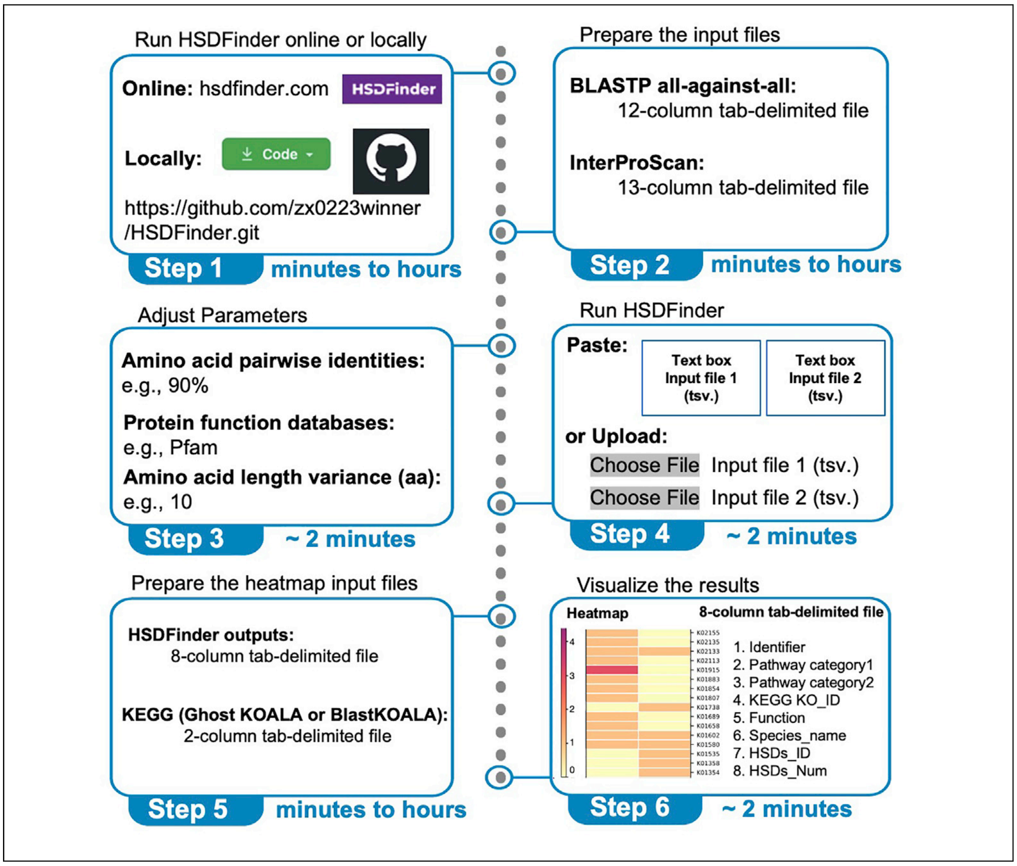
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for HSDFinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes



Although gene duplications have been documented in many species, the precise numbers of highly similar duplicated genes (HSDs) in eukaryotic nuclear genomes remain largely unknown and can be time consuming to explore. We developed HSDFinder to identify, categorize, and visualize HSDs in eukaryotic nuclear genomes using protein family domains and KEGG pathways. In contrast to existing tools, HSDFinder allows users to compare HSDs among different species and visualize results in different KEGG pathway functional categories via heatmap plotting.

Xi Zhang, Yining Hu, David Roy Smith

xzha25@uwo.ca (X.Z.)
dsmit242@uwo.ca (D.R.S.)

Highlights

HSDFinder is a web tool for identifying highly similar duplicated genes (HSDs)

HSDs are annotated and categorized using Pfam domains and KEGG pathways

HSDFinder can generate an HSD heatmap linked to KEGG pathways

HSDs across species can be compared in different KEGG pathway categories

Zhang et al., STAR Protocols 2, 100619
September 17, 2021 © 2021 The Authors.
https://doi.org/10.1016/j.xpro.2021.100619

STAR Protocols

CellPress
OPEN ACCESS

Protocol

Protocol for HSDFinder: Identifying, annotating, categorizing, and visualizing duplicated genes in eukaryotic genomes

Xi Zhang,[1,3,*] Yining Hu,[2] and David Roy Smith[1,4,*]

[1]Department of Biology, Western University, London, ON N6A 5B7, Canada

[2]Department of Computer Science, Western University, London, ON N6A 5B7, Canada

[3]Technical contact

[4]Lead contact

*Correspondence: xzha25@uwo.ca (X.Z.), dsmit242@uwo.ca (D.R.S.)
https://doi.org/10.1016/j.xpro.2021.100619

SUMMARY

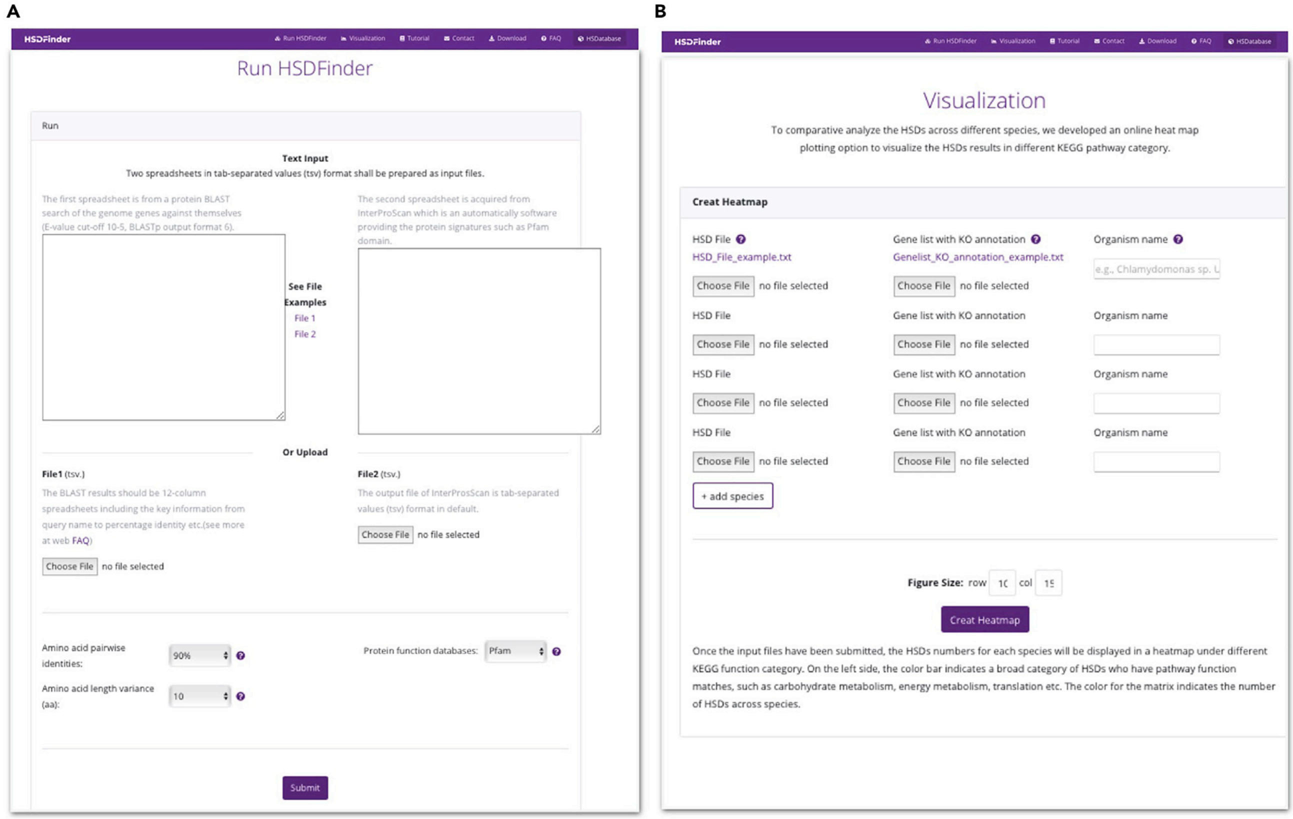
Although gene duplications have been documented in many species, the precise numbers of highly similar duplicated genes (HSDs) in eukaryotic nuclear genomes remain largely unknown and can be time-consuming to explore. We developed HSDFinder to identify, categorize, and visualize HSDs in eukaryotic nuclear genomes using protein family domains and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways. In contrast to existing tools, HSDFinder allows users to compare HSDs among different species and visualize results in different KEGG pathway functional categories via heatmap plotting.

