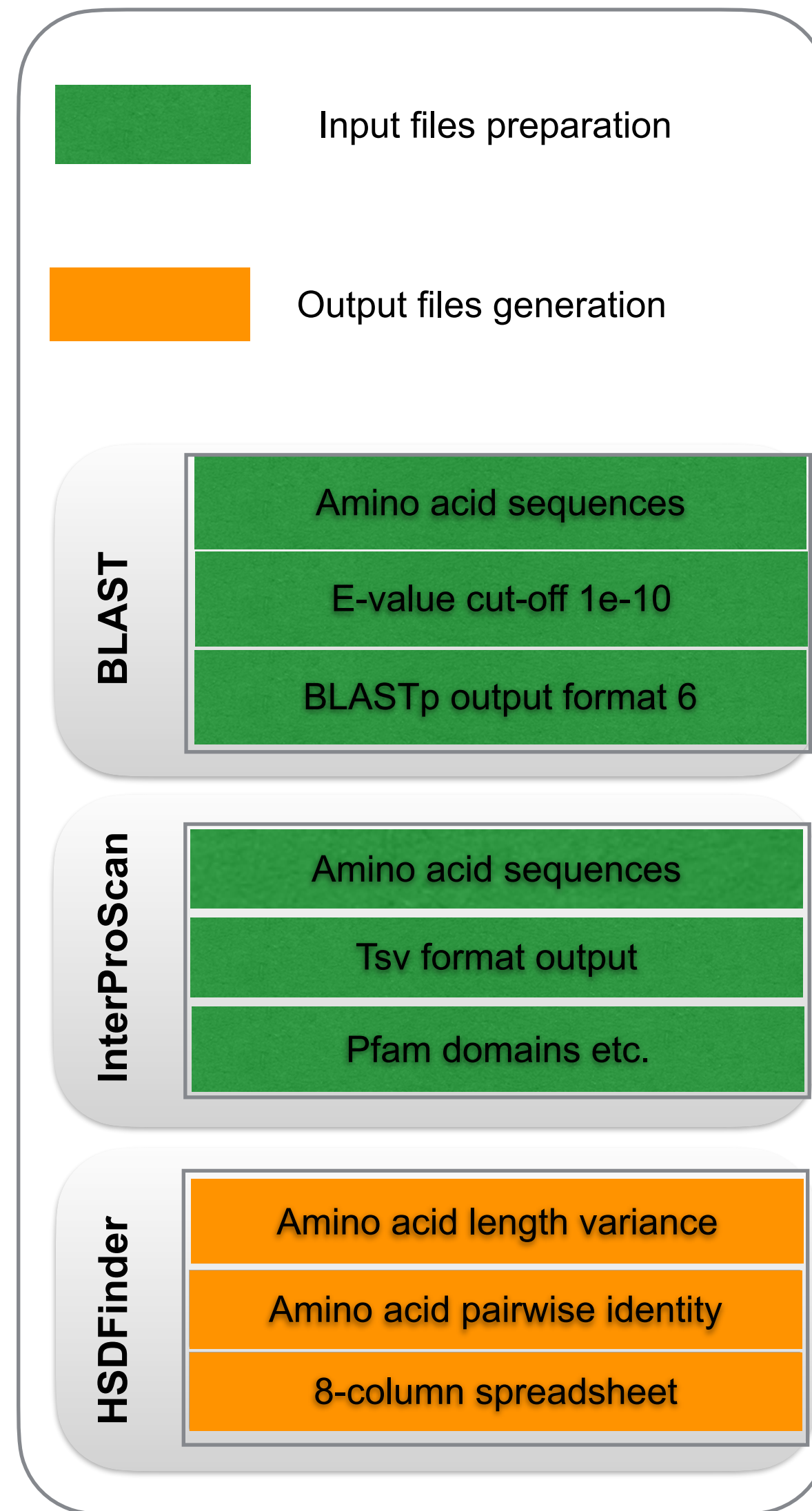


## Part 1: HSDatabase interface and implementation ([www.hsdfinder.com/database/](http://www.hsdfinder.com/database/))

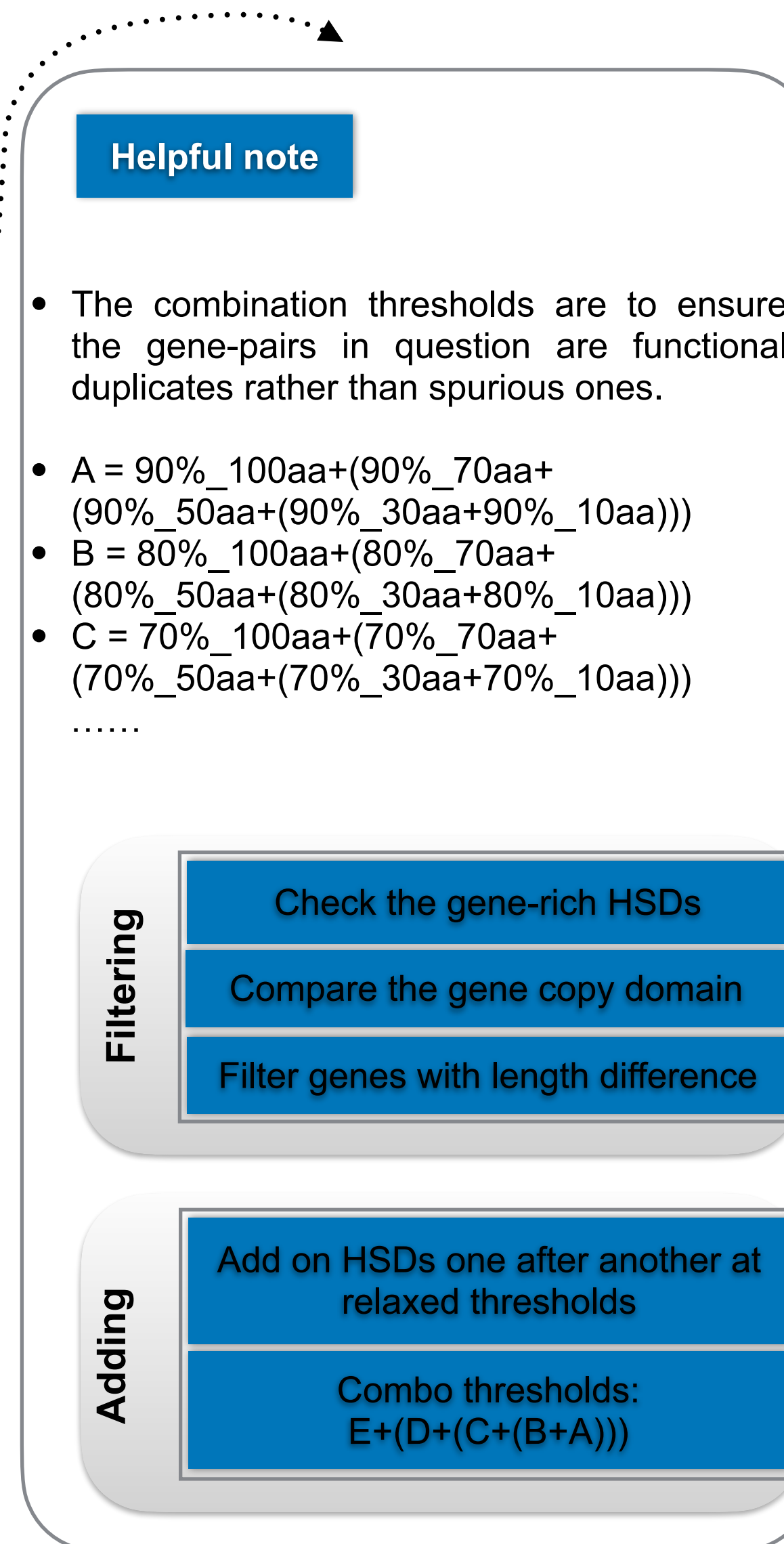


Since the numbers of highly similar duplicates in other eukaryotic genomes are largely unknown, and computational methods for identifying them can be time-consuming and labor-intensive. We created a database to collect HSDs from some model species.

