of the cluster. For example;

|     |   |                               |
|-----|---|-------------------------------|
| 1   | 4   | WP_001229260.1;WP_001229255.1 |
| 2 4 | WP_001164213.1;WP_001164217.1;WP_001164219.1;WP_055027268.1 |                               |

Column 1 represents the cluster number; the lower the number the more frequent it is in the figure. The second column shows the frequency of proteins in the cluster. In first row (Cluster 1), 4 means that two protein accessions WP\_001229260.1 and WP\_001229255.1 are 4 times frequent and in cluster 2 you can see 4 proteins accessions are frequent for 4 times.

6. A description file is generated after clustering, with the suffix “\_outdesc.txt”. **This is the main legend file for interpreting the figure files.** For example;

1(2) WP\_001229260.1 MULTISPECIES: integration host factor subunit alpha  
 1(2) WP\_001229255.1 MULTISPECIES: integration host factor subunit alpha

2(1) WP\_001164213.1 MULTISPECIES: phenylalanine--tRNA ligase subunit alpha  
 2(1) WP\_001164217.1 MULTISPECIES: phenylalanine--tRNA ligase subunit alpha  
 2(1) WP\_001164219.1 phenylalanine--tRNA ligase subunit alpha  
 2(1) WP\_055027268.1 phenylalanine--tRNA ligase subunit alpha

Here you can see in column 1, it starts with cluster number 1 (this is the number shown within genes in the figure files) and “(2)” means the protein WP\_001229260.1 has been found twice in the cluster and the last column is the protein description retrieved from the database. The protein description gives information about the proteins in a cluster (accessions and title from NCBI).

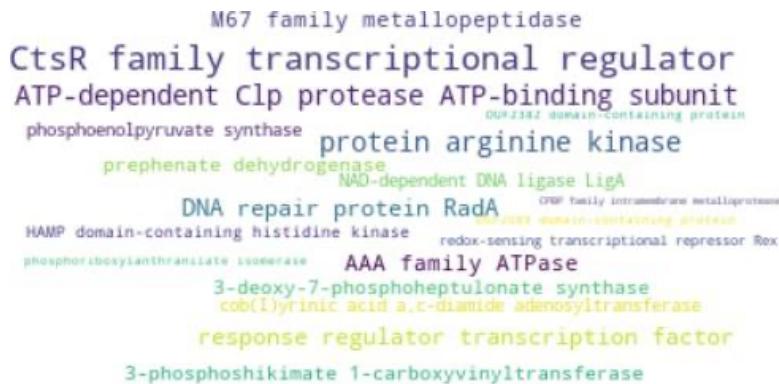
If wanted, this file can be used as an input for the additional bundled python script descriptionCloud.py to save a wordcloud visualization. descriptionCloud.py has additional dependencies: pandas, PIL, wordcloud and matplotlib. If these are not already installed, they are available through anaconda, eg:

```
conda install -c conda-forge wordcloud
```

descriptionCloud.py is run as follows:

```
python descriptionCloud.py -i example_outdesc.txt
```

A png format output image will be created, for example see Figure 5.



**Figure 5. The description wordcloud.** The top 20 most abundant protein descriptions of the flanking genes in an input description file are exhibited in a png format image and the size of the text represents the abundance.

7. Discarded protein ids with improper accession are listed in a text file with “\_NameError.txt” suffix.
8. Discarded protein ids lacking proper information in RefSeq DB are listed in a text file with “\_Insufficient\_Info\_In\_DB.txt” suffix
9. If the -t option used to make a phylogenetic tree, ETE3 saves additional results files to a folder ending with “[output name]\_tree\_”. This includes a Newick format version of the tree.

FlaGs2 also generates some intermediate files to proceed the overall process but only the files mentioned above are useful for interpretation. Intermediate files can be kept with the -k option (warning: these are large files).

## 7. Running online with webFlaGs (Work in progress)

FlaGs2 can be run through the web server at [www.webflags.se](http://www.webflags.se). WebFlaGs is a more user friendly version of FlaGs2 (as it does not require installation) with a web interface where user can paste the input in the specific text area or upload as text file. webFlaGs accepts input datasets that contains up to 200 accessions. If more than 200 queries are required, it is recommended to use the local version of FlaGs. FlaGs2 and webFlaGs both use the same input file format which is autodetected by webFlaGs.

