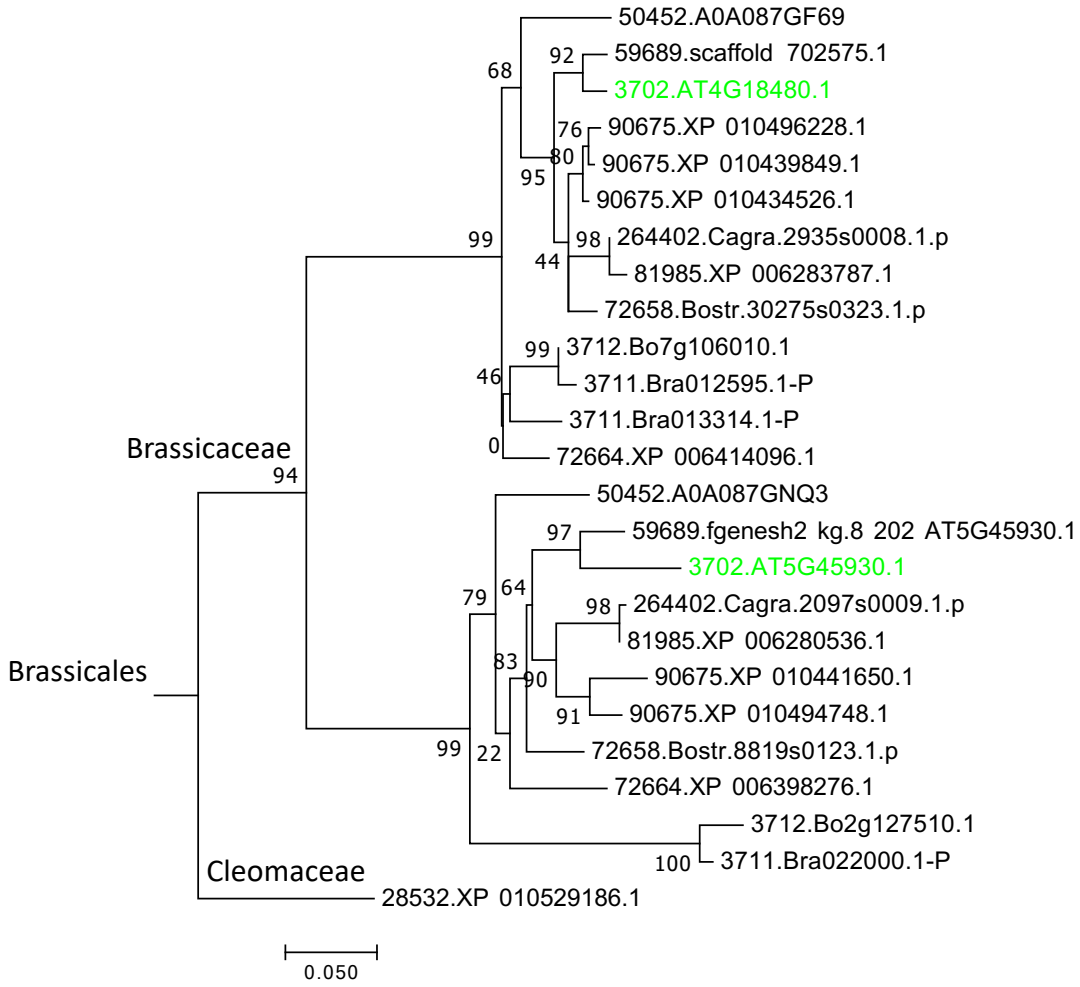
KusakiDB and Hayai-Annotation

Andrea Ghelfi

Kazusa DNA Research Institute

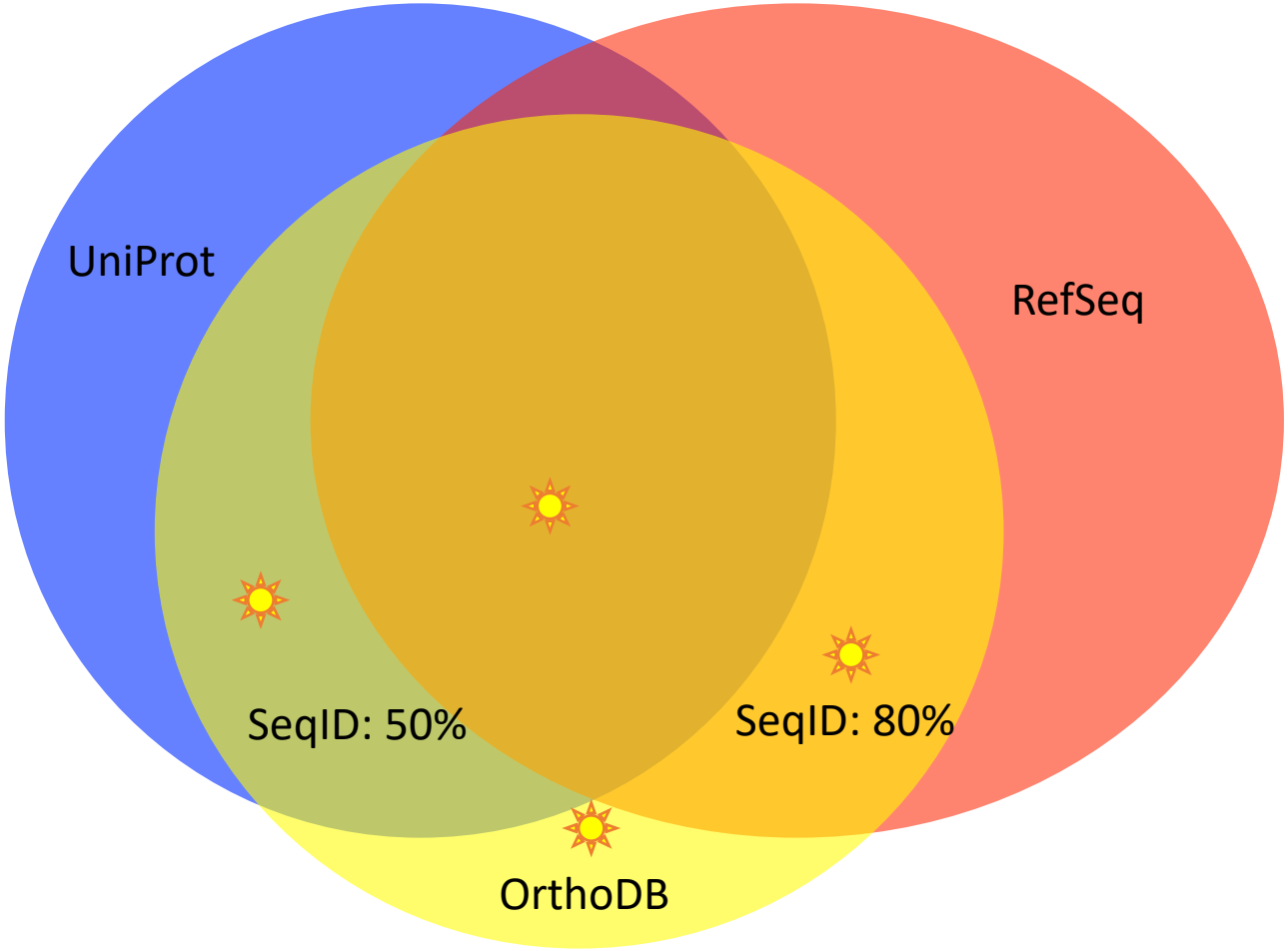
