

# Creating an input file for FlaGs 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.

The screenshot shows the NCBI BLAST search interface. The 'Standard Protein BLAST' form is displayed. The 'Enter Query Sequence' section contains a query sequence: >WP\_016219838.1 hypothetical protein [Dorea sp. 5-2] MVEDILKDKGLSLDKLKLKLSYRSDLGIHLKKNLHYFDRKSLFNLNRINEWYDSCDMLYDIAMDY RIKSI QSAKIYERYYPDHQARKVFDMLGFRALCDNYKEVITLGGCKNIRVADMSGGAHEGDYRGVH VYFQLS. The 'Choose Search Set' section shows the 'Database' dropdown menu open, with 'Reference proteins (refseq\_protein)' selected. Other options include 'Non-redundant protein sequences (nr)', 'Model Organisms (landmark)', 'UniProtKB/Swiss-Prot (swissprot)', 'Patented protein sequences (pataa)', 'Protein Data Bank proteins (pdb)', 'Metagenomic proteins (emv\_nr)', and 'Transcriptome Shotgun Assembly proteins (tsa\_nr)'. The 'Organism' section is set to 'Optional' and 'Exclude' is set to 'Optional'. The 'Program Selection' section is visible at the bottom.

## 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.

The screenshot shows the NCBI BLAST results page. The 'Sequences producing significant alignments' table is displayed. The table has columns for 'Description', 'Max Score', 'Total Score', 'Query Cover', 'E value', 'Per. Ident', and 'Accession'. Several sequences are selected, indicated by checked checkboxes in the first column. The 'GenPept' button is highlighted in the top right corner of the table area.

	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%	WP_016219838.1
<input type="checkbox"/>	hypothetical protein [Clostridium sp. Marseille-P2538]	345	345	100%	5e-119	78.85%	WP_066571301.1
<input checked="" type="checkbox"/>	MULTISPECIES: hypothetical protein [Anaerostipes]	338	338	100%	5e-116	77.88%	WP_118476490.1
<input type="checkbox"/>	hypothetical protein [Lachnospiraceae bacterium A2]	325	325	98%	5e-111	76.59%	WP_016304441.1
<input checked="" type="checkbox"/>	hypothetical protein [Ruminococcus sp. 1xD21-23]	311	311	99%	9e-106	72.95%	WP_150834685.1
<input type="checkbox"/>	hypothetical protein [Lachnospiraceae bacterium MD335]	305	305	99%	4e-103	71.01%	WP_081645688.1
<input checked="" type="checkbox"/>	hypothetical protein [Lachnospiraceae bacterium MD335]	304	304	99%	1e-102	70.53%	WP_162227226.1
<input type="checkbox"/>	hypothetical protein [bacterium 0.1xD8-71]	303	303	99%	2e-102	67.63%	WP_120411987.1
<input checked="" type="checkbox"/>	hypothetical protein [Lachnospiraceae bacterium oral taxon 500]	303	303	99%	3e-102	69.57%	WP_009220694.1
<input type="checkbox"/>	hypothetical protein [Anaerobutyricum hallii]	300	300	97%	5e-101	70.94%	WP_096239383.1
<input type="checkbox"/>	MULTISPECIES: hypothetical protein [unclassified Bacteria (miscellaneous)]	298	298	97%	2e-100	68.32%	WP_129183085.1
<input type="checkbox"/>	hypothetical protein [[Clostridium] hylemonae]	295	295	98%	4e-99	67.80%	WP_1382619

## 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).

Protein   [Help](#)

Advanced

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.

Species: Bacteria (5) [Customize ...](#)

Source databases: RefSeq (5) [Customize ...](#)

Sequence length: Custom range...

Molecular weight: Custom range...

Release date: Custom range...

Revision date: Custom range...

