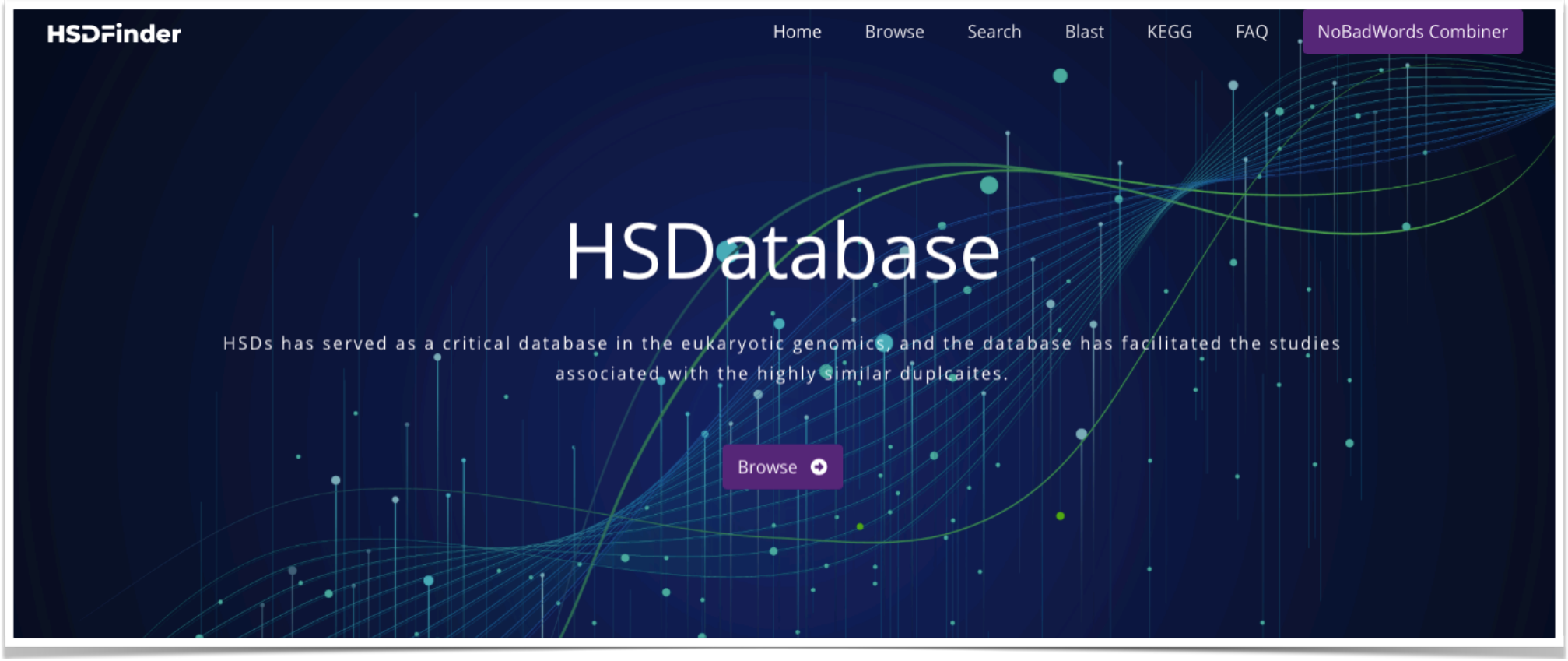
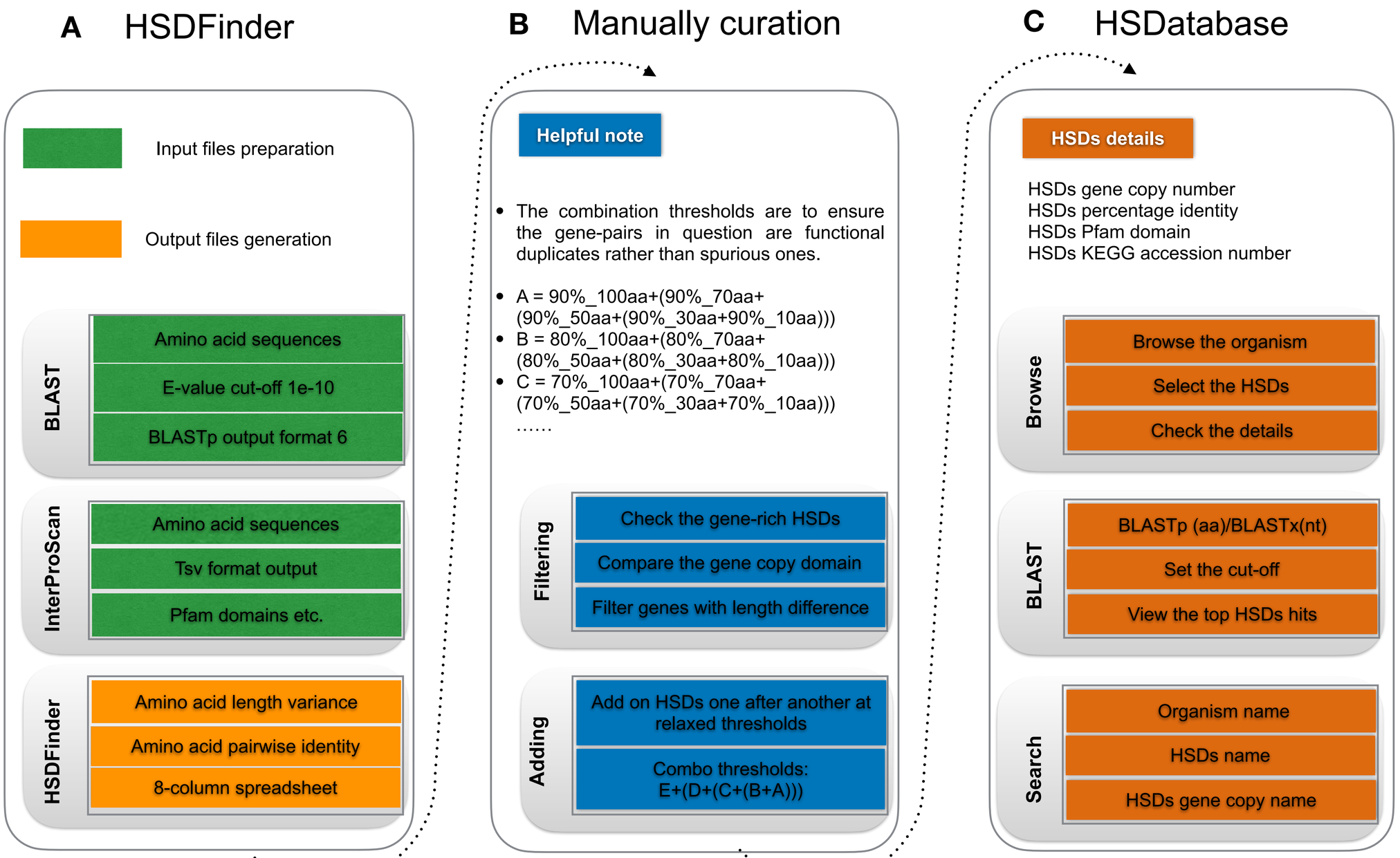
Part 1: HSDatabase interface and implementation (www.hsdfinder.com/database/)



