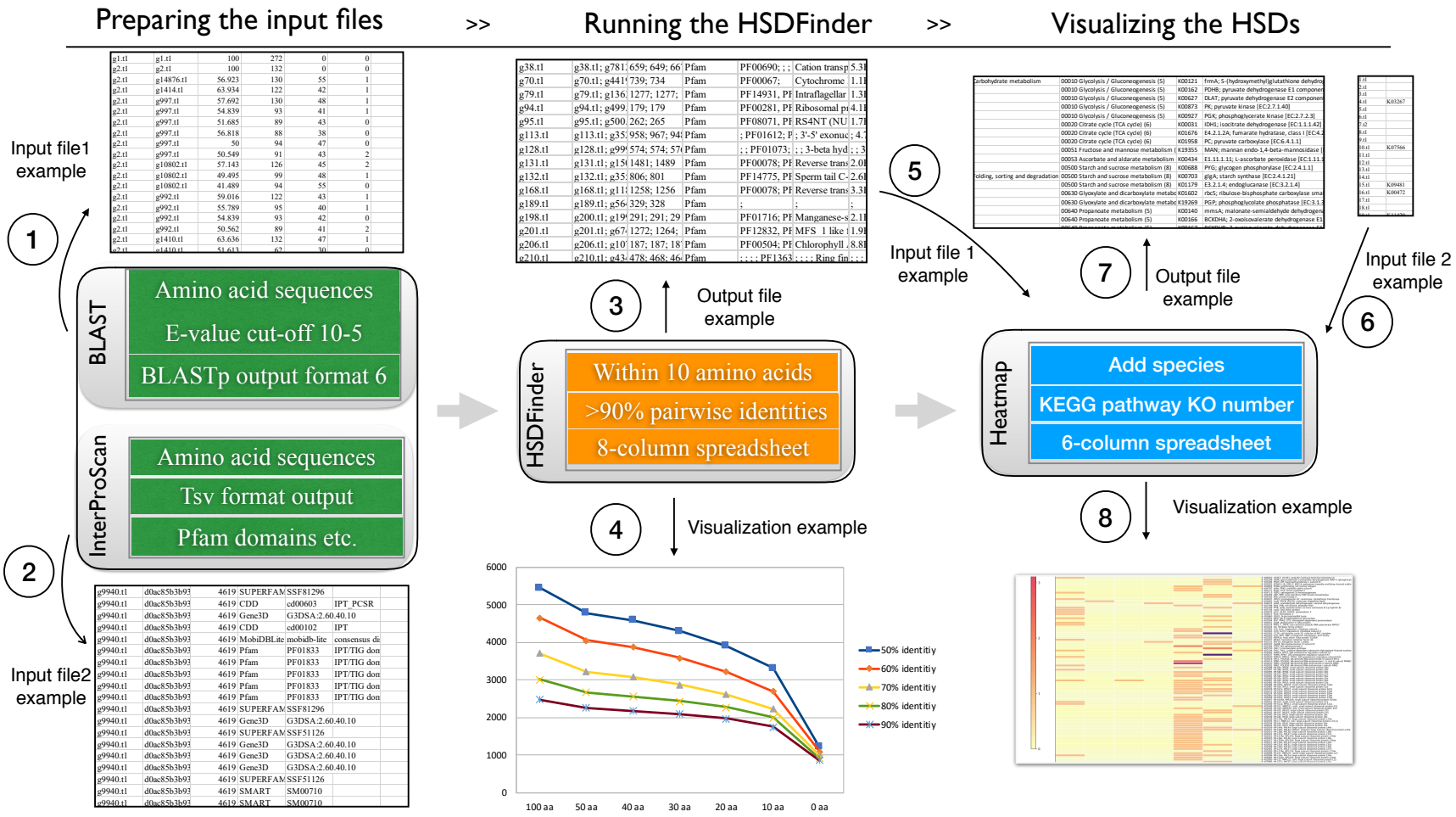
HSDFinder Online server Tutorial: Identification of highly similar duplicates in eukaryotic genomes

1. Upload a protein BLAST search result file of your genome in tab-separated values (tsv) format as the first input file (File 1) of HSDFinder.
2. Upload a InterProScan search result file of your genome in tab-separated values (tsv) format as the second input file (File 2) of HSDFinder.
3. Yielding the output of HSDFinder with three personalized options.
4. Visualizing the HSDFinder outputs via the Excel tools (optional).
5. Upload the results of HSDFinder from your respective genomes.
6. Upload a gene list with KO annotation from KEGG database.
7. The output files of the online Heatmap Visualization tool.
8. The heatmap of HSDs levels across species.

