The 63rd Annual Meeting of JSPP
The 18th Database Workshop
March 24, 2022

Literature-based curation of gene annotation, genomic variations, and transcriptomics information in RAP-DB

Yoshihiro Kawahara[1], Tomoko Hirozane-Kishikawa[1], Xiaohui Wang[1], Ryo Hirata[2]

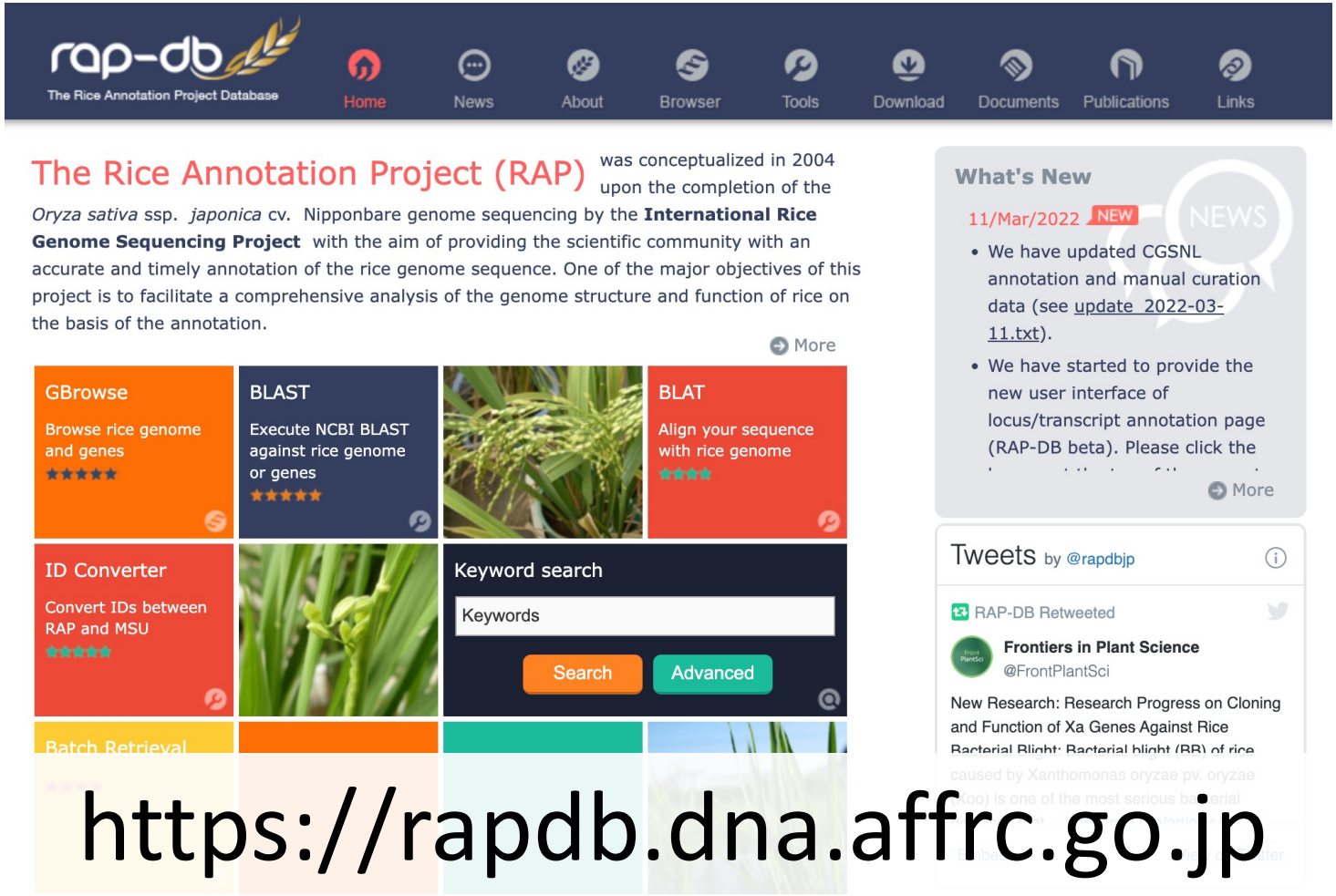
([1]Advanced Analysis Center, NARO (NAAC), [2]IMSBIO Co., Ltd.)

