D4

DIRECT SUBMISSION SYSTEM AND LITERATURE ANNOTATION OF RICE GENES IN ORYZABASE

Yukiko Yamazaki[1], Rie Tsuchiya[1], Takao Asanuma[2], Yo Shidahara[2], Shingo Sakaniwa[1]
(1. National Institute of Genetics, Mishima, Japan, 2. NalaPro Technologies, Tokyo, Japan)

Abstract

Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/) is a comprehensive rice science database [1]. It houses a variety of genetic resources, relevant literatures, gene dictionary, DNA sequences, and basic information such as developmental biology and anatomy. In order to keep the gene dictionary up-to-date, literature annotation has been conducted manually since 1995. However as the publication of journal articles increases year by year after genomic sequences were released, it became more difficult to update the dictionary timely and in high quality without sufficient annotators. To overcome this difficulty, we applied machine learning and text-mining to extract known and unknown genes from journals. The machine extraction followed by manual annotation achieved promising results and increased efficiency in manual annotation.
Furthermore a direct submission system where rice researchers can deposit new genes according to the standardized nomenclature [2] became operational in 2008. Recent advances will be introduced.

1.Kurata, N. and Y. Yamazaki., Oryzabase, An Integrated Biological and Genome Information Database for Rice. *Plant Physiology* (2006) 140, 12-17
2 Susan R. McCouch, Gene Nomenclature System for Rice, *Rice* (2008) 1:72-84

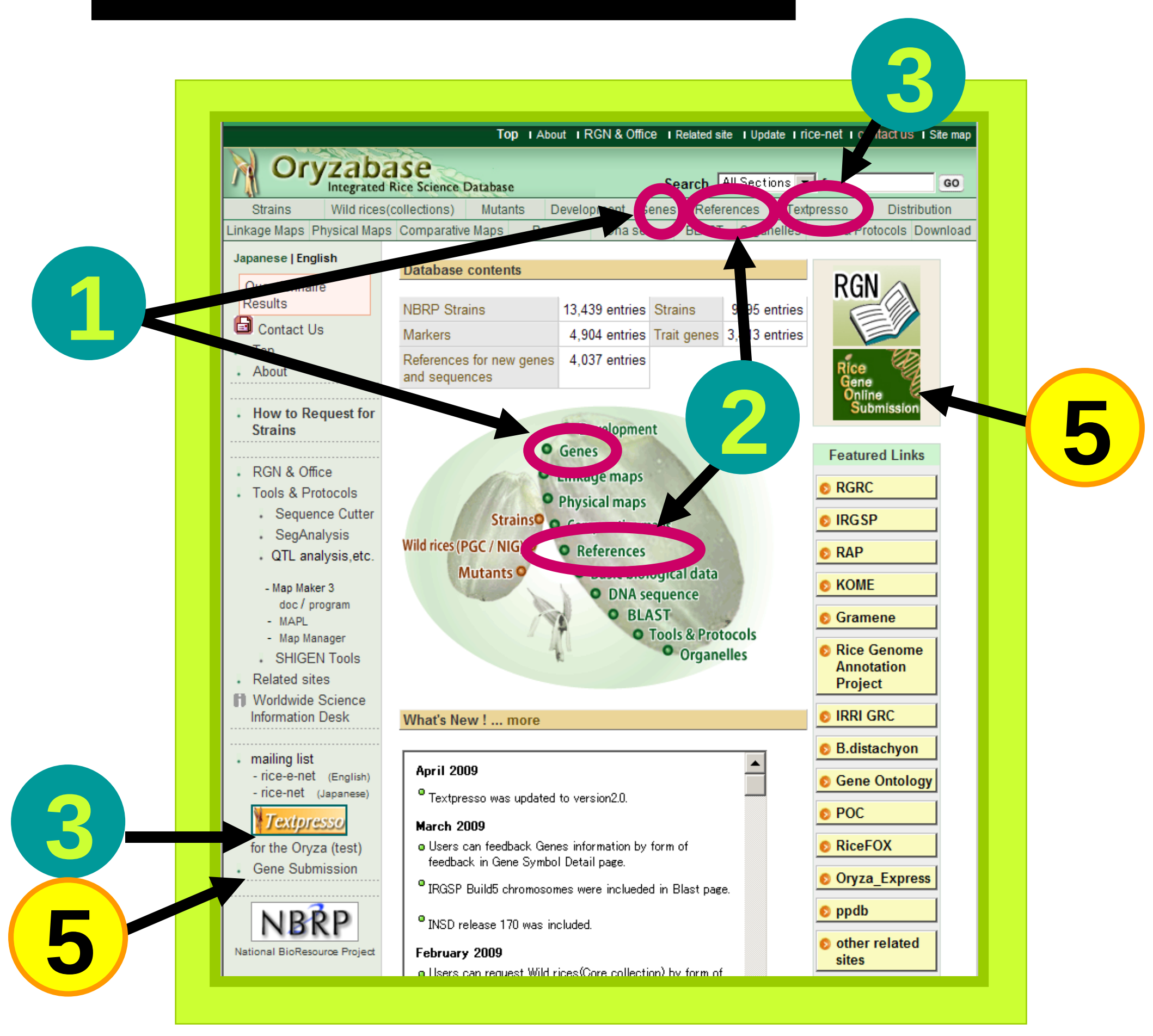
Future plan

Make the gene dictionary more up-to-date, accurate, and comprehensive.