For complete details on the use and execution of this protocol, please refer to Zhang et al. (2021).

BEFORE YOU BEGIN

Gene duplication is a near-ubiquitous phenomenon across the tree of life (Innan and Kondrashov, 2010) and is particularly prevalent in eukaryotic nuclear genomes (Kondrashov, 2012), where it is often linked to adaptation to various environmental conditions. In green algae, for example, large-scale gene duplications have been observed in the nuclear DNA (nucDNA) of the acidophile *Chlamydomonas eustigma,* which harbors ~10 copies of genes encoding arsenate reductase (ArsC) and ~20 copies of genes encoding glutaredoxin (Grx), which together might be contributing to survival in an acidic environments (Hirooka et al., 2017). Similarly, duplications of genes encoding carotene biosynthesis-related protein (CBR) and Lhc-like protein (Lhl4) are associated with adaptation to the highly variable light conditions in the Antarctic alga *Chlamydomonas* sp. ICE-L (Zhang et al., 2020d). More recently, it was discovered that hundreds of highly similar duplicated genes (HSDs) are potentially aiding the survival of the Antarctic green alga *Chlamydomonas* sp. UWO241 via gene dosage (Cvetkovska et al., 2018; Zhang et al., 2021). These HSDs were curated into a filtered gene set with near-identical protein lengths (within 10 amino acids) and ≥90% pairwise identities (Zhang et al., 2021).

It can be time-consuming and computationally challenging to identify, categorize, and visualize gene duplicates in eukaryotic nuclear genomes. Currently, there are very few user-friendly bioinformatics tools for this type of work, especially tools that allow for comparative analyses of duplicates across species. HSDFinder is an open-source, online, and user-friendly bioinformatics tool for efficiently detecting and categorizing HSDs in eukaryotic genomes by integrating data from InterProScan and KEGG. This tool also allows the predicted HSDs to be compared across species, including via high-resolution heatmaps. Some of the limitations of HSDFinder include the



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 10-5, BLASTp output format 6).

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

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)

no file selected

File2 (tsv.)

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

no file selected

Amino acid pairwise identities: 90%

Amino acid length variance (aa): 10

Protein function databases: Pfam

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.

Creat Heatmap

HSD File	Gene list with KO annotation	Organism name
HSD_File_example.txt	Genelist_KO_annotation_example.txt	e.g., Chlamydomonas sp. U
no file selected	no file selected	
HSD File	Gene list with KO annotation	Organism name
no file selected	no file selected	
HSD File	Gene list with KO annotation	Organism name
no file selected	no file selected	
HSD File	Gene list with KO annotation	Organism name
no file selected	no file selected	

Figure Size: row 1(col 1!

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.

Figure 1. The HSDFinder home page

(A) Identifying and annotating HSDs.

(B) Visualizing and categorizing HSDs in a heatmap.

requirement of users to be familiar with the Basic Local Alignment Search Tool (BLAST) package and the dash shell in a Linux/Unix environment as well as the necessity to input files from third-party tools, such as InterProScan and KEGG (BlastKOALA and GhostKOALA).

There exist various strategies for identifying gene duplications in eukaryotic genomes (Lallemand et al., 2020). For detecting all paralogous gene pairs in a genome, for instance, sequence similarity is usually evaluated by three metrics: percent sequence identity, aligned length, and E-value (Lallemand et al., 2020). Note that molecular sequence alignments of duplicate genes are generally carried out using amino acid sequences rather than nucleotide sequences because the former are more conserved than the latter (Koonin, 2005). The protocol presented here describes how to use HSDFinder for comparing and visualizing HSDs among species. The model green algae *Chlamydomonas* sp. UWO241 and *Chlamydomonas reinhardtii* are used as a case-study for this goal.

Overview

HSDFinder groups gene duplicates together based on their pairwise amino acid identity and amino acid length variance. It also provides putative annotations for the identified duplicates via protein functional domains and pathway information based on data from InterProScan and KEGG databases (Figure 1A). Users can employ different thresholds within HSDFinder for filtering duplicates (e.g., from 30%–100% amino acid pairwise identity and from 0–100 amino acid variances in the length of the aligned sequences). There is also an online heatmap plotting option in HSDFinder for categorizing, comparing, and visualizing duplicates under different KEGG pathway functional categories (Figure 1B).

Downloading the software and prerequisites

HSDFinder can either be operated on the web (http://hsdfinder.com) or through a local environment (Linux and Python 3) after downloading the software package from GitHub (https://github.com/zx0223winner/HSDFinder). To run locally, pre-installed Python (preferably Python 3) and Linux (e.g., Ubuntu 20.04 LTS) environments are required. The BLAST and InterProScan software packages as well as the online KEGG pathways tools BlastKOALA and GhostKOALA (Pellerin, 2016) can be accessed via the links from in the Key resources table.

> *Note:* A minimum specification requirement is a machine with 2 cores and 4 GB of RAM, which should allow the HSDs to be identified and visualized within a few minutes.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Chlamydomonas sp. UWO241	GenBank (Zhang et al., 2021)	GCA_016618255
Chlamydomonas reinhardtii	Phytozome 12.1 (Merchant et al., 2007)	JGI 5.5
Volvox carteri	Phtyzome 12.1 (Prochnik et al., 2010)	JGI 2.0
Chlamydomonas eustigma	GenBank (Hirooka et al., 2017)	GCA_002335675.1
Dunaliella salina	Phytozome 12.1 (Polle et al., 2017)	JGI 3.0
Gonium pectorale	GenBank (Hanschen et al., 2016)	GCA_001584585.1
Chlamydomonas sp. ICE-L	GenBank (Zhang et al., 2020d)	GCA_013435795.1
Software and algorithms		
BLAST v2.2.26	(Kent, 2002)	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/
InterProScan v4.7	(Quevillon et al., 2005)	http://www.ebi.ac.uk/interpro/download/
BlastKOALA or GhostKOALA	(Kanehisa and Goto, 2000; Kanehisa et al., 2016)	https://www.kegg.jp
HSDFinder v1.0	(Zhang et al., 2020b)	http://hsdfinder.com; https://github.com/zx0223winner/HSDFinder
HSDatabase v1.0	(Zhang et al., 2020a)	http://hsdfinder.com/hsdatabase
Python 3	N/A	https://www.python.org/downloads/
Django v3.1.5	N/A	https://www.djangoproject.com/download/
pandas v1.2.2	N/A	https://pandas.pydata.org

MATERIALS AND EQUIPMENT

The software implementation was written in Python 3 using the following custom scripts and platforms: *HSDFinder.py*, *operation.py* and *pfam.py*, enabling the duplicates to be filtered and annotated from BLASTP results and protein signature databases (e.g., Pfam); *HSD_to_KEGG.py*, enabling the duplicates to be categorized under KEGG pathway functional categories; Django (3.1.5), a Python-based web platforms, maintaining the web server; and pandas (1.2.2), the software library used for manipulating the data. A full list of utilized packages can be found in the Key resources table. The full HSDFinder source code can be found in the GitHub repository.

The test input data consists of BLASTP results and the protein signature results from InterProScan (Quevillon et al., 2005). The first input document of the BLASTP results was designated as 12 columns (Table 1). The second input document of InterProScan results was designated as 13 columns (Table 2). To create a heatmap of the HSDs under pathway functional categories, the KO accession with each gene model identifier were retrieved from the KEGG database (Kanehisa and Goto, 2000) (Figure 2).

