44

KusakiDB v1.0: New protein database of orthologous genes in plant species

トーゴーの日シンポジウム2020

〇ゲルフィ アンドレア[1], 中村保一[2], 磯部祥子[1]

1. かずさDNA研究所、2. 国立遺伝学研究所

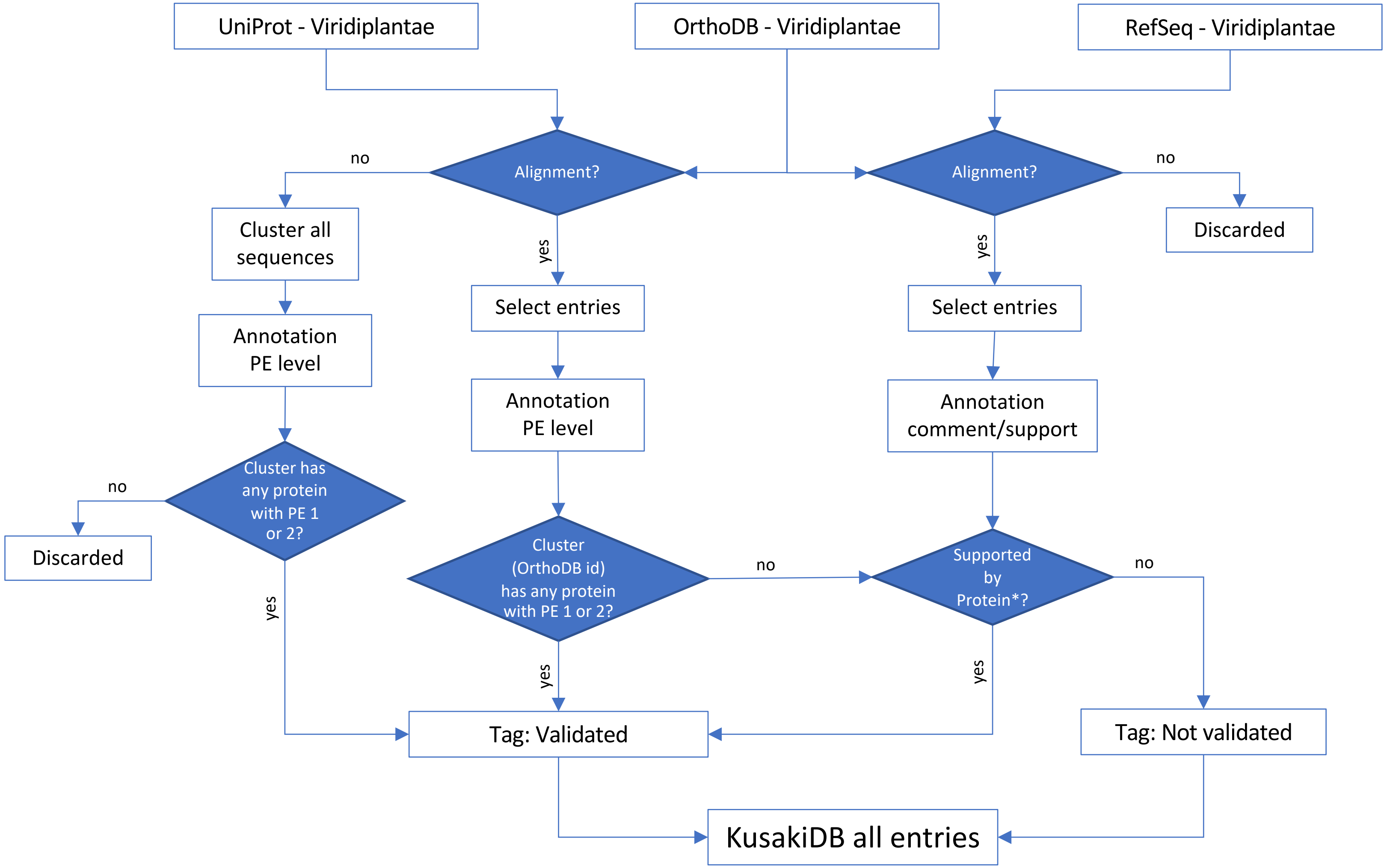
要旨 (Abstract)

KusakiDB is a database of protein orthologous groups (OGs) that provides an assessment and management tool for comparison of OGs in plant species. KusakiDB correlates the information of three important databases, OrthoDB, UniProt and RefSeq. It introduces a validation tag that is based on the existence of at least one protein in each OG which is an attempt to address the problem of error propagation. KusakiDB was used as a database to re-annotate plant gene sequences registered in Plant GARDEN (https://plantgarden.jp) by using Hayai-Annotation (Ghelfi et al., 2019).

背景 (Background)