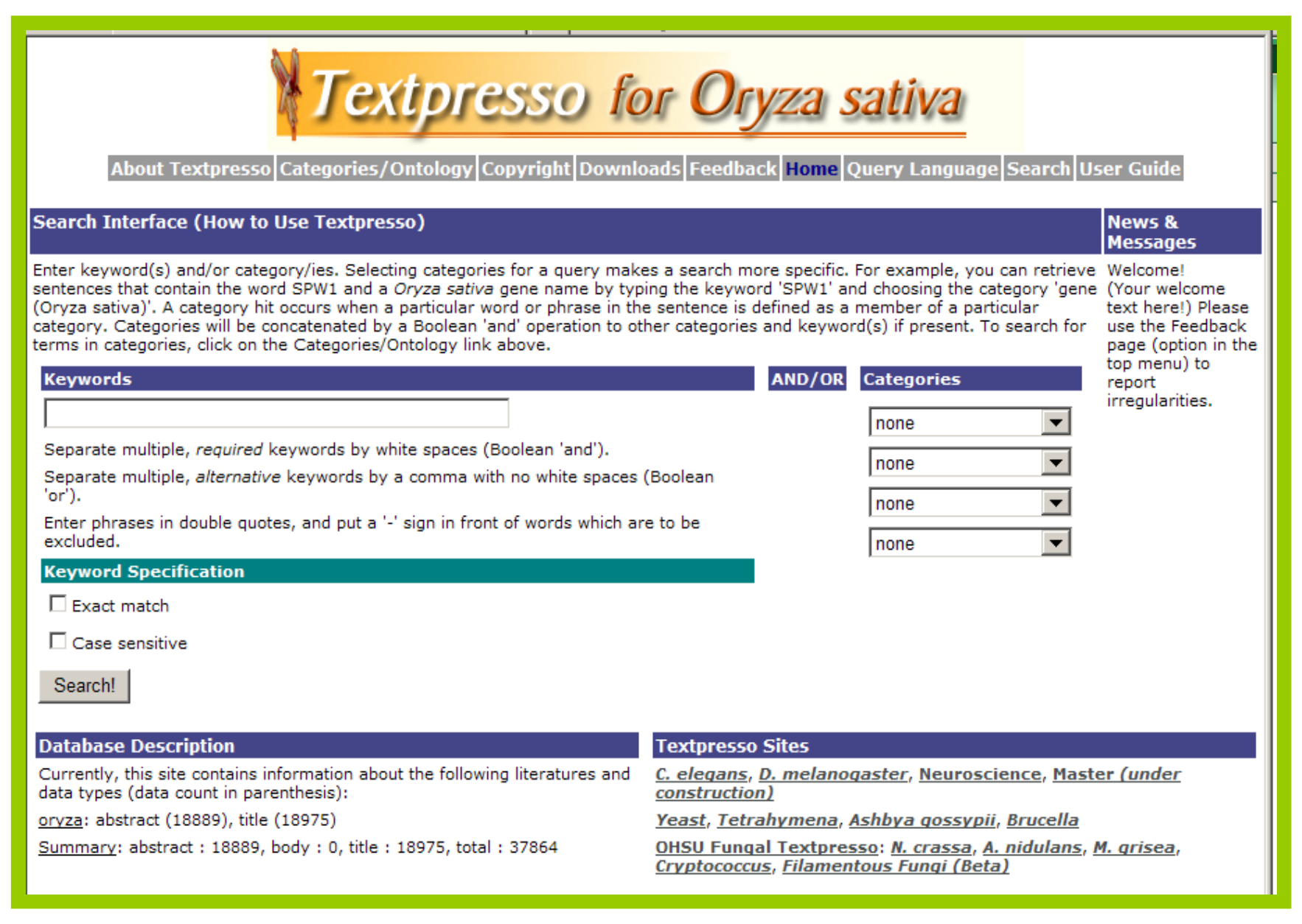
- improve precision of machine extraction
- collaboration with RAP group on gene annotation
- encourage researchers to submit new genes before publication
- encourage researchers to give feedback on Oryzabase genes

Oryzabase HP

www.shigen.nig.ac.jp/rice/oryzabase/



3 Rice Textpresso


Textpresso for Oryza sativa

About Textpresso | Categories/Ontology | Copyright | Downloads | Feedback | Home | Query Language | Search | User Guide

Search Interface (How to Use Textpresso)

Enter keyword(s) and/or category/ies. Selecting categories for a query makes a search more specific. For example, you can retrieve sentences that contain the word SPW1 and a *Oryza sativa* gene name by typing the keyword 'SPW1' and choosing the category 'gene (Oryza sativa)'. A category hit occurs when a particular word or phrase in the sentence is defined as a member of a particular category. Categories will be concatenated by a Boolean 'and' operation to other categories and keyword(s) if present. To search for terms in categories, click on the Categories/Ontology link above.

News & Messages

Welcome! (Your welcome text here!) Please use the Feedback page (option in the top menu) to report irregularities.

Keywords | **AND/OR** | **Categories**

Separate multiple, *required* keywords by white spaces (Boolean 'and').

Separate multiple, *alternative* keywords by a comma with no white spaces (Boolean 'or').

Enter phrases in double quotes, and put a '-' sign in front of words which are to be excluded.

none
none
none
none

Keyword Specification

□ Exact match
□ Case sensitive

Search!

Database Description

Currently, this site contains information about the following literatures and data types (data count in parenthesis):

oryza: abstract (18889), title (18975)

Summary: abstract : 18889, body : 0, title : 18975, total : 37864

Textpresso Sites

C. elegans, D. melanogaster, Neuroscience, Master (under construction)

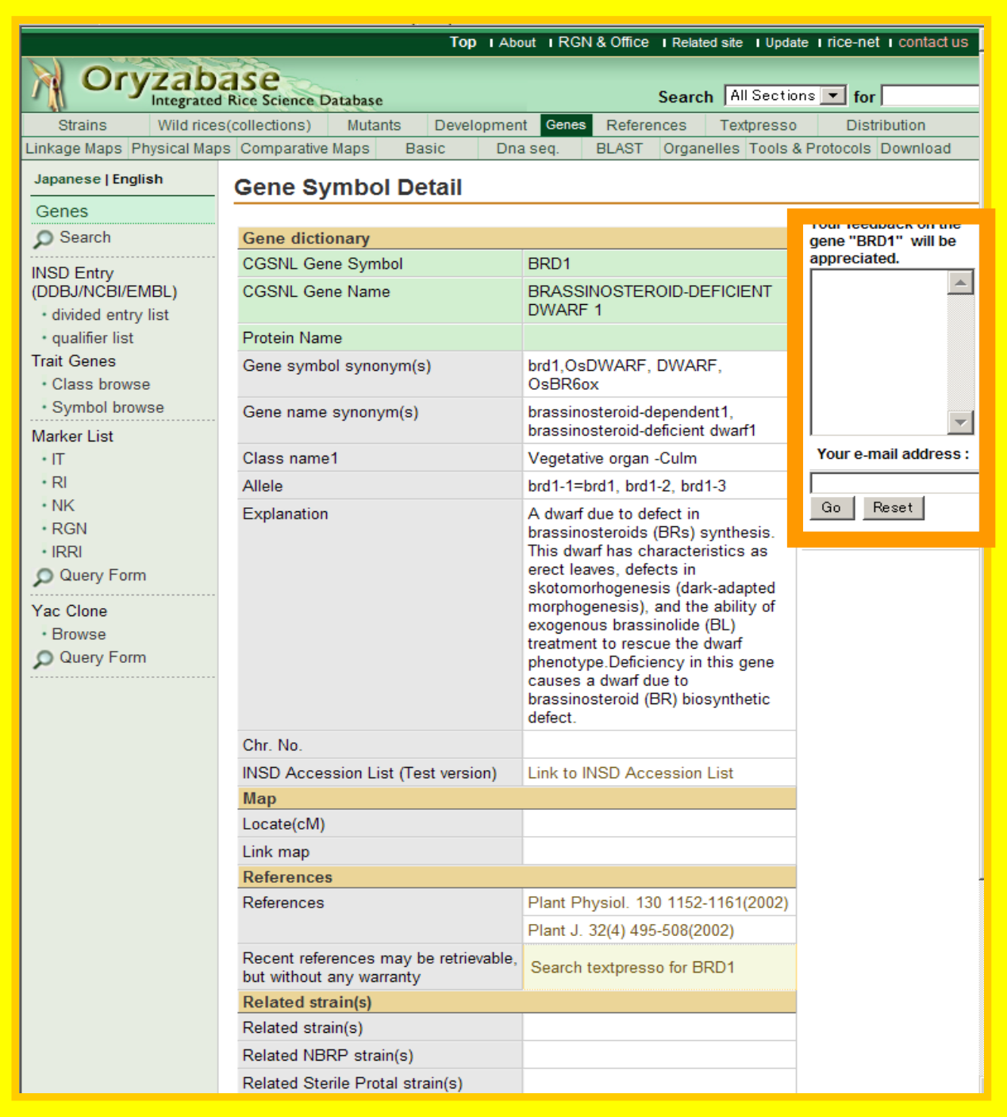
Yeast, Tetrahymena, Ashbya gossypii, Brucella