Workflow of HSDFinder (8 steps)



Preparing the appropriate input files.
First, before running HSDFinder to acquire the HSDs of your interest genome, there are two spreadsheets (File 1 and File 2) in tab-separated values (tsv) format shall be prepared as input files.

Step 1: Upload a protein BLAST search result file of your genome in tab-separated values (tsv) format as the first input file (File 1) of HSDFinder.

- Using BLASTP search of the amino acid sequences of genome against themselves (The BLASTP running parameters: E-value cut-off 10-5, BLASTP output format 6)
- The BLAST results should be 12-column spreadsheets including the key information from query name to percentage identity etc. For the **File 1 example** please check the red arrows below.
- The BLAST Package can be found via https://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
  Please download the appropriate tools based on your computer systems (Windows, MacOS or Linux).

Here, this tutorial will go over how to acquire local BLASTP search results via an example FASTA file (The FASTA file can be found via the link https://github.com/zx0223winner/HSDFinder/blob/master/Example.fasta). You can ignore this step if you know how to acquire the BLASTP searching results by other ways.

```
$ head Example.fasta
>g735.t1
MTMRSPTAAAALALLLLACTAGIQCARPLLQSEDEVDCSNPPPQAPVDLGNASAFVVLAGHEVTTGAGVASRILGDLG
VSPGNSVTGQPDMLEGSVIQKANGASAAGIRDLGIAYNDAAGRTVCPILVAGELGGMTLYPGLYKSSSGLEITDADLTL
SGDGIFIFQMATTFLMTKTMKVTLTNGAQAKNIFWQVGTSATLMDKSVLYGTILADQSITSGTGAVIHGRALARIASVT
MESAVFSLPAE
>g741.t1
MTMRSPTVAAALVLLLLACTAGVQCARPLLQSEDAVDCSNPPPQAPVDLRNATTFVVLAGHEVTTGAGVASRILSDLG
VSPGSSVTGQPDMLEGSVIQKANGASAAGIRDLGIAYNDAAGRTVCPILVAGELGGMTLYPGLYKSSSGLEITDANLTLS
GDGVFIFQMATTFLMTKTMKVTLTNGAQAKNIFWQVGTSATLMDKSVLYGTILADQSITSGTGAVIHGRALARIASVT
MESAVFSLPAE
```

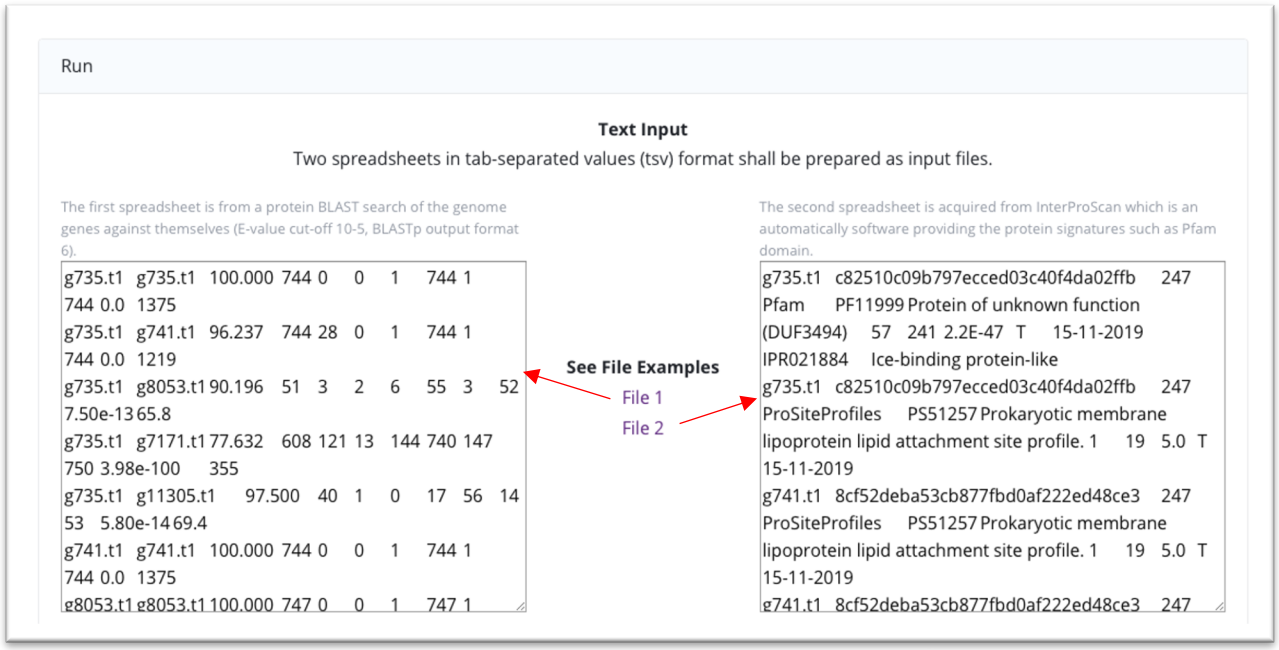
Build a Database via the example FASTA file. Check command line below:

```
$ makeblastdb -in Example.fasta -dbtype prot Example_database_name
```