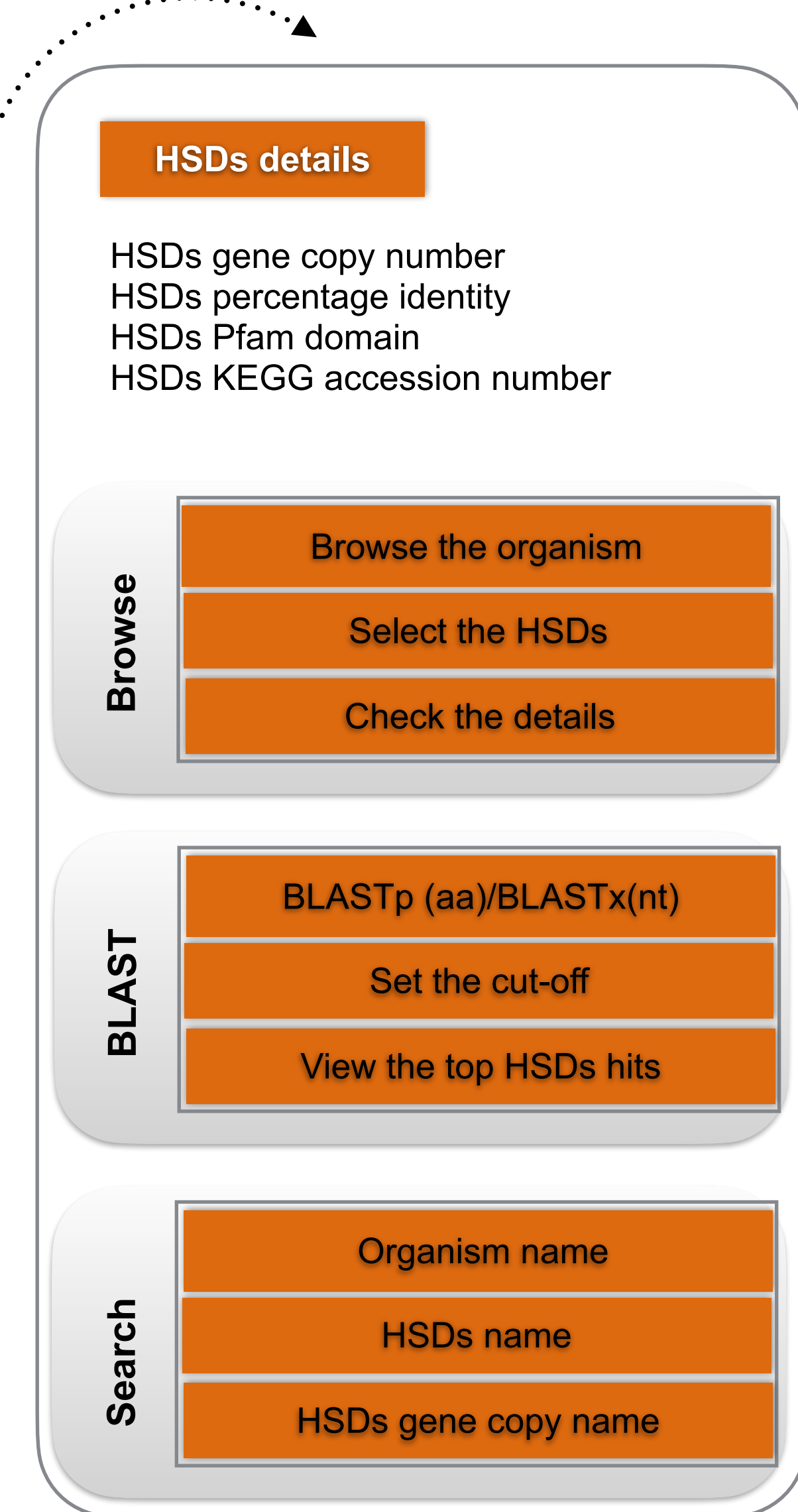
## A HSDFinder



## B Manually curation



## C HSDatabase



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

>>

## 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

## Step 1: Browse Interface

Plant			
Organism	All HSDs	Description	Download
Arabidopsis thaliana	628	Arabidopsis thaliana is a small flowering plant of mustard family, brassicaceae (Cruciferae). It is distributed throughout the world and was first reported in the sixteenth century by Johannes Thal. It has been used for over fifty years to study plant mutations and for classical genetic analysis. It is now being used as a model organism to study different aspects of plant biology.	<a href="#">Download</a>
Coccomyxa subellipsoidea C-169 (Green alga)	79	Coccomyxa is a genus of unicellular green alga. It comprises of both, free living species and those species that form symbiotic relationships with lichens. Coccomyxa subellipsoidea C-169 is a small, elongated, non-motile and well adapted to survive under extremely cold conditions of Antarctica.	<a href="#">Download</a>
Chlamydomonas eustigma (Green alga)	276	Chlamydomonas eustigma is an acidophilic species isolated from acid mine drainage.	<a href="#">Download</a>
Chlamydomonas reinhardtii (Green alga)	54	Chlamydomonas is a genus of unidirectional swimming, green, non-flagellated, and a few occur in the study photosynthesis and chloroplast assembly and motility, phototaxis.	<a href="#">Download</a>
Dunaliella salina (Green alga)	72	Dunaliella salina is a unicellular, halophilic, green alga. It lacks a rigid cell wall and is tolerant to high salt concentration. It is known for its high concentration of carotenoids, including beta-carotene.	<a href="#">Download</a>