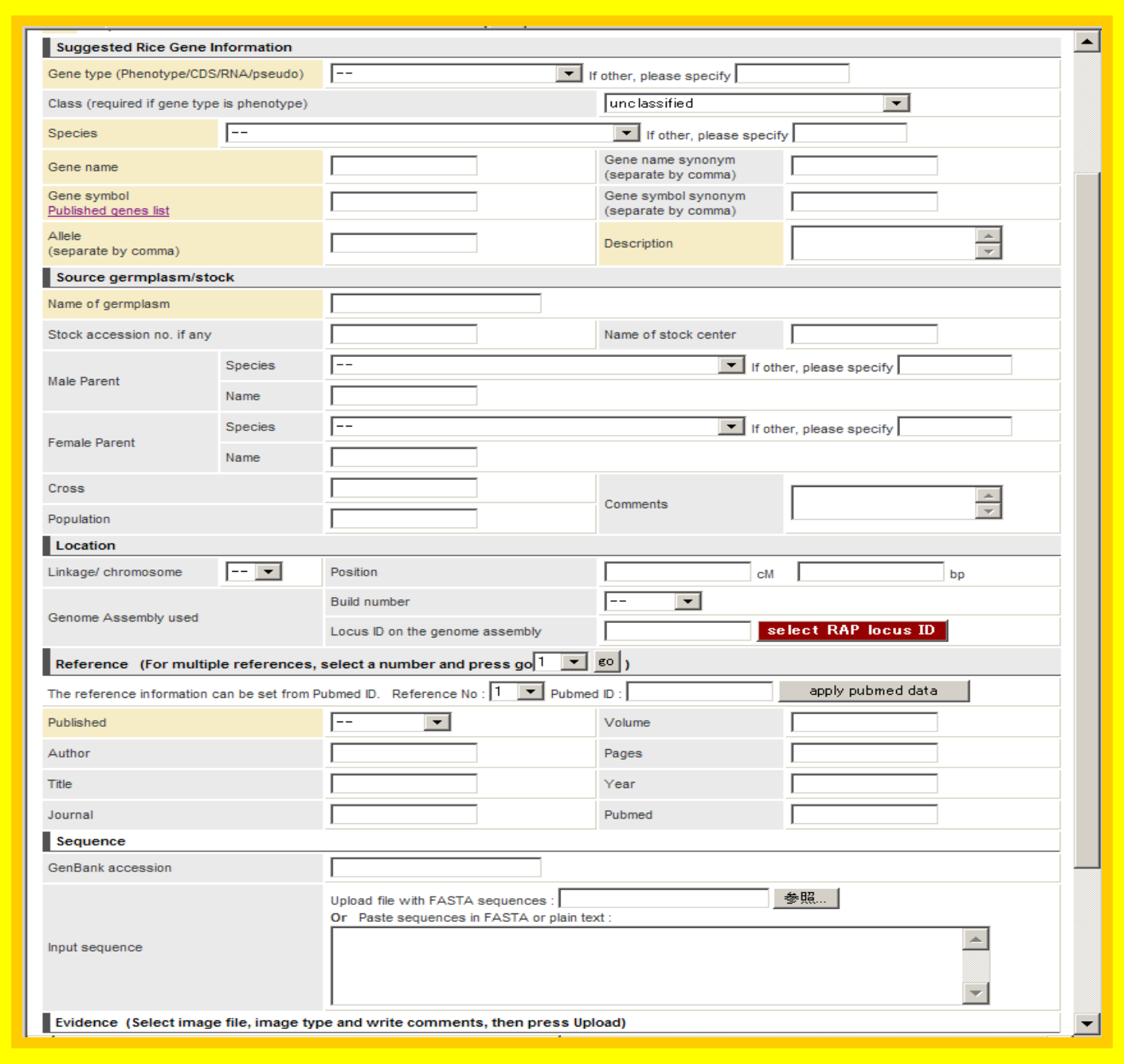
OHSU Fungal Textpresso: N. crassa, A. nidulans, M. grisea, Cryptococcus, Filamentous Fungi (Beta)

abstract (18889), title (18975)