[Clear all](#) [Show additional filters](#)

Summary | Sort by Default order

**Format**

- Summary
- GenPept
- GenPept (full)
- FASTA
- FASTA (text)
- ASN.1
- Revision History
- Accession List
- GI List

1. [hypothetical protein \[Dorea sp. 5-2\]](#)  
Accession: WP\_016219838.1 GI: 510884812  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[Proteins](#) [FASTA](#) [Graphics](#)

2. **208 aa protein**  
Accession: WP\_118476490.1 GI: 1474059822  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

3. **208 aa protein**  
Accession: WP\_150834685.1 GI: 1755181999  
[BioProject](#) [Nucleotide](#) [Taxonomy](#)  
[GenPept](#) [Identical Proteins](#) [FASTA](#) [Graphics](#)

Send to: [Filters: Manage Filters](#)

**Results by taxon**

Top Organisms [\[Tree\]](#)

- Lachnospiraceae bacterium oral taxon 500 (1)
- Lachnospiraceae bacterium MD335 (1)
- Dorea sp. 5-2 (1)
- Anaerostipes (1)
- Ruminococcus sp. 1xD21-23 (1)

**Analyze these sequences**

Run BLAST

Align sequences with COBALT

Identify Conserved Domains with CD-Search

**Find related data**

Database:

The list of accessions can be pasted into a text file to use as FlaGs 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)"**

**Descriptions** | Graphic Summary | Alignments | Taxonomy

**Sequences producing significant alignments**

select all 5 sequences selected

Description	Query cover	E value	Per. Ident	Accession	
<input checked="" type="checkbox"/> <a href="#">hypothetical protein [Dorea sp. 5-2]</a>	100%	3e-150	100.00%	WP_016219838.1	
<input type="checkbox"/> <a href="#">hypothetical protein [Clostridium sp. Marseille-P2538]</a>	0%	5e-119	78.85%	WP_066571301.1	
<input checked="" type="checkbox"/> <a href="#">MULTISPECIES: hypothetical protein [Anaerostipes]</a>	100%	5e-116	77.88%	WP_118476490.1	
<input type="checkbox"/> <a href="#">hypothetical protein [Lachnospiraceae bacterium A2]</a>	98%	5e-111	76.59%	WP_016304441.1	
<input checked="" type="checkbox"/> <a href="#">hypothetical protein [Ruminococcus sp. 1xD21-23]</a>	99%	9e-106	72.95%	WP_150834685.1	
<input type="checkbox"/> <a href="#">hypothetical protein [Lachnospiraceae bacterium MD335]</a>	9%	4e-103	71.01%	WP_081645688.1	
<input checked="" type="checkbox"/> <a href="#">hypothetical protein [Lachnospiraceae bacterium MD335]</a>	9%	1e-102	70.53%	WP_162227226.1	
<input type="checkbox"/> <a href="#">hypothetical protein [bacterium 0.1xD8-71]</a>	303	303	99%	2e-102	WP_1204119
<input type="checkbox"/> <a href="#">hypothetical protein [Lachnospiraceae bacterium oral taxon 500]</a>	303	303	99%	3e-102	WP_0092206

Show

FASTA (complete sequence) [ance tree of results](#) [Multiple alignment](#)

FASTA (aligned sequences)

GenBank (complete sequence)

Hit Table (text)