Fragilariopsis cylindrus (Diatom)	124	Fragilariopsis cylindrus is a pennate diatom. It is found in the sea and antarctic waters, with a potential for biotechnology.	<a href="#">Download</a>
Saccharomyces cerevisiae (Yeast)	1	Saccharomyces cerevisiae is a species of yeast. It is a eukaryotic microorganism, a member of the genus Saccharomyces.	<a href="#">Download</a>

Arabidopsis thaliana			
HSDs 628			
Arabidopsis thaliana is a small flowering plant of mustard family, brassicaceae (Cruciferae). It is distributed throughout the world and was first reported in the sixteenth century by Johannes Thal. It has been used for over fifty years to study plant mutations and for classical genetic analysis. It is now being used as a model organism to study different aspects of plant biology.			
HSD ID	Genes	Number	Download
hsd_id_AT_1	AT1G01100.1; AT5G47700.1; AT4G00810.1	3	<a href="#">Download</a>
hsd_id_AT_2	AT1G01210.1; AT4G07950.1	2	<a href="#">Download</a>
hsd_id_AT_3	AT1G01510.1; AT4G00400.1	2	<a href="#">Download</a>
hsd_id_AT_4	AT1G01620.1; AT4G00430.1; AT3G61430.1; AT4G23400.1	4	<a href="#">Download</a>
hsd_id_AT_5	AT1G02000.1; AT4G00110.1	2	<a href="#">Download</a>
hsd_id_AT_6	AT1G02500.1; AT4G01850.1; AT3G17290.1	3	<a href="#">Download</a>
hsd_id_AT_7	AT1G02520.1; AT1G02530.1	2	<a href="#">Download</a>
hsd_id_AT_8	AT1G02880.3; AT2G44750.2	2	<a href="#">Download</a>
hsd_id_AT_9	AT1G02920.1; AT1G02930.1	2	<a href="#">Download</a>
hsd_id_AT_10	AT1G03130.1; AT4G02770.1	2	<a href="#">Download</a>
hsd_id_AT_11	AT1G03200.1; AT1G03240.1	2	<a href="#">Download</a>
hsd_id_AT_12	AT1G03220.1; AT1G03230.1	2	<a href="#">Download</a>
hsd_id_AT_13	AT1G03495.1; AT1G03496.1	2	<a href="#">Download</a>