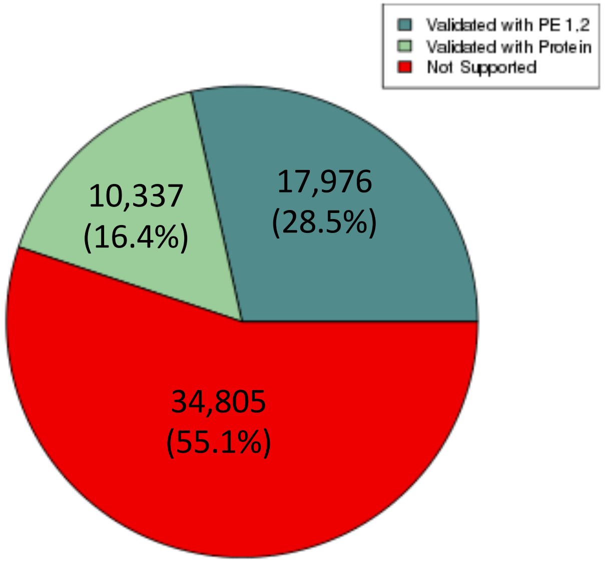
- Plants have quite a low coverage in the major protein databases despite their large number of species (roughly 350,000), and environmental and economic impacts. To explain the latter, two important global meetings can be cited, first the Paris Agreement (2015), which established a process toward stabilizing greenhouse gas concentrations; second, the Global Bioeconomy Summit (Berlin, 2018), that identified Bioeconomy as a transformative strategy for advancing a Sustainable Development Goals.
- Furthermore, agricultural sector is one of the main industrial sectors in bioeconomy (FAO, 2016). In the European Union and US together biology-based industries (non-food) accounted for 21 million jobs and generated more than US$ 2.57 trillion annually (El-Chichakli et al., 2016).
- Besides, misannotation of molecular function in public databases continues to be a significant problem (Schnoes et al., 2009).
- We have developed KusakiDB, which provides a validation tag of orthologous groups, besides presenting an assessment and management tools to evaluate orthologous groups at a family level in 117 plant species. This is an attempt to enrich information regarding physical evidence of a protein or transcript at each OG and implement different methods of comparison of OG within plant species.

KusakiDB v1.0: Overview of the Integration of Three Major Databases



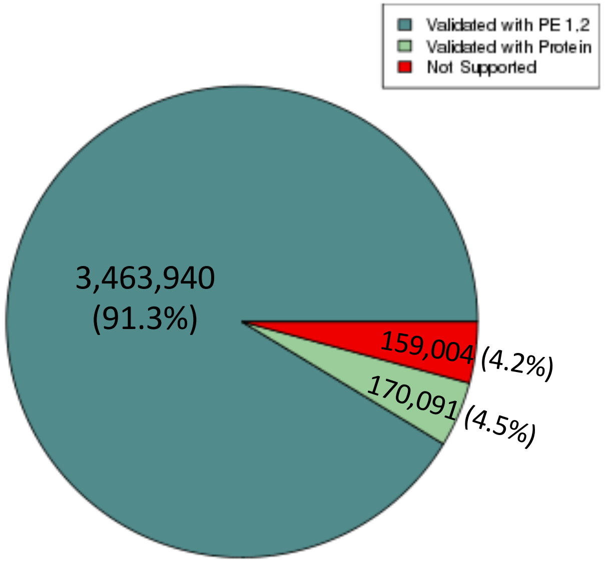
KusakiDB v1.0: Distribution of Entries by Validation Tag (and source of validation)

