

Supplementary Data. Hayai-Annotation v3.0: A functional gene prediction tool that integrates orthologs and gene ontology for network analysis

Andrea Ghelfi^{1*}, Sachiko Isobe²

1: National Institute of Genetics, Yata, 1111, Mishima, Shizuoka, 411-8540, Japan

2: Kazusa DNA Research Institute, Kazusa-Kamatari, 2-6-7, Kisarazu, Chiba, 292-0818, Japan

Corresponding Author

Andrea Ghelfi, E-mail: andreaghelfi@nig.ac.jp

National Institute of Genetics, Yata, 1111, Mishima, Shizuoka, 411-8540, Japan.

Tel.: +81-55-981-6853. Fax.: +81-55-981-6849

* 'To whom correspondence should be addressed'

Table S1 Complete co-occurrence tables of OrthoDB IDs and GO (MF and BP) for wild rice (tabs 1 and 2) and *O. sativa* (tabs 3 and 4). Filename: Table_S1.xlsx

Table S2 Benchmark of GO enrichments using CAFA-evaluation, weighted PR, RC, and F, with *A. thaliana*, comparing Hayai-Annotation v3 and BLAST2GO. Replications 1-3 (A, B, C).

(A)

filename	ns	tau	n	pr_w	rc_w	f_w
blast2go_rep1.tsv	biological_process	0.001	550	0.9574	0.7050	0.8120
blast2go_rep1.tsv	cellular_component	0.001	612	0.9848	0.7188	0.8310
blast2go_rep1.tsv	molecular_function	0.001	605	0.9813	0.7659	0.8603
hayai_rep1.tsv	biological_process	0.001	538	0.9627	0.6948	0.8071
hayai_rep1.tsv	cellular_component	0.001	648	0.9851	0.8649	0.9211
hayai_rep1.tsv	molecular_function	0.001	579	0.9928	0.7451	0.8513

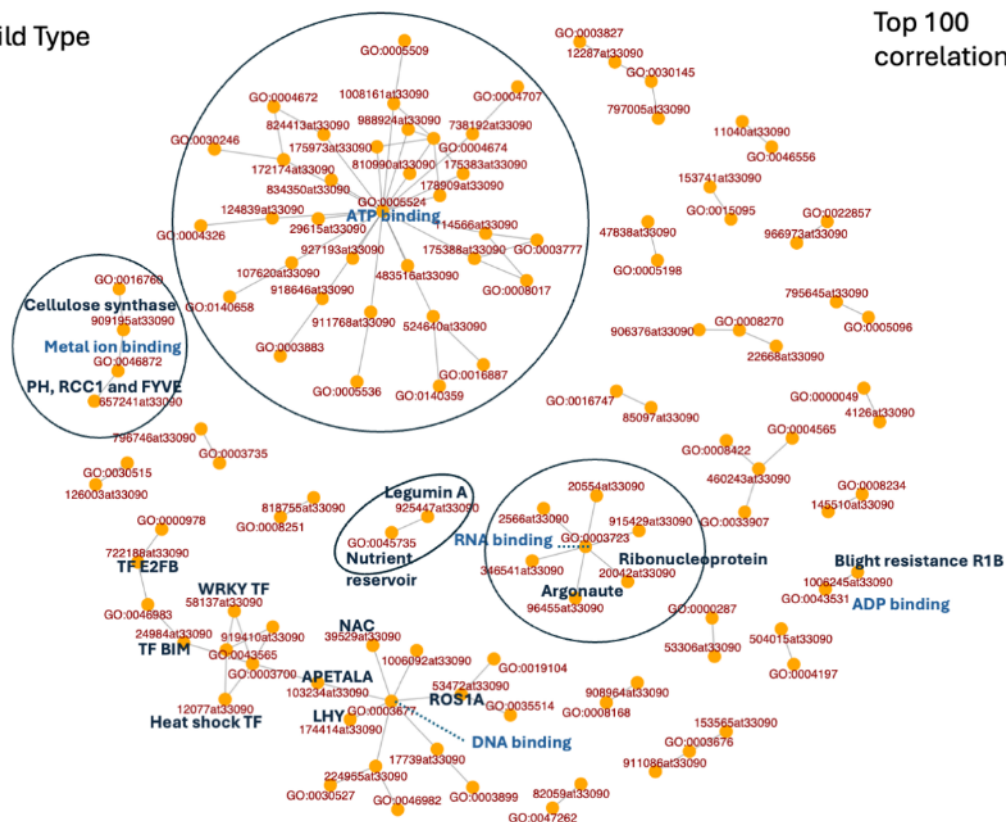
(B)