Table 1. Input file example 1

Query_ID	Seq_ID	Percentage_ identity	Aligned length	Mismatches	Gap_ openings	Query_ start	Query_ end	Sequence_ start	Sequence_ end	E-value	Bit-score
g735.t1	g735.t1	100	744	0	0	1	744	1	744	0	1375
g735.t1	g741.t1	96.237	744	28	0	1	744	1	744	0	1219
g735.t1	g8053.t1	90.196	51	3	2	6	55	3	52	7.50E-13	65.8
g735.t1	g7171.t1	77.632	608	121	13	144	740	147	750	3.98E-100	355
g735.t1	g11305.t1	97.5	40	1	0	17	56	14	53	5.80E-14	69.4
g741.t1	g741.t1	100	744	0	0	1	744	1	744	0	1375
g8053.t1	g8053.t1	100	747	0	0	1	747	1	747	0	1380
g7171.t1	g7171.t1	100	750	0	0	1	750	1	750	0	1386
g11305.t1	g11305.t1	100	1059	0	0	1	1059	1	1059	0	1956
…	…	…	…	…	…	…	…	…	…	…	…

STEP-BY-STEP METHOD DETAILS

Preparing the protein BLAST search result file

Timing: minutes to hours

Upload a protein BLAST-search (BLASTP) result file for your genome of interest in tab-separated values (tsv) format as the first input file (File 1) of HSDFinder. This protocol will go over how to acquire local BLAST-search results via an example FASTA file. The example file can be acquired from GitHub under the tutorial directory (Figure 3A).

Note: You can ignore this step and proceed with your own protein dataset if you know how to acquire the appropriate BLASTP search results.

1. Download the BLAST Package and FASTA file. The BLAST-search result example can be found in the ZIP file in the GitHub "tutorial" directory under the name HSDFinder_example_doc.zip (Figure 3A).
   a. 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 operating systems (Windows, MacOS or Linux)
   b. Unzip the "HSDFinder_example_doc.zip" file, the file named "Chlamydomnas_UWO241_protein.fasta" is the example FASTA file.

```
# display the first ten rows of the FASTA file.
$ head Chlamydomnas_UWO241_protein.fasta
>g1.t1
MAATVENVVERVKSFSSVVRGVKSGKPDGATTQLVQETIEILATYCDFEEVVPVCLKFLDEVL
TAAPQTSTLIRLEGGAK
IFPSIIRNFMGVDASILALCAKVMCKCASGSPAMQHHLVKEKGLPTLLLSCCSAHAGE
PAVVGPLLEVLVALARYSKGAT
ALSNANLVHACKELLVGLMGHWHAFGMVLKLIKSVMKHEGPCLAALKAGEVVRLLLG
VARLVSRMPDQRKLLKRASRTLW
VLSQRSLHPLPEMELNWPHTHTHTHTHTHT
>g2.t1
MMMLAYRFGFTTLMYATVKGHADAMRLLLKHPSADTAAMMMLTDIRGCTALM
FAAQDGHVNAIRMLLDHPSADVAARIAV
RSTVGISALTSAAGFAAGQPTLSRRASPARSCTPLLFLLRRVAVEPQLCDTQ
>g3.t1
MVPTDGARHGWTATSLPAILGAASHAKITVQQLVVGGPPPSCPYGPEIVGRSLSLFSK
SAKTWDRAPGGVVSAFCAATGE
```

Table 2. Input file example 2

Protein accession	Unique code	Sequence length	Protein signature	Signature accession	Signature description	Start location	Stop location	E-value	Status	Date	InterPro accession	InterPro description
g735.t1	c82510c09b797ecced03c40f4da02ffb	247	Pfam	PF11999	Protein of unknown function (DUF3494)	57	241	2.20E-47	T	15-11-2019	IPR021884	Ice-binding protein-like
g735.t1	c82510c09b797ecced03c40f4da02ffb	247	ProSite Profiles	PS51257	Prokaryotic membrane lipoprotein lipid attachment site profile.	1	19	5	T	15-11-2019	N/A	N/A
g741.t1	8cf52deba53cb877fbd0af222ed48ce3	247	ProSite Profiles	PS51257	Prokaryotic membrane lipoprotein lipid attachment site profile.	1	19	5	T	15-11-2019	N/A	N/A
g741.t1	8cf52deba53cb877fbd0af222ed48ce3	247	Pfam	PF11999	Protein of unknown function (DUF3494)	57	241	7.80E-47	T	15-11-2019	IPR021884	Ice-binding protein-like
g8053.t1	3d70a0c7f160037bf79f409bd805d577	248	Pfam	PF11999	Protein of unknown function (DUF3494)	58	244	2.50E-47	T	15-11-2019	IPR021884	Ice-binding protein-like
g7171.t1	9455b619e60679693d39c8191c410d18	249	Pfam	PF11999	Protein of unknown function (DUF3494)	58	244	8.00E-47	T	15-11-2019	IPR021884	Ice-binding protein-like
g11305.t1	299faccc0b8751e2919a8a332d5e123f	352	Pfam	PF11999	Protein of unknown function (DUF3494)	157	348	7.80E-55	T	15-11-2019	IPR021884	Ice-binding protein-like
…	…	…	…	…	…	…	…	…	…	…	…	…

2. Build a database via the example FASTA file.
   a. Using the command line below:

      # The makeblastdb command is used to construct a protein database by taking in the FASTA file with the parameter (-in), setting up the database type (e.g., protein) with the parameter (-dbtype protein), and titling the name of database (e.g., database_name) with parameters (-title database_name).

      # note: if your FASTA data are nucleotides, you can change the database type with the parameter (-dbtype nucl)

      ```
      > makeblastdb -in Chlamydomnas_UWO241_protein.fasta -dbtype prot -title database_name
      ```

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

      # The blastp command is used to do the protein similarity search by searching the query file (Chlamydomnas_UWO241_protein.fasta) against the protein database created from former step with the default parameters, such as '-evalue' indicating the significance of the BLAST hits, '-outfmt' meaning the tabular format of the BLAST result, and '-out' telling the file name of the output file (e.g., BLASTP_UWO241.txt).

      ```
      > blastp -query Chlamydomnas_UWO241_protein.fasta -db database_name -out BLASTP_UWO241.txt -evalue 1e-5 -outfmt 6
      ```

⚠ CRITICAL: Make sure to use the BLASTP option, which allows for greater sensitivity (Figure 1A). The BLAST output parameter has to be format to 6. Users can adjust the parameter of the E-value, but we recommend that it be no greater than 1e-5 (to ensure accurate predictions). Trouble shooting 1.

3. This will give a BLAST result file formed by a 12-column spreadsheet including the key information from query name to percentage identity, etc. (Figure 3B).
4. The 12-column explanation of BLAST search result file at format 6 (Table 1)
   a. query_ID (e.g., g735.t1)