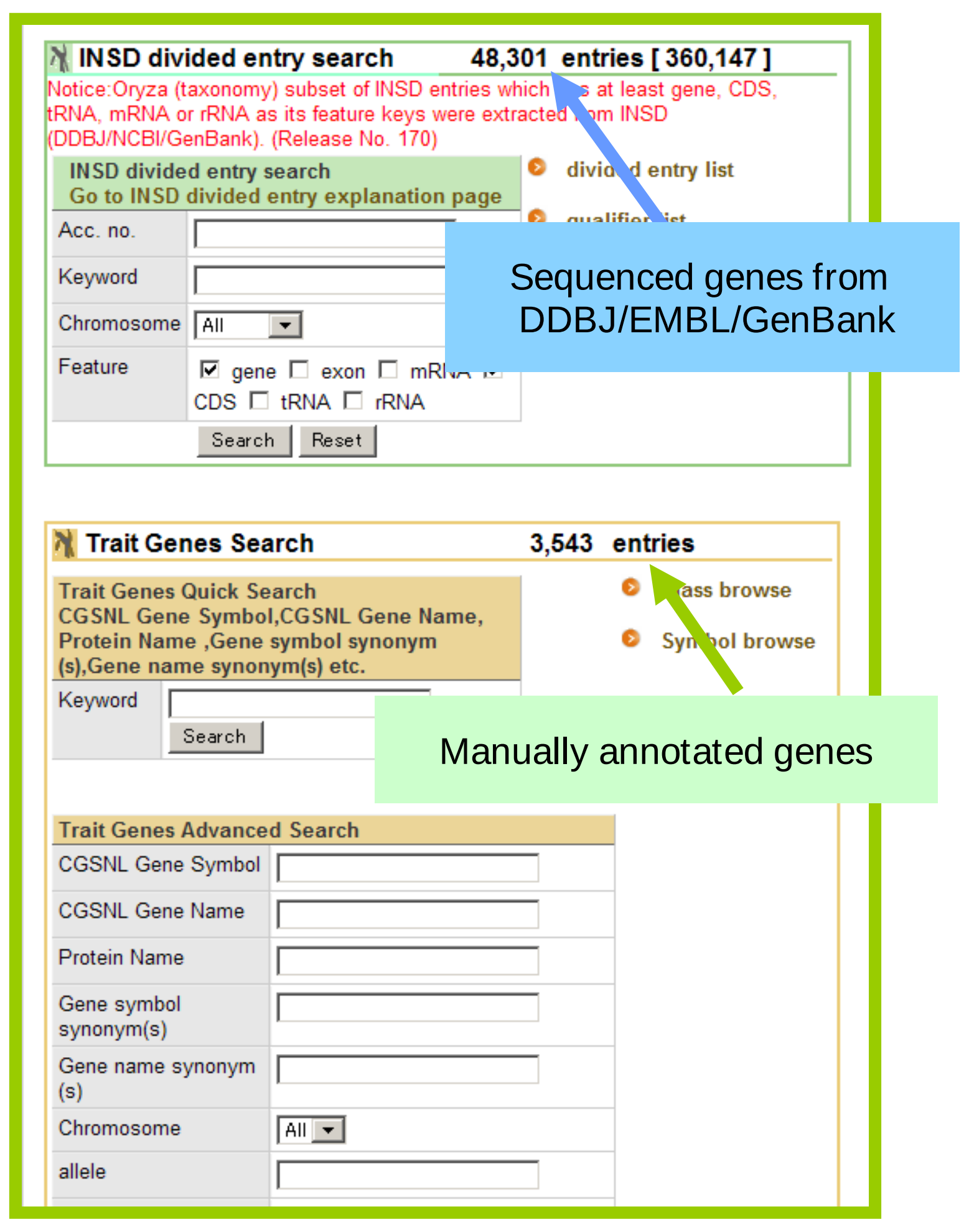
4 Feedback to a gene

5 on-line gene submission system





1 Oryzabase genes


INSD divided entry search 48,301 entries [360,147]

Notice:Oryza (taxonomy) subset of INSD entries which has at least gene, CDS, tRNA, mRNA or rRNA as its feature keys were extracted from INSD (DDBJ/NCBI/GenBank). (Release No. 170)

INSD divided entry search
Go to INSD divided entry explanation page

divided entry list
qualifier list

Acc. no.
Keyword
Chromosome All
Feature ☑ gene □ exon □ mRNA ☑ CDS □ tRNA □ rRNA

Search Reset

Sequenced genes from DDBJ/EMBL/GenBank

Trait Genes Search 3,543 entries

Trait Genes Quick Search
CGSNL Gene Symbol,CGSNL Gene Name, Protein Name ,Gene symbol synonym (s),Gene name synonym(s) etc.

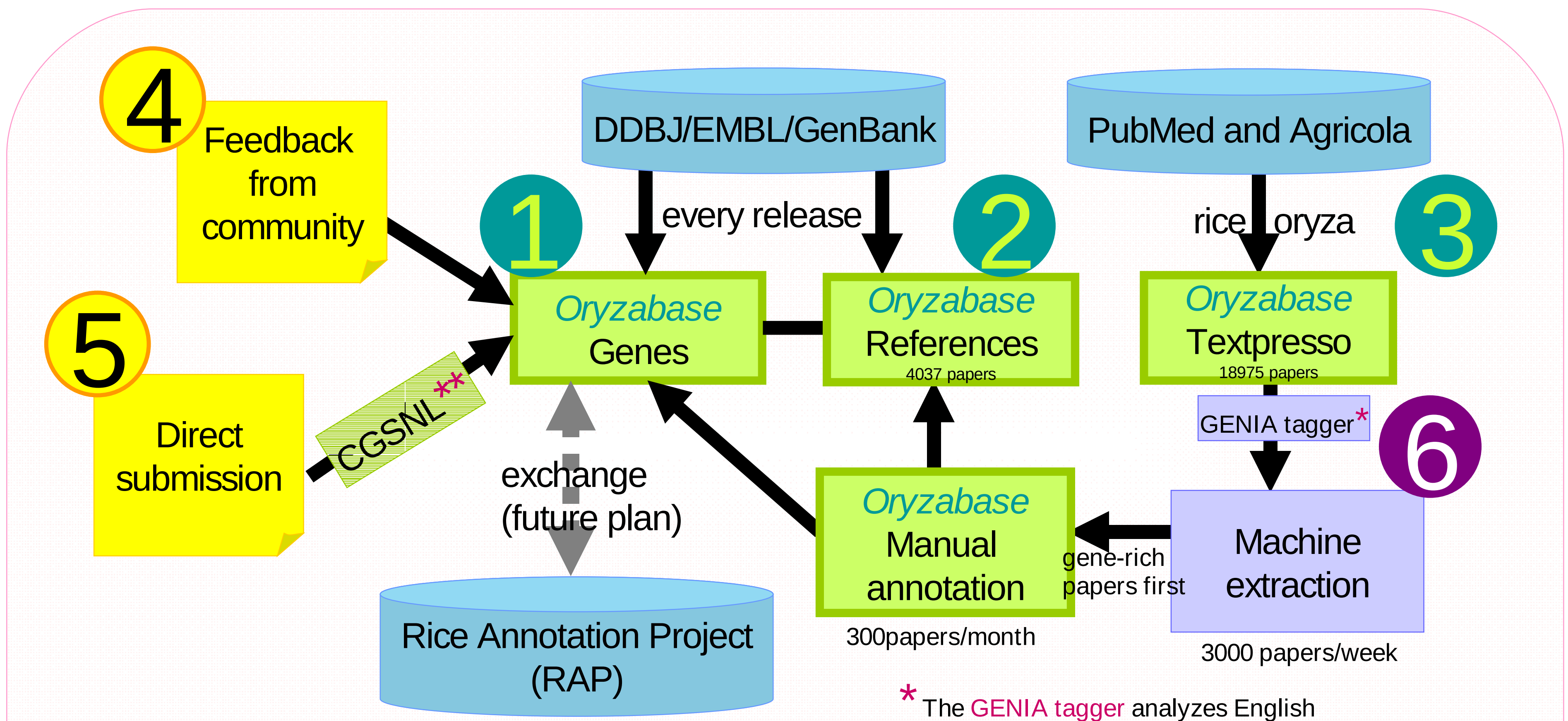
Class browse
Symbol browse

Keyword
Search

Manually annotated genes

Trait Genes Advanced Search
CGSNL Gene Symbol
CGSNL Gene Name
Protein Name
Gene symbol synonym(s)
Gene name synonym (s)
Chromosome All
allele