filename	ns	tau	n	pr_w	rc_w	f_w
blast2go_rep2.tsv	biological_process	0.001	531	0.963	0.6879	0.8025
blast2go_rep2.tsv	cellular_component	0.001	570	0.9854	0.6863	0.8091
blast2go_rep2.tsv	molecular_function	0.001	594	0.9782	0.7538	0.8514
hayai_rep2.tsv	biological_process	0.001	512	0.9713	0.6647	0.7893
hayai_rep2.tsv	cellular_component	0.001	620	0.9894	0.8507	0.9148
hayai_rep2.tsv	molecular_function	0.001	567	0.9909	0.7238	0.8366

(C)

filename	ns	tau	n	pr_w	rc_w	f_w
blast2go_rep3.tsv	biological_process	0.001	545	0.9704	0.7198	0.8265
blast2go_rep3.tsv	cellular_component	0.001	588	0.9821	0.7226	0.8326
blast2go_rep3.tsv	molecular_function	0.001	593	0.9848	0.7521	0.8528
hayai_rep3.tsv	biological_process	0.001	528	0.9777	0.7036	0.8183
hayai_rep3.tsv	cellular_component	0.001	615	0.989	0.8415	0.9093
hayai_rep3.tsv	molecular_function	0.001	575	0.9914	0.7239	0.8368