**Hit Table (CSV)**

Text

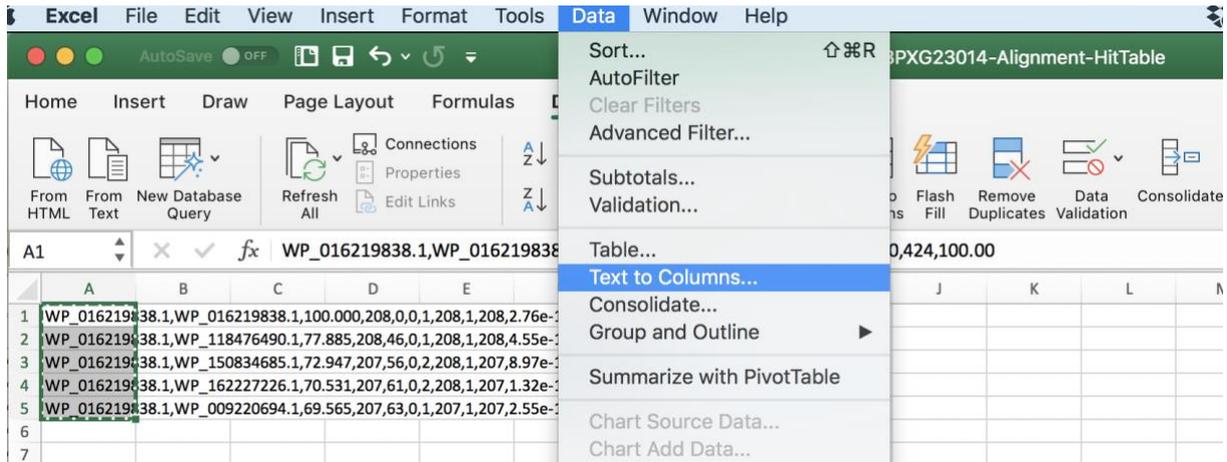
XML

ASN.1

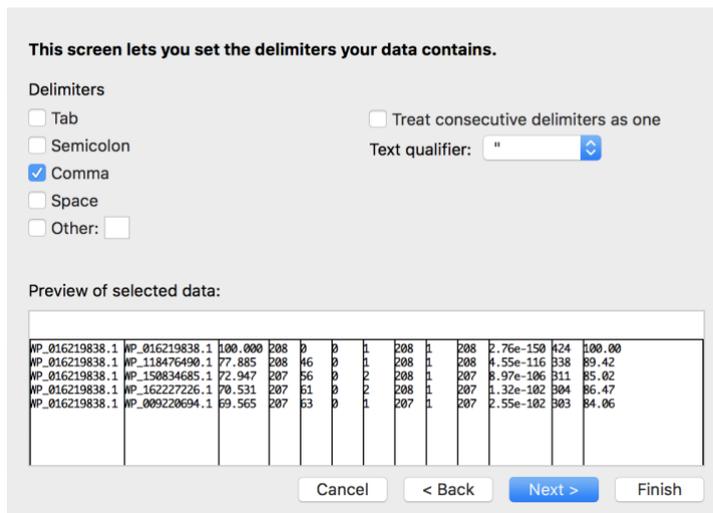
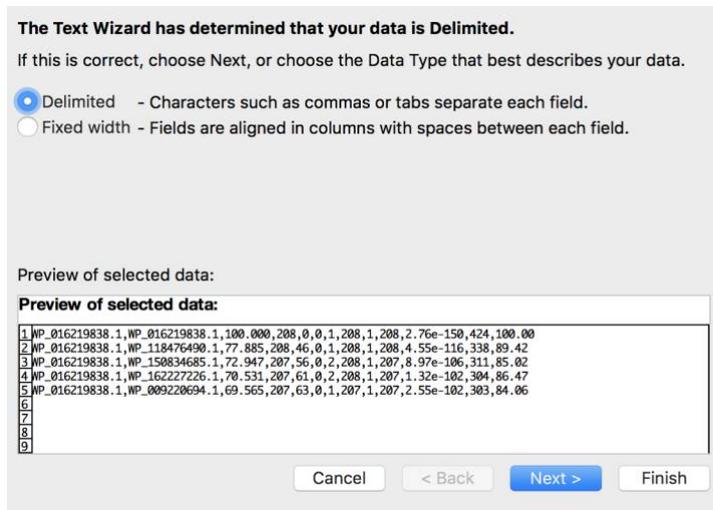
This is a table that can be opened in a spreadsheet program such as Excel, Numbers or Google Sheets

#### 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_016219838.1	WP_016219838.1	100.000	208	0	0	1	208	1	208	2.76e-150	424	100.00
2	WP_016219838.1	WP_118476490.1	77.885	208	46	0	1	208	1	208	4.55e-116	338	89.42
3	WP_016219838.1	WP_150834685.1	72.947	207	56	0	2	208	1	207	8.97e-106	311	85.02
4	WP_016219838.1	WP_162227226.1	70.531	207	61	0	2	208	1	207	1.32e-102	304	86.47
5	WP_016219838.1	WP_009220694.1	69.565	207	63	0	1	207	1	207	2.55e-102	303	84.06

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

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