An additional input allowed by webFlaGs is a single accession or sequence (raw or FASTA format). In this case, a BLASTP search is carried out to retrieve homologues that are then used as input for webFlaGs. The user decides how many hits to return (up to 200) and the E-value cut-off, as well as the query database to be searched. The query database can be either the RefSeq database (excluding eukaryotes), or a variation of the eukaryote-excluded RefSeq database that is reduced to one representative genome per bacterial and archaeal species, plus an additional 711 genomes from microbes represented in the COG database and 63 non-eukaryotic NCBI reference genomes.

FlaGs2 User Manual All viral genomes from RefSeq are included. The list of 24464 included taxa can be found

following the link in the interface or directly at <http://130.239.193.227/html>List.txt>.

In webFlaGs, the user can set the E-value cut-off that is used by Jackhmmer to detect homology among all flanking genes. By default, it is 1e-10. The user can also choose the number of Jackhmmer iterations that allows it to find more distant homologs. By default this is set to 3. Like FlaGs, in webFlaGs the user can define the number of upstream and downstream genes to look for flanking each query in the input list. By default, it is 4.

When running webFlaGs, a valid email address is required, which is only used for sending results (we use our own email address for accessing the NCBI API). The user will receive a download link for a zipped file which contains:

1. The one or two figure files described above based on the output type selected in the webFlaGs webpage by the user. There are three output type options in webFlaGs. If the option ‘With phylogenetic tree and query showed as tree order’ is selected, the user will get the annotated phylogenetic tree output and the to-scale postscript output where queries will be presented as the same order that of the annotated phylogenetic tree output. For option ‘With phylogenetic tree but query showed as input order’, queries in the to-scale postscript output will be presented as the same order as submitted by the user, also the user will get the annotated phylogenetic tree output in svg format. If ‘No phylogenetic tree and query showed as input order’ is selected, the user will get only the to-scale postscript output along with the converted PDF file (see below).
2. A converted PDF version of the to-scale postscript output. This is provided for Windows users who can not easily open postscript files.
3. A description file, which is generated after clustering, with the suffix “\_outdesc.txt” (described above). This is the main legend file for the figures.
4. The query input file with the suffix “\_query.flagsIn”

5. A text file with the suffix “\_QueryStatus.txt” suffix that records the success (or not) of each query, and importantly informs if the input accession failed but was converted or updated to an identical RefSeq sequence that worked.
6. Information about the flanking genes for each of the queries retrieved from databases is stored in a tab delimited file named with “\_operon.tsv” suffix. (described above)
7. (Optional)

If the user selected ‘With phylogenetic tree’ as the Output type option, they will receive the multiple sequence alignment FASTA file with the suffix “\_tree.fasta” which was used for making tree file, and the resulting Newick format version of the tree, having a suffix “.nw”.

## 8. Recommendations and tips

When the results are received, it is recommended that users check the tab delimited results file with the “\_QueryStatus.txt” suffix. This records the success (or not) of each query, and importantly informs whether the input accession failed but was converted or updated to an identical RefSeq sequence that worked. When using the -p option, the query status file is also useful for identifying which genome (identified with a GCF number) was used to retrieve the flanking genes.

When using the -to option, the two output figures show input taxa listed in the same order (by the tree). Therefore, in Adobe Illustrator (or similar) it is very easy to collate these two vector images together, in order to make a figure of the to-scale flanking gene output mapped on to the tree.

FlaGs2 can handle relatively large input datasets, although of course the smaller, the faster. Since FlaGs2 retrieves information from NCBI (multiple times for each input accession), it is dependent on both a stable local internet connection and the stability of the NCBI server. Our example input file “GCF\_accession\_input.txt” with 30 accessions takes around 7 minutes to run with the -a, -t, - to options on a 2015 MacBook Pro with an internet connection of around ~75 mbps. Without the -t option it takes around 6 minutes. We have also tested input files containing around 1000 protein accessions (four flanking genes each side) and they worked successfully over the course of a few hours.

If you have >1000 entries, you may want to reduce the number of input accessions. One way to do this is by reducing redundancy with a tool such as Usearch (Edgar 2010).