Top 100 correlations



(B) *O. sativa*

Top 100
correlations

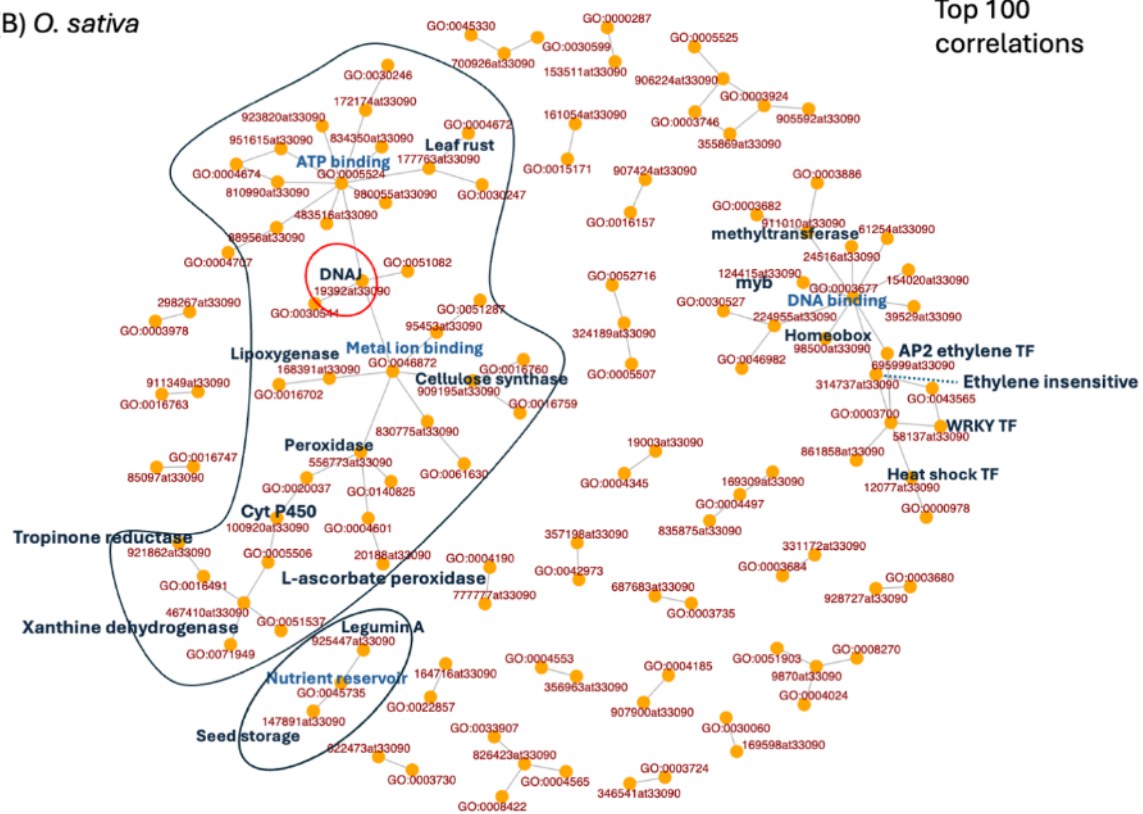
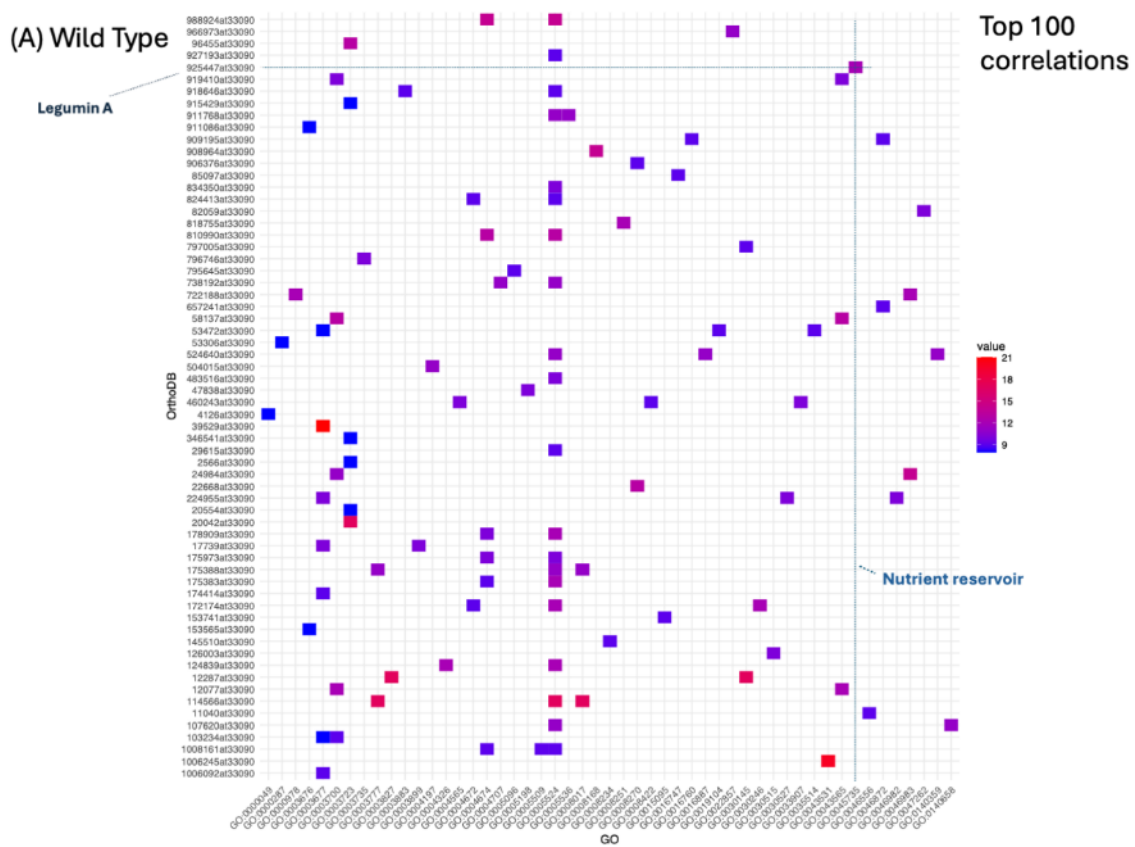


Figure S1. Network analysis performed on the top 100 co-occurrences of ODB IDs and GO molecular function. (A) wild type; (B) cultivated rice.