KusakiDB by OrthoDB unique OG ID



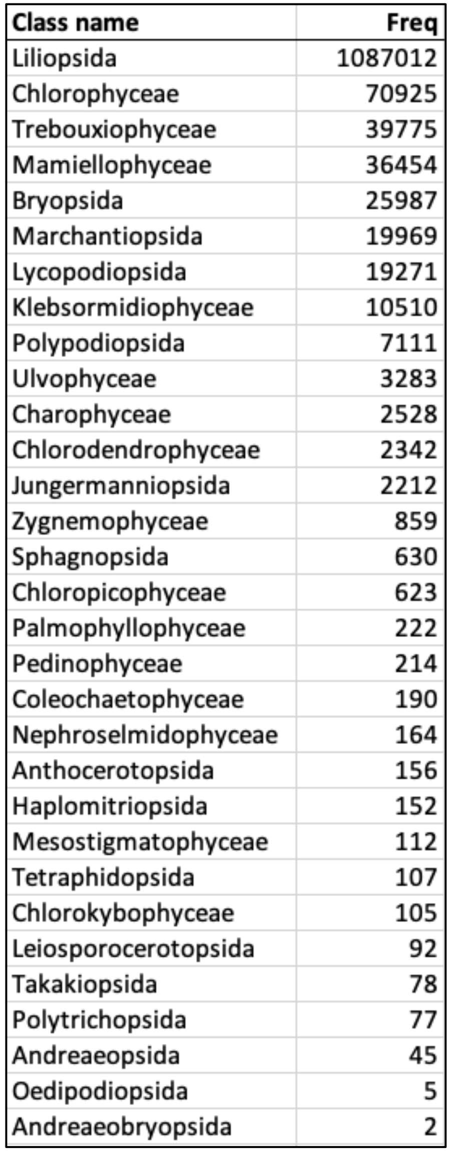
Total: 63,118 clusters

KusakiDB all sequences



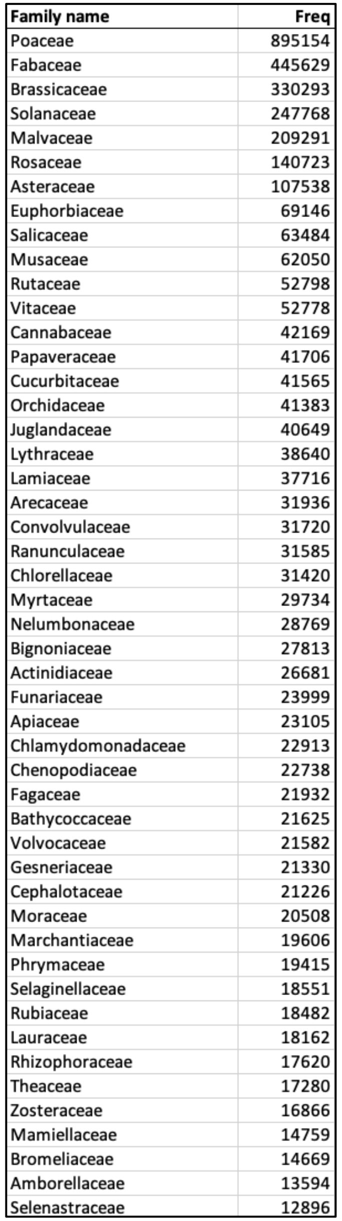
Total: 3,793,035 protein sequences

KusakiDB Distribution of Unique Genes by Taxonomic Level



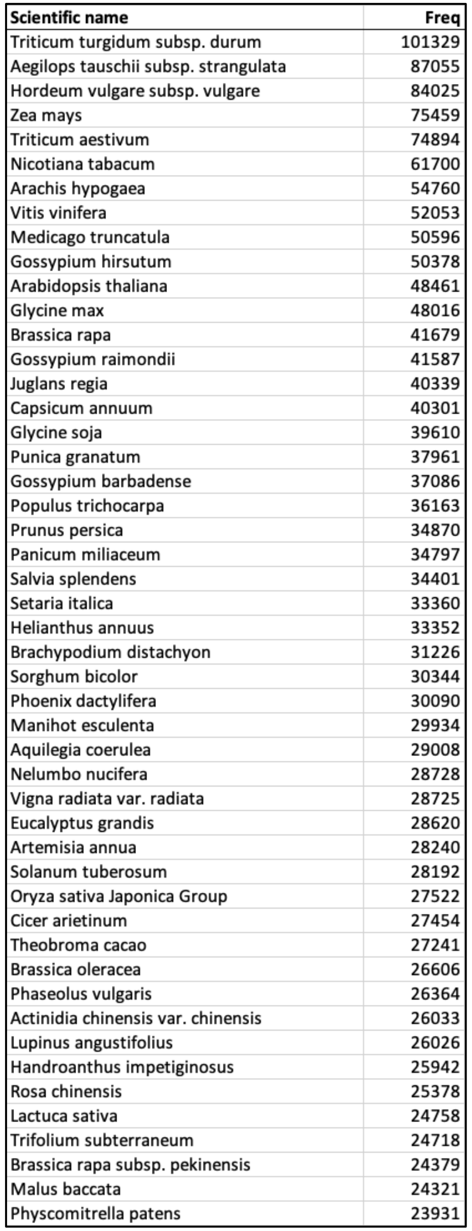
Class name	Freq
Liliopsida	1087012
Chlorophyceae	70925
Trebouxiophyceae	39775
Mamiellophyceae	36454
Bryopsida	25987
Marchantiopsida	19969
Lycopodiopsida	19271
Klebsormidiophyceae	10510
Polypodiopsida	7111
Ulvophyceae	3283
Charophyceae	2528
Chlorodendrophyceae	2342
Jungermanniopsida	2212
Zygnemophyceae	859
Sphagnopsida	630
Chloropicophyceae	623
Palmophyllophyceae	222
Pedinophyceae	214
Coleochaetophyceae	190
Nephroselmidophyceae	164
Anthocerotopsida	156
Haplomitriopsida	152
Mesostigmatophyceae	112
Tetraphidopsida	107
Chlorokybophyceae	105
Leiosporocerotopsida	92
Takakiopsida	78
Polytrichopsida	77
Andreaeopsida	45
Oedipodiopsida	5
Andreaeobryopsida	2

Total: 31 classes



Family name	Freq
Poaceae	895154
Fabaceae	445629
Brassicaceae	330293
Solanaceae	247768
Malvaceae	209291
Rosaceae	140723
Asteraceae	107538
Euphorbiaceae	69146
Salicaceae	63484
Musaceae	62050
Rutaceae	52798
Vitaceae	52778
Cannabaceae	42169
Papaveraceae	41706
Cucurbitaceae	41565
Orchidaceae	41383
Juglandaceae	40649
Lythraceae	38640
Lamiaceae	37716
Arecaceae	31936
Convolvulaceae	31720
Ranunculaceae	31585
Chlorellaceae	31420
Myrtaceae	29734
Nelumbonaceae	28769
Bignoniaceae	27813
Actinidiaceae	26681
Funariaceae	23999
Apiaceae	23105
Chlamydomonadaceae	22913
Chenopodiaceae	22738
Fagaceae	21932
Bathycoccaceae	21625
Volvocaceae	21582
Gesneriaceae	21330
Cephalotaceae	21226
Moraceae	20508
Marchantiaceae	19606
Phrymaceae	19415
Selaginellaceae	18551
Rubiaceae	18482
Lauraceae	18162
Rhizophoraceae	17620
Theaceae	17280
Zosteraceae	16866
Mamiellaceae	14759
Bromeliaceae	14669
Amborellaceae	13594
Selenastraceae	12896

Total: 708 families



Scientific name	Freq
Triticum turgidum subsp. durum	101329
Aegilops tauschii subsp. strangulata	87055
Hordeum vulgare subsp. vulgare	84025
Zea mays	75459
Triticum aestivum	74894
Nicotiana tabacum	61700
Arachis hypogaea	54760
Vitis vinifera	52053
Medicago truncatula	50596
Gossypium hirsutum	50378
Arabidopsis thaliana	48461
Glycine max	48016
Brassica rapa	41679
Gossypium raimondii	41587
Juglans regia	40339
Capsicum annuum	40301
Glycine soja	39610
Punica granatum	37961
Gossypium barbadense	37086
Populus trichocarpa	36163
Prunus persica	34870
Panicum miliaceum	34797
Salvia splendens	34401
Setaria italica	33360
Helianthus annuus	33352
Brachypodium distachyon	31226
Sorghum bicolor	30344
Phoenix dactylifera	30090
Manihot esculenta	29934
Aquilegia coerulea	29008
Nelumbo nucifera	28728
Vigna radiata var. radiata	28725
Eucalyptus grandis	28620
Artemisia annua	28240
Solanum tuberosum	28192
Oryza sativa Japonica Group	27522
Cicer arietinum	27454
Theobroma cacao	27241
Brassica oleracea	26606
Phaseolus vulgaris	26364
Actinidia chinensis var. chinensis	26033
Lupinus angustifolius	26026
Handroanthus impetiginosus	25942
Rosa chinensis	25378
Lactuca sativa	24758
Trifolium subterraneum	24718
Brassica rapa subsp. pekinensis	24379
Malus baccata	24321
Physcomitrella patens	23931