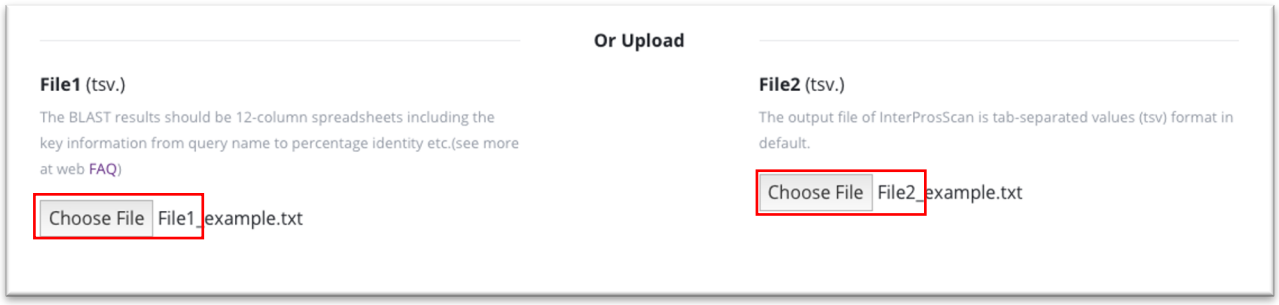
Using BLASTP search option to blast the amino acid sequences against themselves:

```
$ blastp -query IBP.FASTA -db Example_database_name -out File1_example_BALSTP_result.txt -evalue 1e-5 -
outfmt 6
```

This will give a BLAST result file form by 12-column spreadsheets including the key information from query name to percentage identity etc. Check the File 1 example below on the left-hand side.



Users can either paste or upload a BLASTP search result as the input of file 1.



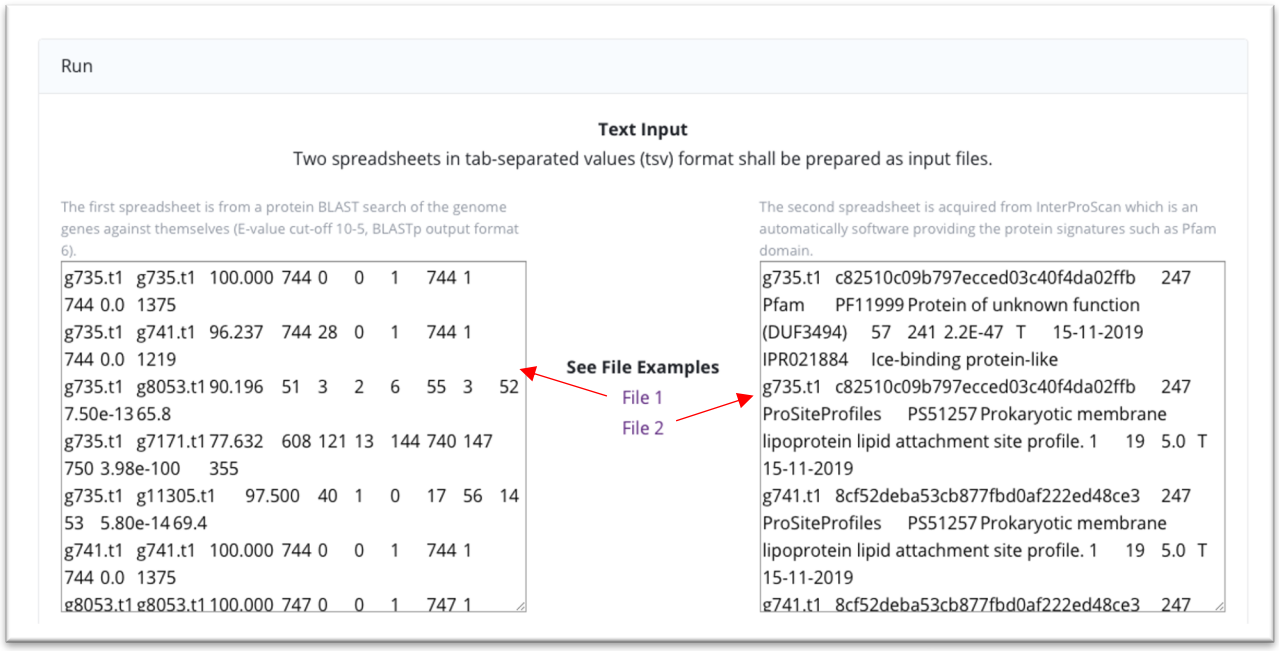
Step 2: Upload a InterProScan search result file of your genome in tab-separated values (tsv) format as the second input file (File 2) of HSDFinder.