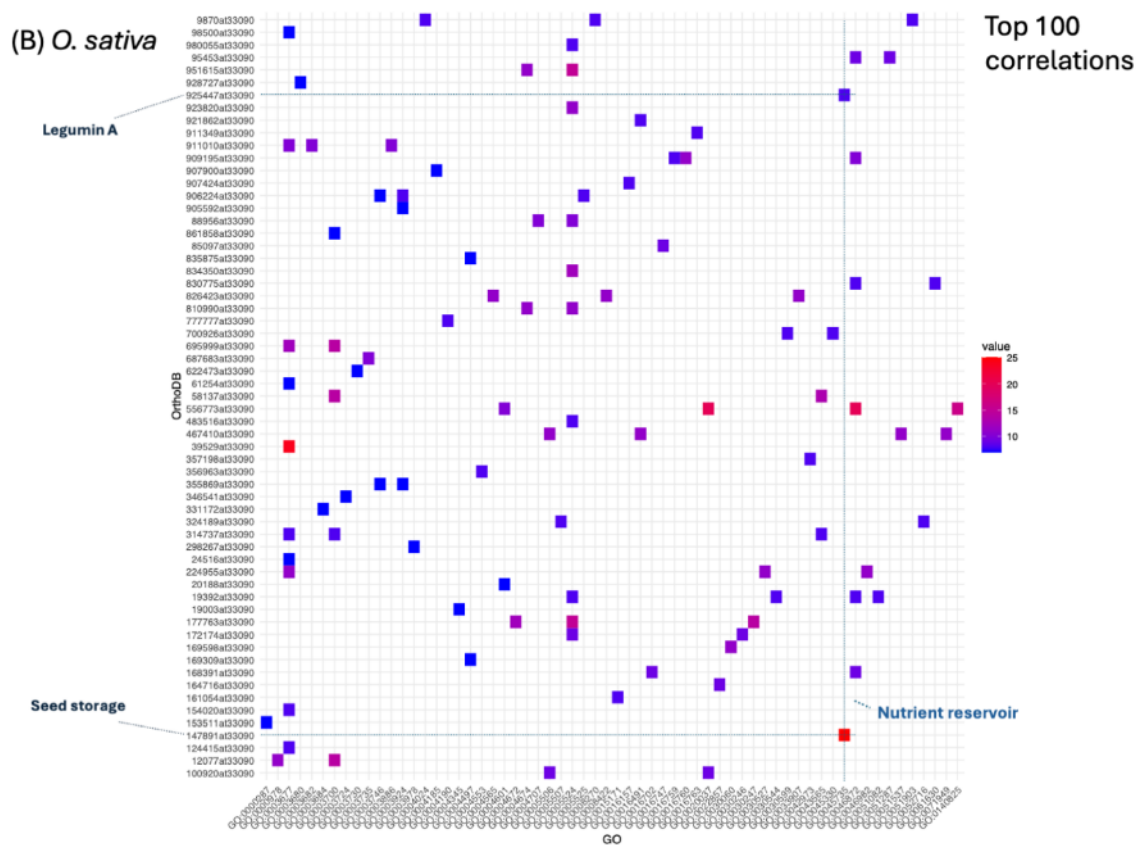
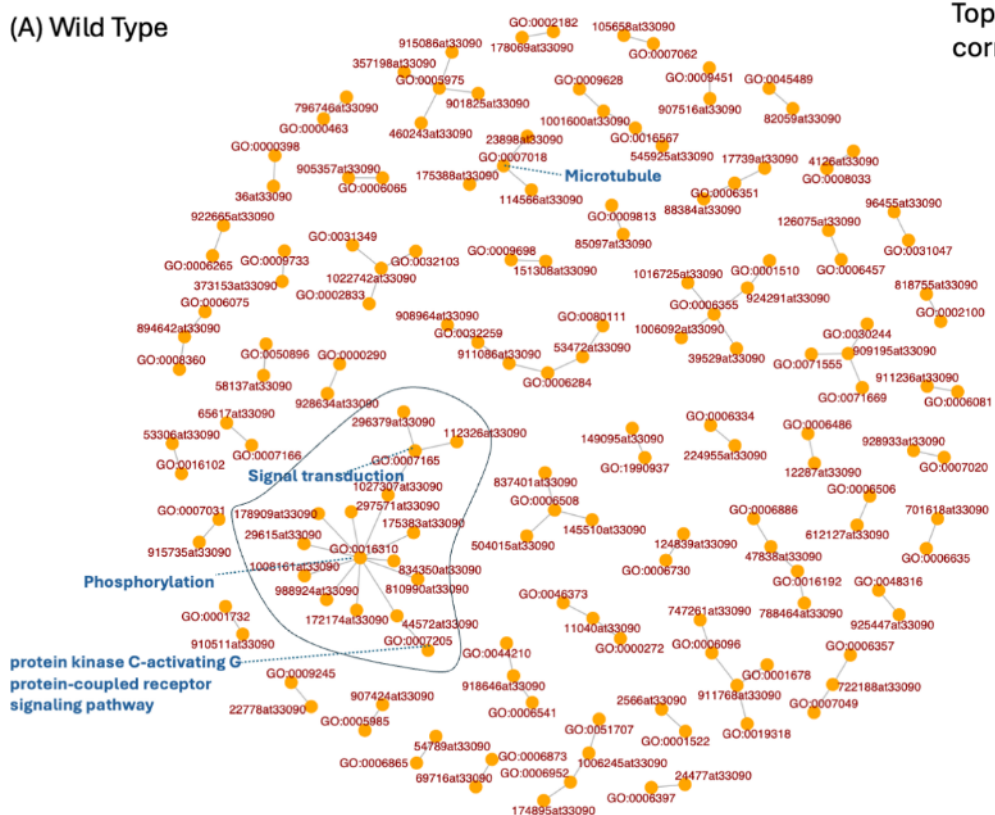