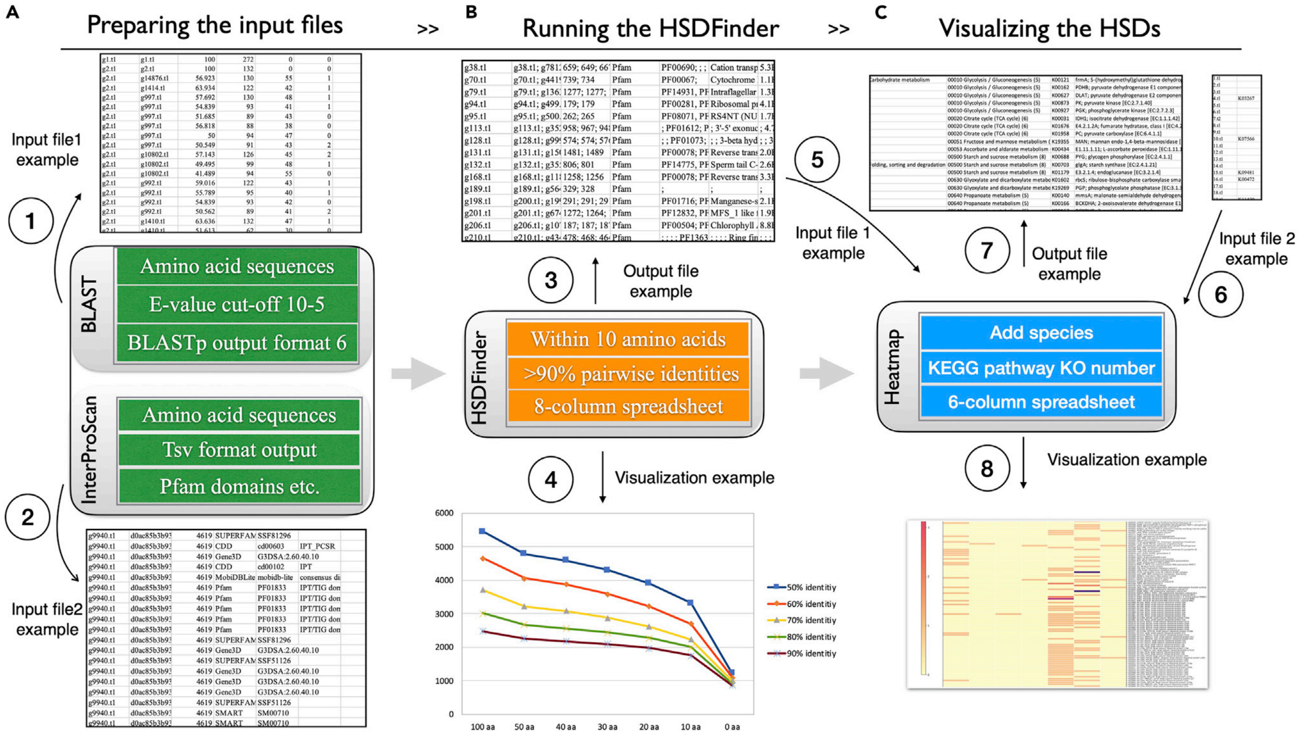


Figure 2. The workflow of the HSDFinder

(A) Two spreadsheets in tab-delimited format are displayed as examples for the input files of HSDFinder. One is acquired from the BLAST results in tabular format (-outfmt 6) and the other is the running result in default mode via InterProScan.

(B) The output of HSDFinder is an 8-column spreadsheet including information on gene copies to Pfam domain descriptions. Users have a choice to set different cut-off values to acquire potential duplicates. A trendline figures has been used as an example to interpret the number of total gene copies based on different cut-off thresholds.

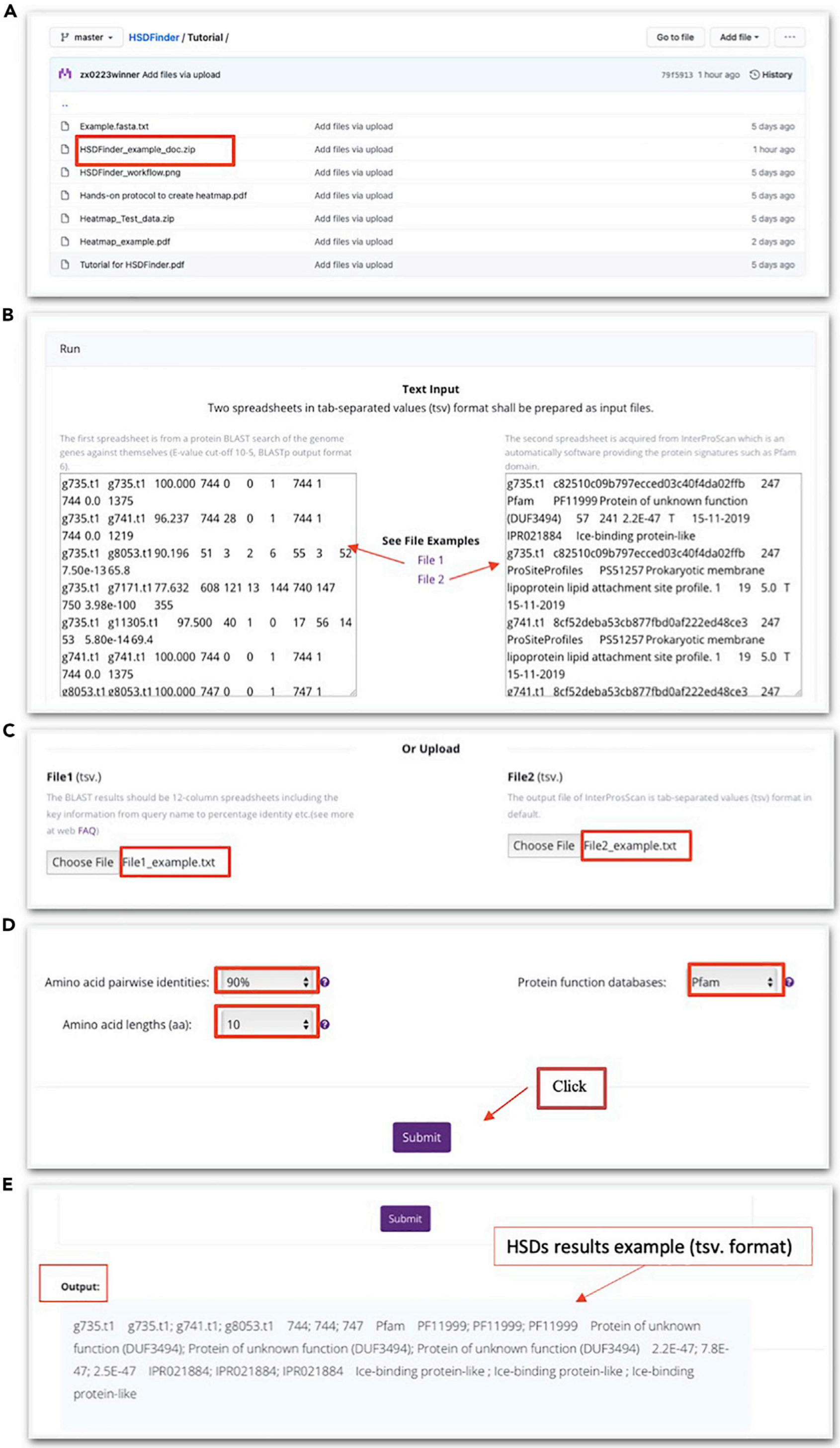
(C) The output file from step B together with a KEGG KO mapper file will be used as the input files to visualize the HSDs distribution across species. To create an appropriate heatmap, at least two species are needed. One of the 6-column output files have been displayed as an example to indicate the HSDs under the KEGG function categories with matching KO number and description. The heatmap example based on four species has been presented here. There is an option for users to download the high resolution heatmap figure and spreadsheet for future analysis. Image adapted from (Zhang et al., 2020b).

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)

5. If the BLAST-result file is too large to be copied and pasted, users have the option to upload a BLASTP-search result as the input of file 1 (Figure 3C). Troubleshooting 2.

Preparing the InterProScan search result file

⏱ Timing: minutes to hours



HSDFinder / Tutorial /

zx0223winner Add files via upload — 79f5913 1 hour ago

..		
Example.fasta.txt	Add files via upload	5 days ago
HSDFinder_example_doc.zip	Add files via upload	1 hour ago
HSDFinder_workflow.png	Add files via upload	5 days ago
Hands-on protocol to create heatmap.pdf	Add files via upload	5 days ago
Heatmap_Test_data.zip	Add files via upload	5 days ago
Heatmap_example.pdf	Add files via upload	2 days ago
Tutorial for HSDFinder.pdf	Add files via upload	5 days ago

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 10-5, 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-13 65.8
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-14 69.4
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
```

See File Examples
File 1
File 2

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

```
g735.t1  c82510c09b797ecced03c40f4da02ffb    247
Pfam     PF11999 Protein of unknown function
(DUF3494)    57   241 2.2E-47  T    15-11-2019
IPR021884    Ice-binding protein-like
g735.t1  c82510c09b797ecced03c40f4da02ffb    247
ProSiteProfiles     PS51257 Prokaryotic membrane
lipoprotein lipid attachment site profile. 1     19   5.0  T
15-11-2019
g741.t1  8cf52deba53cb877fbd0af222ed48ce3    247
ProSiteProfiles     PS51257 Prokaryotic membrane
lipoprotein lipid attachment site profile. 1     19   5.0  T
15-11-2019
g741.t1  8cf52deba53cb877fbd0af222ed48ce3    247
```

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 File1_example.txt

File2 (tsv.)

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

Choose File File2_example.txt

Amino acid pairwise identities: 90%

Amino acid lengths (aa): 10

Protein function databases: Pfam

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
```

Figure 3. Screenshots of specific steps when running HSDFinder

(A) Example GitHub dataset for running HSDFinder.
(B) Examples of text input files.
(C) The upload option to submit HSDFinder.
(D) The three point-and-click features for running HSDFinder.
(E) Example of an output from HSDFinder.