NARO

Rice Annotation Project Database (RAP-DB)

農研機構 NARO

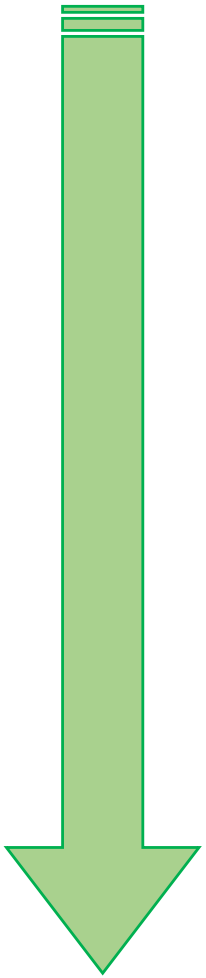
Over the past 15 years, through several updates, RAP-DB has continued to provide rice genome annotation.



https://rapdb.dna.affrc.go.jp

History of RAP-DB

Over the past 15 years, through several updates, RAP-DB has continued to provide rice genome annotation.



- **2005**: International Rice Genome Sequencing Project, ***Nature***

  The map-based sequence of the rice genome

- **2006**: Ohyanagi, H., et al., ***Nucleic Acids Res.***

  The rice annotation project database (RAP-DB): hub for Oryza sativa ssp. japonica genome information

- **2008**: Rice Annotation Project, ***Nucleic Acids Res***.

  The Rice Annotation Project Database (RAP-DB): 2008 update

- **2013**: Kawahara, Y., et al., ***Rice***

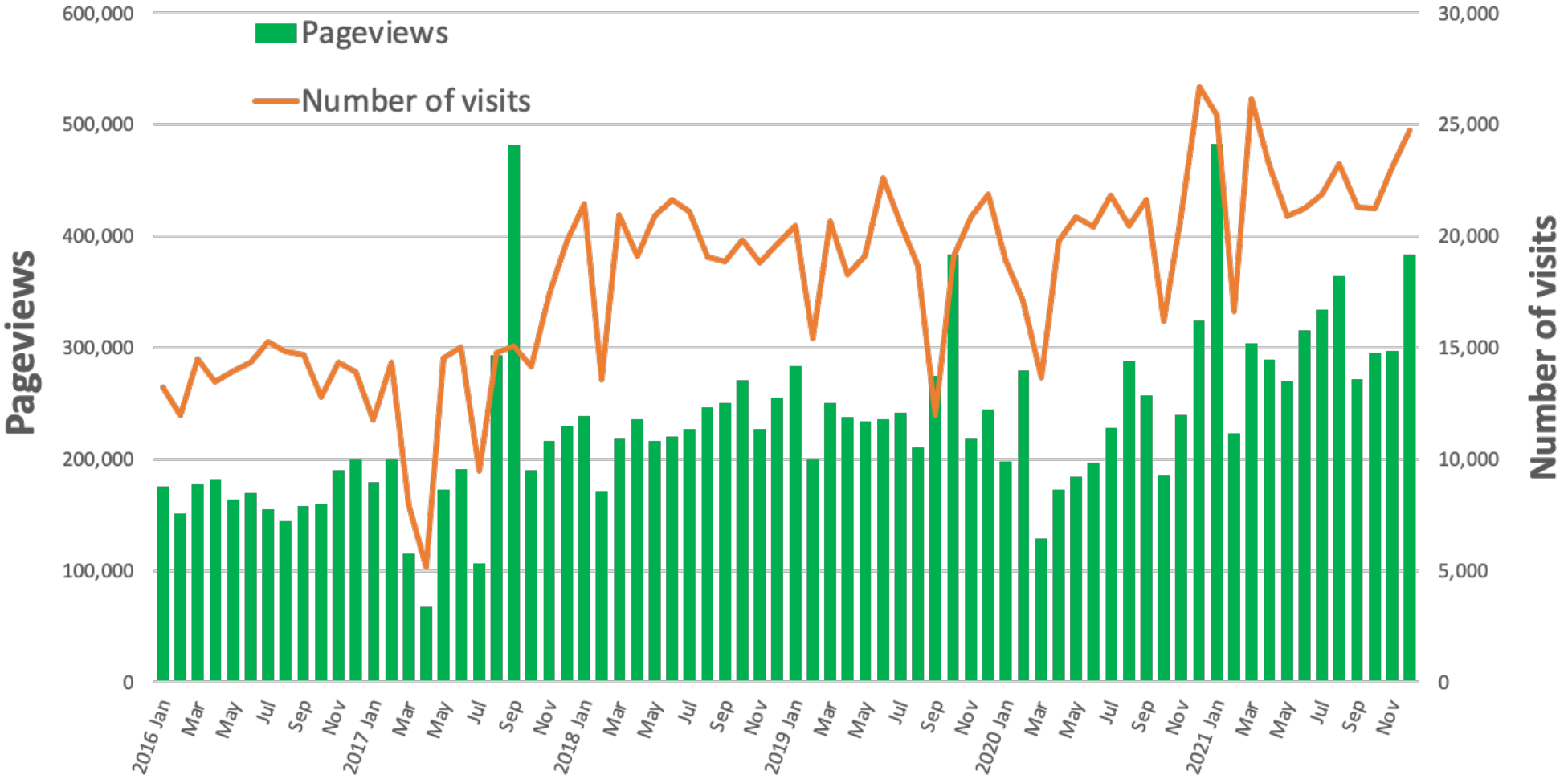
  Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data