Figure S2. Heatmap analysis performed on the top 100 co-occurrences of ODB IDs and GO molecular function. (A) wild type; (B) cultivated rice.

(A) Wild Type

Top 100 correlations



Top 100 correlations

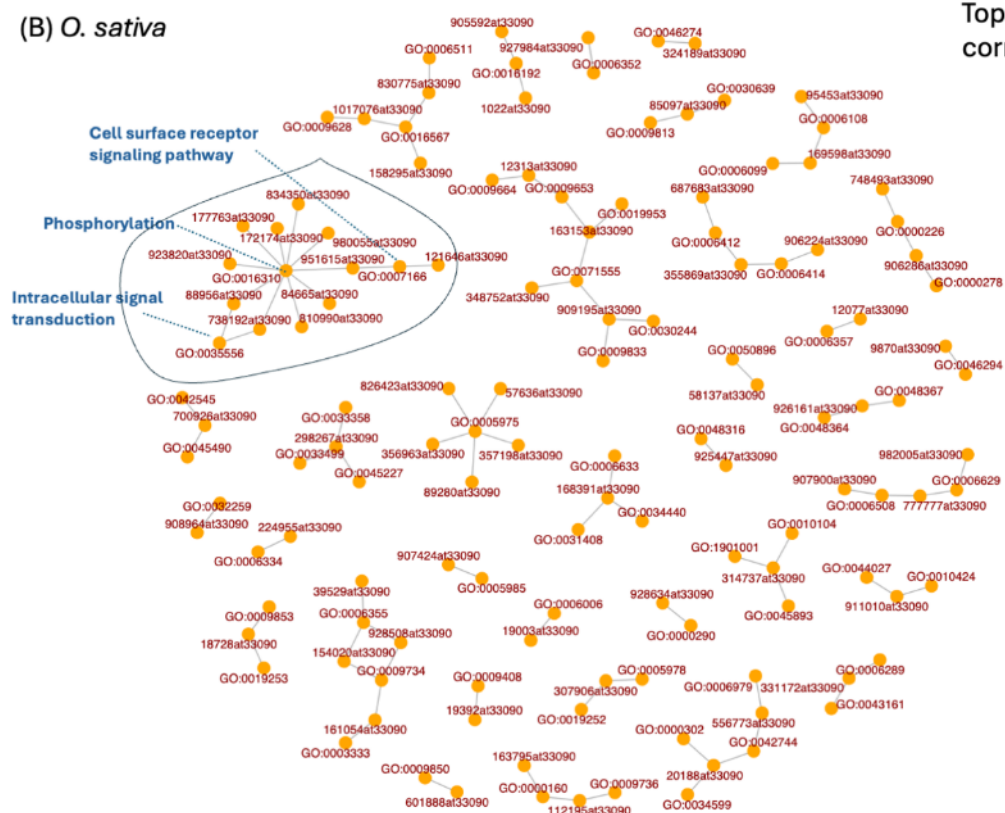
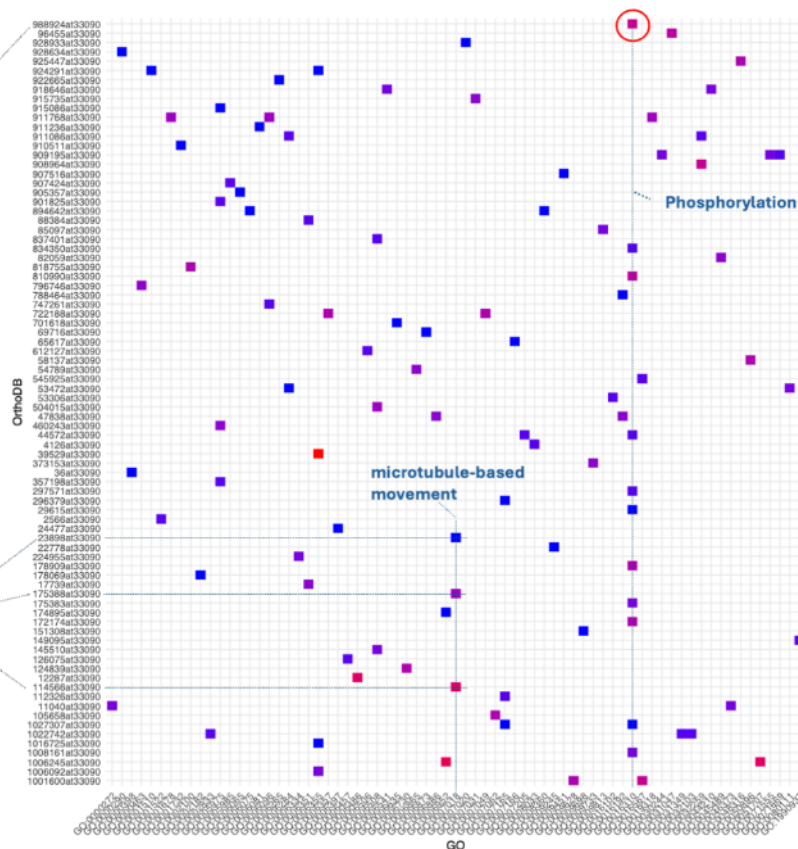


