

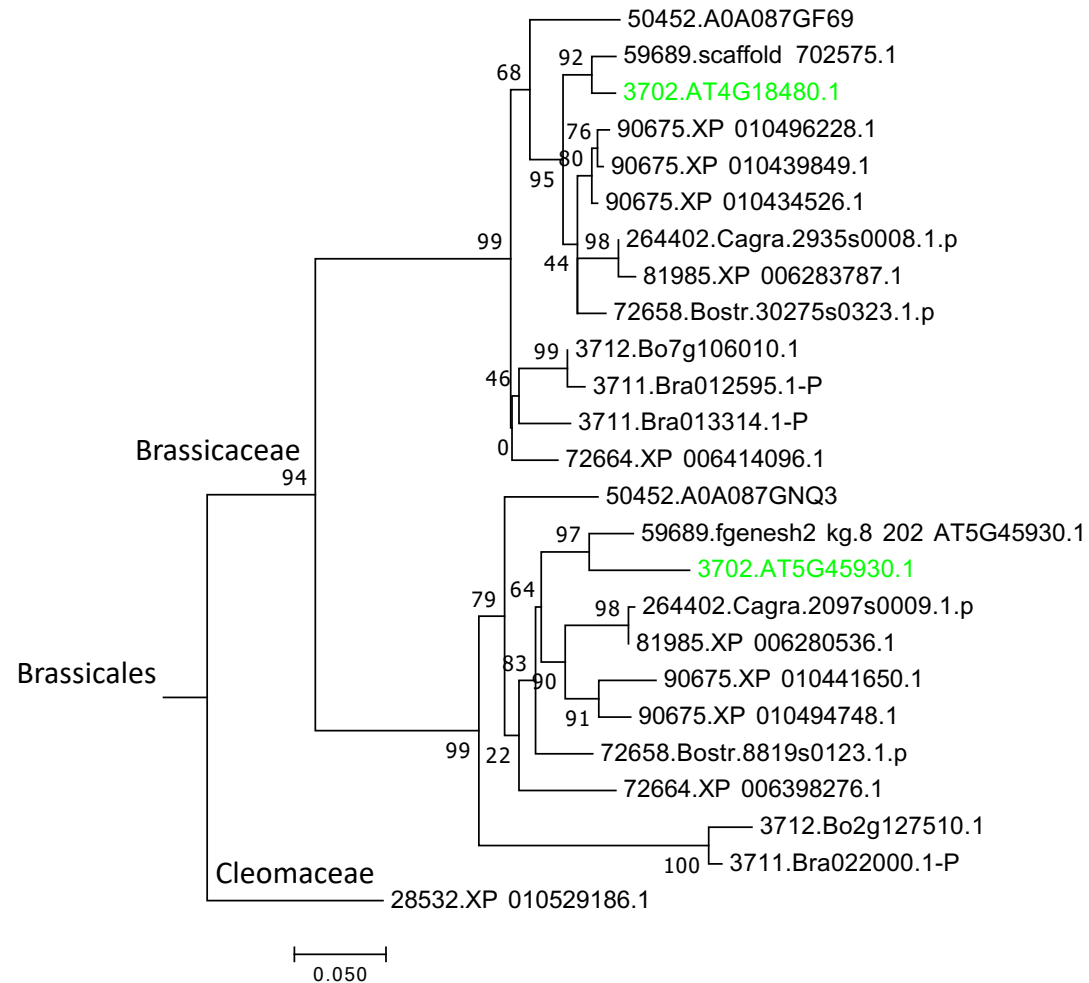
KusakiDB and Hayai-Annotation

Andrea Ghelfi

Kazusa DNA Research Institute

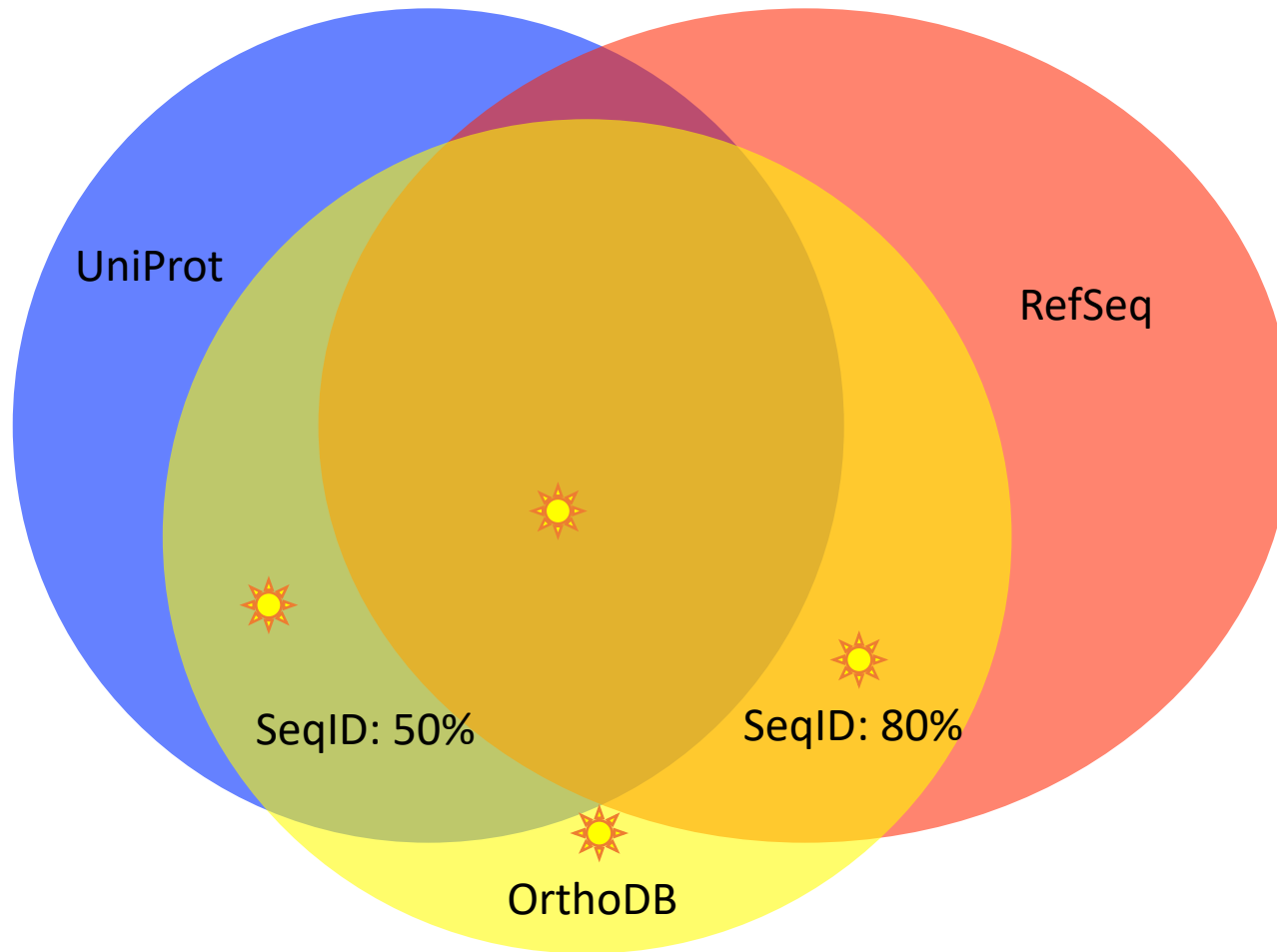
Definition of orthogroups

- We refer as orthogroups to all descendants of a particular single gene of the last common ancestor, in our case Viridiplantae level.
- Important to note that notions of orthologs and paralogs are disjoint because paralogs can be co-orthologs if duplicated after the speciation.