Total: 21,617 species

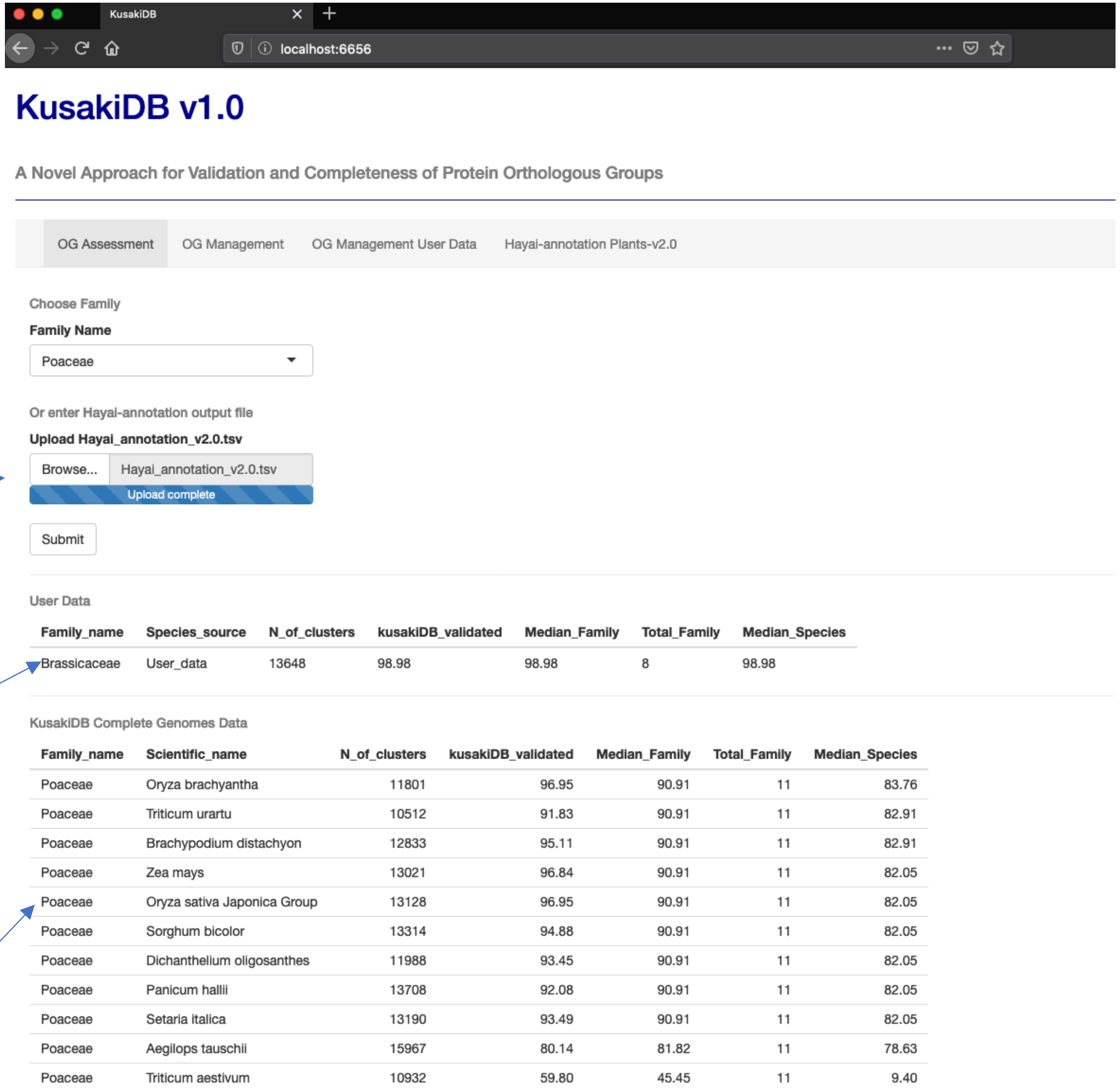
KusakiDB v1.0 Tools: OG assessment

Select a Family from KusakiDB

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

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

Results of KusakiDB data



KusakiDB v1.0

A Novel Approach for Validation and Completeness of Protein Orthologous Groups

Choose Family

Family Name: Poaceae

Or enter Hayai-annotation output file

Upload Hayai_annotation_v2.0.tsv

User Data

Family_name	Species_source	N_of_clusters	kusakiDB_validated	Median_Family	Total_Family	Median_Species
Brassicaceae	User_data	13648	98.98	98.98	8	98.98

KusakiDB Complete Genomes Data

Family_name	Scientific_name	N_of_clusters	kusakiDB_validated	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