- **2013**: Sakai, H., et al., ***Plant & Cell Physiol.***

  Rice Annotation Project Database (RAP-DB): An integrative and interactive database for rice genomics

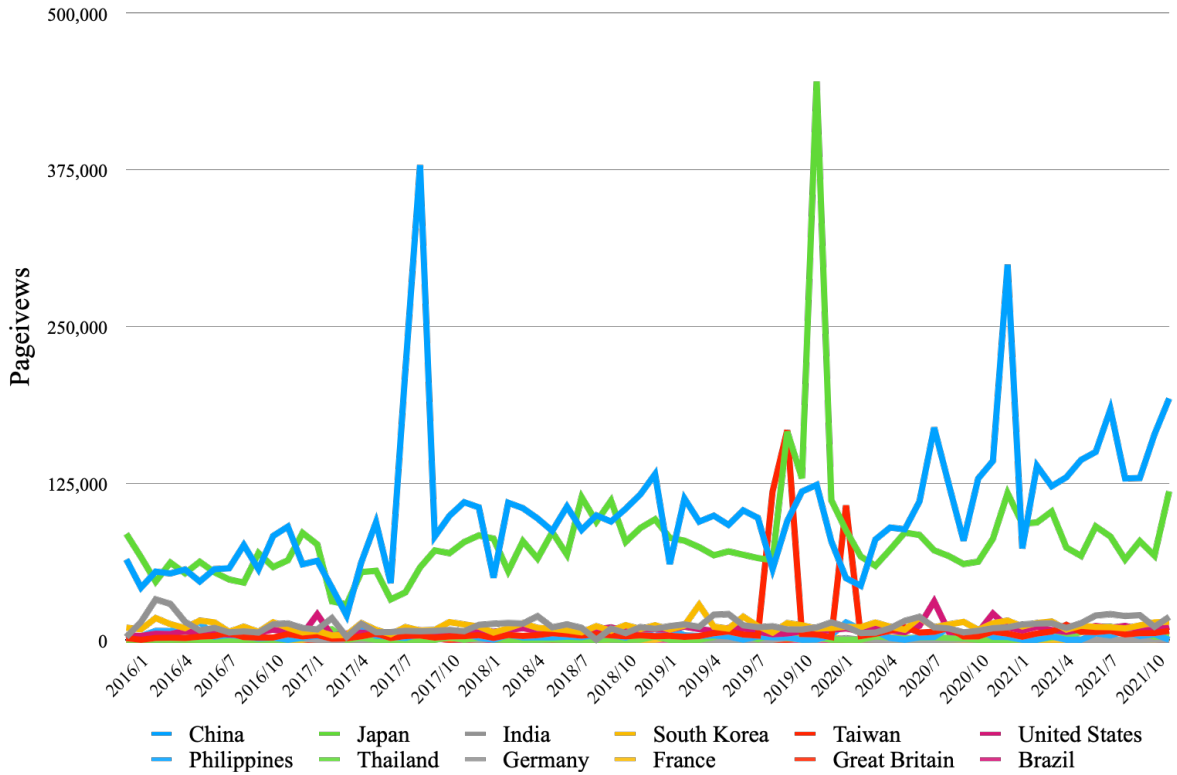
Monthly accesses

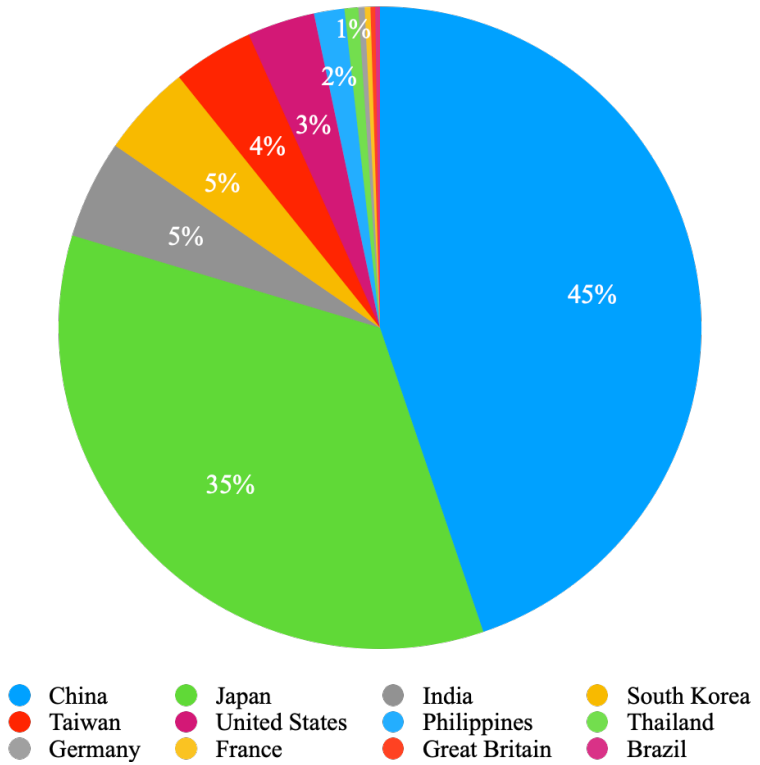
approximately 300,000 pageviews and >20,000 visits per month