Definition of orthogroups

- 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.



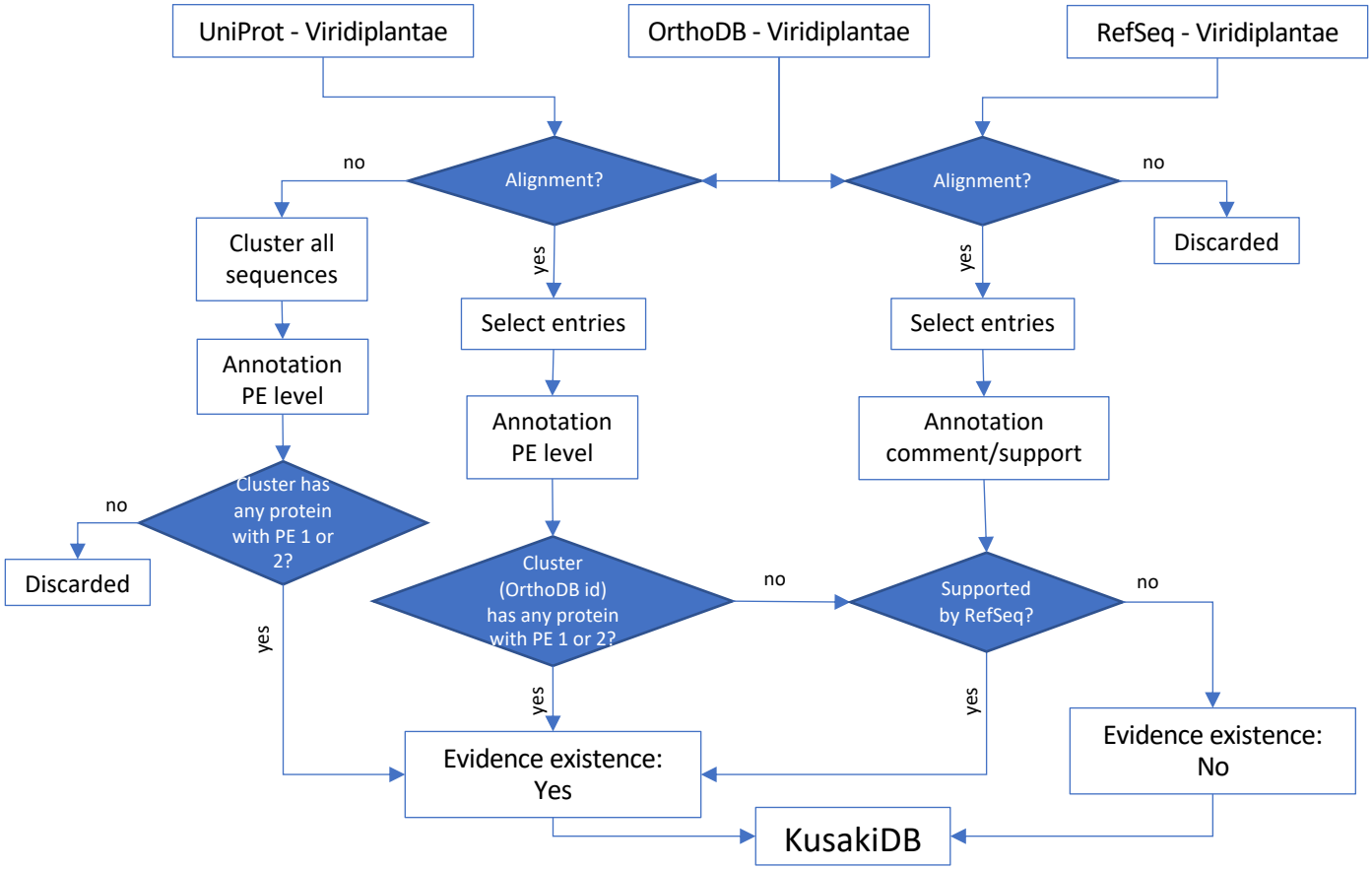
Magnesium-chelatase subunit ChlI-2, chloroplastic