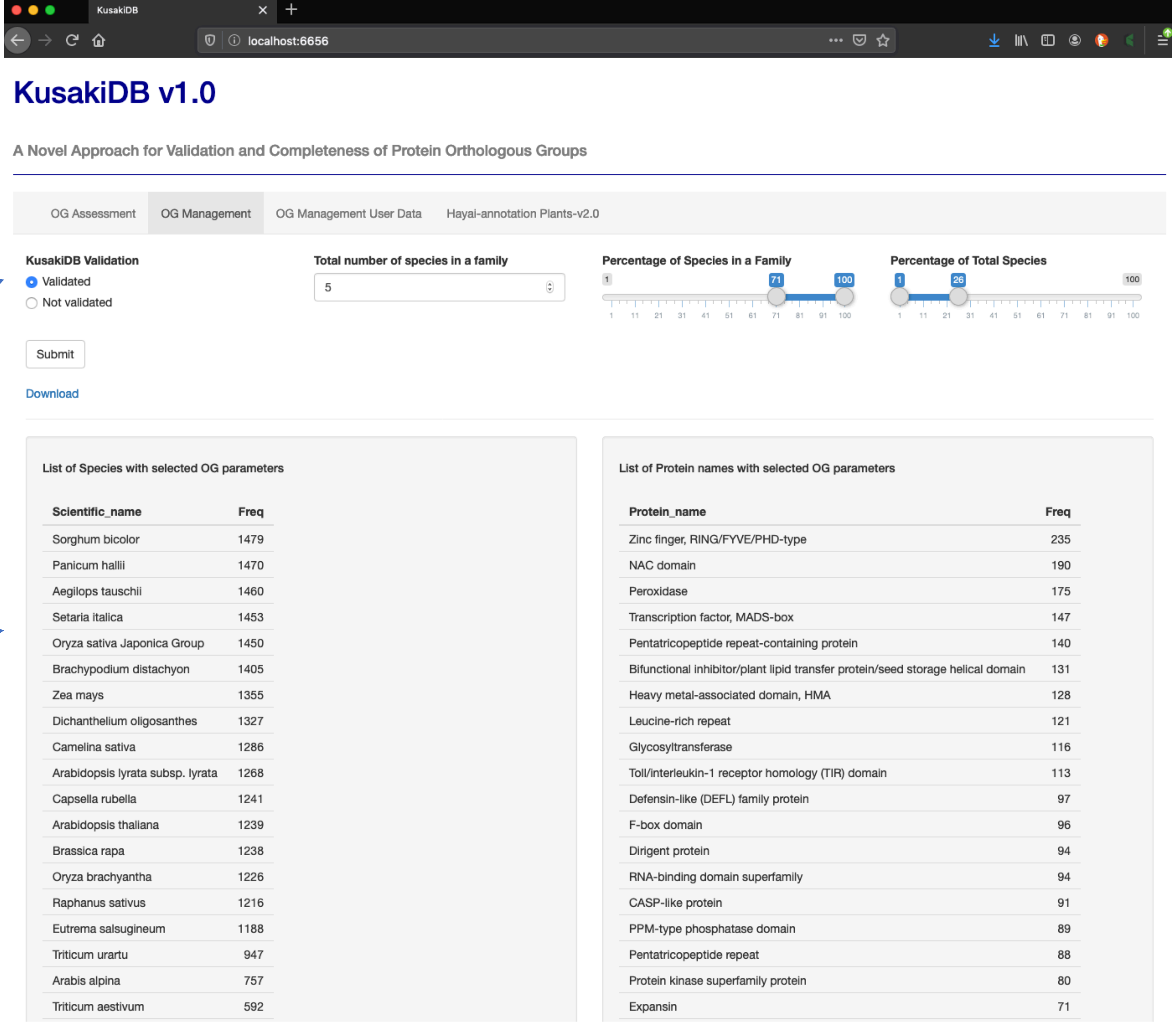
KusakiDB v1.0 Tools: OG management

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

- Validation tag
- Number of species in each family
- Percentage of species in each family
- Percentage within all species

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 v1.0

A Novel Approach for Validation and Completeness of Protein Orthologous Groups

KusakiDB Validation

● Validated

○ Not validated

Total number of species in a family: 5

Percentage of Species in a Family: 71 – 100

Percentage of Total Species: 1 – 26

List of Species with selected OG parameters

Scientific_name	Freq
Sorghum bicolor	1479
Panicum hallii	1470
Aegilops tauschii	1460
Setaria italica	1453
Oryza sativa Japonica Group	1450
Brachypodium distachyon	1405
Zea mays	1355
Dichanthelium oligosanthes	1327
Camelina sativa	1286
Arabidopsis lyrata subsp. lyrata	1268
Capsella rubella	1241
Arabidopsis thaliana	1239
Brassica rapa	1238
Oryza brachyantha	1226
Raphanus sativus	1216
Eutrema salsugineum	1188
Triticum urartu	947
Arabis alpina	757
Triticum aestivum	592

List of Protein names with selected OG parameters

Protein_name	Freq
Zinc finger, RING/FYVE/PHD-type	235
NAC domain	190
Peroxidase	175
Transcription factor, MADS-box	147
Pentatricopeptide repeat-containing protein	140
Bifunctional inhibitor/plant lipid transfer protein/seed storage helical domain	131
Heavy metal-associated domain, HMA	128
Leucine-rich repeat	121
Glycosyltransferase	116
Toll/interleukin-1 receptor homology (TIR) domain	113
Defensin-like (DEFL) family protein	97
F-box domain	96
Dirigent protein	94
RNA-binding domain superfamily	94
CASP-like protein	91
PPM-type phosphatase domain	89
Pentatricopeptide repeat	88
Protein kinase superfamily protein	80
Expansin	71