Figure S3. Network analysis performed on the top 100 co-occurrences of ODB IDs and GO biological process. (A) wild type; (B) cultivated rice.

(A) Wild Type

serine/threonine
-protein kinase

Kinesin-like
protein



Top 100
correlations

(B) *O. sativa*

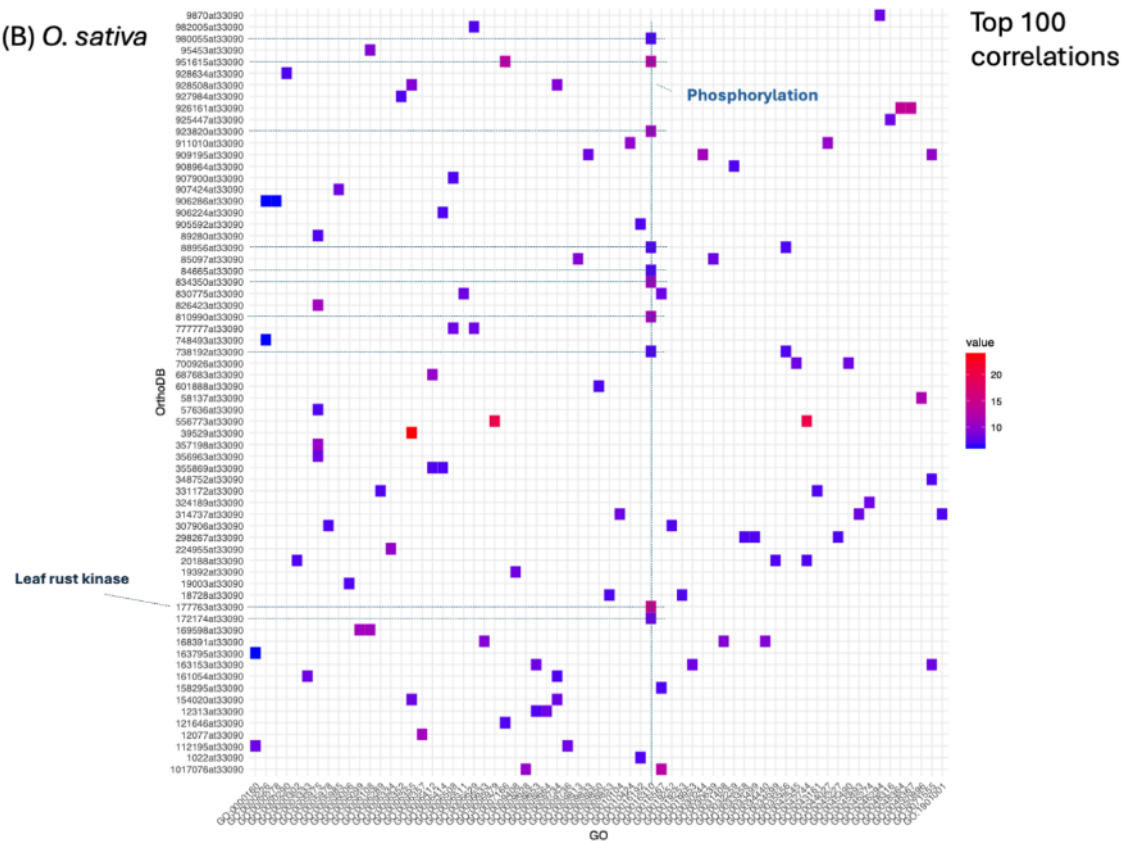


Figure S4. Heat analysis performed on the top 100 co-occurrences of ODB IDs and GO biological process. (A) wild type; (B) cultivated rice.

Hayai-Annotation Plants v3.0
Functional Annotation and Network Analysis of Orthologs and Gene Ontology
Specialized in Plant Species

Query Sequence Type
● Protein
● DNA

Download Sample

Upload FASTA File
Browse... Araport11_gi
Upload complete

Submit

Download Results

Show 10 entries

Query	Accession	Product_Name	OrthoDB	OrthoDB_Desc	Evidence_existence
AT1G06190.5	A0A1P8AM12	Rho termination factor	96820at33090	Rho_N domain-containing protein	1
AT1G06620.1	Q84MB3	1-aminocyclopropane-1-carboxylate oxidase homolog 1	632243at33090	1-aminocyclopropane-1-carboxylate oxidase homolog 1-like	2
AT1G067670.1	Q9FXD3	F12A21.18	833372at33090	uncharacterized protein LOC116128294	4
AT1G12100.1	F4IC43	Bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	824799at33090	14 kDa proline-rich protein DC2.15	3
AT1G49040.1	Q8RKA7	DENN domain and WD repeat-containing protein SCD1	49215at33090	DENN domain and WD repeat-containing protein SCD1	1
AT1G08710.2	Q9CAZ0	F-box protein SKP24	115093at33090	F-box protein SKP24	1
AT1G78300.1	Q01525	14-3-3-like protein GF14 omega	319121at33090	14-3-3 protein	1
AT1G09210.1	Q38658	Calreticulin-2	324299at33090	Calreticulin	1