KusakiDB: Intersections of UniProt, RefSeq and OrthoDB

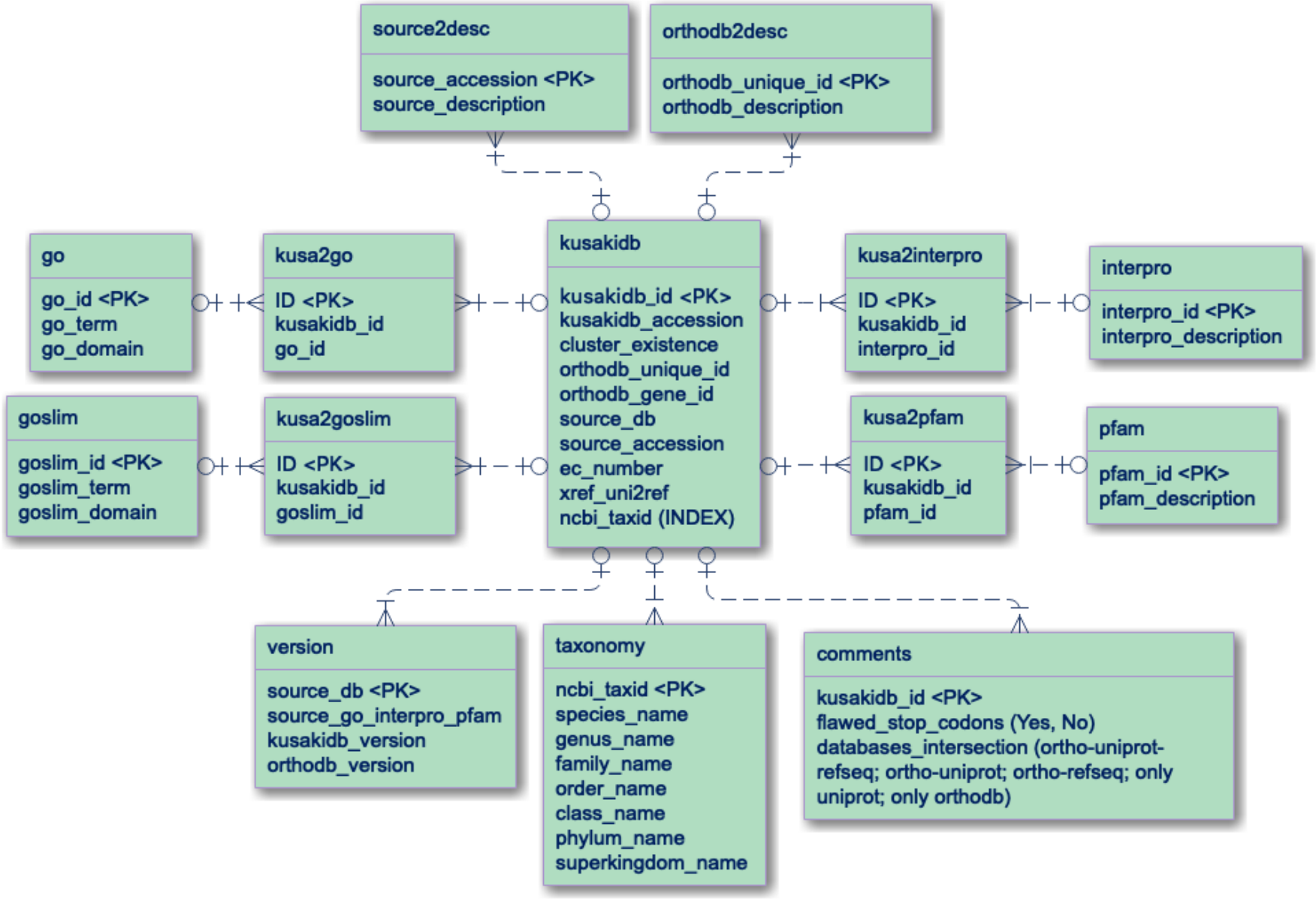


Clustered to 99% sequence identity to remove redunduncies

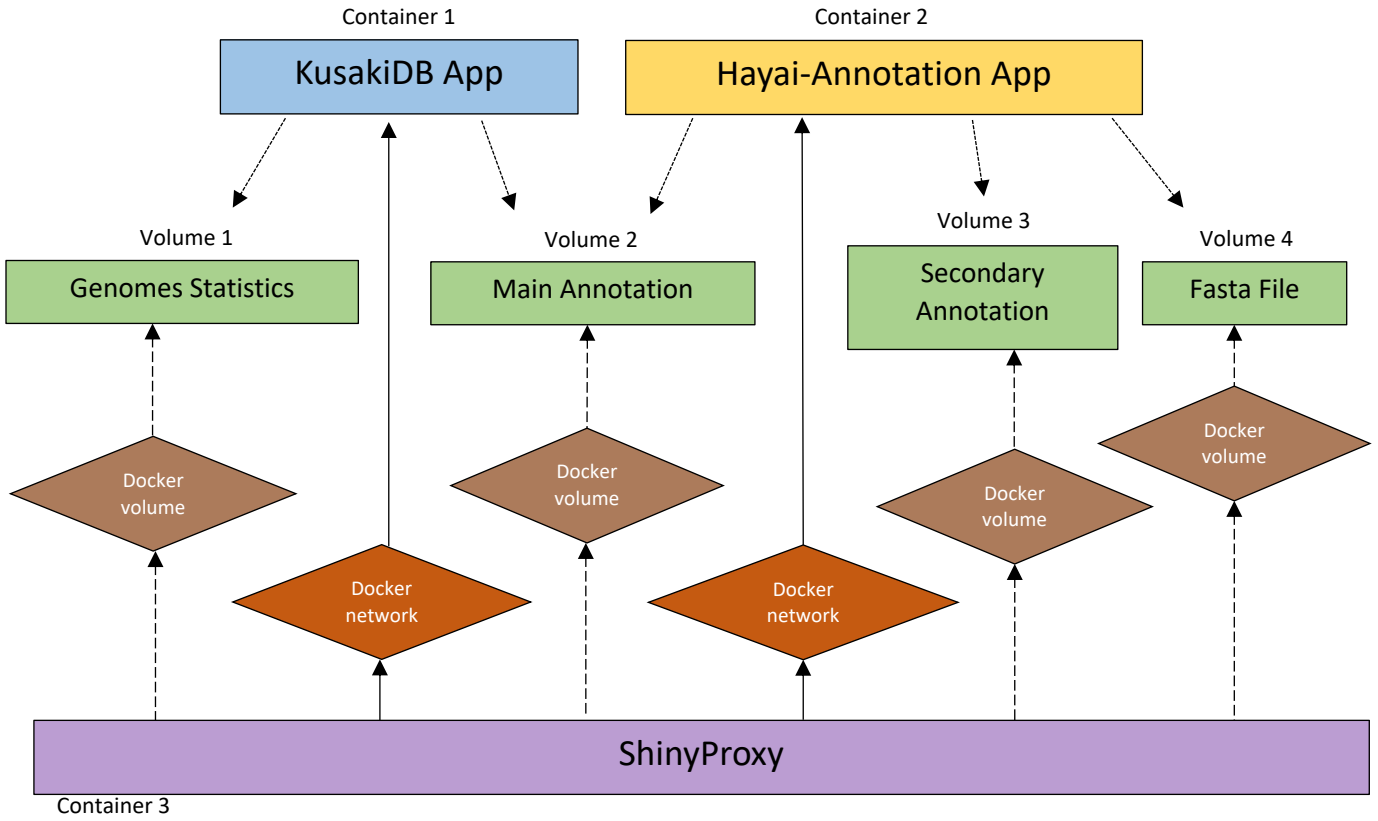
KusakiDB: a novel approach for assessing the existence and completeness of orthogroups in plant species