Upload an InterProScan search result file of your genome in tsv format as the second input file (File 2). User has to download and install the InterProScan individually to acquire the input file for HSDFinder tool. The latest InterProScan documentation can be found via the link https://interproscan-docs.readthedocs.io/en/latest/index.html. But, here, we provide the necessary steps to use InterProScan:

6. Installation requirements
   a. InterProScan is developed to run on Linux and no versions are planned for Windows or Apple (MAC OS X) operating systems.
   b. Software requirements: 64-bit Linux; Perl 5; Python 3; Java JDK/JRE version 11.
   c. Obtaining the core InterProScan software (Direct link: ftp://ftp.ebi.ac.uk/pub/software/unix/iprscan/5/5.51-85.0/interproscan-5.51-85.0-64-bit.tar.gz).
7. Running InterProScan
   a. Once a user has uncompressed the package of InterProScan, it can be run directly from the command line.

      ```
      #If run this script with no arguments, the usage instructions will be presented.
      >./interproscan.sh
      ```

   b. Run the shell script below:

      ```
      # interproscan.sh is the command taking in the input file with parameter (-i) and setting up the format of output file (e.g., tsv format). '-dp' is to ensure all the database matches proceeded in local environment.
      >/interproscan.sh -i proteins_of_your_genome.fasta -f tsv -dp
      ```

8. Output files
   a. InterProScan should run through properly without any warnings and it will create a tsv output file containing several member database matches, including Pfam, etc. For your convenience, the InterProScan search result example can be found in the ZIP file under the GitHub directory of tutorial with the name HSDFinder_example_doc.zip. Troubleshooting 3
9. The 13-column explanation of InterProScan search result file (Table 2)
   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)

   Note: Before clicking the submission button, there are three personalized options designed for HSDFinder (amino acid pairwise identity, amino acid length difference, and protein function database)

Yielding the output of HSDFinder with three personalized options

Timing: minutes to hours

10. By default, HSDFinder will filter duplicates with near-identical protein lengths (within 10 amino acids) and 90% pairwise identities. With such a strict cut-off, the user will capture the most similar duplicated genes within the dataset. But keep in mind that less similar duplicates will not necessarily be identified (Figure 3D).

11. Nevertheless, the user has the option to use different parameters and thresholds (from 30%–100% pairwise identity or from within 10-100 amino acid variance). Note that the false-positive rate of HSDs will increase with larger amino acid variance and smaller amino acid pairwise identity.
12. The output of this step is an 8-column spreadsheet with information on the HSD identifier, gene copy number, and Pfam domain (Figure 3E).
13. The 8-column explanation of HSDFinder result file.
    a. HSDs identifiers: By default, the first gene model of the duplicate gene copy is used as the HSD identifer. (e.g., g735.t1)
    b. Duplicate gene copies (within 10 amino acids, ≥90% pairwise identities) (e.g., g735.t1; g741.t1; g8053.t1)
    c. Amino acid length of duplicate gene copies (aa) (e.g., 744; 744; 747)
    d. Pfam identifier (e.g., PF11999; PF11999; PF11999)
    e. Analysis (e.g., Pfam / PRINTS / Gene3D)
    f. Pfam Description (e.g., Protein of unknown function (DUF3494); Protein of unknown function (DUF3494); Protein of unknown function (DUF3494))
    g. InterPro Entry Identifier (e.g., IPR021884; IPR021884; IPR021884)
    h. InterPro Entry Description (e.g., Ice-binding protein-like; Ice-binding protein-like; Ice-binding protein-like)

⚠ CRITICAL: A HSDFinder result example can be found in the ZIP file under the GitHub directory of tutorial with the name HSDFinder_example_doc.zip (Figure 3A). Troubleshooting 4.

Visualizing the HSDFinder results in Microsoft Excel

⏲ Timing: minutes to hours

14. 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 the psychrophilic green alga *Chlamydomonas* sp. UWO241 (NCBI BioProject: PRJNA547753) (Figures 4A and 4B).

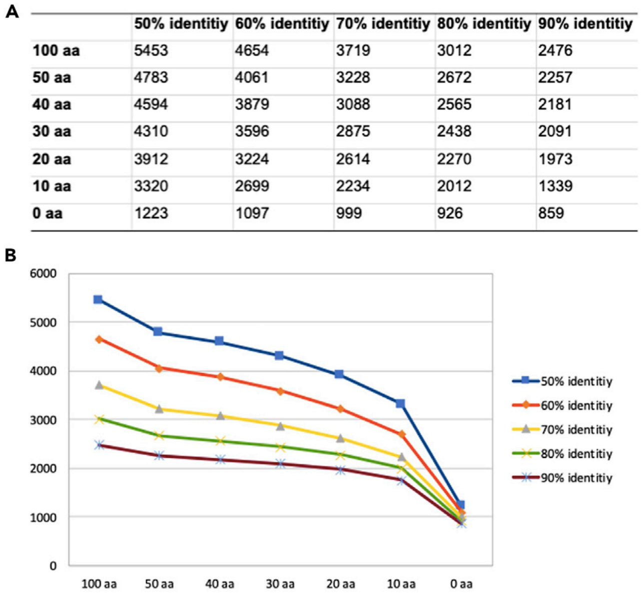
Note: The online heatmap tool is a great choice if you want to compare HSDs (and their associated KEGG pathway categories) among two or more species.

Upload the results of HSDFinder from your respective genomes

⏲ Timing: hours to days

Upload the results of HSDFinder from your respective genomes. Two files are needed to plot heatmap for each species. The first input file is the output of your species of interest after running the HSDFinder; file examples are given to guide the appropriate input file (Figure 5A).

15. Hands-on protocol to create heatmap with Test data. Download the Heatmap_Test_data.zip via the link from GitHub (https://github.com/zx0223winner/HSDFinder).
16. We provide the data from eight algal species (*Chlamydomonas* sp. UWO241, *Chlamydomonas* sp. ICE-L, *Chlamydomonas reinhardtii*, *Chlamydomonas eustigma*, *Gonium pectorale*, *Dunaliella salina*, *Volvox carteri*, and *Fragilariopsis cylindrus*). Users can create a heatmap by selecting some of them.
    a. Each folder represents one species. There are two files in each folder. For *C. reinhardtii*, for example, there is a file named "HSD_Chlamy_90pct_10aa.txt", which contains the *C. reinhardtii* nuclear genome HSDs results (filtering option more than 90% amino acid pairwise identity and within 10 amino acid differences).



	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

Figure 4. Visualizing the HSD results under different thresholds

(A) Table of duplicate gene copy numbers at different thresholds of amino acid pairwise identity and deduced amino acid length.

(B) 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 and the Y-axis tells the number of gene copies. Images adopted from (Zhang et al., 2021) with permission.

b. Another file named "Chlamy_KO_annotation.txt" represents the retrieved results from the KEGG database, documenting the correlation of KO accession with each gene model identifier.