Oryzabase gene annotation flow



**The Committee on Gene Symbolization, Nomenclature and Linkage

*The GENIA tagger analyzes English sentences
and outputs the base forms, part-of-speech tags,
chunk tags, and named entity tags. The tagger is
specifically tuned for biomedical text such as MEDLINE abstracts.
http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/

6-1 Materials used for machine extraction

Oryzabase Gene Dictionary (as referencedata)
Rice DNA marker name (from RGP)
WordNet Dictionary for general words

GENIA Tagger (Corpus)*
Protein
DNA
RNA
Cell Line
Cell Type

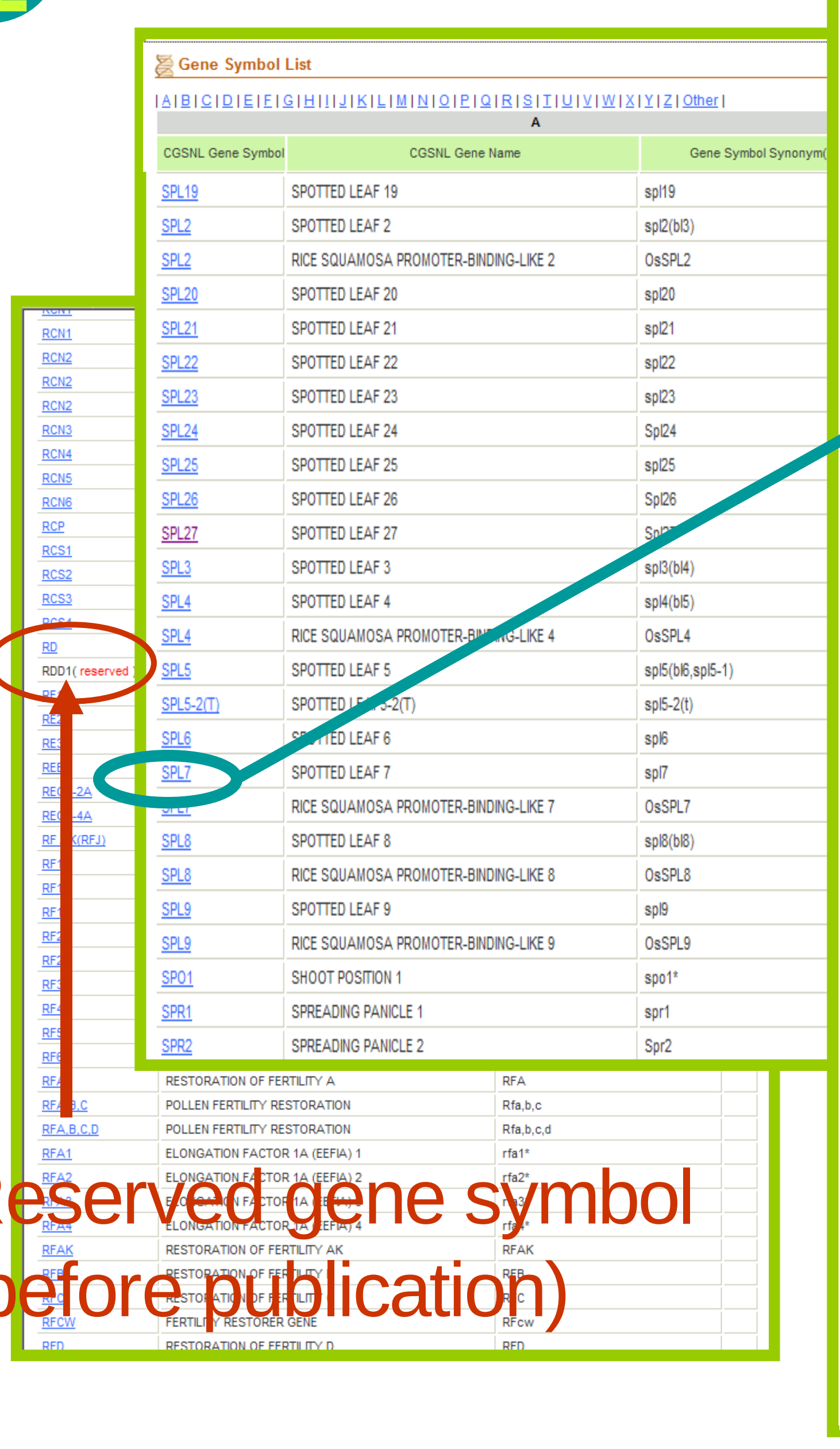
6-2 Result of natural language processing

Journal	Number	Manually extracted genes (A)	Machine extracted genes (B)	Match (C) %	Precision (D=C/B)%	Recall (E=C/A) %	F-value 2x(DxE)/(D+E)
Nature/Science	127	85	206	67	32.5	78.8	46
TAG*	103	47	171	33	19.3	70.2	30.3
Plant Cell etc.	233	226	656	181	27.9	80.8	41.5
Total	463						