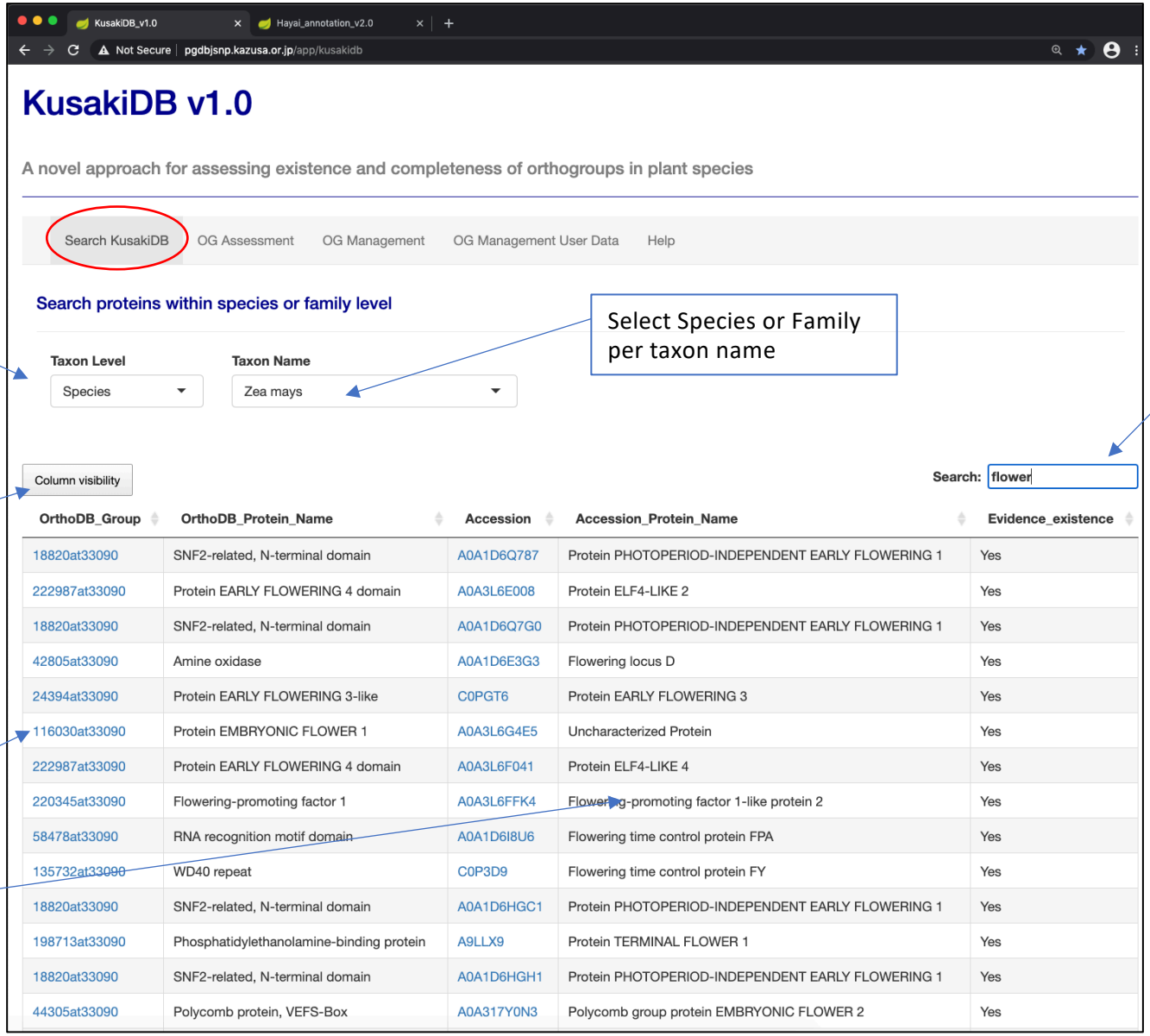
KusakiDB: MySQL Schema



KusakiDB and Hayai-Annotation: Docker Container Structure