Users by country

The most accesses are from China, and that from Japan is second.
This trend has started from mid-2017.





Shown only countries with an average monthly pageviews > 500

Members of RAP-DB team

Leader

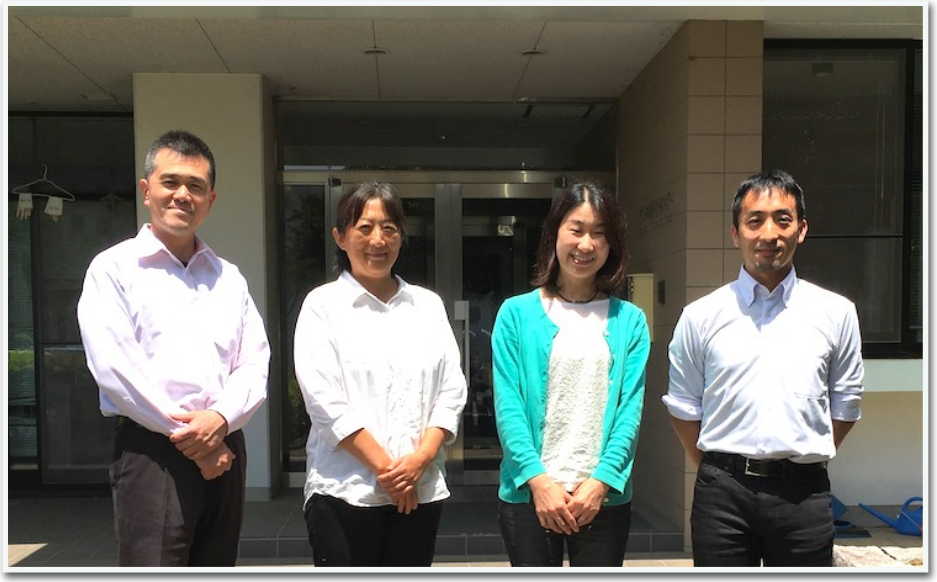
Yoshihiro Kawahara (NAAC)