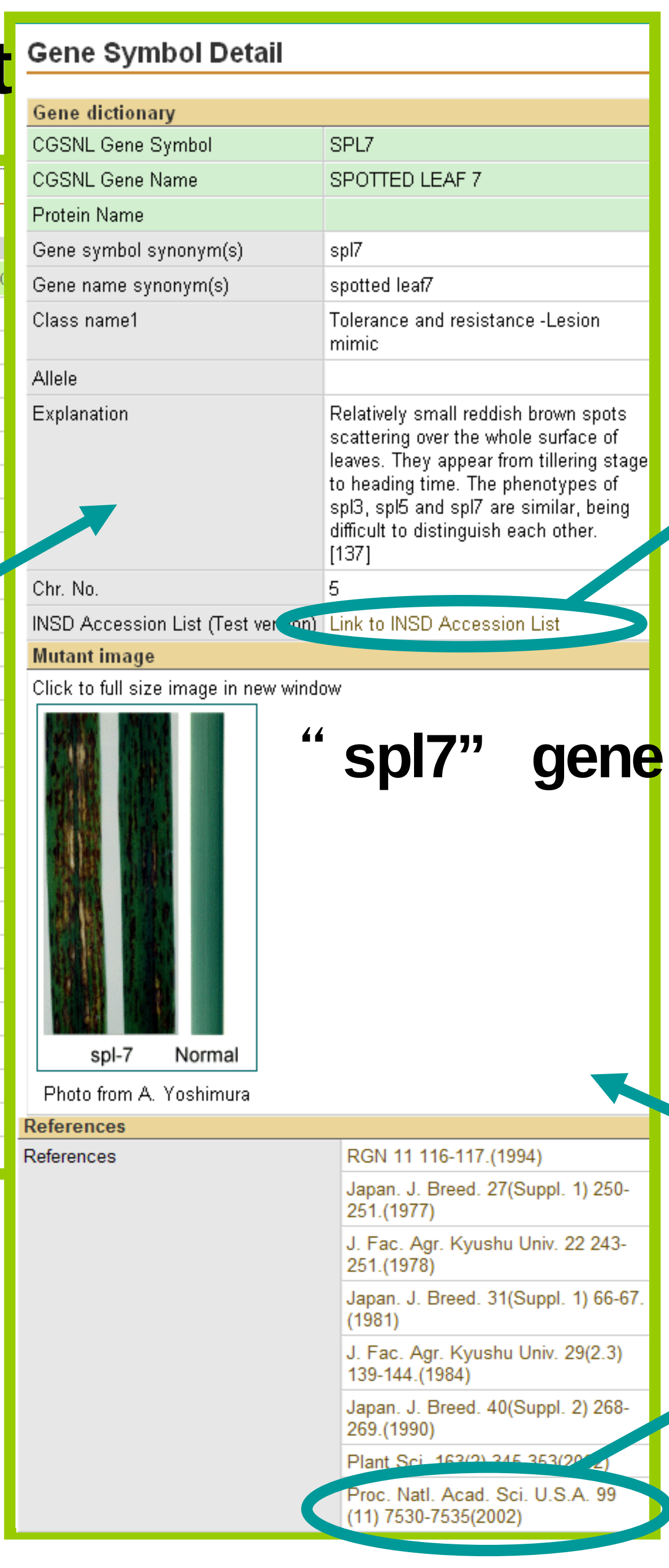
6-3 To improve precision

-- Use correlation of author and gene class
-- Use correlation of author and co-author
-- Improve quality of dictionary
-- Adjust GENIA Corpus to rice related words

1 Oryzabase genes list



Reserved gene symbol
(before publication)


Gene Symbol Detail

Gene dictionary	
CGSNL Gene Symbol	SPL7
CGSNL Gene Name	SPOTTED LEAF 7
Protein Name	
Gene symbol synonym(s)	spl7
Gene name synonym(s)	spotted leaf7
Class name1	Tolerance and resistance -Lesion mimic
Allele	
Explanation	Relatively small reddish brown spots scattering over the whole surface of leaves. They appear from tillering stage to heading time. The phenotypes of spl3, spl5 and spl7 are similar, being difficult to distinguish each other. [137]
Chr. No.	5
INSD Accession List (Test version)	Link to INSD Accession List

Mutant image

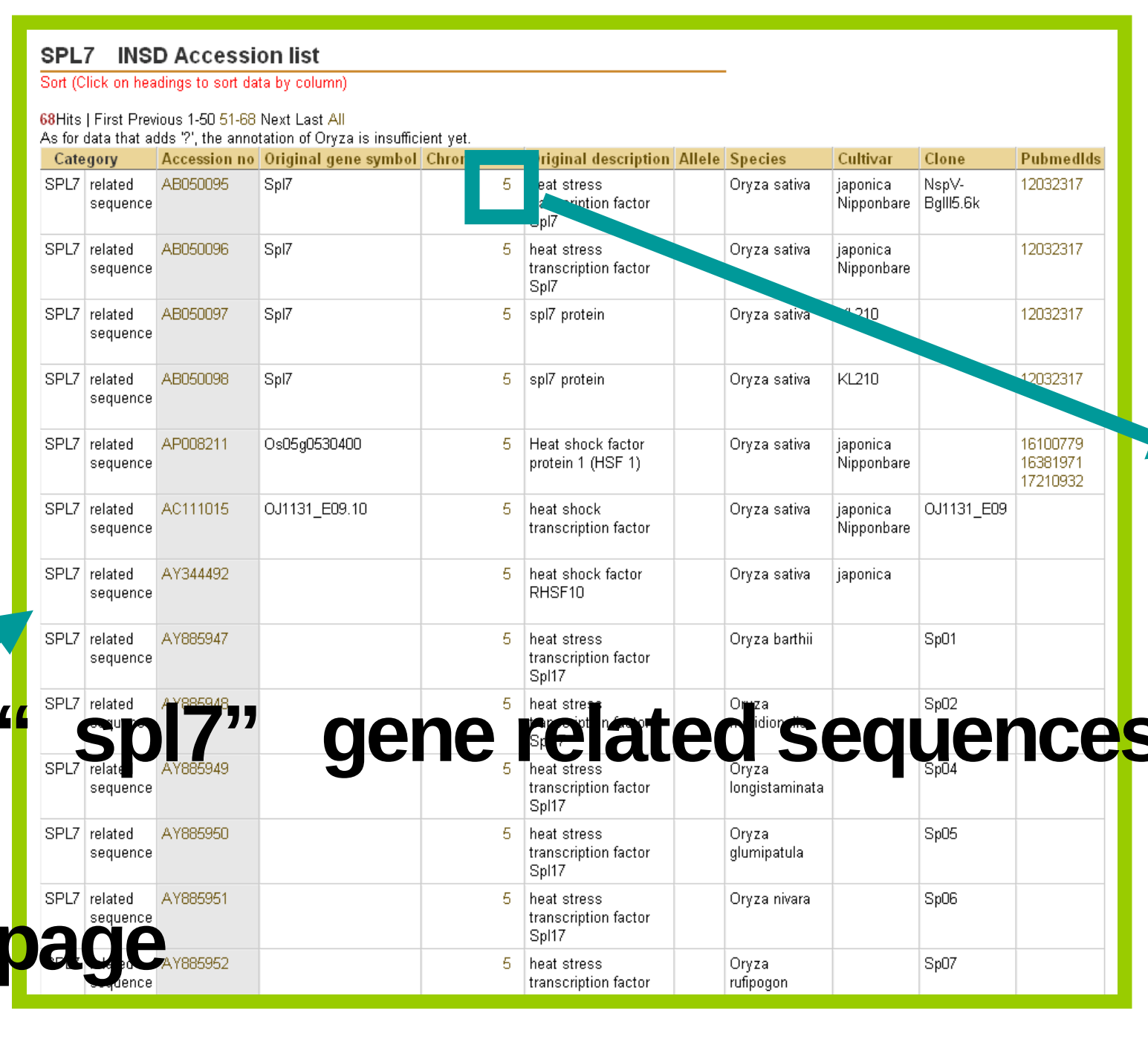
Click to full size image in new window

spl-7 Normal

Photo from A. Yoshimura

References	
References	RGN 11 116-117.(1994) Japan. J. Breed. 27(Suppl. 1) 250-251.(1977) J. Fac. Agr. Kyushu Univ. 22 243-251.(1978) Japan. J. Breed. 31(Suppl. 1) 66-67.(1981) J. Fac. Agr. Kyushu Univ. 29(2.3) 139-144.(1984) Japan. J. Breed. 40(Suppl. 2) 268-269.(1990) Plant Sci. 163(2) 345-353(2002) Proc. Natl. Acad. Sci. U.S.A. 99 (11) 7530-7535(2002)