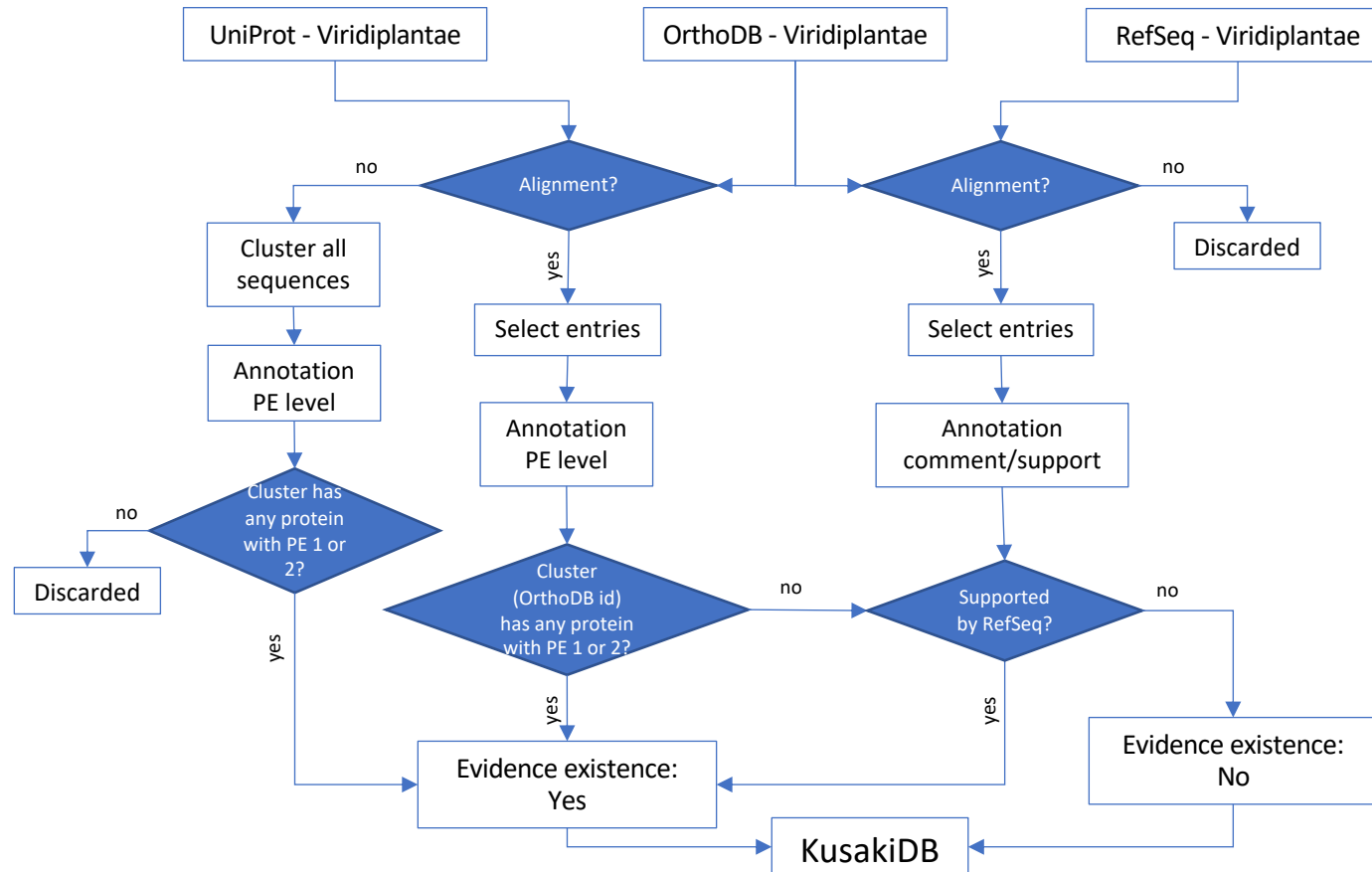
Magnesium-chelatase subunit ChII-2, chloroplastic