Curators

Tomoko Hirozane-Kishikawa (NAAC)
Xiaohui Wang (NAAC)

Programmer

Ryo Hirata (IMSBIO Co., Ltd.)



Adviser

Takeshi Itoh (National Taiwan University)

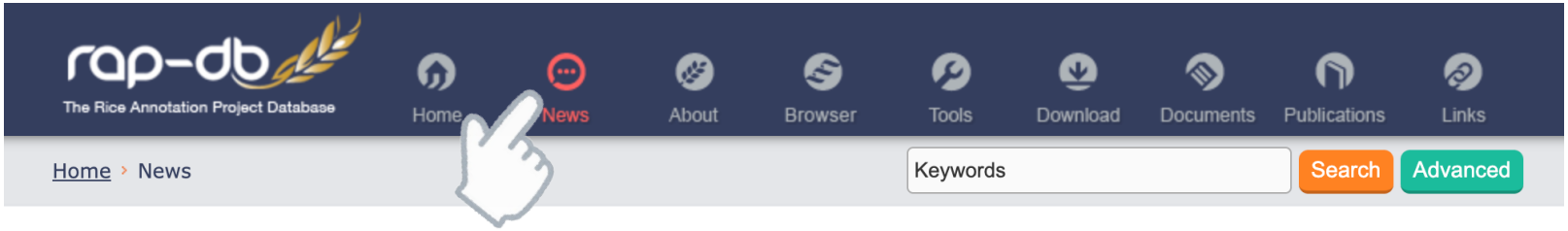
Collaboraters

- Shoko Kawamoto, Yutaka Sato (Oryzabase@NIG)
- Pankaj Jaiswal, Sushma Naithani (Gramene)

This work is supported by

- Grant-in-Aid for Publication of Scientific Research Results (Databases, 2017-2021)
- MAFF Project (DIT2001, 2018-2022)

RAP-DB was updated on March 11



2022

March 11

- We have updated CGSNL annotation and manual curation data (see update_2022-03-11.txt).
- We have started to provide the new user interface of locus/transcript annotation page (RAP-DB beta). Please click the banner at the top of the current annotation pages to see it (e.g. Os01g0911700, Os01t0911700-01).
- Transcriptome information previously provided from TENOR, RiceXPro and RiceFREND has been integrated into RAP-DB beta. In addition, publicly available RNA-Seq data of 565 experiments can be viewed. See the "Expression (TENOR) " section of transcript annotation page and the "Expression (RiceXpro)" and "CoExp Network" section of locus page in RAP-DB beta (e.g. Os06t0133000-01, Os06g0133000). RNA-Seq data is also available in JBrowse.
- Genome-wide variation data among 685 rice varieties is available in the latest version of TASUKE+ (Browser > TASUKE+ (685 varieties)).
- The list of "Agronomically important genes" has been updated (Documents > Agri. Genes). In addition, we have started to provide known functional alleles and mutations for the agronomically important genes (see the "Diversity" section of transcript annotation pages in RAP-DB beta, e.g., Os05t0158500-01).

2021

November 11

- We have updated CGSNL annotation and manual curation data (see update_2021-11-11.txt).
- New track for IsoSeq was added to JBrowse and GBrowse.

May 10

- We have updated CGSNL annotation and manual curation data (see update_2021-05-10.txt).
- Gene expression profiles and co-expression networks in RiceXPro and RiceFREND can now be displayed in the “Expression” and “CoExp Network” tabs for each transcript (e.g. Os09t0565200-02).
- Gene annotations of the rice SDRLK gene subfamily have been updated based on the information reported in Naithani et al. 2021.