" spl7" gene page

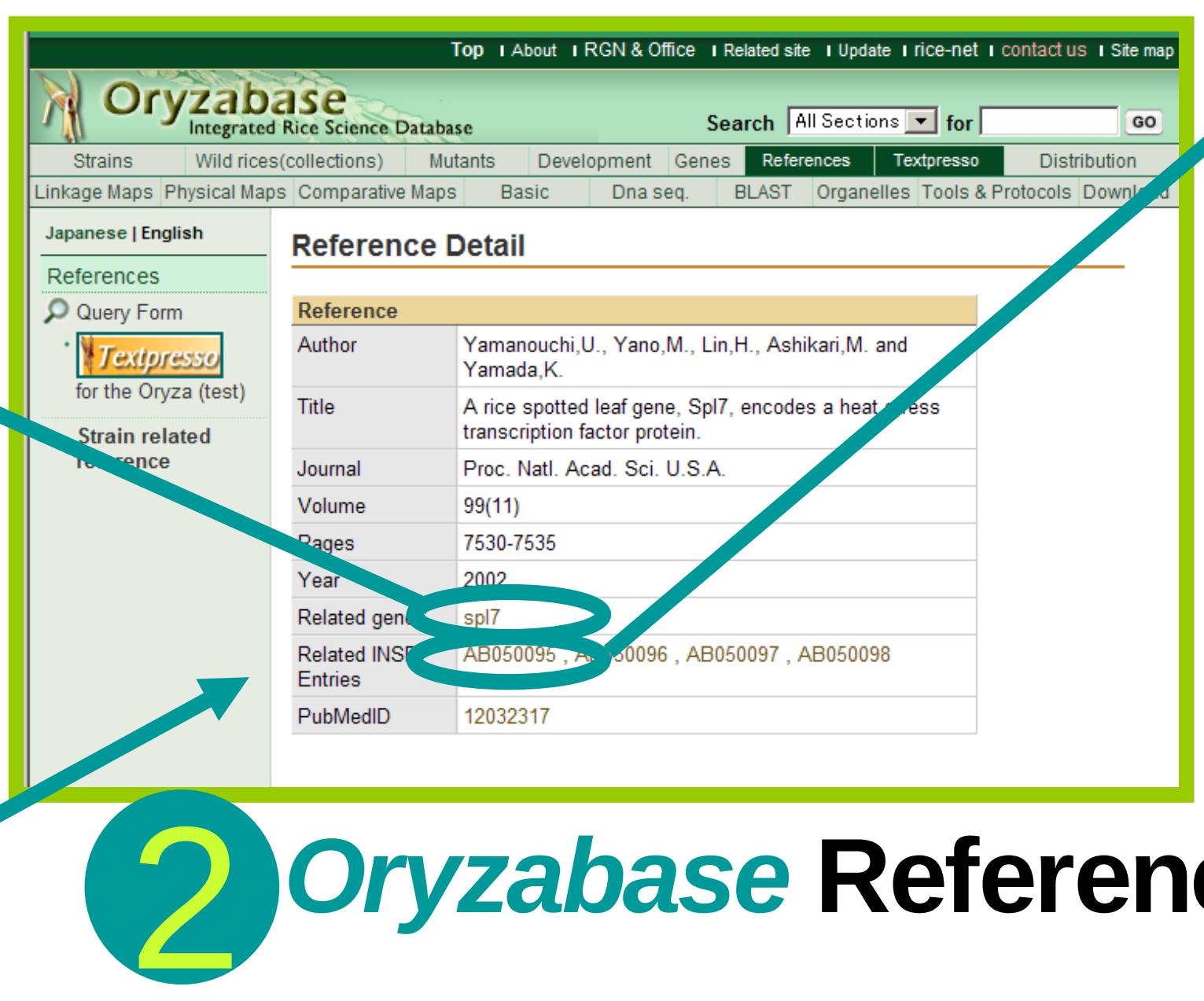

SPL7 INSD Accession list

Sort (Click on headings to sort data by column)

68Hits | First Previous 1-50 51-68 Next Last All
As for data that adds '?', the annotation of Oryza is insufficient yet.

Category		Accession no	Original gene symbol	Chromosome	Original description	Allele	Species	Cultivar	Clone	PubmedIds
SPL7	related sequence	AB050095	Spl7	5	heat stress transcription factor		Oryza sativa	japonica Nipponbare	NspV-BglII5.6k	12032317
SPL7	related sequence	AB050096	Spl7	5	heat stress transcription factor Spl7		Oryza sativa	japonica Nipponbare		12032317
SPL7	related sequence	AB050097	Spl7	5	spl7 protein		Oryza sativa	KL210		12032317
SPL7	related sequence	AY885949		5	heat stress transcription factor Spl17		Oryza longistaminata		Sp04	
SPL7	related sequence	AY885950		5	heat stress transcription factor Spl17		Oryza glumipatula		Sp05	
SPL7	related sequence	AY885951		5	heat stress transcription factor Spl17		Oryza nivara		Sp06	
SPL7	related sequence	AY885952		5	heat stress transcription factor		Oryza rufipogon		Sp07	