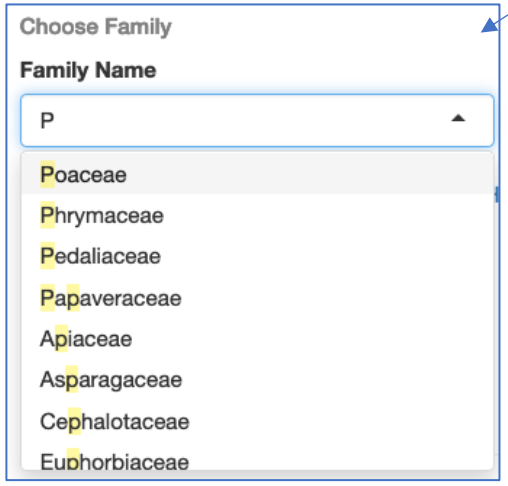


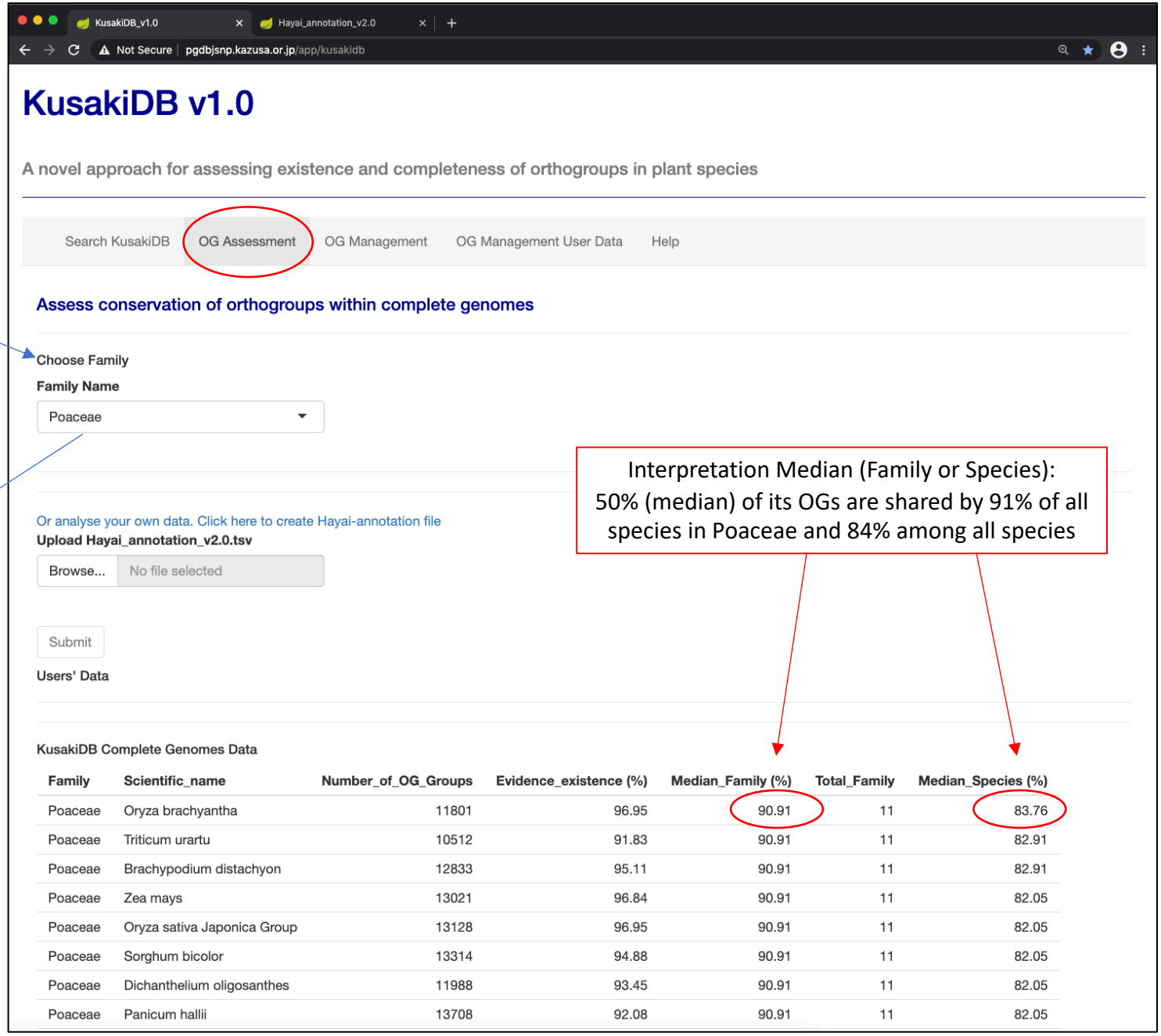
KusakiDB: A novel approach for existence and completeness of protein orthologs