AT1G58430.1	Q9C648	GDSL esterase/lipase At1g58430	345025at33090	GDSL esterase/lipase At2g30310	2
AT1G72290.1	Q9C756	Kunitz trypsin inhibitor 2	365530at33090	Kunitz trypsin inhibitor 2	1

Showing 1 to 10 of 27,512 entries

Previous 1 2 3 4 5 ... 2,752 Next

Figure S5. Hayai-Annotation v3 interface.

Hayai-Annotation v3 generates an output file named ‘output_HayaiAnnotation_v2.zip’. It comprises 7 tables (TSV format) and 8 graphics (PDF format). The files provided correspond to the full annotation of *A. thaliana*.

Hayai_annotation_v3.0.tsv is the main table and contains the full annotations from both methods, DIAMOND and Orthomapper (OrthoLogger).

Four tables aggregate the results for each GO domain (MF, BP, CC) and InterPro annotation layers; they are named, respectively, ‘Hayai_annotation_GO_MF.tsv’, ‘Hayai_annotation_GO_BP.tsv’, ‘Hayai_annotation_GO_CC.tsv’ and ‘Hayai_annotation_Interpro.tsv’.

Two tables show the co-occurrences of ODB ID and GO (MF and BP, independently), named ‘Correlations_OrthoDB_GO_MF.tsv’ and ‘Correlations_OrthoDB_GO_BP.tsv’. The count corresponds to the number of genes for each co-occurrence.

Four graphics are generated, two regarding the network and two the heatmap, based on the top 100 co-occurrences of ODB ID and GO (MF and BP, independently). The filenames are: ‘Graph_Network_OrthoDB_GO_MF.pdf’, ‘Graph_Network_OrthoDB_GO_BP.pdf’, ‘Graph_Heatmap_OrthoDB_GO_MF.pdf’ and ‘Graph_Heatmap_OrthoDB_GO_BP.pdf’.

Using the top 50 counts of the aggregated results, four graphics are generated for the distribution of GO (MF, BP and CC) and InterPro.