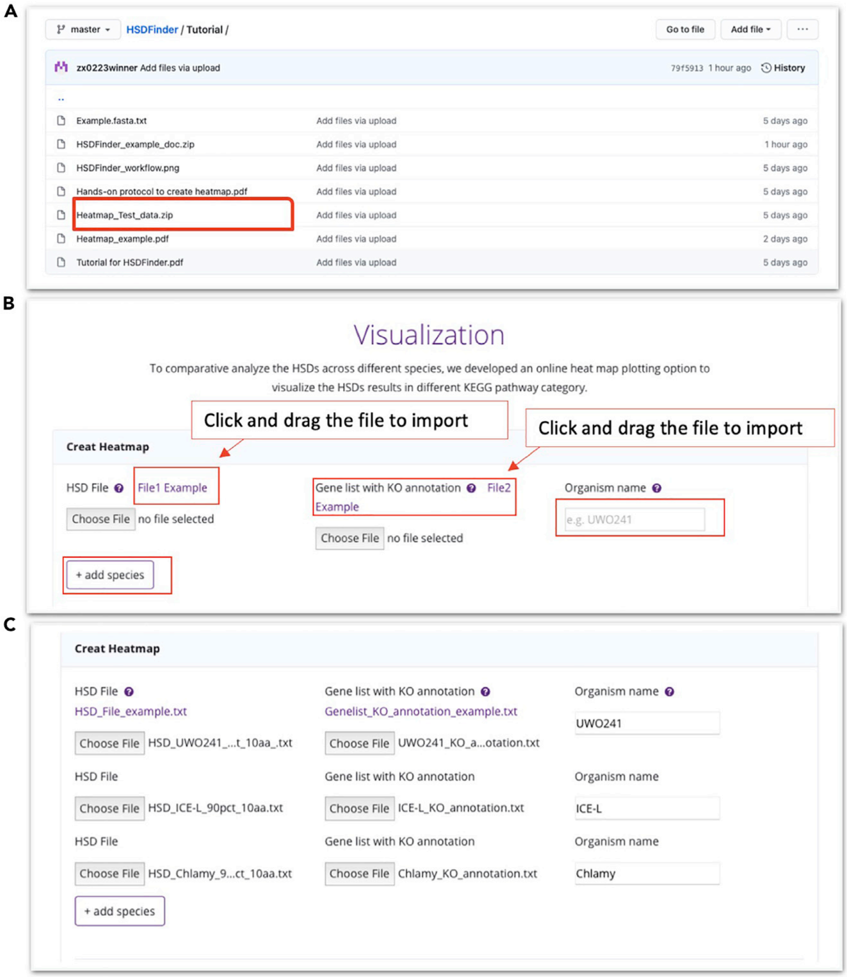
c. The user can upload the respective files to the web server to create the heatmap (Figures 5B and 5C).

Upload a gene list with KO annotation from KEGG database

Timing: minutes to hours

The second file is retrieved from the KEGG database documenting the correlation of KO accession with each gene model identifier. User has to use the Ghost KOALA or BlastKOALA analysis tool of KEGG to acquire the KO annotation file of the genome (https://www.kegg.jp/ghostkoala/). But, here, we provide the necessary steps to guide using the tools:

17. BlastKOALA accepts a smaller dataset and is suitable for annotating high-quality genomes.
    a. Upload query amino acid sequences in FASTA format.
    b. Enter taxonomy group of your genome.
    c. Enter KEGG GENES database file to be searched.
    d. Enter your email address. An email will be sent to you for confirmation of your input data. Make sure to click on the link in the email to initiate your job and then you will receive another email once it is finished.
18. GhostKOALA accepts a larger dataset (e.g., 300 Mb) and is suitable for annotating metagenomes.
    a. Upload query amino acid sequences in FASTA format.
    b. Enter KEGG GENES database file to be searched.
    c. Enter your email address. Same as above (1d).


A
B
C
master
Go to file
Add file
History
Click and drag the file to import
Click and drag the file to import

HSDFinder / Tutorial /

zx0223winner Add files via upload — 79f5913 1 hour ago

Example.fasta.txt	Add files via upload	5 days ago
HSDFinder_example_doc.zip	Add files via upload	1 hour ago
HSDFinder_workflow.png	Add files via upload	5 days ago
Hands-on protocol to create heatmap.pdf	Add files via upload	5 days ago
Heatmap_Test_data.zip	Add files via upload	5 days ago
Heatmap_example.pdf	Add files via upload	2 days ago
Tutorial for HSDFinder.pdf	Add files via upload	5 days ago

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.

Creat Heatmap

HSD File — File1 Example — Choose File no file selected
Gene list with KO annotation — File2 Example — Choose File no file selected
Organism name — e.g. UWO241
+ add species

Creat Heatmap

HSD File	Gene list with KO annotation	Organism name
HSD_File_example.txt	Genelist_KO_annotation_example.txt	
Choose File HSD_UWO241_...t_10aa_.txt	Choose File UWO241_KO_a...otation.txt	UWO241
Choose File HSD_ICE-L_90pct_10aa.txt	Choose File ICE-L_KO_annotation.txt	ICE-L
Choose File HSD_Chlamy_9...ct_10aa.txt	Choose File Chlamy_KO_annotation.txt	Chlamy

+ add species

Figure 5. Screenshots of specific steps when visualizing HSDs in a heatmap

(A) Test data from GitHub used to visualize HSDs across species.

(B) The input file options for the visualization tool.

(C) Example of submitted files.

19. From the KEGG email link, the user can download the gene list with the KO annotations. The format of the output file can be referred to in Table 3. Explanation of the 2-column input file for KO accession (Table 3).
    a. Gene identifier (e.g., g10.t1)
    b. KO accession (e.g., K09481)
20. Use the Ghost KOALA or BlastKOALA analysis tool of KEGG to acquire the KO annotation file of your genome (https://www.kegg.jp/ghostkoala/). An example of a KO annotation file is given under the GitHub directory of tutorial with the name Heatmap_Test_data.zip. Troubleshooting 5

21. Fill in the organism's name. This is the identifier used to compare HSDs among different species. To add more species, use the "+add species" button and select the respective files. Troubleshooting 6.

Note: For the best visualization results, select at least two species. However, the result can still be visualized using a single species. Additional examples of KO annotation files are provided under the GitHub directory of tutorial with the name Heatmap_Test_data.zip.

Table 3. Example of KO accessions with each gene model identifier retrieved from the KEGG database.

Gene identifier	KO accession
g10.t1	K07566
g11.t1	N/A
g12.t1	N/A
g13.t1	N/A
g14.t1	N/A
g15.t1	K09481
g16.t1	K00472

⚠ CRITICAL: Make sure you have an organism name for the files you chose to upload (Figure 5B).

Output files of the online heatmap visualization tool

⏲ Timing: minutes

22. Once the files are uploaded, there is an option to designate the figure size. The 'row' option can change the width of the heatmap image, and the 'col' option can change the length of the heatmap image (Figure 6A).
23. Tap the "Create Heatmap" button and a pending image will jump out. It usually takes less than one minute to run with three to five species selected (Figure 6B).