Since the numbers of highly similar duplicates in other eukaryotic genomes are largely unknown, and computational methods for identifying them can be time-consuming and labor-intensive. We created a database to collect HSDs from some model species.



We followed the workflow to identify the HSDs by hsdfinder, then added the HSDs number based on a combined threshold, and lastly, stored them in the database. It is our hope to build a comparative analysis framework across species, especially for those extremophiles, to understand the role of gene duplication in different survival environments.

>> HSDatabase

Step 1:

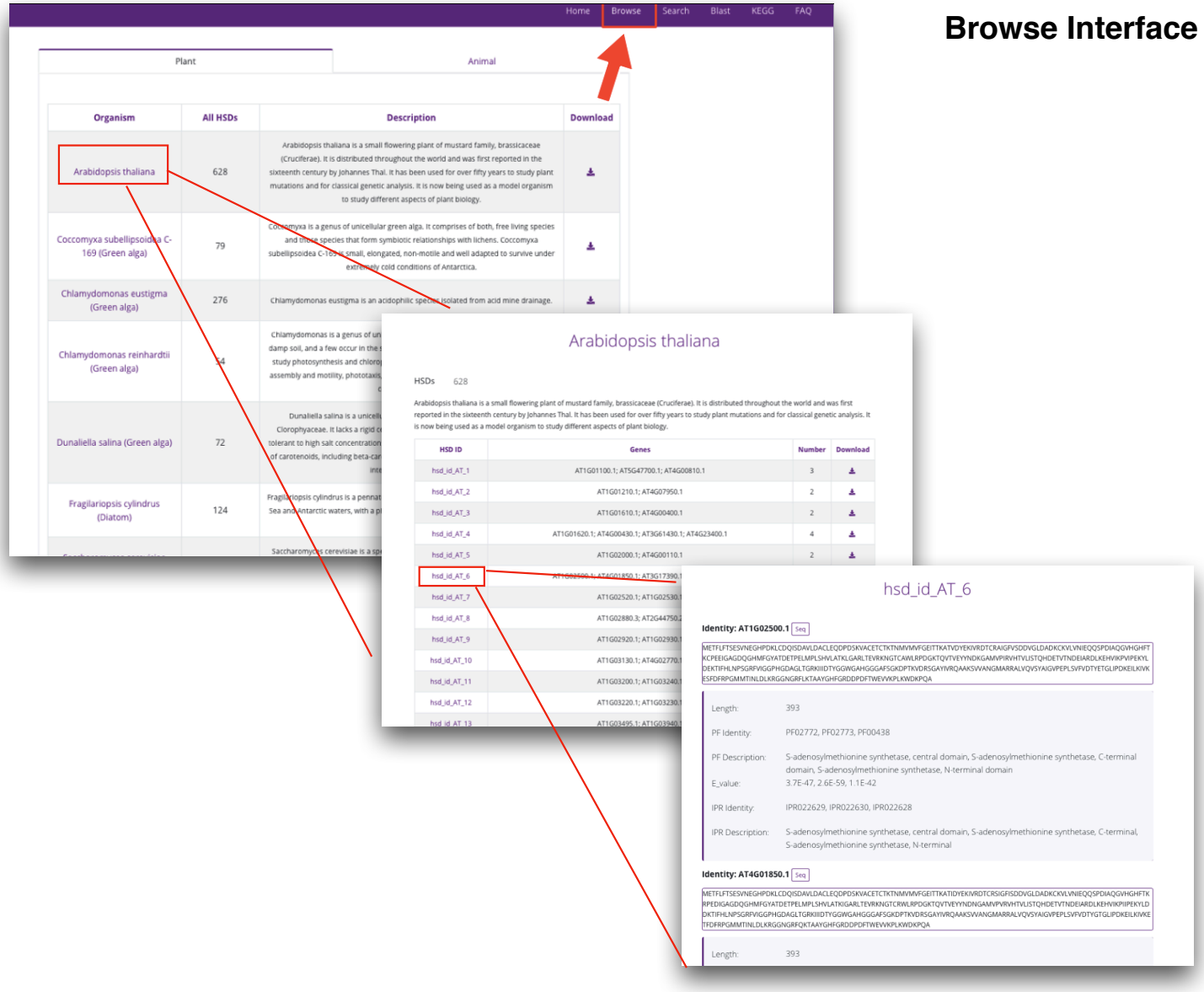
Browse Interface

HSDs details

HSDs gene copy number
HSDs percentage identity
HSDs Pfam domain
HSDs KEGG accession number

Browse

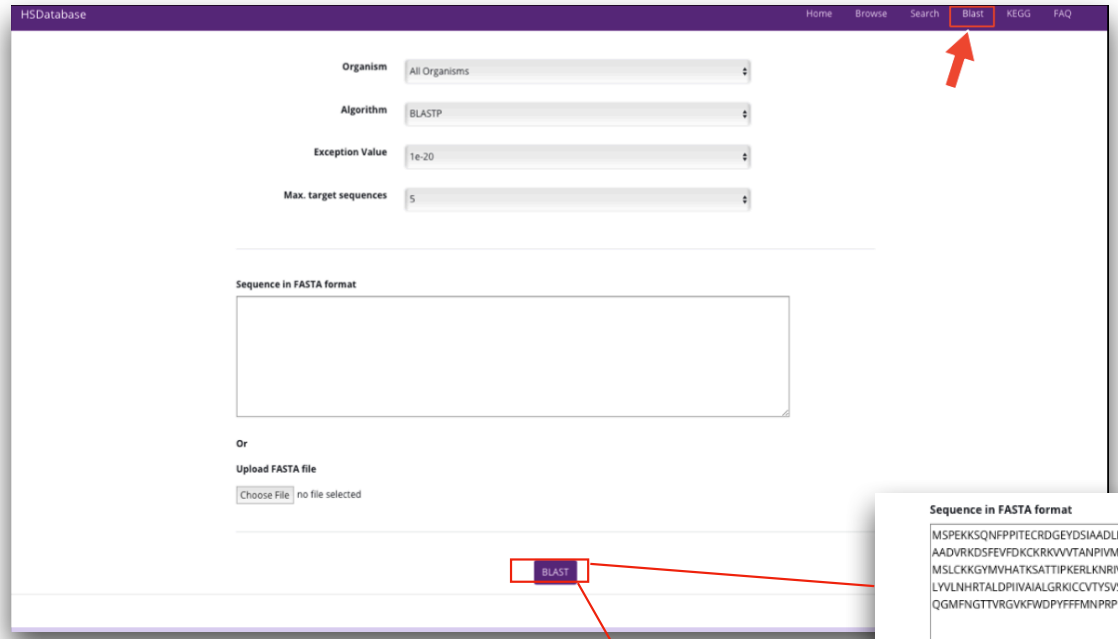
- Browse the organism
- Select the HSDs
- Check the details