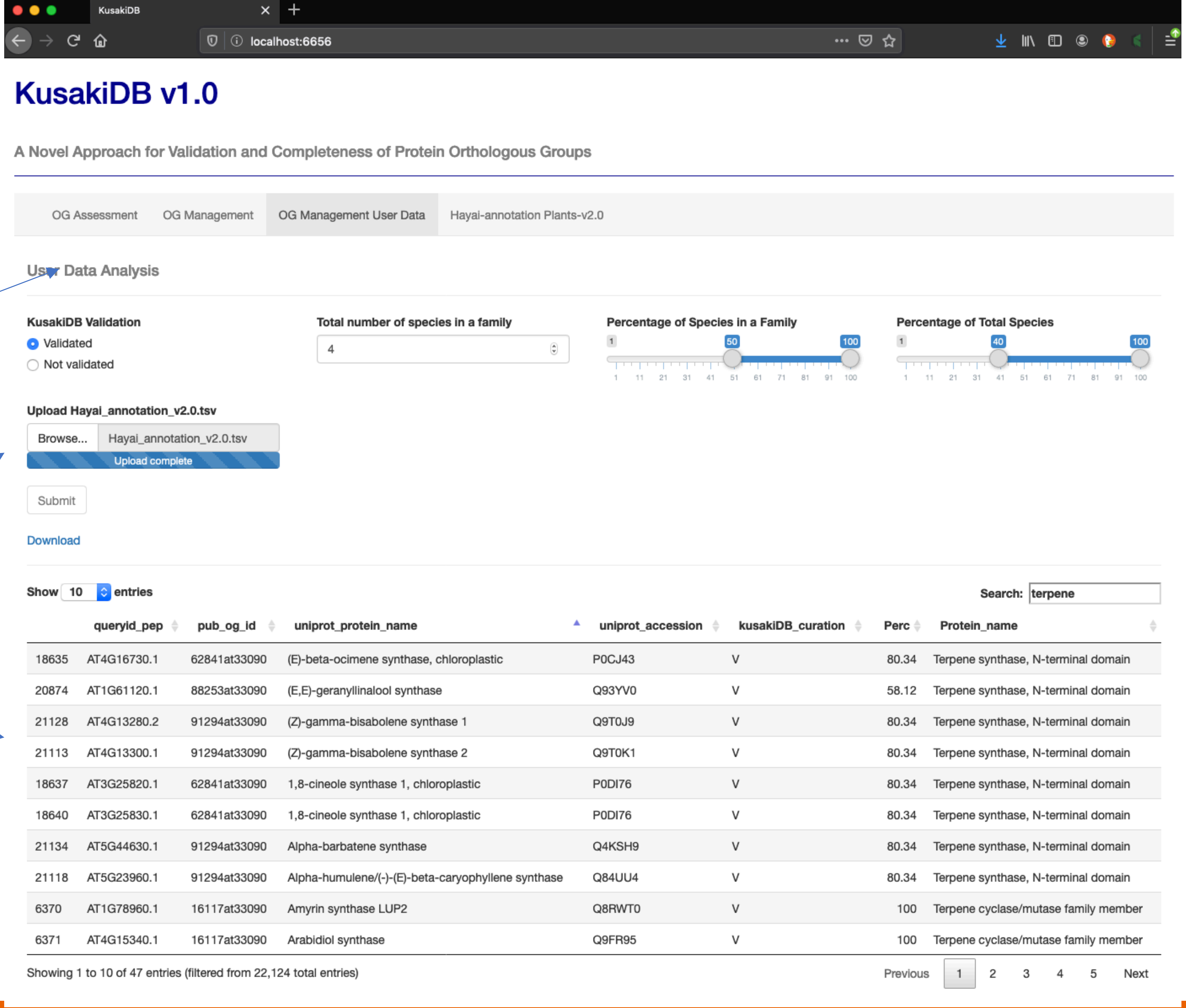
KusakiDB v1.0 Tools: OG management user data

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

- Validation tag
- 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 Validation and Completeness of Protein Orthologous Groups