KusakiDB Tools: OG Assessment

Select a Family from KusakiDB



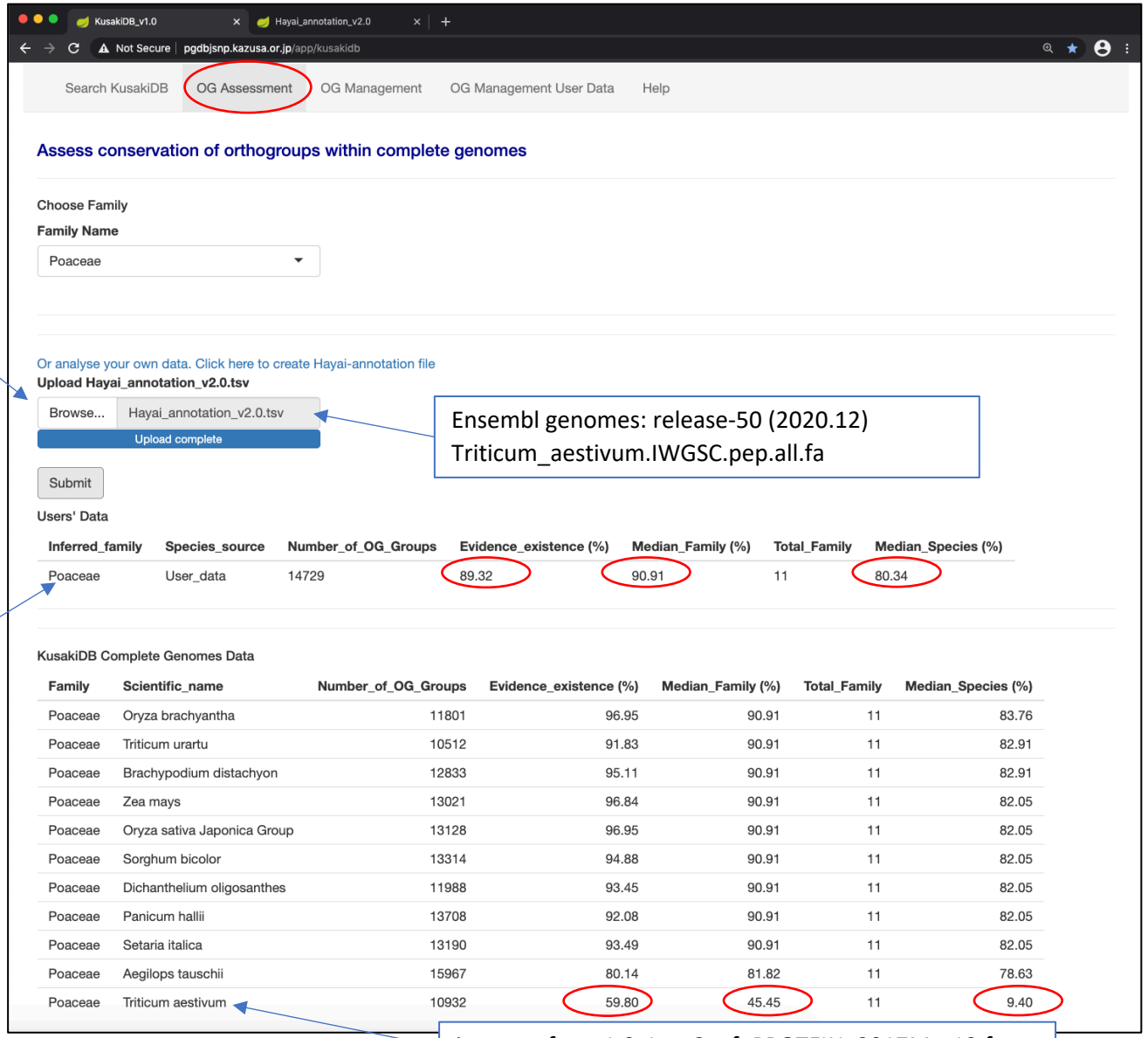


Family	Scientific_name	Number_of_OG_Groups	Evidence_existence (%)	Median_Family (%)	Total_Family	Median_Species (%)
Poaceae	Oryza brachyantha	11801	96.95	90.91	11	83.76
Poaceae	Triticum urartu	10512	91.83	90.91	11	82.91
Poaceae	Brachypodium distachyon	12833	95.11	90.91	11	82.91
Poaceae	Zea mays	13021	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