- InterProScan which is easy-to-use software providing the protein signatures such as Pfam domain. The latest InterProScan software can be found via the link https://www.ebi.ac.uk/interpro/download/

- After finishing downloading, you can test the InterProScan software locally with your interest genome (protein sequence in FASTA file). Check the commands below: More usage instructions can be found via the link https://interproscan-docs.readthedocs.io/en/latest/HowToRun.html

  ```
  ./interproscan.sh -i proteins_of_your_genome.fasta -f tsv -dp
  ```

- The output file of InterProsScan is tab-separated values (tsv) format in default. Please find the example (File 2) on the right-hand side (see below).

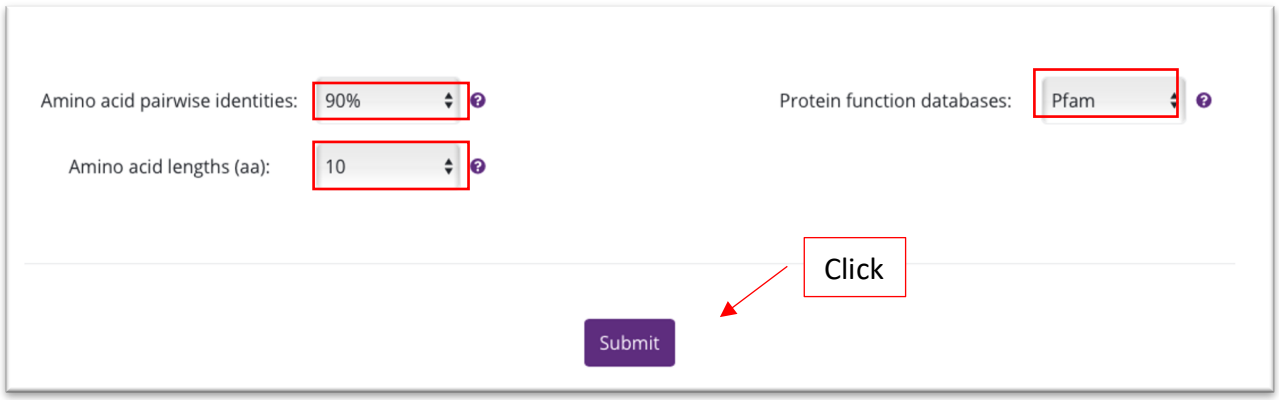


Runing the HSDFinder

Beforing clicking the submission button, there are three personalized options designing for HSDFinder (Amino acid pairwise identity, amino acid length difference, and protein function database)

Step 3: Yielding the output of HSDFinder with three personalized options.

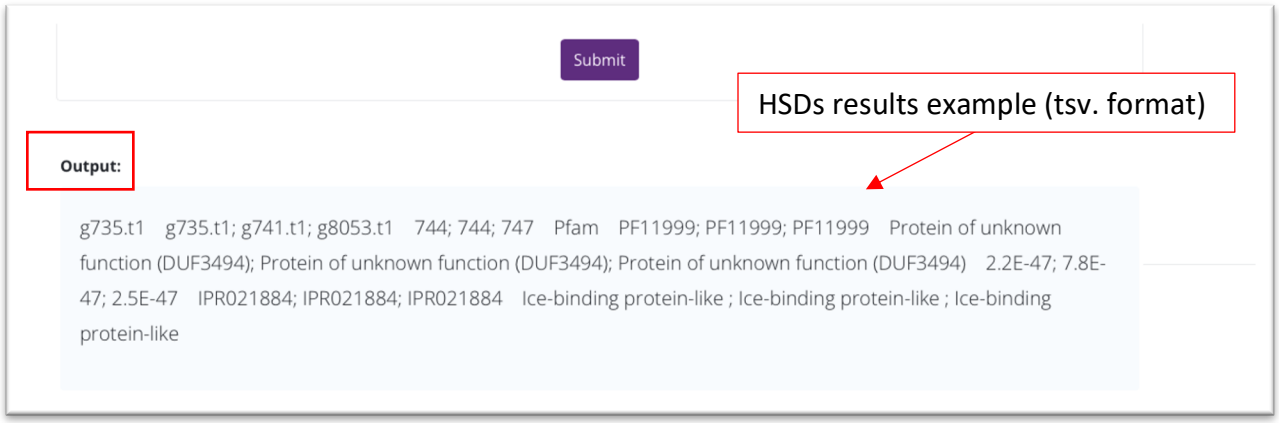
- The HSDFinder is set default to filter those with near-identical protein lengths (within 10 amino acids) and >90% pairwise identities. By setting such a strict cut-off, user should be able to screen out most of highly identical duplicates, but there are few genuine duplicates might be falling out of the threshold.