hsd_id_AT_6	
Identity: AT1G02500.1 <a href="#">Seq</a>	
<p>                     METFLTSTSEVNEHPDKLDCQSDAWLADCLCEQDPSKACCTCTKTNMMVFGEITTKATVDYKEKVDTCRAIGFVSDVGLDADCKVLYNEQSPDIAQGVGHGHTKCPBEIGAGDQGHMFQYATDETPELMPLSHLATKLGARLTVKNGTCWLRIPDGKTQTVVEYNDGAMNPRVHTVLTQHDVETVNDIARDLKEHVKVPKPKYLDKTFPHLNPGRVIGGPHGDAGLGRKIDITGGWGAHGGGAFSGKDPTRVSGVAYNRQAQSVANGMARRALVQVSYAGVPEPLSVFVDTGTGLPKELKVKESDFRPGMMTNLDLKRGGNGRFLKTAAYGHGRDQDPTVEVWVPLKWKPKQA                 </p>	
Length:	393
PF Identity:	PF02772, PF02773, PF00438
PF Description:	S-adenosylmethionine synthetase, central domain, S-adenosylmethionine synthetase, C-terminal domain, S-adenosylmethionine synthetase, N-terminal domain
E-value:	3.7E-47, 2.6E-59, 1.1E-42
IPR Identity:	IPR022629, IPR022630, IPR022628
IPR Description:	S-adenosylmethionine synthetase, central domain, S-adenosylmethionine synthetase, C-terminal, S-adenosylmethionine synthetase, N-terminal
Identity: AT4G01850.1 <a href="#">Seq</a>	
<p>                     METFLTSTSEVNEHPDKLDCQSDAWLADCLCEQDPSKACCTCTKTNMMVFGEITTKATVDYKEKVDTCRAIGFVSDVGLDADCKVLYNEQSPDIAQGVGHGHTKCPBEIGAGDQGHMFQYATDETPELMPLSHLATKLGARLTVKNGTCWLRIPDGKTQTVVEYNDGAMNPRVHTVLTQHDVETVNDIARDLKEHVKVPKPKYLDKTFPHLNPGRVIGGPHGDAGLGRKIDITGGWGAHGGGAFSGKDPTRVSGVAYNRQAQSVANGMARRALVQVSYAGVPEPLSVFVDTGTGLPKELKVKESDFRPGMMTNLDLKRGGNGRFLKTAAYGHGRDQDPTVEVWVPLKWKPKQA                 </p>	
Length:	393

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

## Step 2: BLAST search interface

HSDatabase

Home Browse Search **Blast** KEGG FAQ

Organism: All Organisms

Algorithm: BLASTP

Exception Value: 1e-20

Max. target sequences: 5

Sequence in FASTA format

Or

Upload FASTA file

Choose File no file selected

**BLAST**

Sequence in FASTA format

```
MSPEKKSQNFPPITECRDGEYDSIAADLDGTLILLSRSSFPYFMLVAEAGSLRLGLILLSPPVIISYLVSESGLQILIFSAGLKIRDIELVSRVLPFRY
AADVRKDSFEVFDKCKRKVVYANPVMVEAFVKDYLGGDKVLGTEIVNPKTNRATGFVKPGVLVGDILKRLAILKEFGNESPDILGLDRTSDHDF
MSLCKKGYMVHATKSATTIPKRLKNRIVFHDGRLAQRPTPLNAILYLWLPFGRLSIIRVYNLPLPERFVRYTYEMLGHILTRHGRPPPPSGTLGN
LYVLNHRITLDPIVAIALGRKICCVTYSVSRLSLMLSPIPAVALTRDRAADAANMRKLEKGDLCVPEGTTCCREYLLRFSALFAELSDRIVPAMNCK
QGMFNGTTRVGKFWDPYFFFMNPRPSYEATFLDRLPEEMTVNGGKTPIEVANYQKVGAVLGFECTELTRKDKYLLGGNDGKVESINNTKK
```

Or

Upload FASTA file

Choose File no file selected

**BLAST**

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

BLAST

BLASTp (aa)/BLASTx(nt)

Set the cut-off

View the top HSDs hits

User can also select the BLAST option to search against HSDatabase 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

Home Browse Search Blast KEGG FAQ

Search by HSD ID or Gene ID  Select Category ✓ All  
Plant, Chromista, Fungi  
Animal

370 result(s) found

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	
hsd_id_UWO241_105	g2595.t1; g15288.t1; g9227.t1; g8647.t1	4	

Home Browse Search Blast **KEGG**

KEGG

Organism

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	ALDO; fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	AT2G36460.1, AT3G52930.1	hsd_id_AT_319
	00010 Glycolysis / Gluconeogenesis [PATH:ko00010]	K00134	GAPDH, 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