The heatmap of HSD levels across species

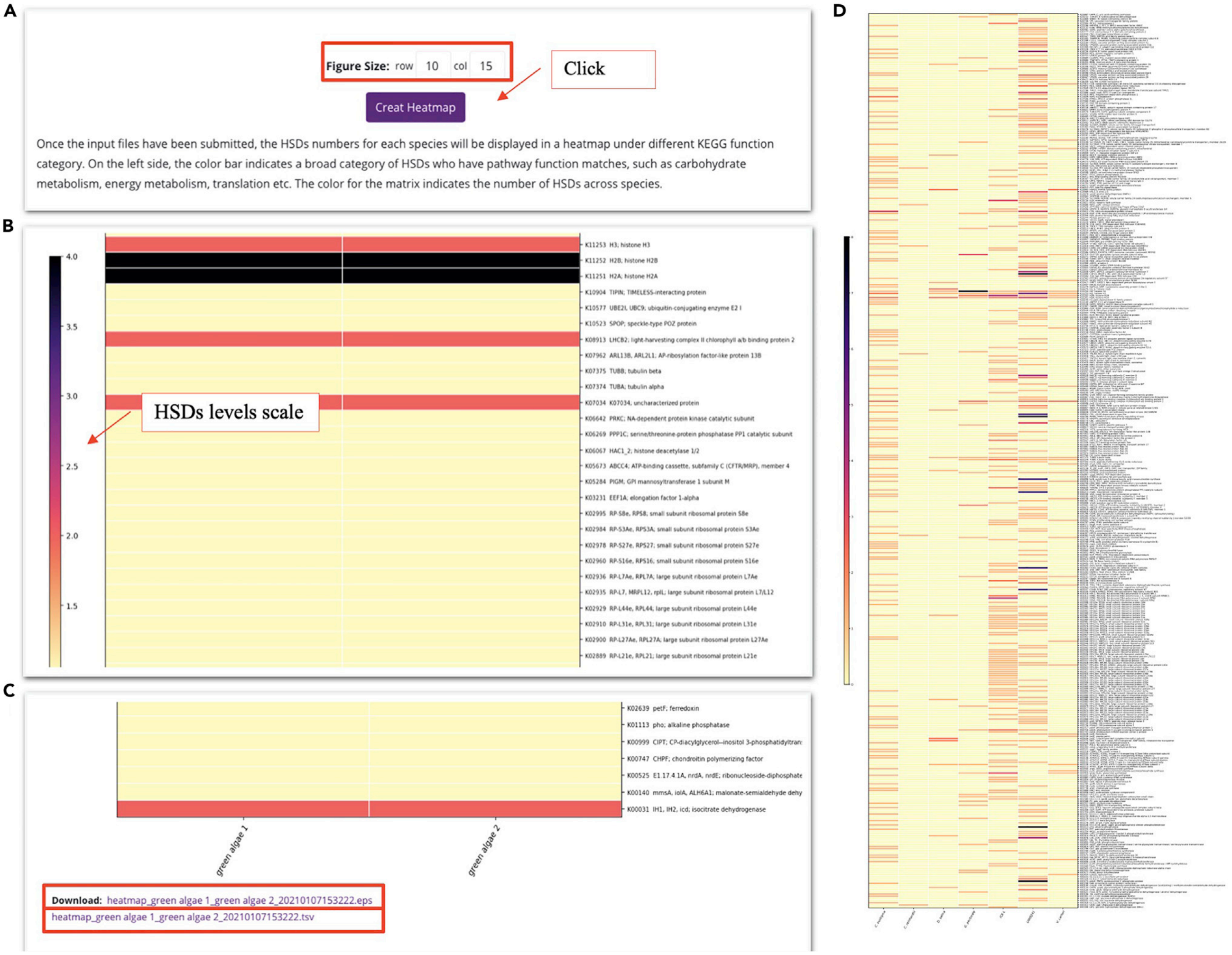
⏲ Timing: minutes

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

24. Below the image, there is an option to download the high-resolution image file and the tab-delimited file for future analysis (Figure 6C).
25. The 8-column explanation of the tab-delimited file (.tsv) file (Table 4).
    a. Identifier (e.g., 0)
    b. Pathway category1 (e.g., 09101 Carbohydrate metabolism)
    c. Pathway category2 (e.g., 00010 Glycolysis / Gluconeogenesis [PATH: ko00010])
    d. KEGG KO_ID (e.g., K13979)
    e. Function (e.g., yahK; alcohol dehydrogenase (NAP+))
    f. Species_name (e.g., UWO241)
    g. HSDs_ID (e.g., g1713.t1)
    h. HSDs_Num (e.g., 1)
26. If you are not satisfied with the heatmap figure size (e.g., the image texts are overlapped), you can always rerun with more appropriate 'width and length' options.

EXPECTED OUTCOMES

HSDFinder is a free, easy-to-use automated online bioinformatics software tool for identifying duplicated genes in nuclear genomes. It offers high resolution heatmap plots (.eps) and the tab-delimited file (tsv.) to visualize and categorize HSDs across species. By comparing the duplicates in different species, a user can easily find out what kind of genes and associated pathways are duplicated within their genome(s) of interest.


A
Figure Size: row 10 col 15
Click
Creat Heatmap
B
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
C
K02639 petF; ferredoxin
K01113 pho; alkaline phosphatase
K00999 CIPT; CP-diacylglycerol--inositol 3-phosphatidyltran:
K00747 CHPF; chondroitin polymerizing factor
K00525 E1.17.4.1A, nrdA, nrdE; ribonucleoside-diphosphate
K00140 mmsA, iolA, ALH6A1; malonate-semialdehyde dehy
K00031 IH1, IH2, icd; isocitrate dehydrogenase
green algae 1
green algae 2
D

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.

Download: heatmap_green algae 1_green algae 2_20210107153222.eps
heatmap_green algae 1_green algae 2_20210107153222.tsv

Figure 6. Screenshots of heatmap example

(A) Option for choosing the size of the heatmap.

(B) The scale and metrics in the heatmap.

(C) High-resolution image and spreadsheet of the heatmap result files.

(D) Example of the heatmap file (.eps) visualizing the HSDs across seven green algal species. Figure 6D was adapted with permission from (Zhang et al. 2021).

It is our aim to build a comprehensive analysis of HSDs in the eukaryotic nuclear genomes. The predicted HSDs results generated by HSDFinder are documented in HSDatabase (Zhang et al., 2020a), which currently contains a total of 28,214 HSDs from fifteen eukaryotic nuclear genomes (http://hsdfinder.com/database/).

The outcome of identified and annotated HSDs

HSDFinder generates one output file: 8-column spreadsheet integrating information on HSD identifier, gene copy number, and Pfam domain. More details have been discussed in yielding the output of HSDFinder with three personalized options and visualizing the HSDFinder results in microsoft excel.

The outcome of categorized and visualized HSDs

HSDFinder generates two output files: 8-column tab-delimited file (.tsv) for HSDs of different species categorized under different KEGG functional categories and high resolution heatmap file (.eps) visualizing the HSDs across your genome(s) of interest (Figure 6D).

Table 4. Example of an 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+)	UWO241	g1713.t1	1
1	09101 Carbohydrate metabolism	00020 itrate cycle (TA cycle) [PATH: ko00020]	K00031	IH1, IH2, icd; isocitrate dehydrogenase	UWO241	g3379.t1	1
2	09101 Carbohydrate metabolism	00030 Pentose phosphate pathway [PATH: ko00030]	K00036	G6P, zwf; glucose-6-phosphate 1-dehydrogenase	UWO241	g852.t1	1
3	09101 Carbohydrate metabolism	00051 Fructose and mannose metabolism [PATH: ko00051]	K19355	MAN; mannan endo-1, 4-beta-mannosidase	UWO241	g3766.t1	1
4	09101 Carbohydrate metabolism	00053 Ascorbate and aldarate metabolism [PATH: ko00053]	K00434	E1.11.1.11; L-ascorbate peroxidase	UWO241	g15878.t1	1
5	09103 Lipid metabolism	00073 utin, suberine and wax biosynthesis [PATH: ko00073]	K13356	FAR; alcohol-forming fatty acyl-CoA reductase	UWO241	g6944.t1	1
6	09108 Metabolism of cofactors and vitamins	00130 Ubiquinone and other terpenoid-quinone biosynthesis [PATH: ko00130]	K17872	NC1, ndbB; demethylphylloquinone reductase	UWO241	g269.t1, g13422.t1	2

An Online heatmap visualization tool has been detailed in Output files of the online Heatmap Visualization tool and The heatmap of HSDs levels across species.

LIMITATIONS

The web tool is limited to presenting the HSDs in a heatmap plot; however, this plot is a straightforward way to visualize the levels of HSDs across species. HSDFinder can categorize and visualize the HSDs under KEGG pathway categories but the specific pathway function items are too detailed to incorporate into the plot. Therefore, an alternative plot method should be used to simplify the description. The web tool is limited to using the InterProScan and KEGG database to annotate the duplicates. Users might have to try different thresholds to filter and identify HSDs. In our experiences from analyzing eukaryotic green algal nuclear genomes, the default settings of HSDFinder were able to detect a significant proportion of complete duplicated genes, but many fragmented and partial duplicates were missed.

Available non-redundant protein sequence databases, such as the NCBI NR database, SwissProt (Consortium, 2019), and TrEMBL (Boeckmann et al., 2003), can also be used to annotate HSDs. We developed a tool called NoBadWordsCombiner v1.0 (http://hsdfinder.com/combiner/) (Zhang et al., 2020c), which can automatically merge the annotations from SwissProt (Consortium, 2019), TrEMBL (Boeckmann et al., 2003) and NCBI databases. More importantly, it can strengthen the duplicated genes' definition by minimizing protein function descriptions containing 'bad words', such as hypothetical and uncharacterized proteins. The web tool is also relying on third-party tools to generate the input files. Users have to be familiar with the basic BLAST package and dash shell in Linux/Unix environment. It is our hope to visualize the duplicates across species with fewer middle steps, but we provide the build-in reference files for each input file as well as a step-by-step protocol to guide the heatmap plot with example data.