KusakiDB OG Assessment: Evaluating Gene Structural Prediction

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



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

iwgsc_refseqv1.0_LowConf_PROTEIN_2017Mar13.fa

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

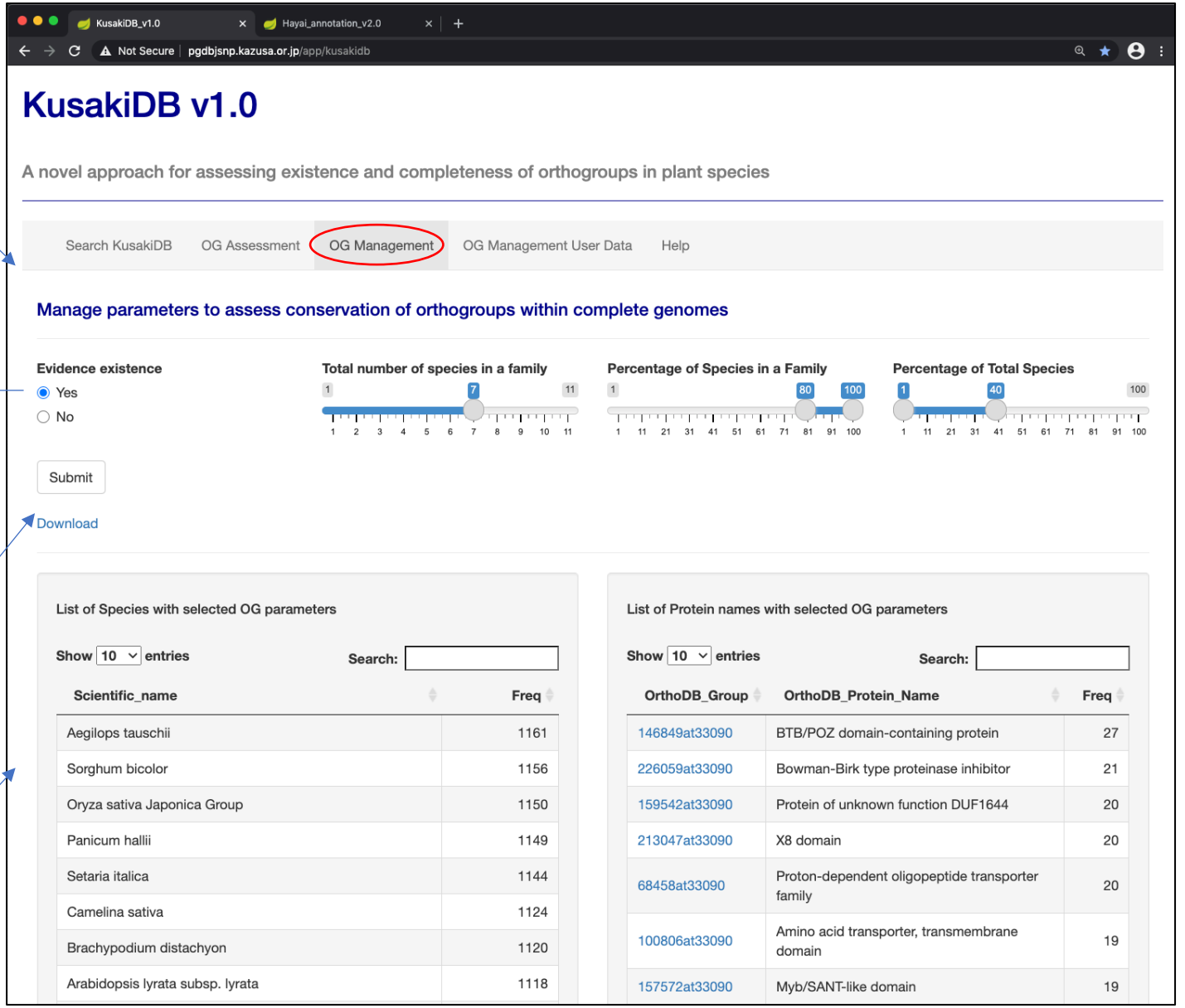
Evidence existence

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

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



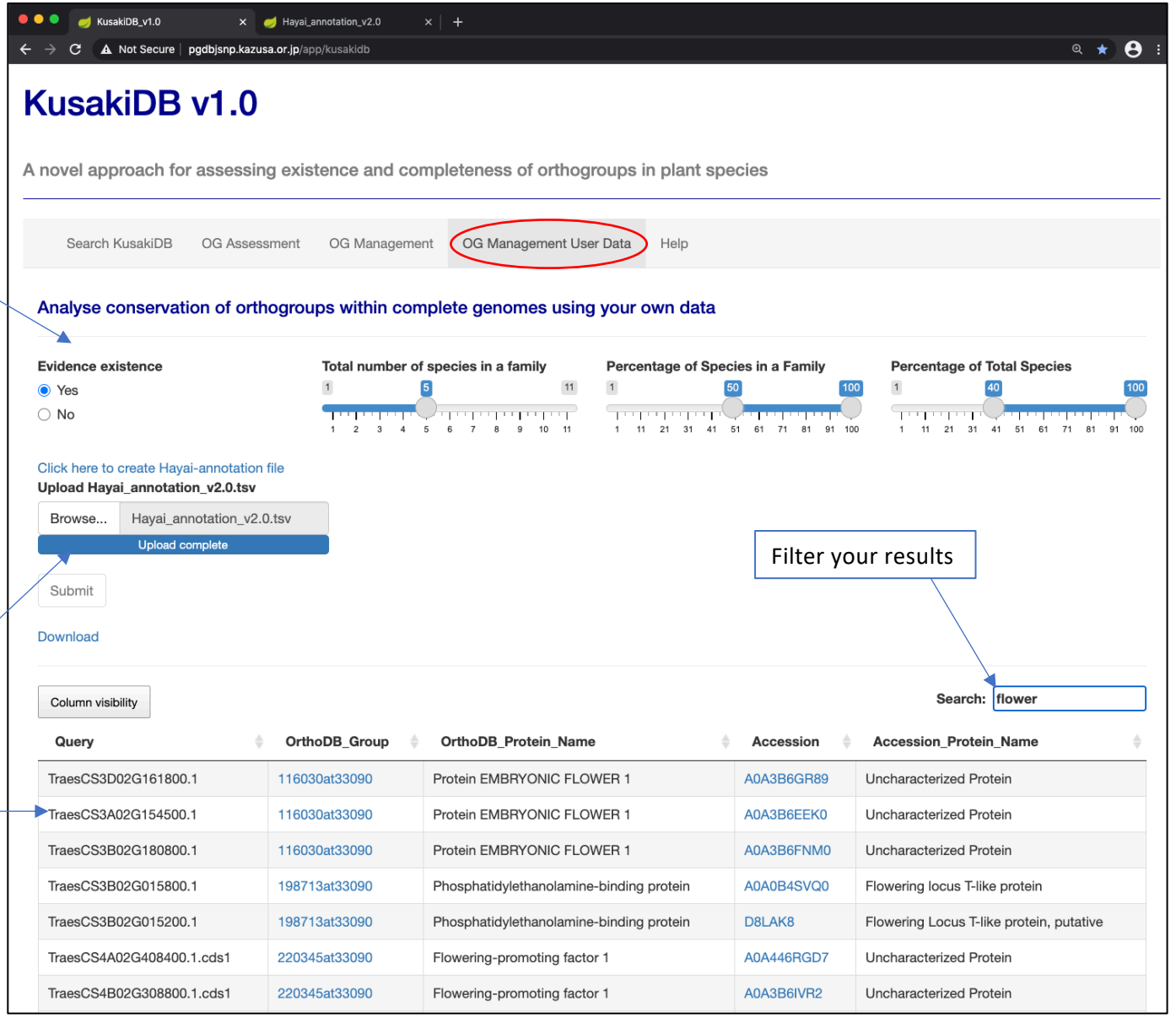
KusakiDB OG Management User's Data: Selecting genes from your data by conservation level