KusakiDB: Intersections of UniProt, RefSeq and OrthoDB

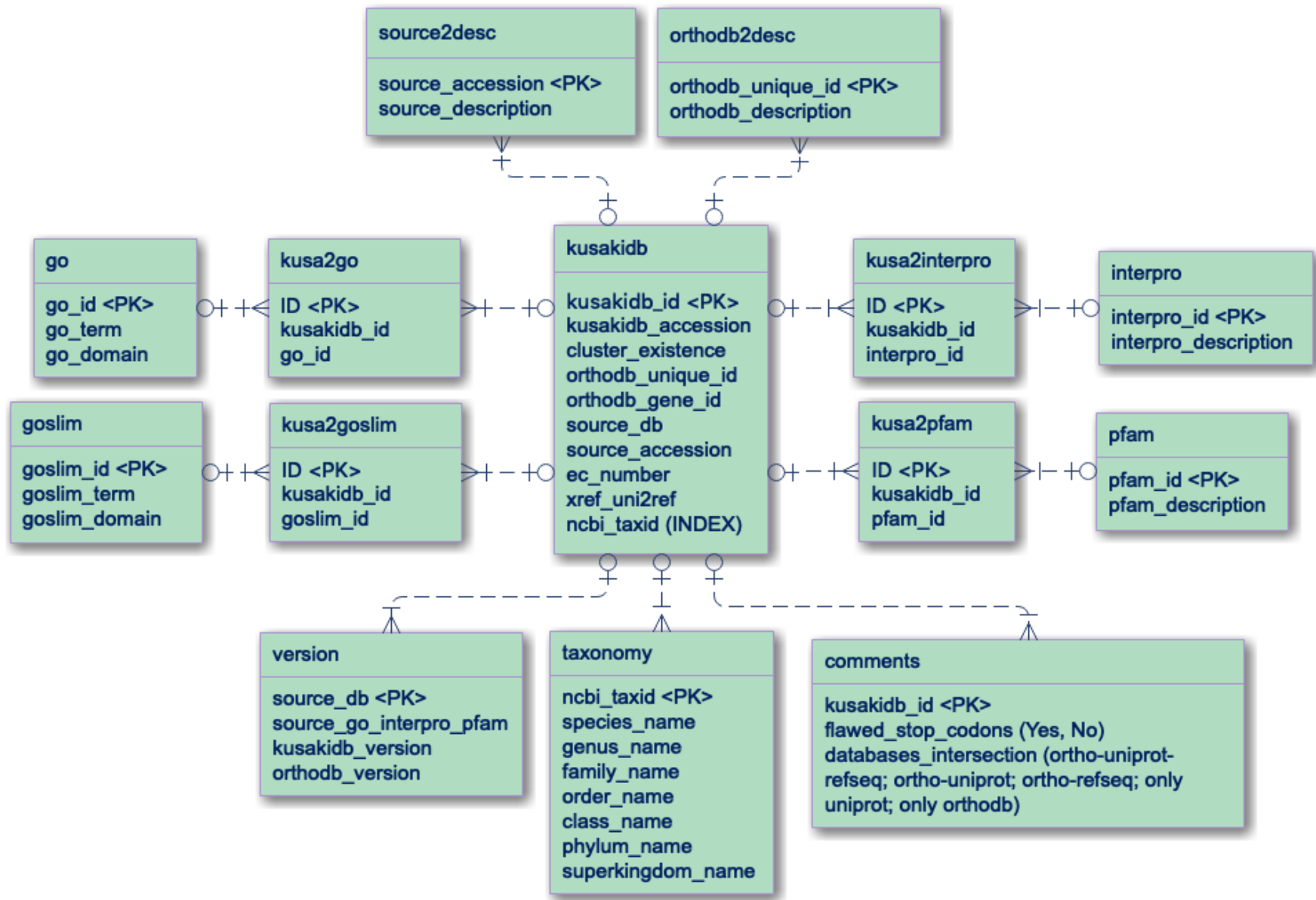


Clustered to 99% sequence identity to remove redundancies

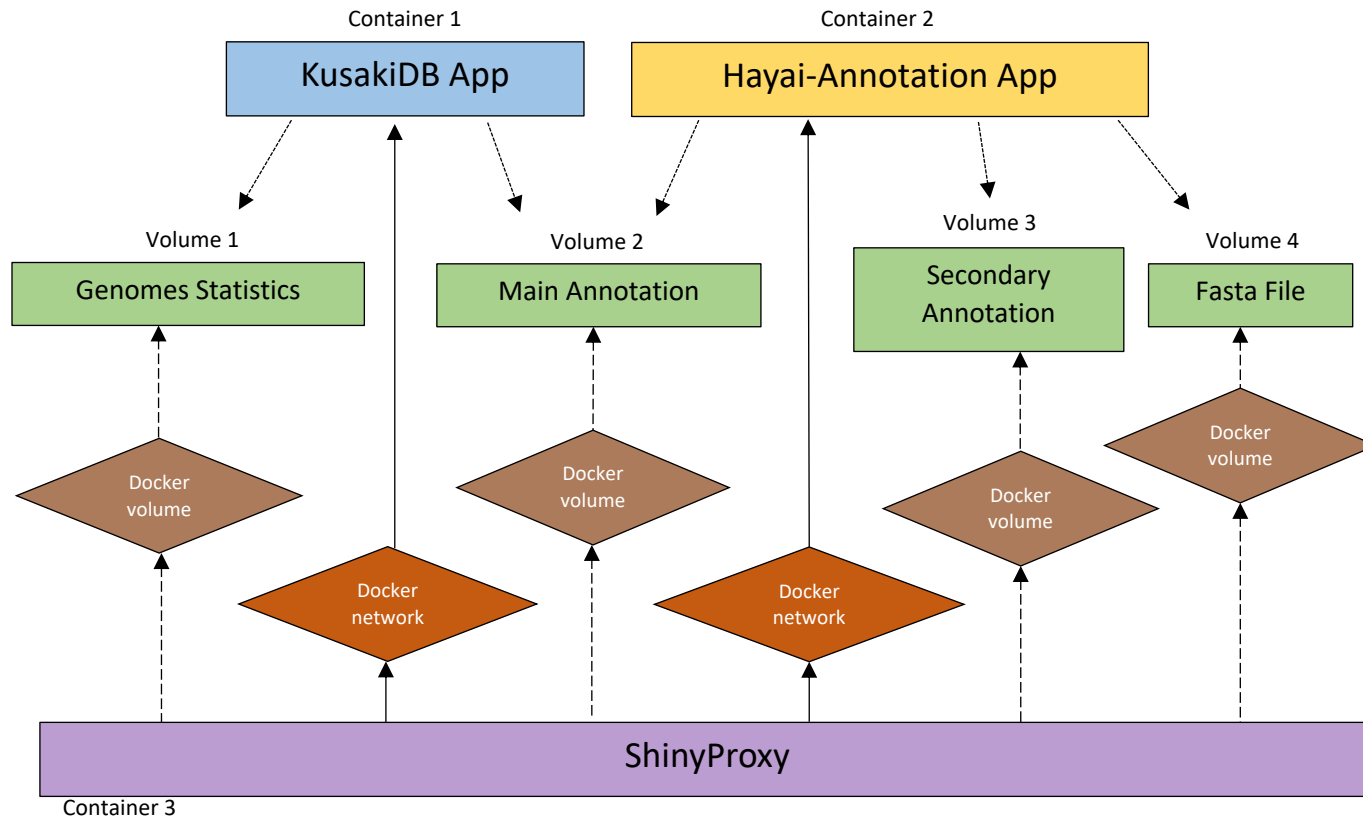
KusakiDB: a novel approach for assessing the existence and completeness of orthogroups in plant species