2020

December 02

- We have updated CGSNL annotation and manual curation data (see update_2020-12-02.txt).
- “FAQ” is now available. (Documents > FAQ)

Take-home messages

RAP-DB …

- is the only primary database for rice that continues to update gene annotation data by manual curation.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integrated into RAP-DB.

- provides genome-wide variations among >600 rice varieties through TASUKE+.

- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

Take-home messages

RAP-DB …

- is the only primary database for rice that continues to update ge Functional annotation ion.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integr Gene expression

- provides genome-wide variations among >600 rice varieties thro TASUKE+ Variations (alleles & mutations)

- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

Take-home messages

RAP-DB …

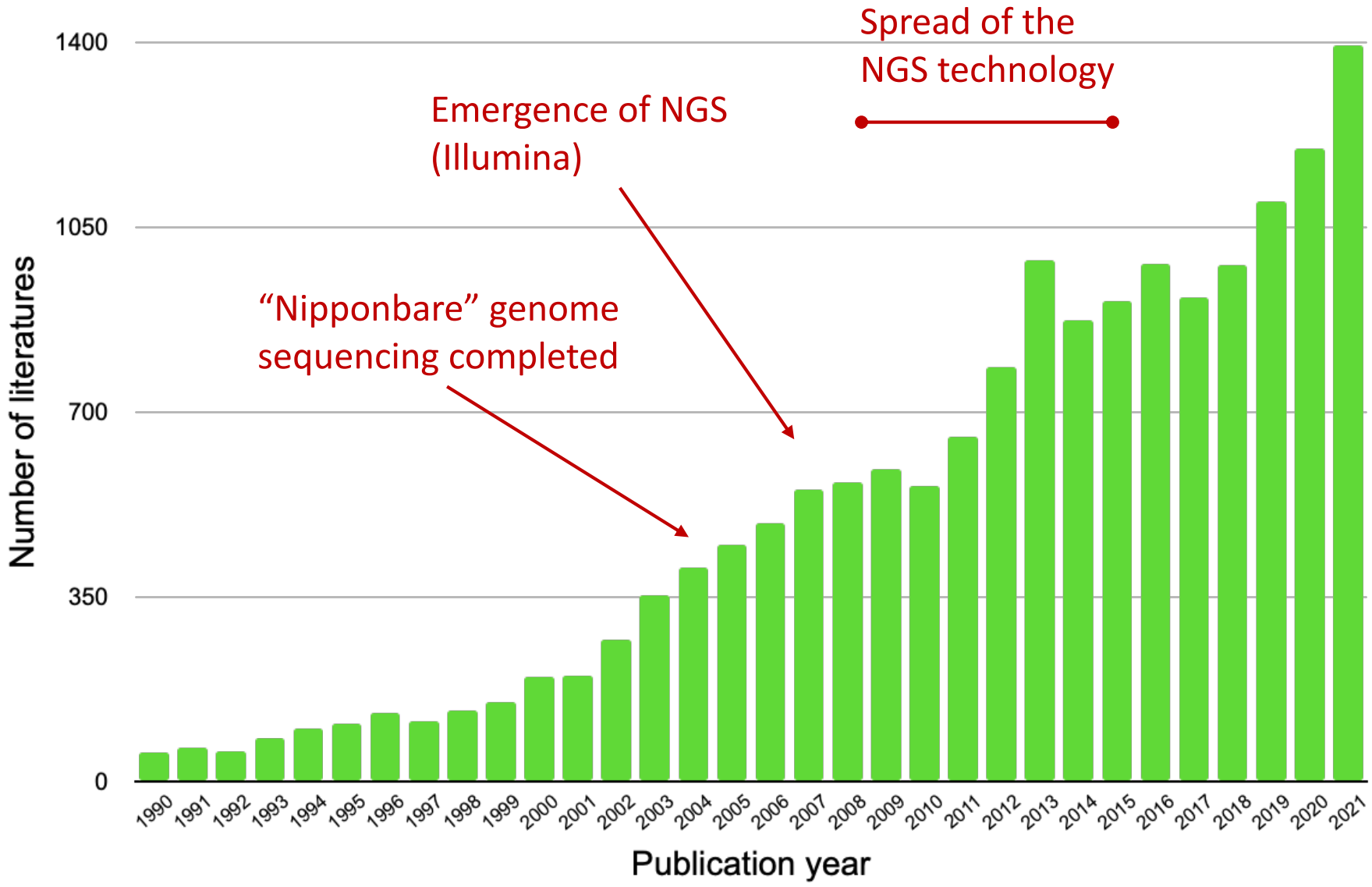
- is the only primary database for rice that continues to update gene annotation data by manual curation.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integrated into RAP-DB.