The -vb (verbose) option is very useful for following the progress of the run, and we recommend its use unless you want to limit what is displayed via STDOUT.

FlaGs2 is designed for use with prokaryotic and bacteriophage genomes where gene clusters can be conserved over vast evolutionary distances, and the space between genes is relatively small. While it will run with eukaryotic nuclear genomes, the running time is much longer and the resulting images can be very wide. For this kind of data, we recommend initial testing with small datasets and few flanking genes (the -g option). FlaGs2 does however work very well with eukaryotic organellar genomes that are more bacteria-like in their organisation and conservation.

As mentioned above, the ‘\_outdesc.txt’ file can be used as an input for the additional bundled python script descriptionCloud.py to save a wordcloud visualisation.

## 9. References

- Eddy SR. 2011. Accelerated Profile HMM Searches. PLoS Comput Biol 7:e1002195.  
Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461.  
Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. Mol Biol Evol 33:1635-1638.

## Appendix 1: Creating an input file for FlaGs2 from an NCBI BlastP or PSI-Blast search

### 1. Run the search against RefSeq proteins:

Go to <https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins>, paste in your query sequence and select the refseq\_protein database. Set any advanced parameters and/or organism limits you like, then click Blast.

### 2. When the results appear, select the proteins of interest

Use the check boxes next to each hit to select the results you're interested in, or leave “select all” checked.

| Descriptions                                      |  | Graphic Summary | Alignments | Taxonomy | Download  | Manage Columns | Show 100                 | ?                  |            |                                |
|---|--|-----------------|------------|----------|-----------|----------------|--------------------------|--------------------|------------|--------------------------------|
| <b>Sequences producing significant alignments</b> |  |                 |            |          |           |                |                          |                    |            |                                |
| <input type="checkbox"/> select all               | 5 sequences selected   |                 |            |          | GenPept   | Graphics       | Distance tree of results | Multiple alignment |            |                                |
|   |  | Description     |            |          | Max Score | Total Score    | Query Cover              | E value            | Per. Ident | Accession                      |
| <input checked="" type="checkbox"/>               | hypothetical protein [Dorea sp. 5-2]                                       |                 |            |          | 424       | 424            | 100%                     | 3e-150             | 100.00%    | <a href="#">WP_016219838.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [Clostridium sp. Marseille-P2538]                     |                 |            |          | 345       | 345            | 100%                     | 5e-119             | 78.85%     | <a href="#">WP_066571301.1</a> |
| <input checked="" type="checkbox"/>               | MULTISPECIES: hypothetical protein [Anaerostipes]                          |                 |            |          | 338       | 338            | 100%                     | 5e-116             | 77.88%     | <a href="#">WP_118476490.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [Lachnospiraceae bacterium A2]                        |                 |            |          | 325       | 325            | 98%                      | 5e-111             | 76.59%     | <a href="#">WP_016304441.1</a> |
| <input checked="" type="checkbox"/>               | hypothetical protein [Ruminococcus sp. 1xD21-23]                           |                 |            |          | 311       | 311            | 99%                      | 9e-106             | 72.95%     | <a href="#">WP_150834685.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [Lachnospiraceae bacterium MD335]                     |                 |            |          | 305       | 305            | 99%                      | 4e-103             | 71.01%     | <a href="#">WP_081645688.1</a> |
| <input checked="" type="checkbox"/>               | hypothetical protein [Lachnospiraceae bacterium MD335]                     |                 |            |          | 304       | 304            | 99%                      | 1e-102             | 70.53%     | <a href="#">WP_162227226.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [bacterium 0.1xD8-71]                                 |                 |            |          | 303       | 303            | 99%                      | 2e-102             | 67.63%     | <a href="#">WP_120411987.1</a> |
| <input checked="" type="checkbox"/>               | hypothetical protein [Lachnospiraceae bacterium oral taxon 500]            |                 |            |          | 303       | 303            | 99%                      | 3e-102             | 69.57%     | <a href="#">WP_009220694.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [Anaerobutyricum hallii]                              |                 |            |          | 300       | 300            | 97%                      | 5e-101             | 70.94%     | <a href="#">WP_096239383.1</a> |
| <input type="checkbox"/>                          | MULTISPECIES: hypothetical protein [unclassified Bacteria (miscellaneous)] |                 |            |          | 298       | 298            | 97%                      | 2e-100             | 68.32%     | <a href="#">WP_129183085.1</a> |
| <input type="checkbox"/>                          | hypothetical protein [Clostridium] hylemonae                               |                 |            |          | 295       | 295            | 98%                      | 4e-99              | 67.80%     | <a href="#">WP_1382619</a>     |