Here is the interface of HSDatabse, by choosing the browse option tab, and selecting the arabidopsis, we collected the detailed entries about the HSDs incluing it is number , function domian, and pathway.

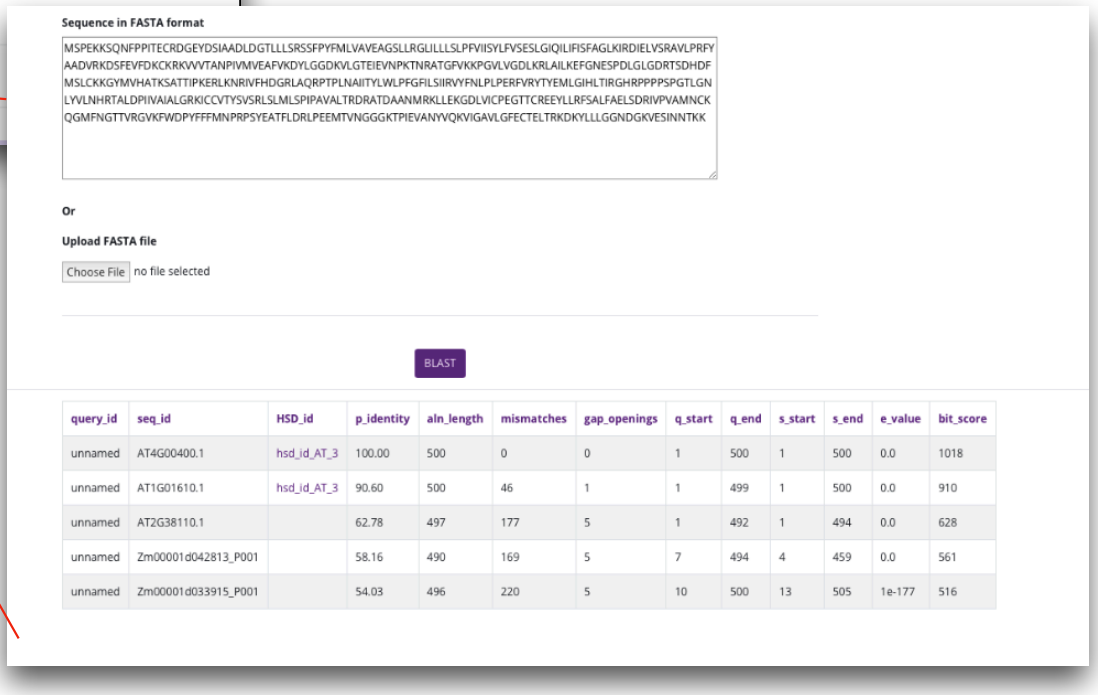
Step 2:

BLAST search interface



BLAST

- BLASTp (aa)/BLASTx(nt)
- Set the cut-off
- View the top HSDs hits

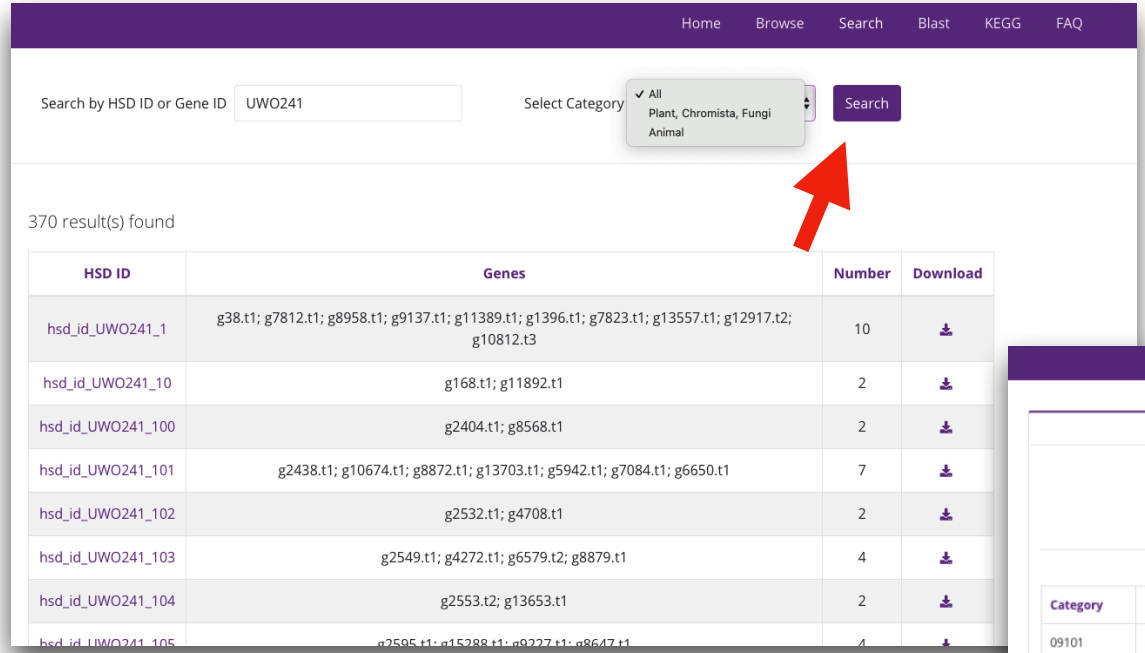


query_id	seq_id	HSD_id	p_identity	aln_length	mismatches	gap_openings	q_start	q_end	s_start	s_end	e_value	bit_score
unnamed	AT4G00400.1	hsd_id_AT_3	100.00	500	0	0	1	500	1	500	0.0	1018
unnamed	AT1G01610.1	hsd_id_AT_3	90.60	500	46	1	1	499	1	500	0.0	910
unnamed	AT2G38110.1		62.78	497	177	5	1	492	1	494	0.0	628
unnamed	Zm00001d042813_P001		58.16	490	169	5	7	494	4	459	0.0	561
unnamed	Zm00001d033915_P001		54.03	496	220	5	10	500	13	505	1e-177	516