KusakiDB: MySQL Schema



KusakiDB and Hayai-Annotation: Docker Container Structure



KusakiDB: A novel approach for existence and completeness of protein orthologs

KusakiDB v1.0

A novel approach for assessing existence and completeness of orthogroups in plant species

Search KusakiDB OG Assessment OG Management OG Management User Data Help

Search proteins within species or family level

Taxon Level: Species Taxon Name: Zea mays

Search: flower

Column visibility

OrthoDB_Group	OrthoDB_Protein_Name	Accession	Accession_Protein_Name	Evidence_existence
18820at33090	SNF2-related, N-terminal domain	A0A1D6Q787	Protein PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1	Yes
222987at33090	Protein EARLY FLOWERING 4 domain	A0A3L6E008	Protein ELF4-LIKE 2	Yes
18820at33090	SNF2-related, N-terminal domain	A0A1D6Q7G0	Protein PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1	Yes
42805at33090	Amine oxidase	A0A1D6E3G3	Flowering locus D	Yes
24394at33090	Protein EARLY FLOWERING 3-like	C0PGT6	Protein EARLY FLOWERING 3	Yes
116030at33090	Protein EMBRYONIC FLOWER 1	A0A3L6G4E5	Uncharacterized Protein	Yes
222987at33090	Protein EARLY FLOWERING 4 domain	A0A3L6F041	Protein ELF4-LIKE 4	Yes
220345at33090	Flowering-promoting factor 1	A0A3L6FFK4	Flowering-promoting factor 1-like protein 2	Yes
58478at33090	RNA recognition motif domain	A0A1D6I8U6	Flowering time control protein FPA	Yes
135732at33090	WD40 repeat	C0P3D9	Flowering time control protein FY	Yes
18820at33090	SNF2-related, N-terminal domain	A0A1D6HGC1	Protein PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1	Yes
198713at33090	Phosphatidylethanolamine-binding protein	A9LLX9	Protein TERMINAL FLOWER 1	Yes
18820at33090	SNF2-related, N-terminal domain	A0A1D6HGH1	Protein PHOTOPERIOD-INDEPENDENT EARLY FLOWERING 1	Yes
44305at33090	Polycomb protein, VEFS-Box	A0A317Y0N3	Polycomb group protein EMBRYONIC FLOWER 2	Yes

Select Taxon Level:
Species or Family

Select Species or Family
per taxon name

Filter your results

Column visibility

Evidence_existe...