User Data Analysis

KusakiDB Validation

● Validated

○ Not validated

Total number of species in a family: 4

Percentage of Species in a Family: 50 – 100

Percentage of Total Species: 40 – 100

Upload Hayai_annotation_v2.0.tsv

Browse... Hayai_annotation_v2.0.tsv

Upload complete

Show 10 entries

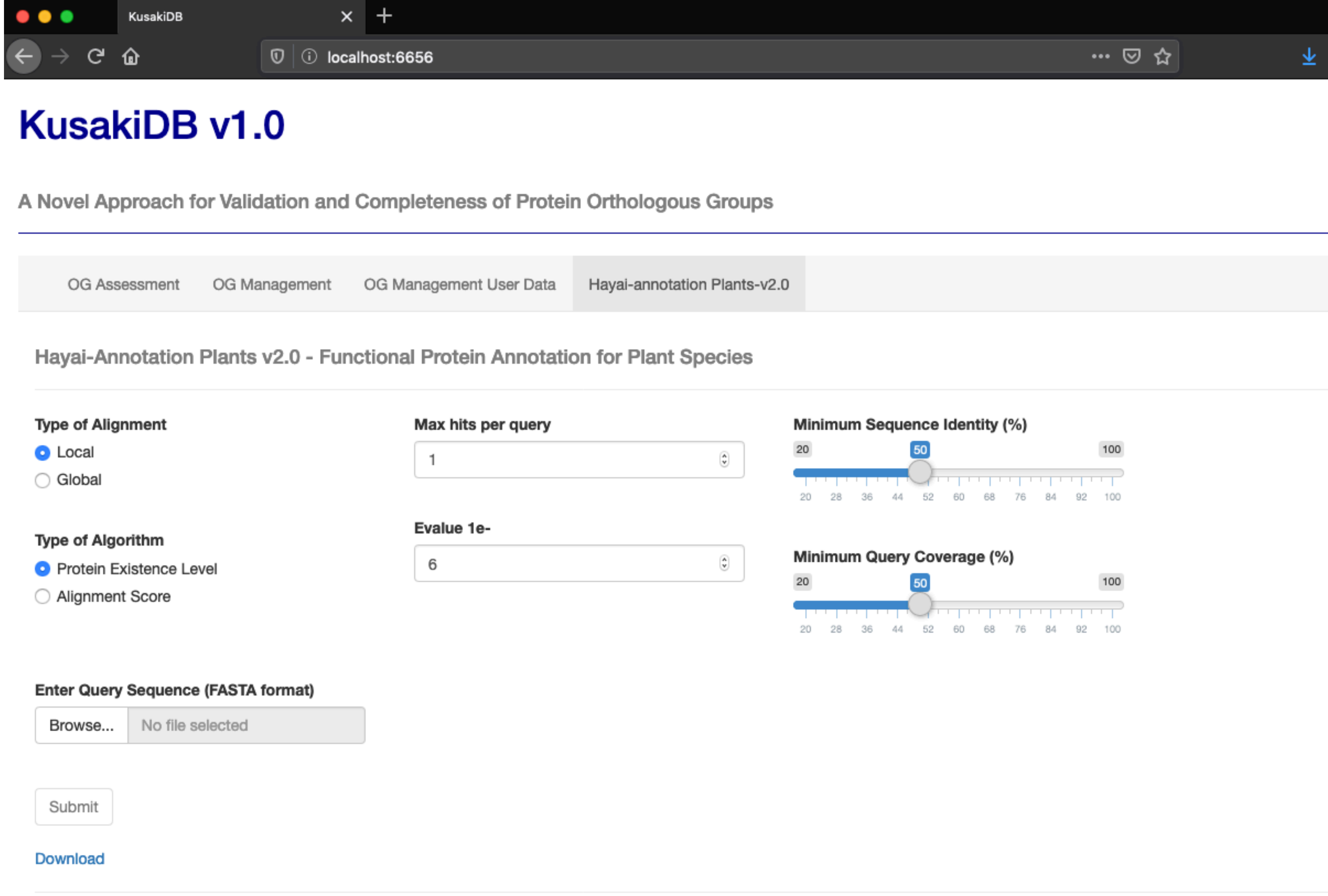
Search: terpene

	queryid_pep	pub_og_id	uniprot_protein_name	uniprot_accession	kusakiDB_curation	Perc	Protein_name
18635	AT4G16730.1	62841at33090	(E)-beta-ocimene synthase, chloroplastic	P0CJ43	V	80.34	Terpene synthase, N-terminal domain
20874	AT1G61120.1	88253at33090	(E,E)-geranyllinalool synthase	Q93YV0	V	58.12	Terpene synthase, N-terminal domain
21128	AT4G13280.2	91294at33090	(Z)-gamma-bisabolene synthase 1	Q9T0J9	V	80.34	Terpene synthase, N-terminal domain
21113	AT4G13300.1	91294at33090	(Z)-gamma-bisabolene synthase 2	Q9T0K1	V	80.34	Terpene synthase, N-terminal domain
18637	AT3G25820.1	62841at33090	1,8-cineole synthase 1, chloroplastic	P0DI76	V	80.34	Terpene synthase, N-terminal domain
18640	AT3G25830.1	62841at33090	1,8-cineole synthase 1, chloroplastic	P0DI76	V	80.34	Terpene synthase, N-terminal domain
21134	AT5G44630.1	91294at33090	Alpha-barbatene synthase	Q4KSH9	V	80.34	Terpene synthase, N-terminal domain
21118	AT5G23960.1	91294at33090	Alpha-humulene/(-)-(E)-beta-caryophyllene synthase	Q84UU4	V	80.34	Terpene synthase, N-terminal domain
6370	AT1G78960.1	16117at33090	Amyrin synthase LUP2	Q8RWT0	V	100	Terpene cyclase/mutase family member
6371	AT4G15340.1	16117at33090	Arabidiol synthase	Q9FR95	V	100	Terpene cyclase/mutase family member

Showing 1 to 10 of 47 entries (filtered from 22,124 total entries)

KusakiDB v1.0 Tools: Hayai-annotation v2.0

Interface of Hayai-annotation v2.0



KusakiDB v1.0

A Novel Approach for Validation and Completeness of Protein Orthologous Groups

Hayai-Annotation Plants v2.0 - Functional Protein Annotation for Plant Species

Type of Alignment
- ● Local
- ○ Global

Type of Algorithm
- ● Protein Existence Level
- ○ Alignment Score

Max hits per query: 1

Evalue 1e-: 6

Minimum Sequence Identity (%): 50 (20–100)