- Nevertheless, the users always have an option to try different parameters from 30% to 100% identity or from within 10 aa to 100 aa variances to acquire the duplicates in need.

But the user should be careful with the false-positive rate of HSDs will increase (not genuine duplicates) with larger amino acid variance and lower amino acid pairwise identity.

- The output of this step will be an 8-column spreadsheet integrating with the information of HSD identifier, gene copies number and Pfam domain.



Output:

HSDFinder generates one output files: 8-column spreadsheet integrating with the information of HSD identifier, gene copies number and Pfam domain.

Example of the 8-column spreadsheet:

g735.t1 g735.t1; g741.t1; g8053.t1 744; 744; 747 Pfam PF11999; PF11999; PF11999 Protein of unknown function (DUF3494); Protein of unknown function (DUF3494); Protein of unknown function (DUF3494) 2.2E-47; 7.8E-47; 2.5E-47 IPR021884; IPR021884; IPR021884 Ice-binding protein-like; Ice-binding protein-like ; Ice-binding protein-like

Column explanation:

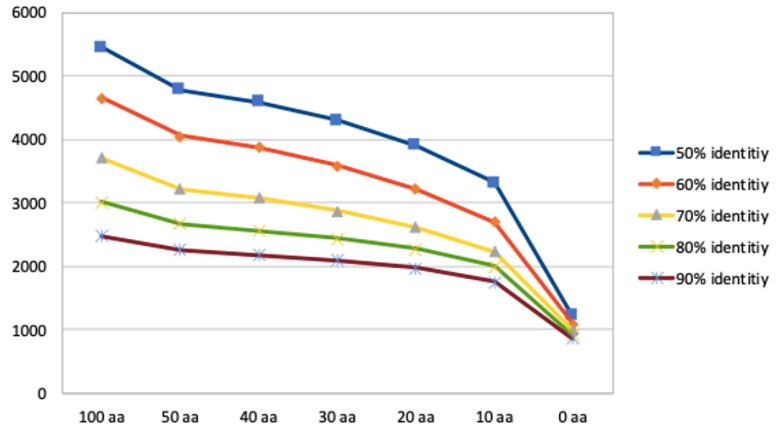
1. Highly Similar Duplicates (HSDs) identifiers: The first gene model of the duplicate gene copies is used as the HSD identifers in default.
2. Duplicate gene copies (within 10 amino acids, ≥90% pairwise identities):g735.t1; g741.t1; g8053.t1
3. Amino acid length of duplicate gene copies (aa):744; 744; 747
4. Pfam identifier:PF11999; PF11999; PF11999
5. Pfam Description: Protein of unknown function (DUF3494); Protein of unknown function (DUF3494); Protein of unknown function (DUF3494)
6. InterPro Entry Identifier: IPR021884; IPR021884; IPR021884
7. InterPro Entry Description: Ice-binding protein-like ; Ice-binding protein-like ; Ice-binding protein-like

Step 4: Visualizing the HSDFinder outputs via the Excel tools (Optional)

Additionally, the user can conveniently set different values to create a trendline graph of the gene copies numbers under different criteria. Check the example we used below. The genome datasets are from a psychrophilic green alga Chlamydomonas sp. UWO241 (NCBI BioProject: PRJNA547753).

	50% identitiy	60% identitiy	70% identitiy	80% identitiy	90% identitiy
100 aa	5453	4654	3719	3012	2476
50 aa	4783	4061	3228	2672	2257
40 aa	4594	3879	3088	2565	2181
30 aa	4310	3596	2875	2438	2091
20 aa	3912	3224	2614	2270	1973
10 aa	3320	2699	2234	2012	1339
0 aa	1223	1097	999	926	859

The table of total gene copies number at different thresholds of amino acid pairwise identity and deduced amino acid length.



The line graph of duplicates set to different thresholds of amino acid pairwise identity and deduced amino acid length. The X-axis indicates the deduced amino acid length (aa) of each duplicate, the Y-axis tells the number of gene copies.