Search

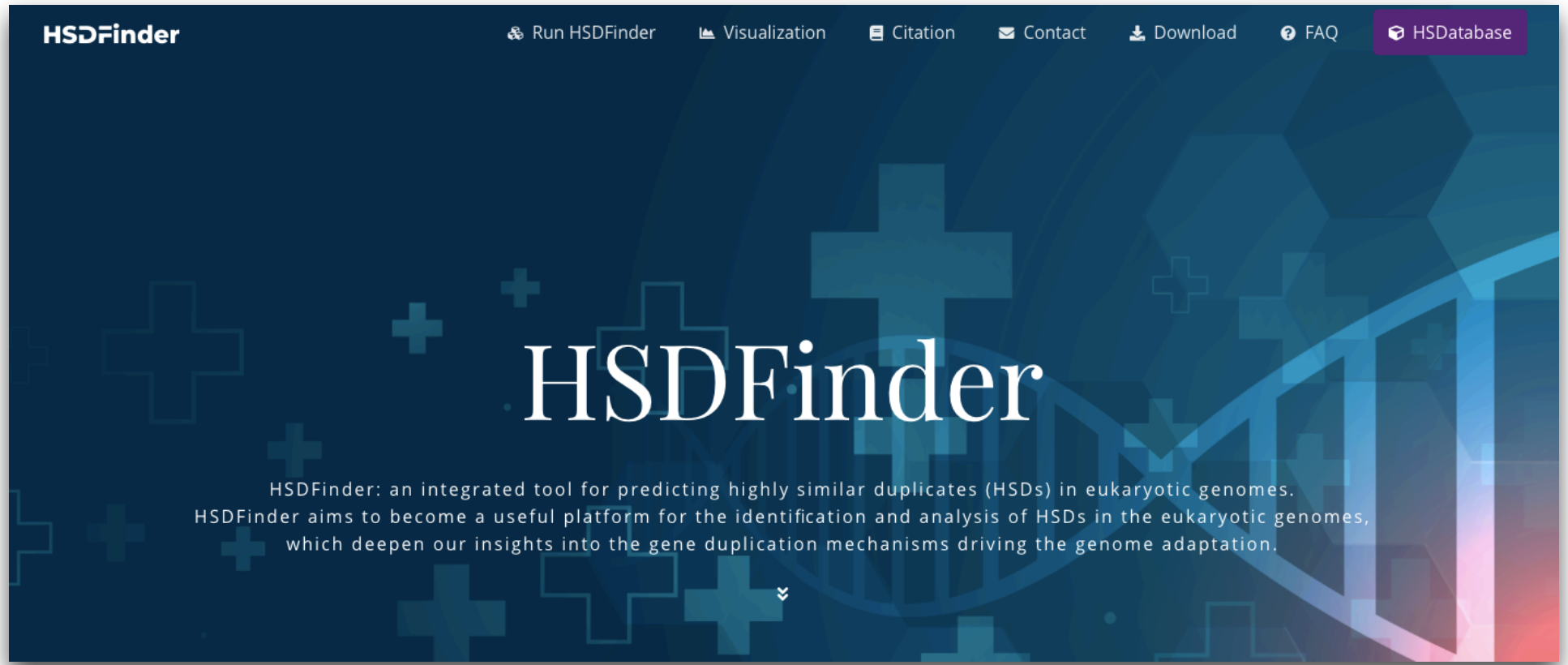
Organism name

HSDs name

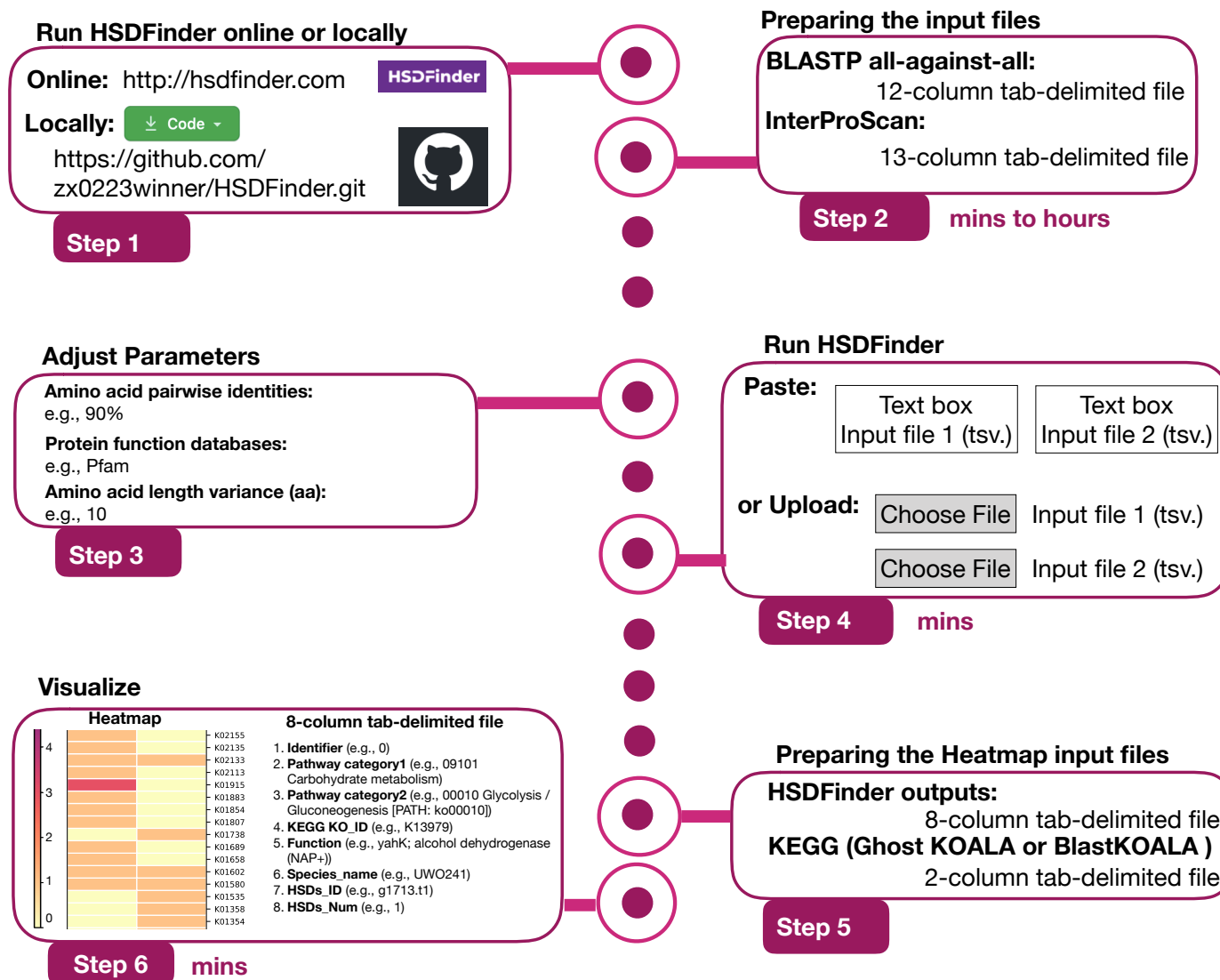
HSDs gene copy name

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](http://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>

**HSDFinder**

**Locally:** [Code](#)

[https://github.com/  
zx0223winner/HSDFinder.git](https://github.com/zx0223winner/HSDFinder.git)



**Step 1**

zx0223winner / HSDFinder

[Code](#) [Issues](#) [Pull requests](#) [Actions](#) [Projects](#) [Wiki](#) [Security](#) [Insights](#) [Settings](#)

master 1 branch 0 tags

Go to file

Add file

Code

zx0223winner Add files via upload

7d70d26 3 days ago

74 commits

Tutorial	Add files via upload	3 days ago
HSDFinder.py	Add files via upload	5 months ago
HSD_to_KEGG.py	Add files via upload	5 months ago
KO database and its category.keg	Add files via upload	5 months ago
NoBadWordsCombiner.py	Add files via upload	5 months ago
Readme.md	Update Readme.md	4 days ago
operation.py	Add files via upload	5 months ago
pfam.py	Add files via upload	5 months ago

## Run HSDFinder

Run

### Text Input

Two spreadsheets in tab-separated values (tsv) format shall be prepared as input files.