Species

Family

kusakiDB_id

Results: in blue
hyperlink to source
OrthoDB, UniProt
and RefSeq

KusakiDB Tools: OG Assessment

Select a Family
from KusakiDB

KusakiDB v1.0

A novel approach for assessing existence and completeness of orthogroups in plant species

Search KusakiDB **OG Assessment** OG Management OG Management User Data Help

Assess conservation of orthogroups within complete genomes

Choose Family

Family Name

Poaceae

Or analyse your own data. Click here to create Hayai-annotation file

Upload Hayai_annotation_v2.0.tsv

Browse... No file selected

Submit

Users' Data

KusakiDB Complete Genomes Data

Family	Scientific_name	Number_of_OG_Groups	Evidence_existence (%)	Median_Family (%)	Total_Family	Median_Species (%)
Poaceae	Oryza brachyantha	11801	96.95	90.91	11	83.76
Poaceae	Triticum urartu	10512	91.83	90.91	11	82.91
Poaceae	Brachypodium distachyon	12833	95.11	90.91	11	82.91
Poaceae	Zea mays	13021	96.84	90.91	11	82.05
Poaceae	Oryza sativa Japonica Group	13128	96.95	90.91	11	82.05
Poaceae	Sorghum bicolor	13314	94.88	90.91	11	82.05
Poaceae	Dichanthelium oligosanthes	11988	93.45	90.91	11	82.05
Poaceae	Panicum hallii	13708	92.08	90.91	11	82.05

Interpretation Median (Family or Species):
50% (median) of its OGs are shared by 91% of all species in Poaceae and 84% among all species

Choose Family

Family Name

P

Poaceae

Phrymaceae

Pedaliaceae

Papaveraceae

Apiaceae

Asparagaceae

Cephalotaceae

Euborbiaceae

KusakiDB OG Assessment: Evaluating Gene Structural Prediction

Upload your own data
annotated by Hayai-
annotation v2.0

KusakiDB_v1.0 x HayaiAnnotation_v2.0 x +

Not Secure pgdbjsnp.kazusa.or.jp/app/kusakidb

Search KusakiDB **OG Assessment** OG Management OG Management User Data Help

Assess conservation of orthogroups within complete genomes

Choose Family

Family Name

Poaceae

Or analyse your own data. Click here to create Hayai-annotation file

Upload Hayai_annotation_v2.0.tsv

Browse... Hayai_annotation_v2.0.tsv

Upload complete

Submit

Ensembl genomes: release-50 (2020.12)
Triticum_aestivum.IWGSC.pep.all.fa

Users' Data

Inferred_family	Species_source	Number_of_OG_Groups	Evidence_existence (%)	Median_Family (%)	Total_Family	Median_Species (%)
Poaceae	User_data	14729	89.32	90.91	11	80.34

KusakiDB Complete Genomes Data

Family	Scientific_name	Number_of_OG_Groups	Evidence_existence (%)	Median_Family (%)	Total_Family	Median_Species (%)
Poaceae	Oryza brachyantha	11801	96.95	90.91	11	83.76
Poaceae	Triticum urartu	10512	91.83	90.91	11	82.91
Poaceae	Brachypodium distachyon	12833	95.11	90.91	11	82.91
Poaceae	Zea mays	13021	96.84	90.91	11	82.05
Poaceae	Oryza sativa Japonica Group	13128	96.95	90.91	11	82.05
Poaceae	Sorghum bicolor	13314	94.88	90.91	11	82.05
Poaceae	Dichanthelium oligosanthes	11988	93.45	90.91	11	82.05
Poaceae	Panicum hallii	13708	92.08	90.91	11	82.05
Poaceae	Setaria italica	13190	93.49	90.91	11	82.05
Poaceae	Aegilops tauschii	15967	80.14	81.82	11	78.63
Poaceae	Triticum aestivum	10932	59.80	45.45	11	9.40

iwgsc_refseqv1.0_LowConf_PROTEIN_2017Mar13.fa

User's data result:
KusakiDB predicts family and
calculate number of validated
OGs, median of OG within
same family and median of OG
within all species

KusakiDB OG Management:

Selecting Genes by Conservation Level

Users can select the parameters to compare OGs among all species in KusakiDB, such as:

- Evidence existence
- Number of species in each family
- Percentage of species in each family
- Percentage within all species

Pop-up

Evidence existence

☒ Yes

☐ No

Evidence existence

Yes: existence of at least one protein, in an OG, at a transcript or protein level

KusakiDB v1.0

A novel approach for assessing existence and completeness of orthogroups in plant species

Search KusakiDB

OG Assessment

OG Management

OG Management User Data

Help

Manage parameters to assess conservation of orthogroups within complete genomes

Evidence existence

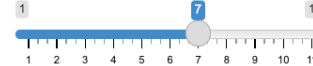
☒ Yes

☐ No

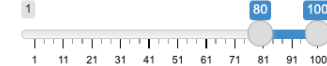
Submit

Download

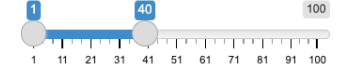
Total number of species in a family



Percentage of Species in a Family



Percentage of Total Species



List of Species with selected OG parameters

Show 10 entries

Search:

Scientific_name

Freq

Aegilops tauschii	1161
Sorghum bicolor	1156
Oryza sativa Japonica Group	1150
Panicum hallii	1149
Setaria italica	1144
Camelina sativa	1124
Brachypodium distachyon	1120
Arabidopsis lyrata subsp. lyrata	1118

List of Protein names with selected OG parameters

Show 10 entries

Search:

OrthoDB_Group

OrthoDB_Protein_Name

Freq

146849at33090	BTB/POZ domain-containing protein	27
226059at33090	Bowman-Birk type proteinase inhibitor	21
159542at33090	Protein of unknown function DUF1644	20
213047at33090	X8 domain	20
68458at33090	Proton-dependent oligopeptide transporter family	20
100806at33090	Amino acid transporter, transmembrane domain	19
157572at33090	Myb/SANT-like domain	19

User can download the results

The results with the selected parameters are shown in two table:

- List of species and number of OGs
- List of protein names and correspondent frequency

KusakiDB OG Management User's Data: Selecting genes from your data by conservation level

Users can select the parameters to compare OGs among all species in KusakiDB, such as:

- Evidence existence
- Number of species in each family
- Percentage of species in each family
- Percentage within all species

Users can upload the functional annotation performed by Hayai-annotation v2.0

The results show a “filter” of the genes that are selected under the conditions regarding the conservation level of each OG.

KusakiDB v1.0

A novel approach for assessing existence and completeness of orthogroups in plant species

Search KusakiDB OG Assessment OG Management **OG Management User Data** Help

Analyse conservation of orthogroups within complete genomes using your own data

Evidence existence
☒ Yes
☐ No

Total number of species in a family
Slider: 1 to 11, value set to 5

Percentage of Species in a Family
Slider: 1 to 100, value set to 50

Percentage of Total Species
Slider: 1 to 100, value set to 40

[Click here to create Hayai-annotation file](#)
Upload Hayai_annotation_v2.0.tsv
Browse... Hayai_annotation_v2.0.tsv
Upload complete
Submit

[Download](#)

Filter your results
Search:

Column visibility

Query	OrthoDB_Group	OrthoDB_Protein_Name	Accession	Accession_Protein_Name
TraesCS3D02G161800.1	116030at33090	Protein EMBRYONIC FLOWER 1	A0A3B6GR89	Uncharacterized Protein
TraesCS3A02G154500.1	116030at33090	Protein EMBRYONIC FLOWER 1	A0A3B6EEK0	Uncharacterized Protein
TraesCS3B02G180800.1	116030at33090	Protein EMBRYONIC FLOWER 1	A0A3B6FNM0	Uncharacterized Protein
TraesCS3B02G015800.1	198713at33090	Phosphatidylethanolamine-binding protein	A0A0B4SVQ0	Flowering locus T-like protein
TraesCS3B02G015200.1	198713at33090	Phosphatidylethanolamine-binding protein	D8LAK8	Flowering Locus T-like protein, putative
TraesCS4A02G408400.1.cds1	220345at33090	Flowering-promoting factor 1	A0A446RGD7	Uncharacterized Protein
TraesCS4B02G308800.1.cds1	220345at33090	Flowering-promoting factor 1	A0A3B6IVR2	Uncharacterized Protein

Hayai-Annotation Plants v2: Full Annotation Engine

Hayai_Annot

KusakiDB

Ribosomal

Ribosomal

Ribosomal

Ribosomal

kusakiDB/

Evolution

Evidence

Frontiers

+

Not Secure pgdbjsnp.kazusa.or.jp/app/hayai2

Update

Hayai-Annotation Plants v2.0 - Functional Annotation System Specialized in Plant Species

Database: KusakiDB v1.0
Reference: Hayai-Annotation Plants

Hayai Annotation

Search Proteins

Perform Full Annotation

Query Sequence Type

☒ Protein

☐ DNA

Minimum Sequence Identity (%)

20

50

100

20 28 36 44 52 60 68 76 84 92 100

Minimum Query Coverage (%)

20

75

100

20 28 36 44 52 60 68 76 84 92 100

Maximum E-value

1

6

100

1 11 21 31 41 51 61 71 81 91 100

Minimum Target Coverage (%)

20

75

100

20 28 36 44 52 60 68 76 84 92 100

[Sample protein fasta file](#)

Upload FASTA File

Browse...