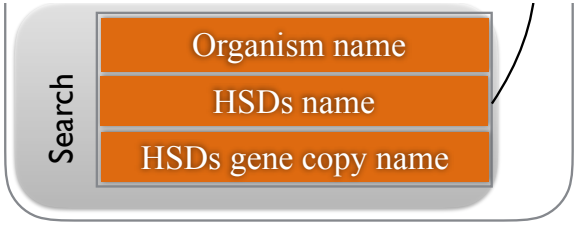
User can also select the BLAST option to search against HSDatabse by using your interest gene or sequences. the most similar and identical sequences are arranged in the top.

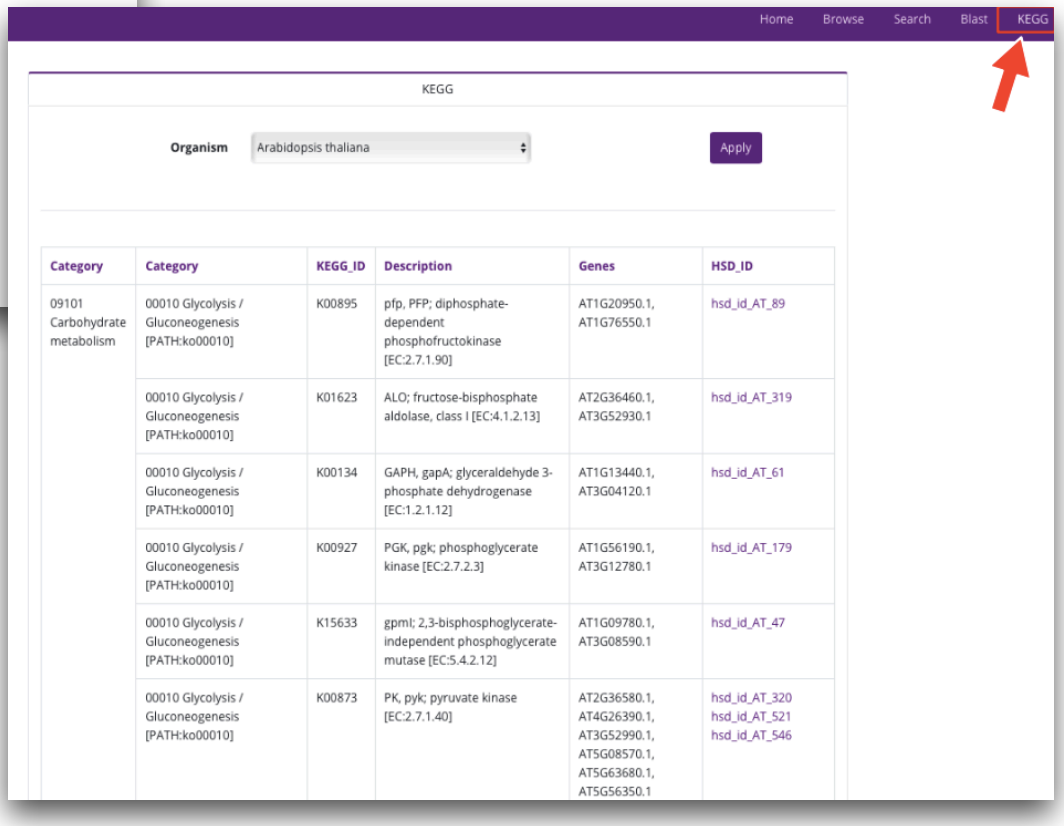
Step 3:

Search and KEGG pathway interface



HSD ID	Genes	Number	Download
hsd_id_UWO241_1	g38.t1; g7812.t1; g8958.t1; g9137.t1; g11389.t1; g1396.t1; g7823.t1; g13557.t1; g12917.t2; g10812.t3	10	
hsd_id_UWO241_10	g168.t1; g11892.t1	2	
hsd_id_UWO241_100	g2404.t1; g8568.t1	2	
hsd_id_UWO241_101	g2438.t1; g10674.t1; g8872.t1; g13703.t1; g5942.t1; g7084.t1; g6650.t1	7	
hsd_id_UWO241_102	g2532.t1; g4708.t1	2	
hsd_id_UWO241_103	g2549.t1; g4272.t1; g6579.t2; g8879.t1	4	
hsd_id_UWO241_104	g2553.t2; g13653.t1	2	