In the future, HSDFinder will be further improved, including continuous updating by considering more scientific discoveries in the field of gene duplication. It will also be expanded to consider other types of genomic data, such as prokaryotic and organelle genomes.

TROUBLESHOOTING

Problem 1

Why does BLASTP need to be chosen as an option? What E-value shall I choose? (step 2)

Potential solution

Make sure to use the BLASTP option due to amino acid substitutions occurring less frequently than nucleotide substitutions. We recommend the E-value to be no larger than 1e-5 to ensure accurate prediction.

Problem 2

Can I submit the input files by using the copy-and-paste txt blank and upload option at the same time? (step 5)

Potential solution

No. The HSDFinder will prioritize using the uploaded files as the input files. If you submit a file by mistake, simply re-fresh (reload) the browser page.

Problem 3

Is it difficult to run the InterProScan? How does it work in HSDFinder? (step 8)

Potential solution

No. It is very straightforward and easy to use. The real example of the InterProScan result has been provided at GitHub in the HSDFinder_example_doc.zip file named "Input_2_InterProScan_UWO241.txt". It is a tab-delimited file including the protein signatures, such as Pfam domain and InterPro annotations.

Problem 4

What does the standard HSDFinder output look like? (step 13)

Potential solution

The example file has been provided with the name "Output_HSDFinder_UWO241_90%_10aa.txt" from GitHub. That indicates the duplicates are detected via the threshold of at least 90% amino acid identity and within 10 amino acid variances.

Problem 5

Why is the KEGG KO annotation file needed and what does it look like? (step 20)

Potential solution

The example file has been provided with the name "Heatmap_KEGG_KO_UWO241.txt" from GitHub. The file documents the correlation of KO accession with each gene model identifier, which can be used to categorize the identified HSDs under different functional categories.

Problem 6

Where to find the data of other species to test the HSDFinder? (step 21)

Potential solution

We have provided the dataset for other species to create the heatmap, which is under the GitHub directory of tutorial with the name Heatmap_Test_data.zip.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact David Roy Smith (dsmit242@uwo.ca) and technical contact Xi Zhang (xzha25@uwo.ca)

Materials availability

This study did not generate new unique reagents.

Data and code availability

The datasets of eukaryotes supporting the conclusions of this article are available from JGI (https://phytozome.jgi.doe.gov/pz/portal.html) or NCBI (https://www.ncbi.nlm.nih.gov) database. The HSDFinder source code has been deposited at https://github.com/zx0223winner/HSDFinder. The web server of HSDFinder is freely available at http://hsdfinder.com. The predicted HSDs of fifteen eukaryotes are documented in HSDatabase, which can be accessed via http://hsdfinder.com/database/.

ACKNOWLEDGMENTS

The authors gratefully acknowledge funding of Discovery Grants from the Natural Sciences and Engineering Research Council of Canada (NSERC).

AUTHOR CONTRIBUTIONS

The study was conceptualized by X.Z. and D.R.S. The data were analyzed by X.Z. and Y.H. implemented the HSDFinder website. X.Z. and D.R.S. drafted the manuscript, and all authors commented to produce the manuscript for peer review.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.-C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'donovan, C., and Phan, I. (2003). The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Consortium, U. (2019). UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. *47*, D506–D515.

Cvetkovska, M., Szyszka-Mroz, B., Possmayer, M., Pittock, P., Lajoie, G., Smith, D.R., and Hüner, N.P. (2018). Characterization of photosynthetic ferredoxin from the Antarctic alga *Chlamydomonas* sp. UWO241 reveals novel features of cold adaptation. New Phytol. *219*, 588–604.

Hanschen, E.R., Marriage, T.N., Ferris, P.J., Hamaji, T., Toyoda, A., Fujiyama, A., Neme, R., Noguchi, H., Minakuchi, Y., and Suzuki, M. (2016). The *Gonium pectorale* genome demonstrates co-option of cell cycle regulation during the evolution of multicellularity. Nat. Commun. *7*, 1–10.

Hirooka, S., Hirose, Y., Kanesaki, Y., Higuchi, S., Fujiwara, T., Onuma, R., Era, A., Ohbayashi, R., Uzuka, A., and Nozaki, H. (2017). Acidophilic green algal genome provides insights into adaptation to an acidic environment. Proc. Natl. Acad. Sci. U S A *114*, 8304–8313.

Innan, H., and Kondrashov, F. (2010). The evolution of gene duplications: classifying and distinguishing between models. Nat. Rev. Genet. *11*, 97–108.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kanehisa, M., Sato, Y., and Morishima, K. (2016). BlastKOALA and GhostKOALA: KEGG tools for functional characterization of genome and metagenome sequences. J. Mol. Biol. *428*, 726–731.

Kent, W.J. (2002). BLAT—the BLAST-like alignment tool. Genome Res. *12*, 656–664.

Kondrashov, F.A. (2012). Gene duplication as a mechanism of genomic adaptation to a changing environment. Proc. R. Soc. B. *279*, 5048–5057.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. Annu. Rev. Genet. *39*, 309–338.

Lallemand, T., Leduc, M., Landès, C., Rizzon, C., and Lerat, E. (2020). An overview of duplicated gene detection methods: Why the duplication mechanism has to be accounted for in their choice. Genes *11*, 1046.

Merchant, S.S., Prochnik, S.E., Vallon, O., Harris, E.H., Karpowicz, S.J., Witman, G.B., Terry, A., Salamov, A., Fritz-Laylin, L.K., and Maréchal-Drouard, L. (2007). The *Chlamydomonas* genome reveals the evolution of key animal and plant functions. Science *318*, 245–250.

Pellerin, E. (2016). Veritas Genetics launches $999 whole genome and sets new standard for genetic testing. https://www.veritasgenetics.com/documents/VG-launches-999-whole-genome.pdf.

Polle, J.E., Barry, K., Cushman, J., Schmutz, J., Tran, D., Hathwaik, L.T., Yim, W.C., Jenkins, J., McKie-Krisberg, Z., and Prochnik, S. (2017). Draft nuclear genome sequence of the halophilic and beta-carotene-accumulating green alga *Dunaliella salina* strain CCAP19/18. Genome Announc. *5*, 01105–01117.

Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., Miller, S.M., Nishii, I., Ferris, P., Kuo, A., Mitros, T., and Fritz-Laylin, L.K. (2010). Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science *329*, 223–226.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005). InterProScan: protein domains identifier. Nucleic Acids Res. *33*, 116–120.

Zhang, X., Cvetkovska, M., Morgan-Kiss, R., Hüner, N.P., and Smith, D.R. (2021). Draft genome sequence of the Antarctic green alga Chlamydomonas sp. UWO241. iScience, 102084.

Zhang, X., Hu, Y., and Smith, D.R. (2020a). HSDatabase - a database of highly similar duplicate genes in eukaryotic genomes. http://hsdfinder.com/database/.

Zhang, X., Hu, Y., and Smith, D.R. (2020b). HSDFinder- an integrated tool to predict highly similar duplicates in eukaryotic genomes. https://github.com/zx0223winner/HSDFinder.

Zhang, X., Hu, Y., and Smith, D.R. (2020c). NoBadWordsCombiner-a tool to integrate the gene function information together without 'bad words' from Nr-NCBI, UniProtKB/Swiss-Prot, KEGG, Pfam databases. https://github.com/zx0223winner/HSDFinder/blob/master/NoBadWordsCombiner.py.

Zhang, Z., Qu, C., Zhang, K., He, Y., Zhao, X., Yang, L., Zheng, Z., Ma, X., Wang, X., and Wang, W. (2020d). Adaptation to extreme Antarctic environments revealed by the genome of a sea ice green alga. Curr. Biol. *30*, 1–12.