### 3a: If you have <100 sequences selected, you can click “Genpept”

This takes you to a multi-protein summary page. You can change the format to “Accession list” (making sure the number of results shown per page is set high enough).

The screenshot shows the NCBI Protein search interface. At the top, there's a search bar with 'Protein' selected, an 'Advanced' link, a 'Search' button, and a 'Help' link. A red banner at the top of the results page informs users about COVID-19 and provides links to CDC and NIH websites. On the left, there's a sidebar with filters for species (Bacteria, Customize), source databases (RefSeq, Customize), sequence length (Custom range), molecular weight (Custom range), release date (Custom range), revision date (Custom range), and a 'Clear all' link. Below these are 'Show additional filters' and a 'Format' dropdown menu containing options like Summary, GenPept, GenPept (full), FASTA, FASTA (text), ASN.1, Revision History, Accession List, and GI List. The main content area displays a list of protein accessions with their details. One entry is expanded to show 'MULTISPECIES: hypothetical protein [Anaerostipes]' with an accession of WP\_118476490.1, a length of 208 aa, and a taxon of Dorea sp. 5-2. Another entry shows 'hypothetical protein [Ruminococcus sp. 1xD21-23]' with an accession of WP\_150834685.1, a length of 208 aa, and a taxon of Ruminococcus sp. 1xD21-23. To the right, there are sections for 'Results by taxon' (listing Top Organisms, Lachnospiraceae bacterium oral taxon 500, Dorea sp. 5-2, Anaerostipes, Ruminococcus), 'Analyze these sequences' (Run BLAST, Align sequences with COBALT, Identify Conserved Domains with CD-Search), and 'Find related data' (Database: Select). A 'Send to:' link and a 'Filters: Manage Filters' link are also present.

The list of accessions can be pasted into a text file to use as FlaGs2 input. You're done and ready to run FlaGs! You can stop reading here.

```
WP_016219838.1
WP_118476490.1
WP_150834685.1
WP_162227226.1
WP_009220694.1
```

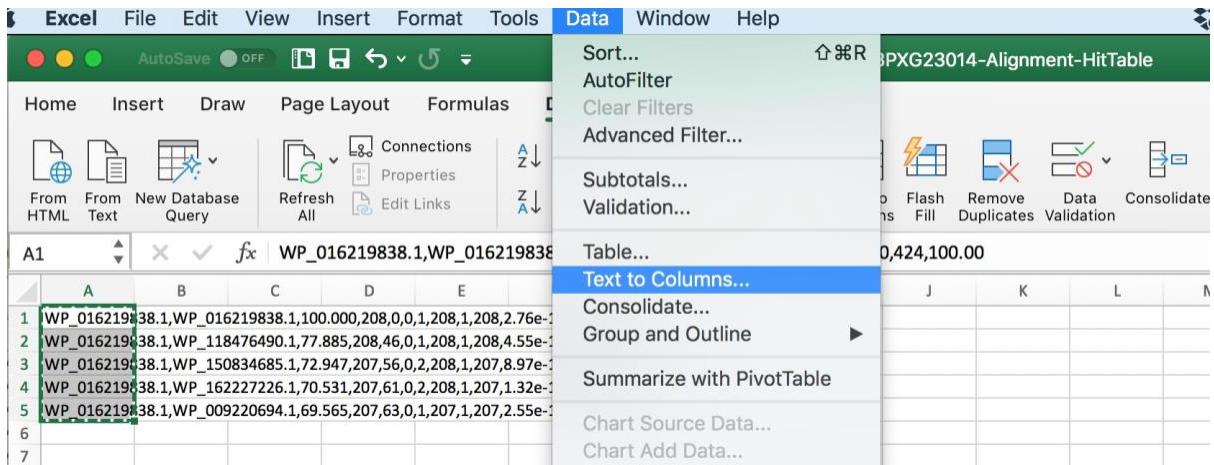
**3b: If you have >100 sequences selected (or just want to try a different way), you can click “Download”, then “Hit Table (CSV)”**