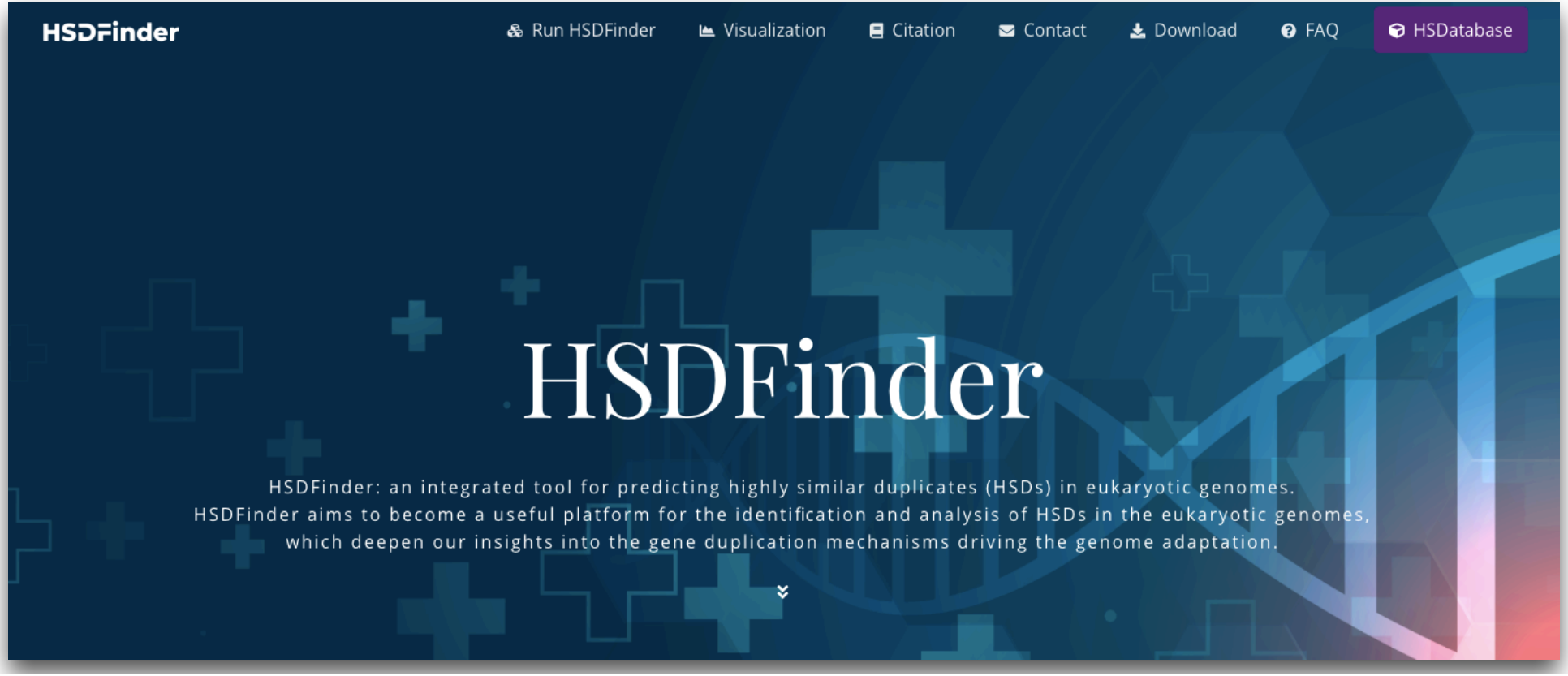




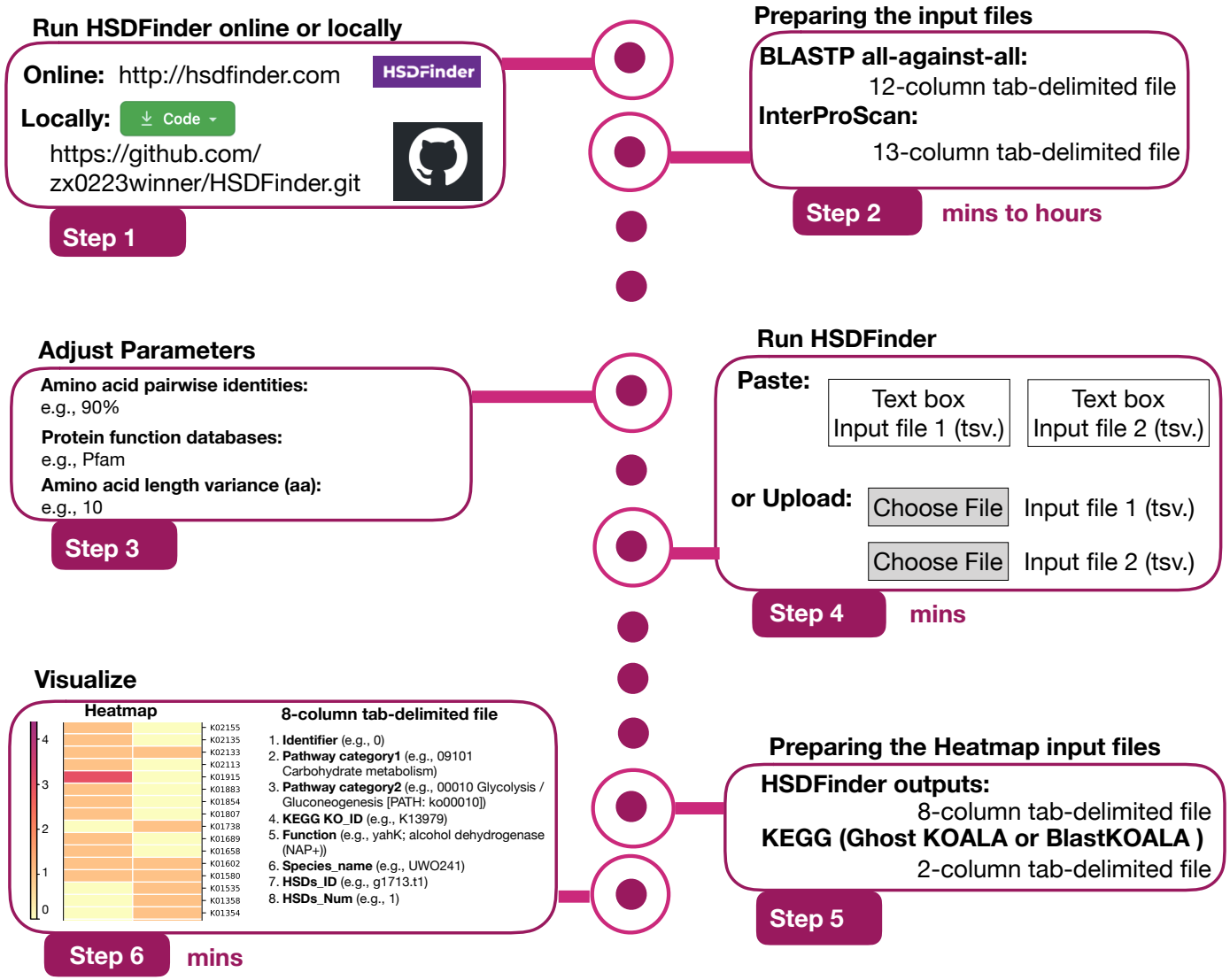
Category	Category	KEGG_ID	Description	Genes	HSD_ID
09101 Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K00895	pfp, PFP; diphosphate-dependent phosphofructokinase [EC:2.7.1.90]	AT1G20950.1, AT1G76550.1	hsd_id_AT_89
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K01623	ALO; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	AT2G36460.1, AT3G52930.1	hsd_id_AT_319
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K00134	GAPH, gapA; glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]	AT1G13440.1, AT3G04120.1	hsd_id_AT_61
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K00927	PGK, pgk; phosphoglycerate kinase [EC:2.7.2.3]	AT1G56190.1, AT3G12780.1	hsd_id_AT_179
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K15633	gpmI; 2,3-bisphosphoglycerate-independent phosphoglycerate mutase [EC:5.4.2.12]	AT1G09780.1, AT3G08590.1	hsd_id_AT_47
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K00873	PK, pyk; pyruvate kinase [EC:2.7.1.40]	AT2G36580.1, AT4G26390.1, AT3G52990.1, AT5G08570.1, AT5G63680.1, AT5G56350.1	hsd_id_AT_320 hsd_id_AT_521 hsd_id_AT_546

As for the Search and KEGG INTERFACE, user can browse the duplicate gene item by searching the name as the keyword and see where the KEGG functional categories is falling in.