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

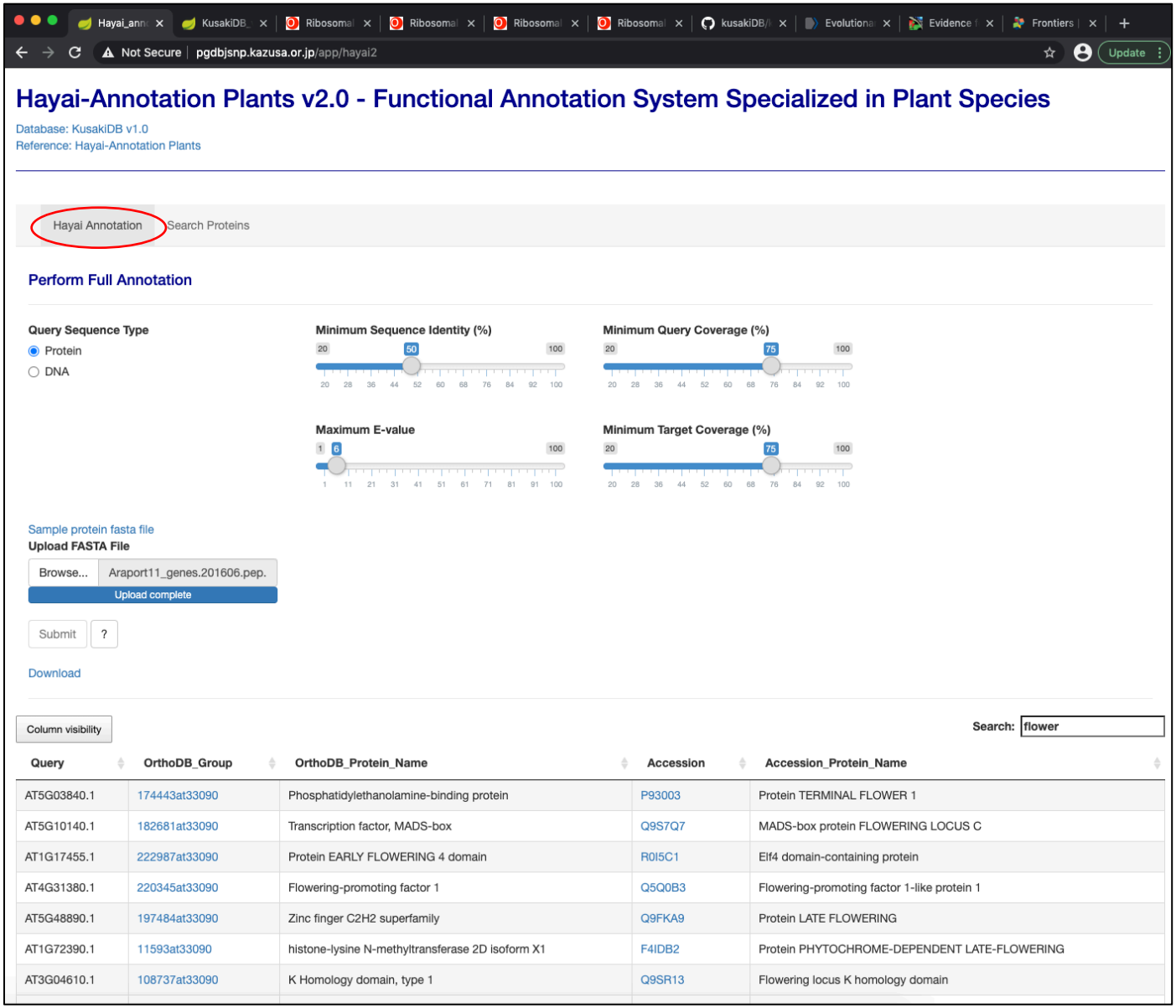
- Evidence existence
- Number of species in each family
- Percentage of species in each family
- Percentage within all species

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.

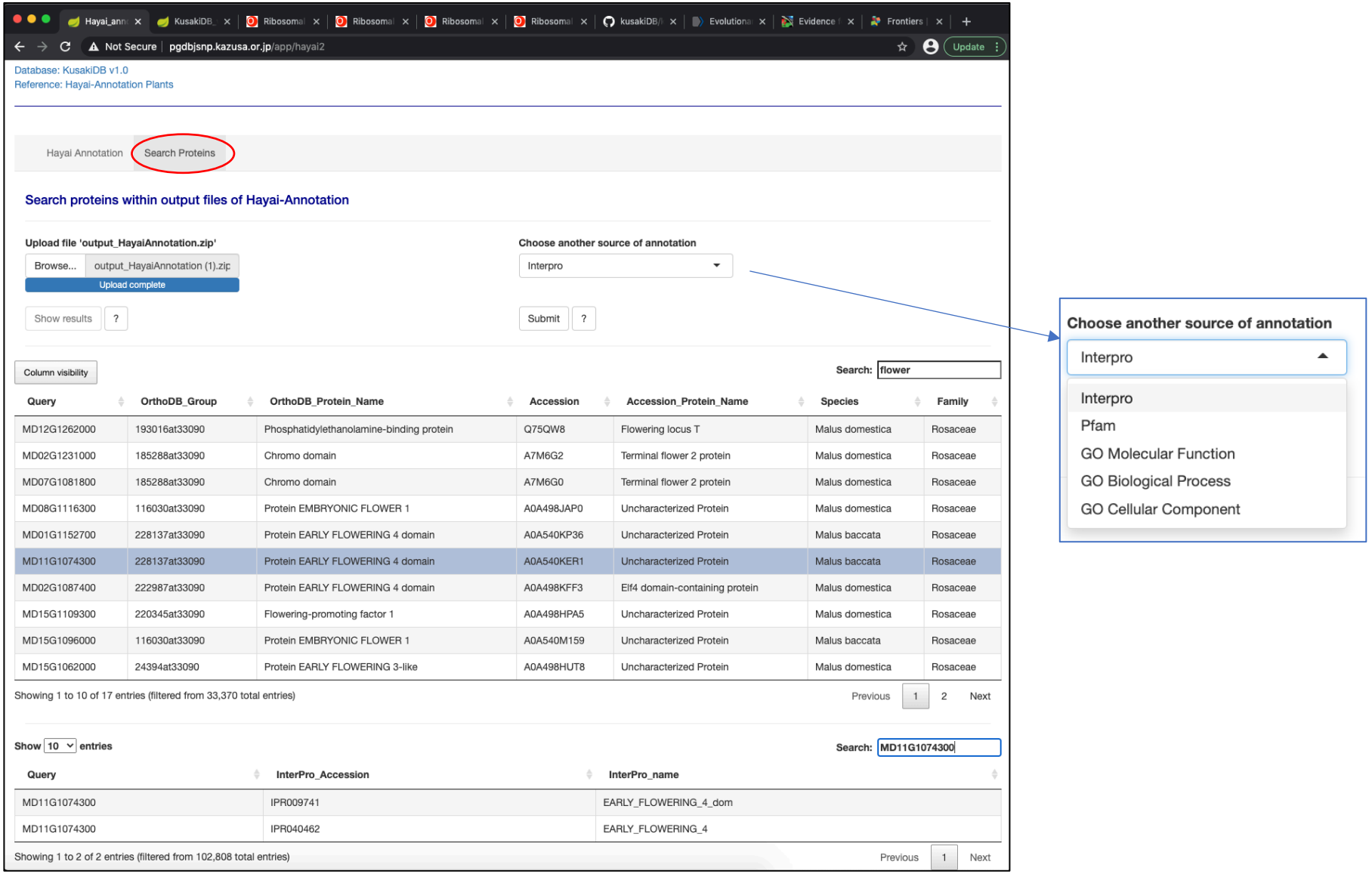


Hayai-Annotation Plants v2: Full Annotation Engine



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



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