The screenshot shows the NCBI Blast results page. At the top, there are tabs for 'Descriptions', 'Graphic Summary', 'Alignments', and 'Taxonomy'. Below these is a table titled 'Sequences producing significant alignments'. The table has a header row with columns for 'Description' and 'Query', 'E-value', 'Per. Ident.', and 'Accession'. There are also buttons for 'Download', 'Manage Columns', and 'Show 100'. The table lists several hits, each with a checkbox next to it. Some checkboxes are checked, such as for 'hypothetical protein [Dorea sp. 5-2]', 'MULTISPECIES: hypothetical protein [Anaerostipes]', 'hypothetical protein [Lachnospiraceae bacterium A2]', 'hypothetical protein [Ruminococcus sp. 1xD21-23]', 'hypothetical protein [Lachnospiraceae bacterium MD335]', and 'hypothetical protein [Lachnospiraceae bacterium MD335]'. Other checkboxes are unchecked. The table also includes rows for 'FASTA (complete sequence)', 'FASTA (aligned sequences)', 'GenBank (complete sequence)', 'Hit Table (text)', 'Text', 'XML', and 'ASN.1'. The bottom of the table shows a footer with the URL 'nlm.nih.gov/Blast.cgi#protein [Lachnospiraceae bacterium oral taxon 500]' and a green 'E' icon.

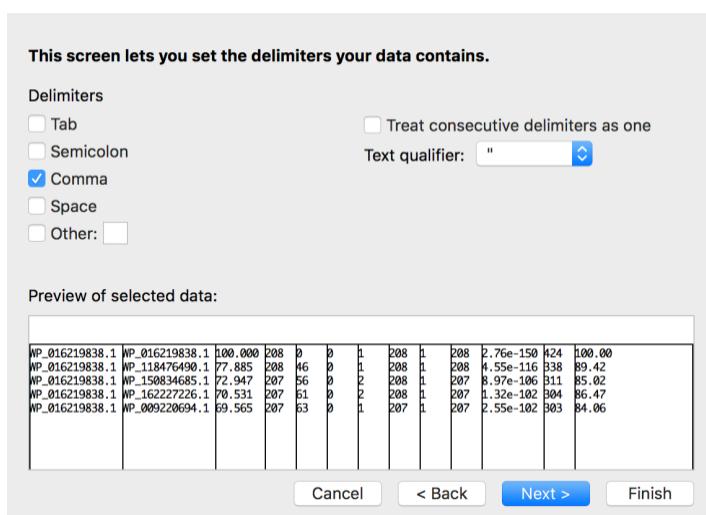
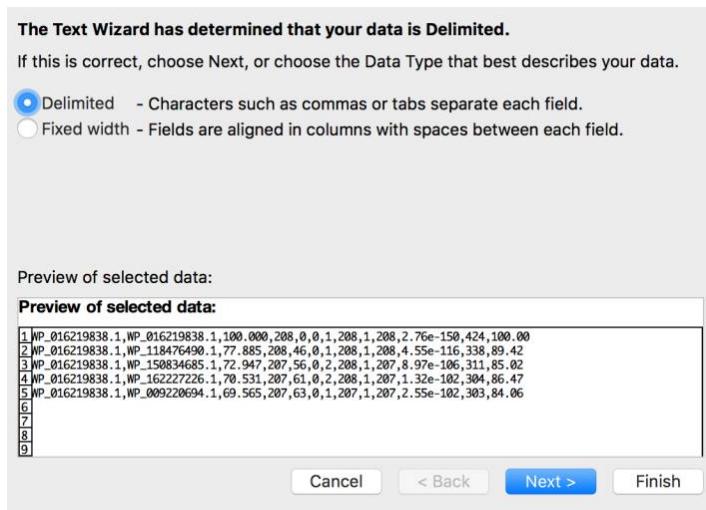
This is a table that can be opened in a spreadsheet program such as Excel, Numbers or Google Sheets