Part 2: HSDFinder interface and implementation (www.hsdfinder.com)



In our analysis of the UWO241 genome we struggled to find adequate bioinformatics tools for identifying and categorizing highly similar duplicate genes (HSDs). In fact, for the most part, we were forced to use basic BLAST algorithms. SO, We designed our own easy-to-use automated Software tool, called HSDFinder.



It is designed for identifying HSDs in eukaryotic genomes with high accuracy and reliability using Pfam domains and KEGG pathways. HSDFinder also offers an online heatmap plotting option to visualize the results in different KEGG pathway functional categories. Ultimately, we feel that this software will be of great benefit to anyone analyzing eukaryotic genomes, even those with few bioinformatics backgrounds.

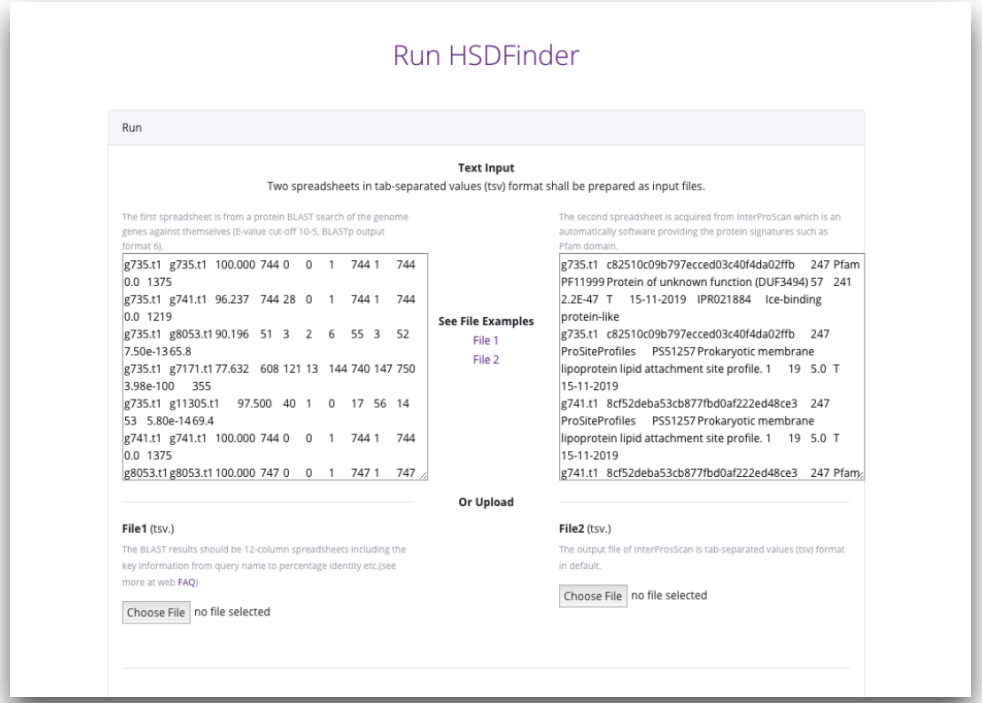
Run HSDFinder online or locally

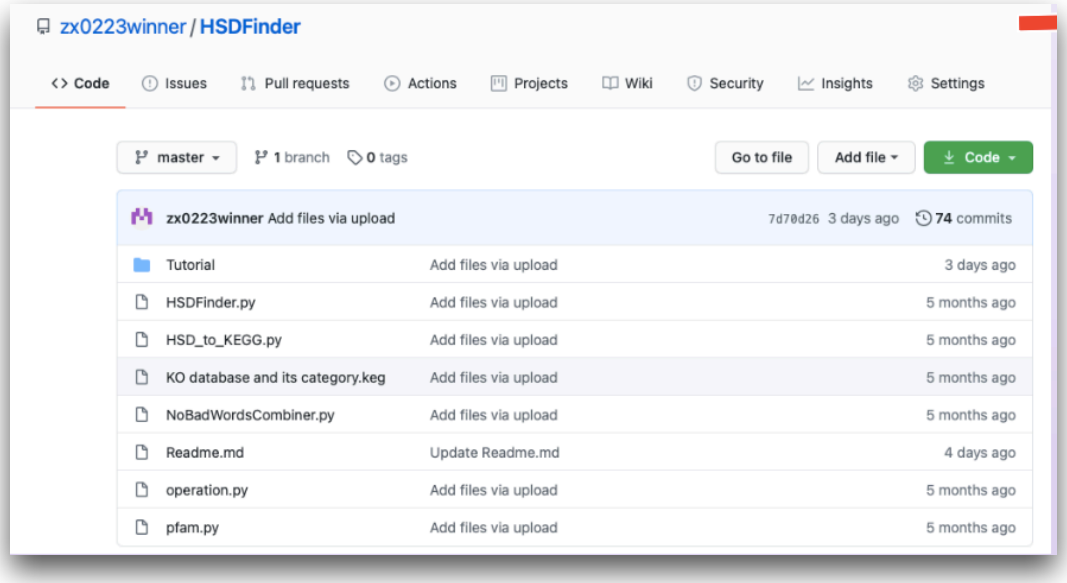
Online: http://hsdfinder.com

Locally: Code

https://github.com/zx0223winner/HSDFinder.git

Step 1

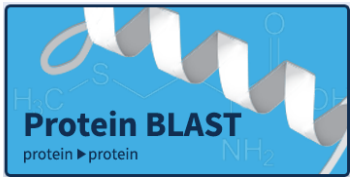




This tool can either be running on the web or locally. For those experienced users, if you have big datasets, you can download the source code of from GitHub and run it in local linux environemnt. We offered the tutorial to go over each steps and file examples. For those new to Bioinformatics tools, the online version is very user fridendly and each step have the build-in examples.