Visualizing the HSDs across species

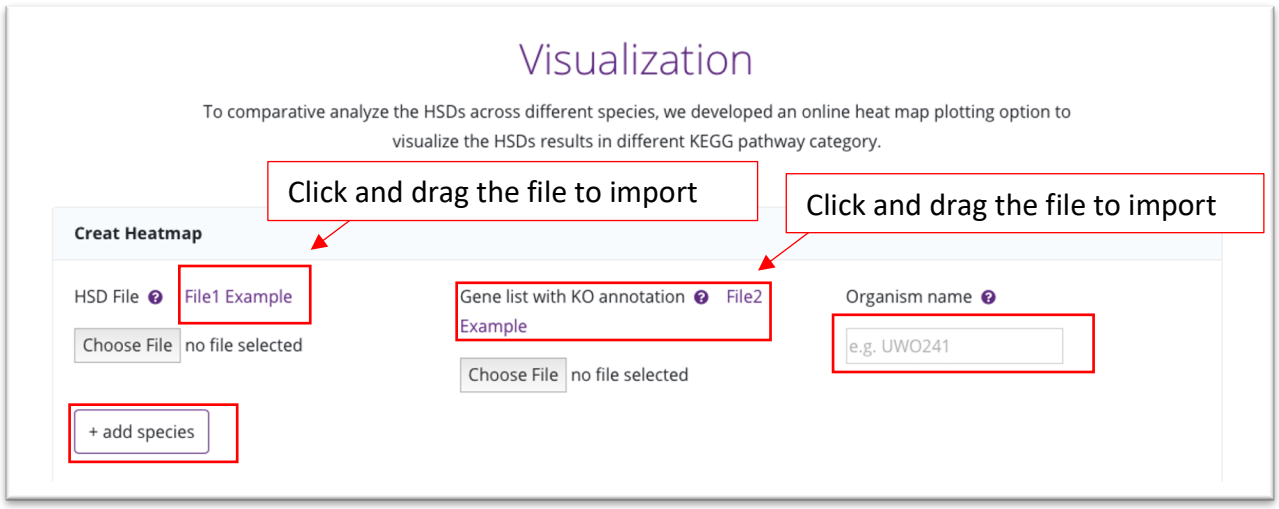
If you have more than one species needing to compare the levels of HSDs in each genome, the following option of Online Heatmap tool is a great choice. User can use the online heatmap plotting option to visualize the HSDs results in different KEGG pathway category.

Step 5: Upload the results of HSDFinder from your respective genomes.

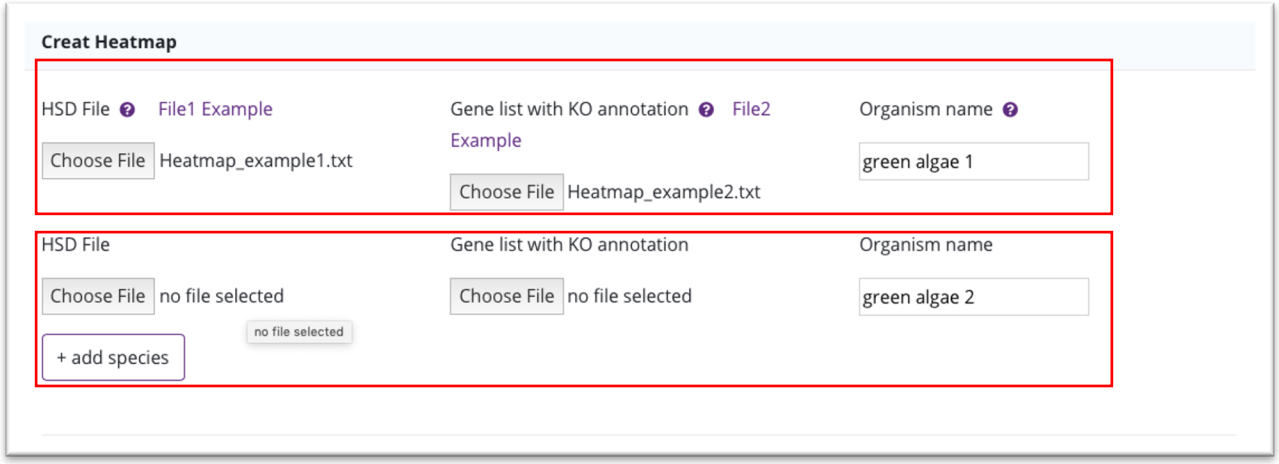
There will be two files needed to plot the heatmap for each species. First input file is the output of your interest species after running the HSDFinder (the step 3), file examples are given to guide the appropriate input file (see the File 1 example below in the image).

Step 6: Upload a gene list with KO annotation from KEGG database.

- The second file is retrieved from the KEGG database documented the correlation of KO accession with each gene model identifier.
- Using the Ghost KOALA analysis tool of KEGG to acquire the KO annotation file of your genome. https://www.kegg.jp/ghostkoala/ File examples are given to guide the appropriate input file (see the File 2 example below).
- Fill in the organism name with respective to your species, this is the identifier to compare levels of HSDs in different species.