- provides genome-wide variations among >600 rice varieties through TASUKE+.

- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

The number of literatures on “rice genes” continues to grow



The literatures on “rice genes” were searched by Pubmed using “rice” and “gene” as keywords

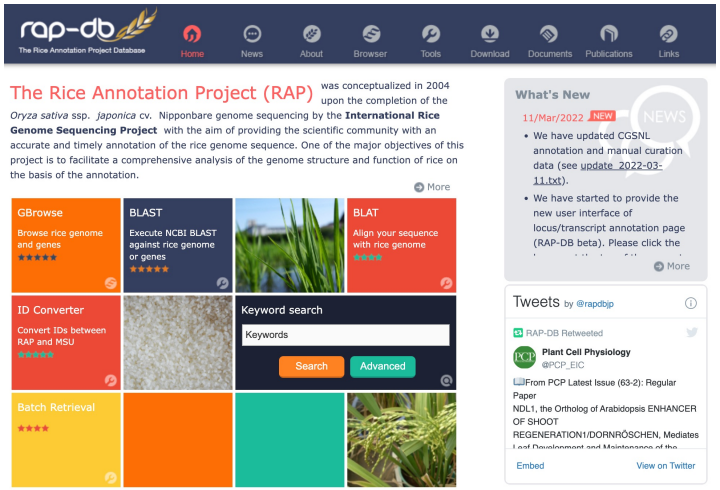
Literature-based curation of gene function and structures

農研機構 NARO

Collect literatures describing rice genes and its function

PubMed.gov

Oryzabase sponsored by NBRP
Integrated Rice Science Database

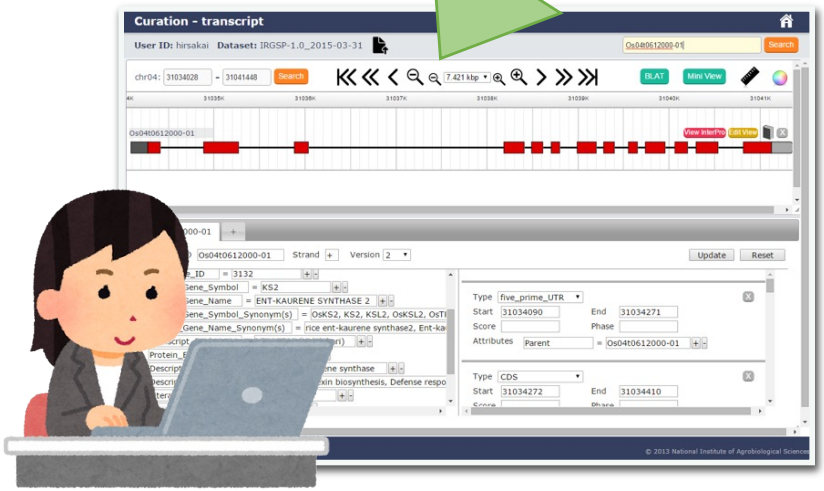


Gene annotation in RAP-DB is updated twice a year.

So far, >4,000 genes have been manually curated (May 2009–March 2022).