The 12-column explanation of BLAST search result file at format 6

a. Query_ID (e.g., g735.t1)
b. Seq_ID (e.g., g741.t1)
c. Percentage_identity (e.g., 96.237)
d. Aligned length (e.g., 744)
e. Mismatches (e.g., 28)
f. Gap_openings (e.g., 0)
g. Query_start (e.g., 1)
h. Query_end (e.g., 744)
i. Sequence_start (e.g., 1)
j. Sequence_end (e.g., 744)
k. E-value (e.g., 0.0)
l. Bit-score (e.g., 1219)



Preparing the input files

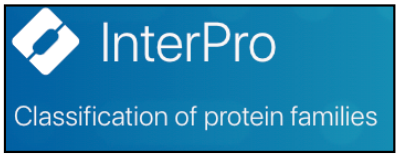
BLASTP all-against-all:
12-column tab-delimited file

InterProscan:
13-column tab-delimited file

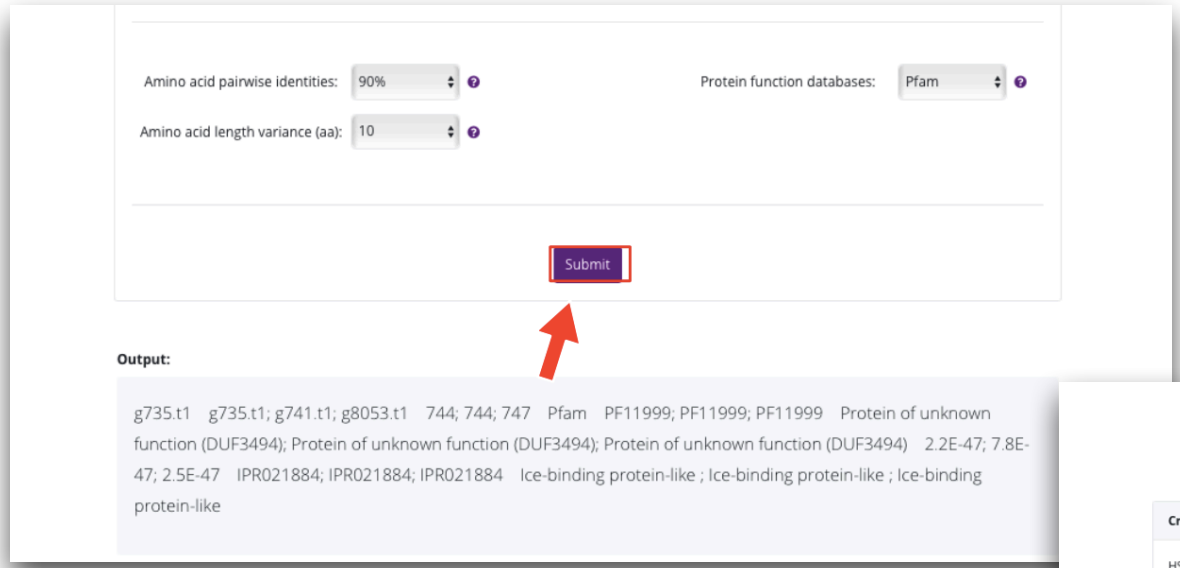
Step 2

The 13-column explanation of InterProScan search result file

a. Protein accession (e.g., g735.t1)
b. Sequence unique code (e.g., c82510c09b797ecced03c40f4da02ffb)
c. Sequence length (e.g., 247)
d. Protein signature (e.g., Pfam)
e. Signature accession (e.g., PF11999)
f. Signature description (e.g., Protein of unknown function (DUF3494))
g. Start location
h. Stop location
i. E-value (or score) (e.g., 2.2E-47)
j. Status - is the status of the match (T: true)
k. Date - is the date of the run (e.g., 15-11-2019)
l. InterPro annotations - accession (e.g., IPR021884)
m. InterPro annotations - description (e.g., Ice-binding protein-like)

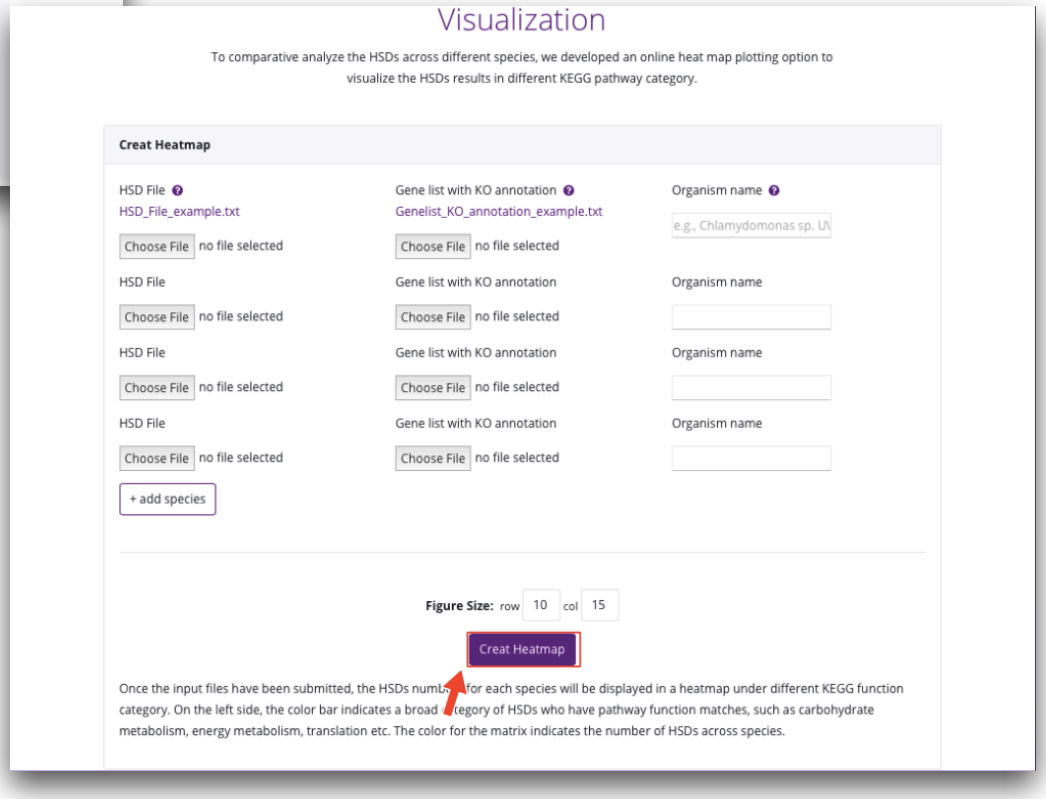


The limitation of the tool is requiring the external software to prepare the input files. But the two software are also easy-to -use and straightforward.