" spl7" gene related sequences

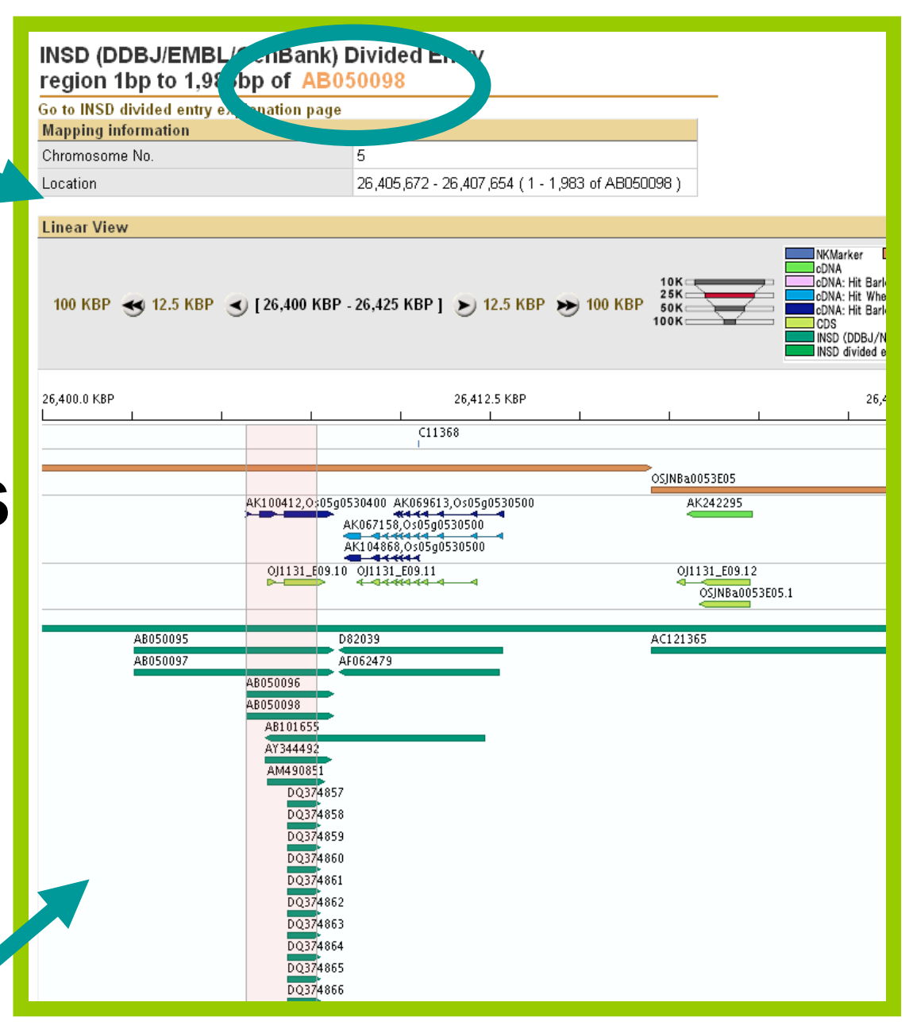

Top | About | RGN & Office | Related site | Update | rice-net | contact us | Site map
Oryzabase
Integrated Rice Science Database
Search All Sections for GO
Strains | Wild rices(collections) | Mutants | Development | Genes | References | Textpresso | Distribution
Linkage Maps | Physical Maps | Comparative Maps | Basic | Dna seq. | BLAST | Organelles | Tools & Protocols | Download
Japanese | English
References
Query Form
Textpresso for the Oryza (test)
Strain related reference

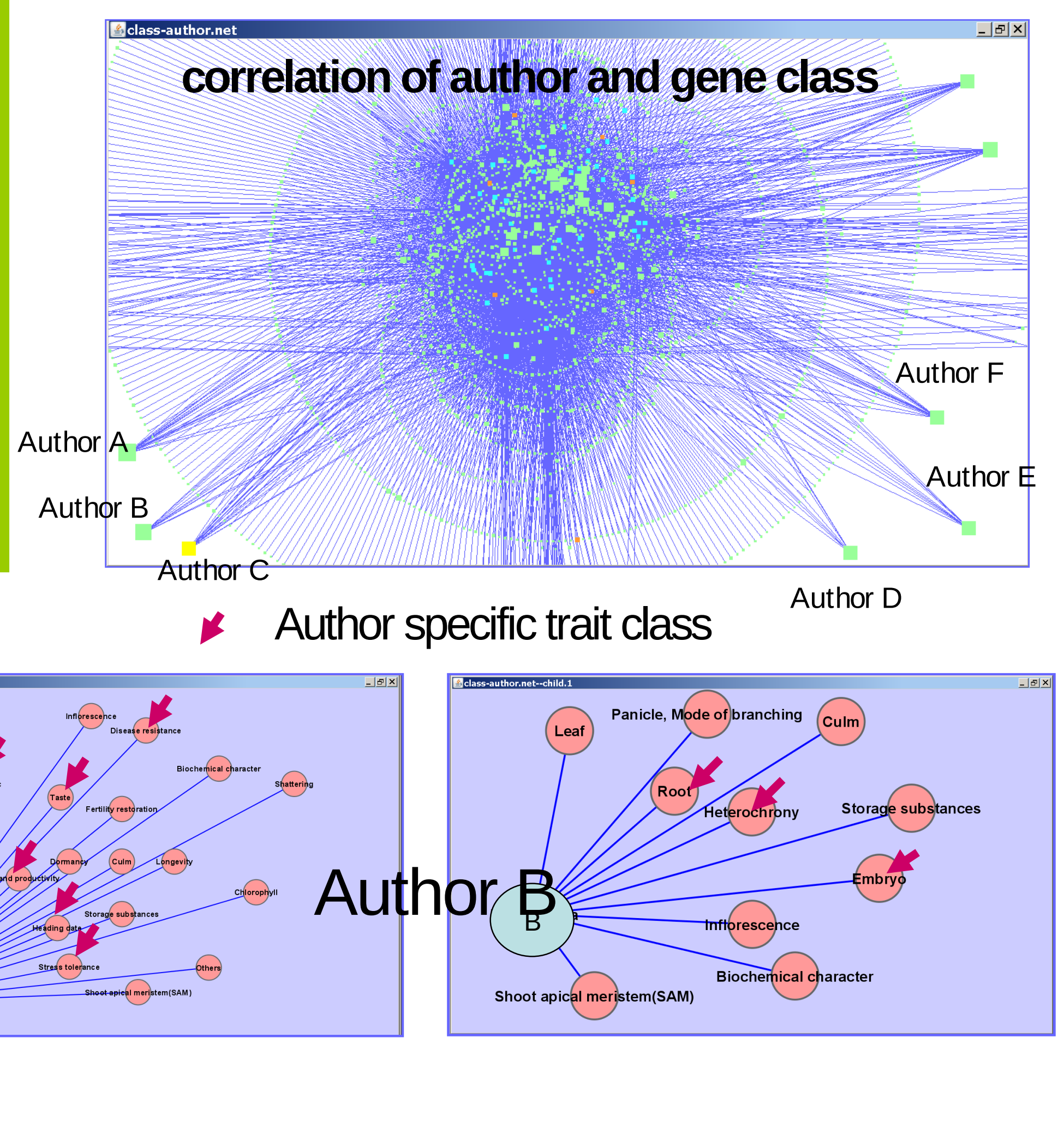
Reference Detail

Reference	
Author	Yamanouchi,U., Yano,M., Lin,H., Ashikari,M. and Yamada,K.
Title	A rice spotted leaf gene, Spl7, encodes a heat stress transcription factor protein.
Journal	Proc. Natl. Acad. Sci. U.S.A.
Volume	99(11)
Pages	7530-7535
Year	2002
Related genes	spl7
Related INSD Entries	AB050095 , AB050096 , AB050097 , AB050098
PubMedID	12032317

2 Oryzabase Reference

Oryzabase Genome viewer


INSD (DDBJ/EMBL/GenBank) Divided Entry
region 1bp to 1,9..bp of AB050098
Go to INSD divided entry explanation page


correlation of author and gene class
Author A
Author B
Author C
Author D
Author E
Author F
Author specific trait class
Author A
Author B
Panicle, Mode of branching
Culm
Leaf
Root
Heterochrony
Storage substances
Embryo
Inflorescence
Biochemical character
Shoot apical meristem(SAM)