The first spreadsheet is from a protein BLAST search of the genome genes against themselves (E-value cut-off 10<sup>-5</sup>, BLASTp output format 6).

g735.t1	g735.t1	100.000	744	0	0	1	744	1	744
0.0	1375								
g735.t1	g741.t1	96.237	744	28	0	1	744	1	744
0.0	1219								
g735.t1	g8053.t1	90.196	51	3	2	6	55	3	52
7.50e-1365.8									
g735.t1	g7171.t1	77.632	608	121	13	144	740	147	750
3.98e-100	355								
g735.t1	g11305.t1	97.500	40	1	0	17	56	14	53
5.80e-1469.4									
g741.t1	g741.t1	100.000	744	0	0	1	744	1	744
0.0	1375								
g8053.t1	g8053.t1	100.000	747	0	0	1	747	1	747

### See File Examples

File 1

File 2

The second spreadsheet is acquired from InterProScan which is an automatic software providing the protein signatures such as Pfam domain.

g735.t1	c82510c09b797ecccd03c40f4da02ffb	247	Pfam
PF11999	Protein of unknown function (DUF3494)	57	241
2.2E-47	T	15-11-2019	IPRO21884
ice-binding			
protein-like			
g735.t1	c82510c09b797ecccd03c40f4da02ffb	247	
ProSiteProfiles	P551257	Prokaryotic membrane	
lipoprotein lipid attachment site profile. 1	19	5.0	T
15-11-2019			
g741.t1	8cf52deba53cb877fbd0af222ed48ce3	247	
ProSiteProfiles	P551257	Prokaryotic membrane	
lipoprotein lipid attachment site profile. 1	19	5.0	T
15-11-2019			
g741.t1	8cf52deba53cb877fbd0af222ed48ce3	247	Pfam

### Or Upload

#### File1 (tsv)

The BLAST results should be 12-column spreadsheets including the key information from query name to percentage identity etc.(see more at web FAQ)

Choose File no file selected

#### File2 (tsv)

The output file of InterProScan is tab-separated values (tsv) format in default.

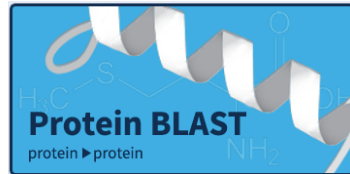
Choose File no file selected

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 fridently 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.

Amino acid pairwise identities: 90%

Protein function databases: Pfam

Amino acid length variance (aa): 10

**Submit**

Output:

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

### Step 3

### Adjust Parameters and Run

### Step 4 and 5

### Visualize and categorize the results

#### Visualization

To comparative analyze the HSDs across different species, we developed an online heat map plotting option to visualize the HSDs results in different KEGG pathway category.

#### Create Heatmap

HSD File

HSD\_File\_example.txt

**Choose File** no file selected

HSD File

**Choose File** no file selected

HSD File

**Choose File** no file selected

HSD File

**Choose File** no file selected

**+ add species**

Gene list with KO annotation

Genelist\_KO\_annotation\_example.txt

**Choose File** no file selected

Gene list with KO annotation

**Choose File** no file selected

Gene list with KO annotation

**Choose File** no file selected

Gene list with KO annotation

**Choose File** no file selected

Organism name

e.g., Chlamydomonas sp. U

Organism name

Organism name

Organism name

Figure Size: row 10 col 15

**Create Heatmap**

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.

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 duplicates. 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.

## 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	HSD_s_Num
0	09101 Carbohydrate metabolism	00010 Glycolysis / Gluconeogenesis [PATH: ko00010]	K13979 yahK; alcohol dehydrogenase (NAP+)	UWO 241	g1713.t	1	1
1	09101 Carbohydrate metabolism	00020 itrate cycle (TA cycle) [PATH: ko00020]	K00031 IH1, IH2, icd; isocitrate dehydrogenase	UWO 241	g3379.t	1	1
2	09101 Carbohydrate metabolism	00030 Pentose phosphate pathway [PATH: ko00030]	K00036 G6P, zwf; glucose-6-phosphate 1-dehydrogenase	UWO 241	g852.t1	1	1
3	09101 Carbohydrate metabolism	00051 Fructose and mannose metabolism [PATH: ko00051]	K19355 MAN; mannan endo-1,4-beta-mannosidase	UWO 241	g3766.t	1	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	1
5	09103 Lipid metabolism	00073 utin, suberine and wax biosynthesis [PATH: ko00073]	K13356 FAR; alcohol-forming fatty acyl-CoA reductase	UWO 241	g6944.t	1	1
6	09108 Metabolism of cofactors and vitamins	00130 Ubiquinone and other terpenoid-quinone biosynthesis [PATH: ko00130]	K17872 NC1, ndbB; demethylphyloquinone reductase	UWO 241	g269.t1, g13422.t1	2	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 dupciates, the manually curation can be done to filter the dataset when necessary.