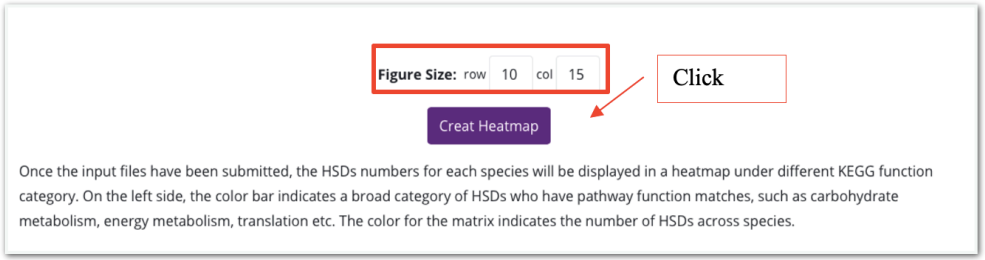
Step 3 **Adjust Parameters and Run**

Step 4 and 5 **Visualize and categorize the results**

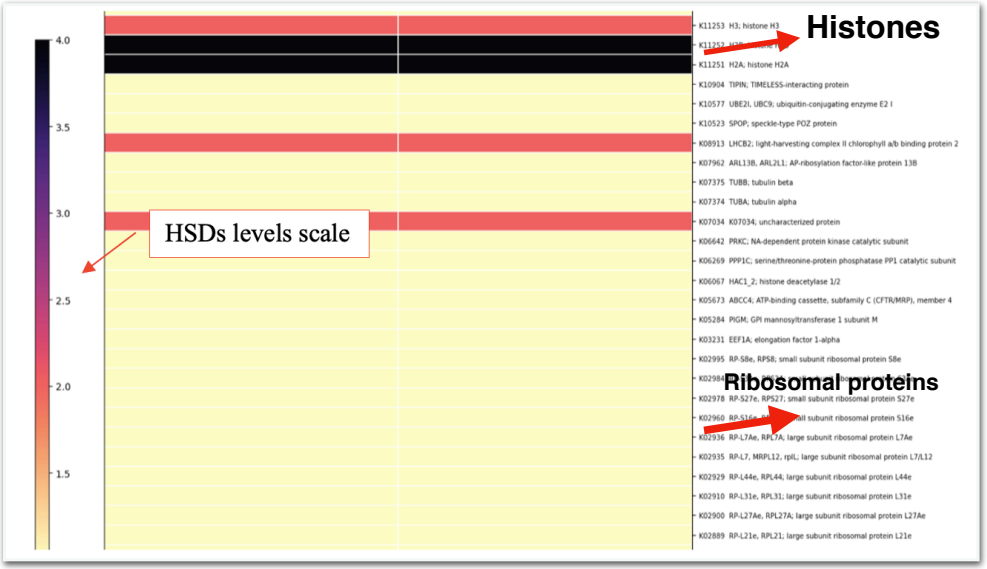


Once the input files have prepared, user can adjust the parameters such as aa pairwise, aa aligned length variance, to set different threshold for filtering the duplciates. we set the default 90% and 10 aa to best filter the HSDs according to our experience on green algal genomes. Then there is an online heatmap plotting option for users to compare duplicates in different species.

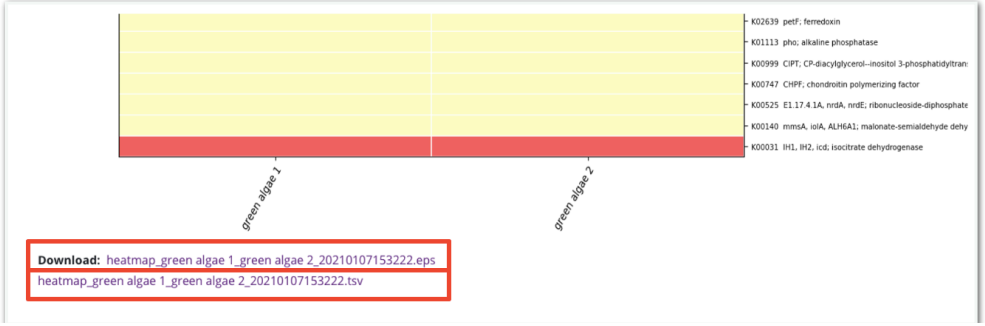
A



B



C



Step 6

Decipher the results

- **High resolution heat map image (eps.)**
- **Categorized spreadsheet (tsv.)**

Example of the 8-column tab-delimited file (.tsv) for HSDs of different species categorized under different KEGG functional categories.

Identifier	Pathway Category1	Pathway Category2	KO_ID	Function	Species_name	HSDs_ID	HSDs_Num
0	09101 Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis [PATH: ko00010]	K13979 yahK; alcohol dehydrogenase (NAP+)		UWO 241	g1713.t1	1
1	09101 Carbohydrate metabolism	00020 itrate cycle (TA cycle) [PATH: ko00020]	K00031 IH1, IH2, icd; isocitrate dehydrogenase		UWO 241	g3379.t1	1
2	09101 Carbohydrate metabolism	00030 Pentose phosphate pathway [PATH: ko00030]	K00036 G6P, zwf; glucose-6-phosphate 1-dehydrogenase		UWO 241	g852.t1	1
3	09101 Carbohydrate metabolism	00051 Fructose and mannose metabolism [PATH: ko00051]	K19355 MAN; mannan endo-1,4-beta-mannosidase		UWO 241	g3766.t1	1
4	09101 Carbohydrate metabolism	00053 Ascorbate and aldarate metabolism [PATH: ko00053]	K00434 E1.11.1.11; L-ascorbate peroxidase		UWO 241	g15878.t1	1
5	09103 Lipid metabolism	00073 utin, suberine and wax biosynthesis [PATH: ko00073]	K13356 FAR; alcohol-forming fatty acyl-CoA reductase		UWO 241	g6944.t1	1
6	09108 Metabolism of cofactors and vitamins	00130 Ubiquinone and other terpenoid-quinone biosynthesis [PATH: ko00130]	K17872 NC1, ndbB; demethylphylloquinone reductase		UWO 241	g269.t1 , g13422.t1	2

The outputs of the heatmap include one high resolution image, a 8-column spreadsheet categorizing the duplicates under different KEGG pathway functional categories. The tool presented here is the primary selection of duplciates, the manually curation can be done to filter the dataset when necessary.