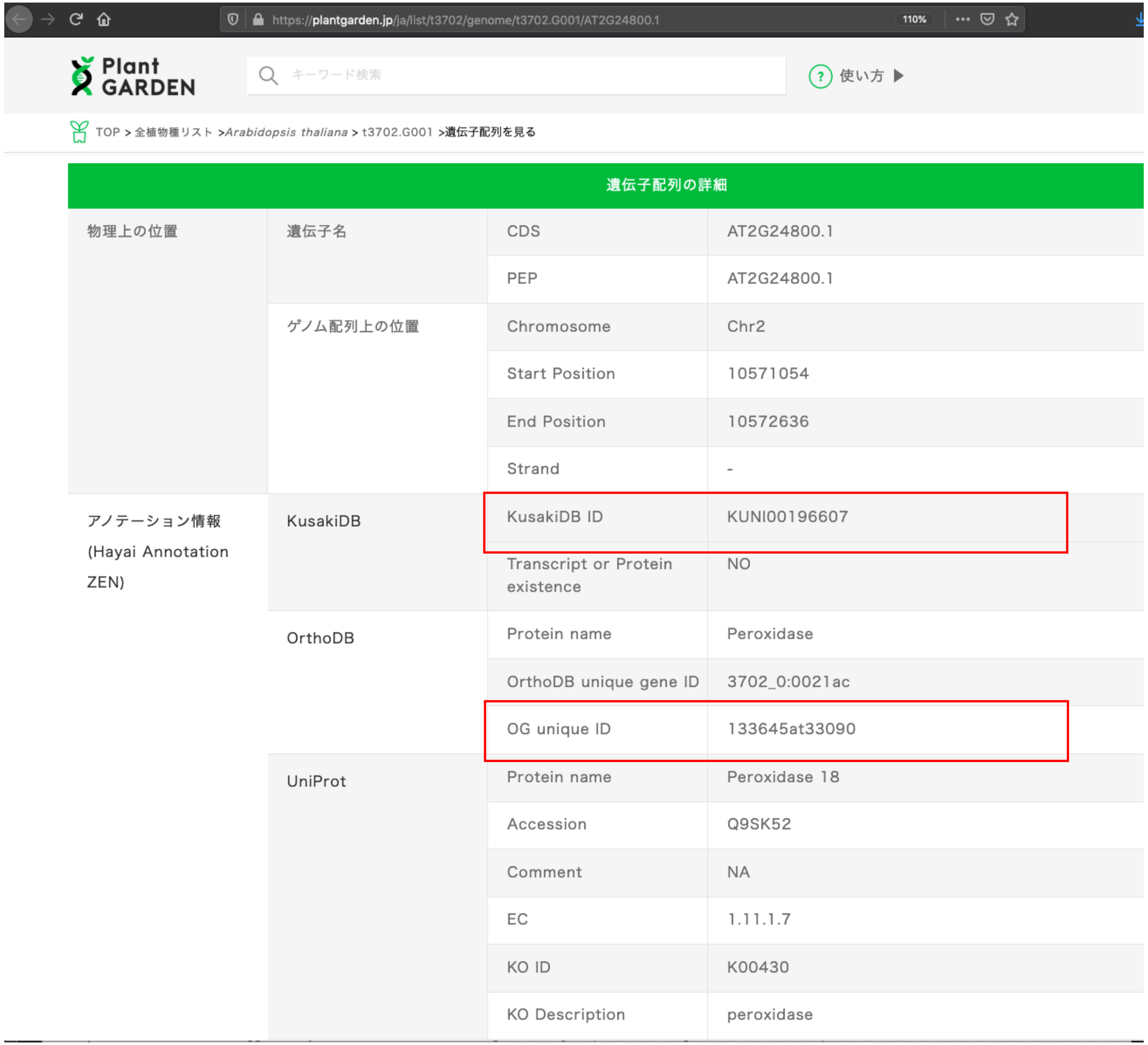
Minimum Query Coverage (%): 50 (20–100)

Enter Query Sequence (FASTA format)

Browse... No file selected

KusakiDB v1.0 Relevance: Source of Orthologous Groups for Plant GARDEN

Re-annotation of genes registered in Plant GARDEN was performed using KusakiDB as database and Hayai-Annotation as an annotation program. Thus all entries at Plant GARDEN have a correspondent OG associated with each gene.



https://plantgarden.jp/ja/list/t3702/genome/t3702.G001/AT2G24800.1
110%
Plant GARDEN
キーワード検索
使い方

TOP > 全植物種リスト >*Arabidopsis thaliana* > t3702.G001 >遺伝子配列を見る

遺伝子配列の詳細

物理上の位置	遺伝子名	CDS	AT2G24800.1
		PEP	AT2G24800.1
	ゲノム配列上の位置	Chromosome	Chr2
		Start Position	10571054
		End Position	10572636
		Strand	-
アノテーション情報 (Hayai Annotation ZEN)	KusakiDB	KusakiDB ID	KUNI00196607
		Transcript or Protein existence	NO
	OrthoDB	Protein name	Peroxidase
		OrthoDB unique gene ID	3702_0:0021ac
		OG unique ID	133645at33090
	UniProt	Protein name	Peroxidase 18
		Accession	Q9SK52
		Comment	NA
		EC	1.11.1.7
		KO ID	K00430
		KO Description	peroxidase

今後の連携への取り組み (Future assignments)

- Provide an API interface for RDF users in order to promote an easier integration of data among other databases.
- Implement conservation analysis of OGs for glycogenes identified in Plant GARDEN (using AMAI v0.2) and available at GlyCosmos (https://glycosmos.org/plantgardens/index).
- Implement further analysis of OG conservation in Plant GARDEN.

まとめ(Summary)

- KusakiDB v1.0 was developed to offer a new method to evaluate the ”real” existence of a protein through OGs existence (KusakiDB validation tag).
- Besides, if an OG is assigned as ”Not validated” means that until now no one found a protein or transcript. If the gene has a potential interest some researches may focus their attention in order to properly identify that transcript or protein.
- KusakiDB v1.0 may provide some tools for researchers in order to find, for example, proteins that are conserved only within a family, or in other words, find genes that have higher evolutionary rates.