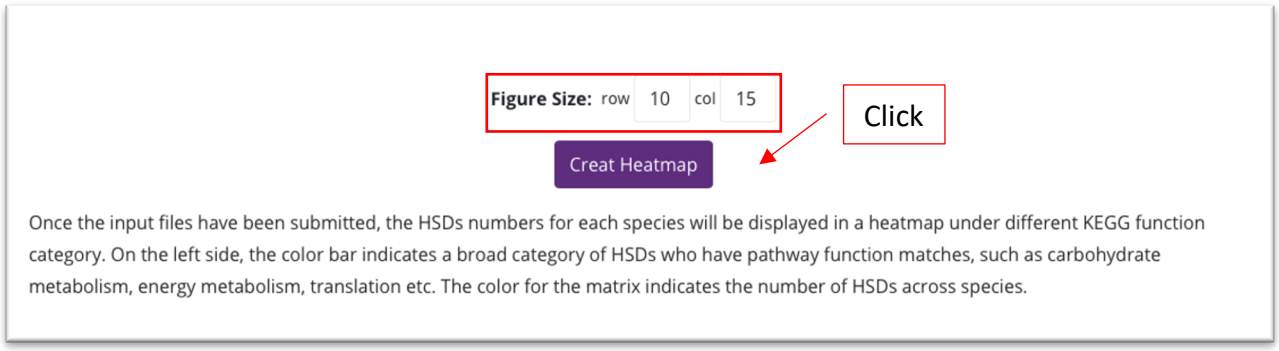


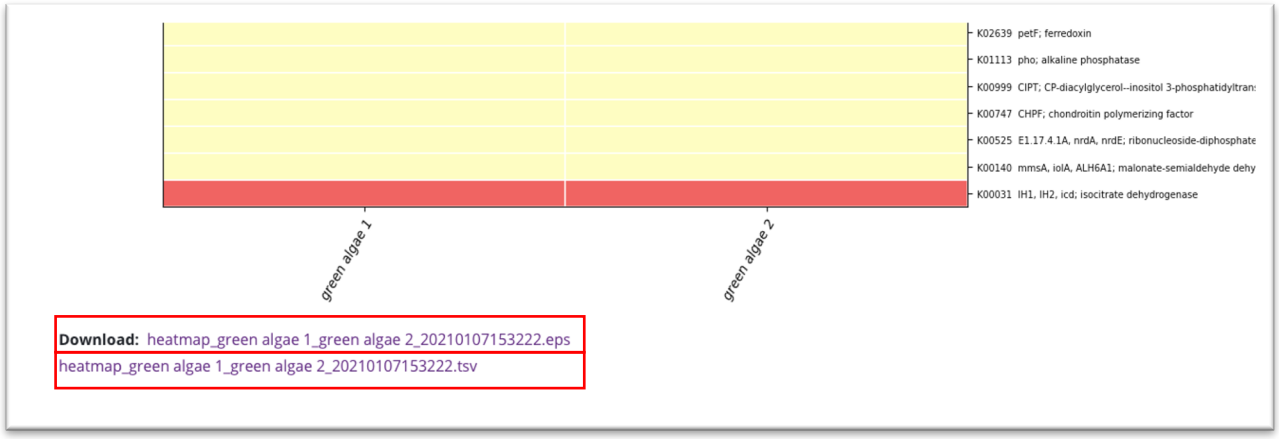
To add more species, you can simply touch the +add species button and fill in the respective files (Repeat step 5 and 6). To best present your visualization result, at least two species shall be submitted. You can still visualize the result via a single species though, if you simply want to try.



Step 7: The output files of the online Heatmap Visualization tool.