Araport11_genes.201606.pep.

Upload complete

Submit

?

[Download](#)

Column visibility

Search: flower

Query	OrthoDB_Group	OrthoDB_Protein_Name	Accession	Accession_Protein_Name
AT5G03840.1	174443at33090	Phosphatidylethanolamine-binding protein	P93003	Protein TERMINAL FLOWER 1
AT5G10140.1	182681at33090	Transcription factor, MADS-box	Q9S7Q7	MADS-box protein FLOWERING LOCUS C
AT1G17455.1	222987at33090	Protein EARLY FLOWERING 4 domain	R0I5C1	Elf4 domain-containing protein
AT4G31380.1	220345at33090	Flowering-promoting factor 1	Q5Q0B3	Flowering-promoting factor 1-like protein 1
AT5G48890.1	197484at33090	Zinc finger C2H2 superfamily	Q9FKA9	Protein LATE FLOWERING
AT1G72390.1	11593at33090	histone-lysine N-methyltransferase 2D isoform X1	F4IDB2	Protein PHYTOCHROME-DEPENDENT LATE-FLOWERING
AT3G04610.1	108737at33090	K Homology domain, type 1	Q9SR13	Flowering locus K homology domain

Hayai-Annotation Plants v2: Search Proteins Engine

Database: KusakiDB v1.0
Reference: Hayai-Annotation Plants

Hayai Annotation **Search Proteins**

Search proteins within output files of Hayai-Annotation

Upload file 'output_HayaiAnnotation.zip'
Browse... output_HayaiAnnotation (1).zip
Upload complete

Choose another source of annotation
Interpro

Show results ? Submit ?

Column visibility Search: flower

Query	OrthoDB_Group	OrthoDB_Protein_Name	Accession	Accession_Protein_Name	Species	Family
MD12G1262000	193016at33090	Phosphatidylethanolamine-binding protein	Q75QW8	Flowering locus T	Malus domestica	Rosaceae
MD02G1231000	185288at33090	Chromo domain	A7M6G2	Terminal flower 2 protein	Malus domestica	Rosaceae
MD07G1081800	185288at33090	Chromo domain	A7M6G0	Terminal flower 2 protein	Malus domestica	Rosaceae
MD08G1116300	116030at33090	Protein EMBRYONIC FLOWER 1	A0A498JAP0	Uncharacterized Protein	Malus domestica	Rosaceae
MD01G1152700	228137at33090	Protein EARLY FLOWERING 4 domain	A0A540KP36	Uncharacterized Protein	Malus baccata	Rosaceae
MD11G1074300	228137at33090	Protein EARLY FLOWERING 4 domain	A0A540KER1	Uncharacterized Protein	Malus baccata	Rosaceae
MD02G1087400	222987at33090	Protein EARLY FLOWERING 4 domain	A0A498KFF3	Elf4 domain-containing protein	Malus domestica	Rosaceae
MD15G1109300	220345at33090	Flowering-promoting factor 1	A0A498HPA5	Uncharacterized Protein	Malus domestica	Rosaceae
MD15G1096000	116030at33090	Protein EMBRYONIC FLOWER 1	A0A540M159	Uncharacterized Protein	Malus baccata	Rosaceae
MD15G1062000	24394at33090	Protein EARLY FLOWERING 3-like	A0A498HUT8	Uncharacterized Protein	Malus domestica	Rosaceae

Showing 1 to 10 of 17 entries (filtered from 33,370 total entries)

Previous 1 2 Next

Show 10 entries Search: MD11G1074300

Query	InterPro_Accession	InterPro_name
MD11G1074300	IPR009741	EARLY_FLOWERING_4_dom
MD11G1074300	IPR040462	EARLY_FLOWERING_4

Showing 1 to 2 of 2 entries (filtered from 102,808 total entries)

Previous 1 Next

Choose another source of annotation

Interpro

Interpro

Pfam

GO Molecular Function

GO Biological Process

GO Cellular Component

Software Availability

- KusakidB
 - <http://pgdbjsnp.kazusa.or.jp/app/kusakidb>
 - <https://github.com/aghelfi/kusakiDB>
 - <https://hub.docker.com/r/ghelfi/kusakidb>
- Hayai-annotation
 - <http://pgdbjsnp.kazusa.or.jp/app/hayai2>
 - https://hub.docker.com/r/ghelfi/hayai_annotation_v2