#### 4. Open your table, and split the text into columns

In Excel and Google Sheets this is done through the Data > Text to columns menu option



Follow the wizard to split by commas



Click finish and you will see your list of accessions in the second column

|   | A         | B         | C       | D   | E  | F | G | H   | I | J   | K         | L   | M      |
|---|-----------|-----------|---------|-----|----|---|---|-----|---|-----|-----------|-----|--------|
| 1 | WP_016219 | WP_016219 | 100.000 | 208 | 0  | 0 | 1 | 208 | 1 | 208 | 2.76e-150 | 424 | 100.00 |
| 2 | WP_016219 | WP_118476 | 77.885  | 208 | 46 | 0 | 1 | 208 | 1 | 208 | 4.55e-116 | 338 | 89.42  |
| 3 | WP_016219 | WP_150834 | 72.947  | 207 | 56 | 0 | 2 | 208 | 1 | 207 | 8.97e-106 | 311 | 85.02  |
| 4 | WP_016219 | WP_162227 | 70.531  | 207 | 61 | 0 | 2 | 208 | 1 | 207 | 1.32e-102 | 304 | 86.47  |
| 5 | WP_016219 | WP_009220 | 69.565  | 207 | 63 | 0 | 1 | 207 | 1 | 207 | 2.55e-102 | 303 | 84.06  |
| 6 |           |           |         |     |    |   |   |     |   |     |           |     |        |

The list of accessions can be copied and pasted into a text file to use as FlaGs2 input. You're done and ready to run FlaGs2!

```
WP_016219838.1
WP_118476490.1
WP_150834685.1
WP_162227226.1
WP_009220694.1
```

## Appendix 2: Creating an input file for FlaGs2 using BlastP and our reduced protein database

The **eukaryote-excluded RefSeq** database is reduced to one representative genome per bacterial and archaeal species, all viral RefSeq genomes, plus an additional 711 genomes from microbes represented in the COG database and 63 non-eukaryotic NCBI reference genomes.

Go to <http://130.239.193.227/html/webFlaGsBP.html>, paste in or upload a text file that contains your query as single protein sequence or NCBI protein accession.

The user decides the E-value cut-off and how many hits to return (default 200, maximum 1000). A valid email address is required for sending the user a download link for a zipped file which contains a text file with the suffix “\_query.txt” that comprises a list of NCBI protein accessions processed out from BlastP search against reduced eukaryote-excluded RefSeq database. This file can be used directly as an input for FlaGs2 and webFlaGs2 (if total number of NCBI protein accessions is not more than 200).

The screenshot shows the FlaGs2 web interface. At the top, there is a navigation bar with icons for Home, web, Flags, Download, Manual, and Contact. Below the navigation bar, the title "FlaGs" is displayed in red. A message states: "Our reduced BLAST database contains around 49 million protein sequences which were found in 13118 bacterial, 483 archaeal and 9868 viral genomes. To check the list of genomes [click here](#)." Below this, there is a text input field with placeholder text: "Paste your query in the box below:" followed by "Input can be a single protein sequence or NCBI accession". A "Show Example" link is also present. A large text area contains a sample protein sequence: "HENRPSSIIINVAPAWSKGELKRLGEALFSGTATPEQHARYNEVMLWHNELAAEVAATIYTDDWQACPSLFIDITARPKTDLILQKLQRGRSMSLDEVQDLAGVRIDADIDLKVTALAEELIAHFGEKSRIRDHRRENPHSGYRAVHVWLRLPAGRVEQIIRTVPQSEWANTYERLGRTYGRGIRYGASEDDEARELVERMHINDAALASSEGSTVELAQLEDELIQQLKERLAEMRPLARASGPYIMHRRVSKALDTFRKAETTRRLNQHRSSTLQMLREMRSHLTDGS". Below the input fields, there is a section for uploading a query with a "Browse..." button and a message "No file selected.". Under the "Parameters:" section, there are four input fields: "Your email address\* (results will be sent here):", "Project title (eg. MyProject1):", "BlastP E-value (Default 1e-3):", and "Number of hits returned (Default 200, max 1000)". At the bottom, there are "Submit" and "Reset" buttons.