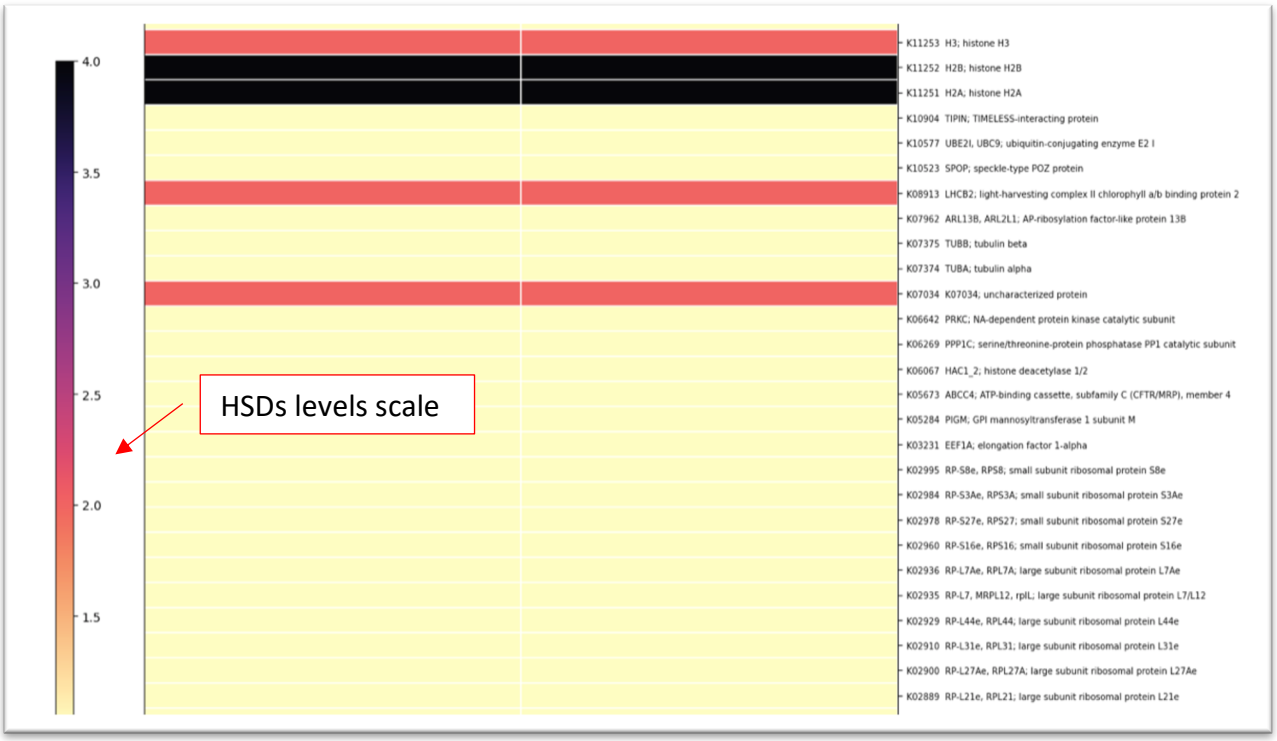
User can feel free to set the Figure size via the options of row and col. After clicking the Create Heatmap button. There will be a heatmap below as well as two documents generated. One is the high resolution heatmap in eps. format (image). Another one is the HSDs content details in tsv. format (text), which user can be conveniently used for future analysis. (This spreadsheet file includes information about HSDs under different KEGG function category such as carbohydrate metabolism, energy metabolism, translation etc.)





Step 8: The heatmap of HSDs levels across species.

Once the input files have been submitted, the HSDs numbers for each species will be displayed in a heatmap under different KEGG function category. On the left side, the color bar indicates a broad category of HSDs who have pathway function matches, such as carbohydrate metabolism, energy metabolism, translation etc. The color for the matrix indicates the number of HSDs across species.


4.0
3.5
3.0
2.5
2.0
1.5
HSDs levels scale
K11253 H3; histone H3
K11252 H2B; histone H2B
K11251 H2A; histone H2A
K10904 TIPIN; TIMELESS-interacting protein
K10577 UBE2I, UBC9; ubiquitin-conjugating enzyme E2 I
K10523 SPOP; speckle-type POZ protein
K08913 LHCB2; light-harvesting complex II chlorophyll a/b binding protein 2
K07962 ARL13B, ARL2L1; AP-ribosylation factor-like protein 13B
K07375 TUBB; tubulin beta
K07374 TUBA; tubulin alpha
K07034 K07034; uncharacterized protein
K06642 PRKC; NA-dependent protein kinase catalytic subunit
K06269 PPP1C; serine/threonine-protein phosphatase PP1 catalytic subunit
K06067 HAC1_2; histone deacetylase 1/2
K05673 ABCC4; ATP-binding cassette, subfamily C (CFTR/MRP), member 4
K05284 PIGM; GPI mannosyltransferase 1 subunit M
K03231 EEF1A; elongation factor 1-alpha
K02995 RP-S8e, RPS8; small subunit ribosomal protein S8e
K02984 RP-S3Ae, RPS3A; small subunit ribosomal protein S3Ae
K02978 RP-S27e, RPS27; small subunit ribosomal protein S27e
K02960 RP-S16e, RPS16; small subunit ribosomal protein S16e
K02936 RP-L7Ae, RPL7A; large subunit ribosomal protein L7Ae
K02935 RP-L7, MRPL12, rplL; large subunit ribosomal protein L7/L12
K02929 RP-L44e, RPL44; large subunit ribosomal protein L44e
K02910 RP-L31e, RPL31; large subunit ribosomal protein L31e
K02900 RP-L27Ae, RPL27A; large subunit ribosomal protein L27Ae
K02889 RP-L21e, RPL21; large subunit ribosomal protein L21e