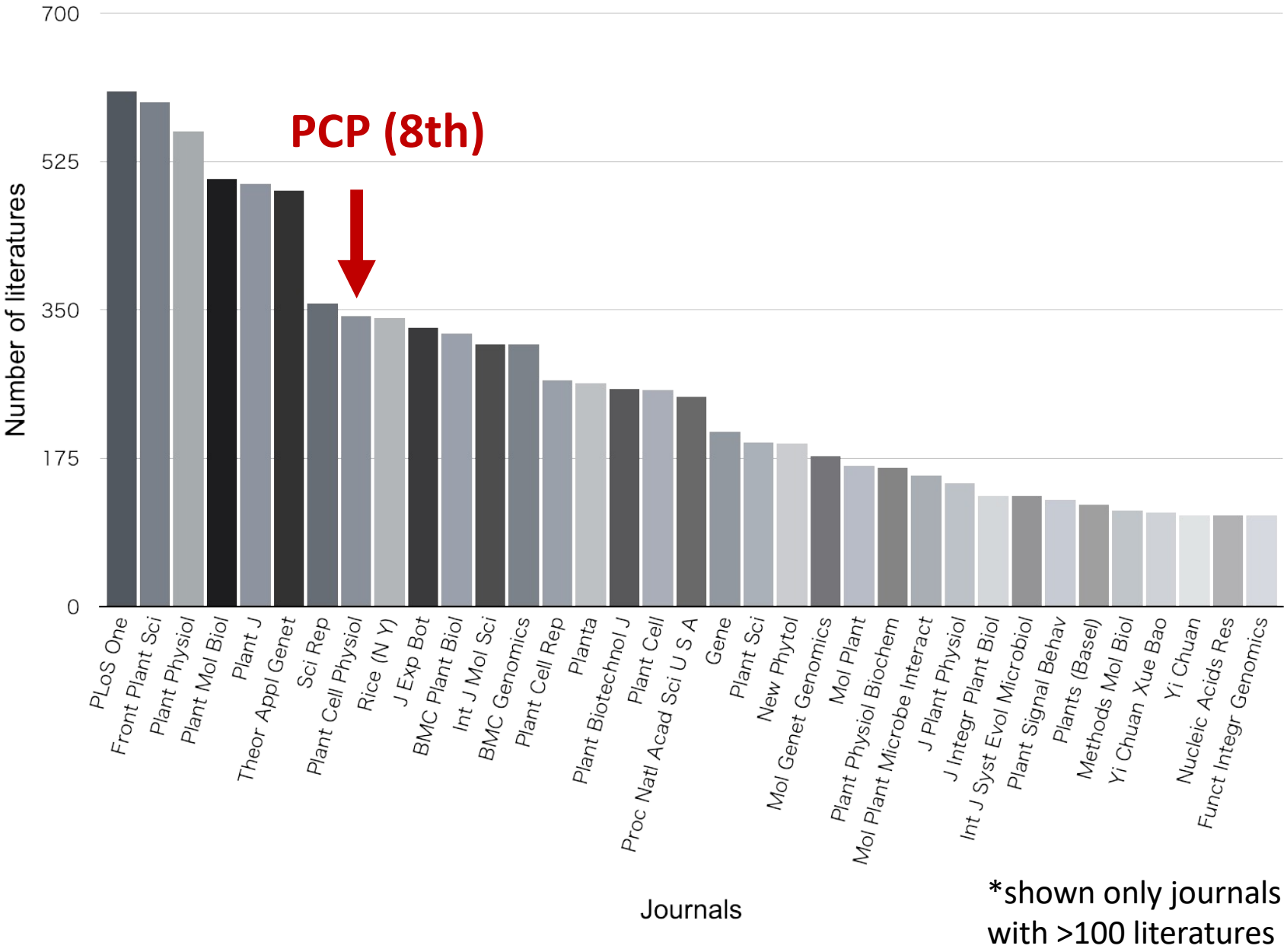
Extract structural and functional information on rice genes through a careful review of the literature.

Curators



Edit rice gene annotation through in-house curation system.11

The number of literatures on “rice gene” by journal



*shown only journals with >100 literatures

Take-home messages

RAP-DB …

- is the only primary database for rice that continues to update gene annotation data by manual curation.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integrated into RAP-DB.

- provides genome-wide variations among >600 rice varieties through TASUKE+.

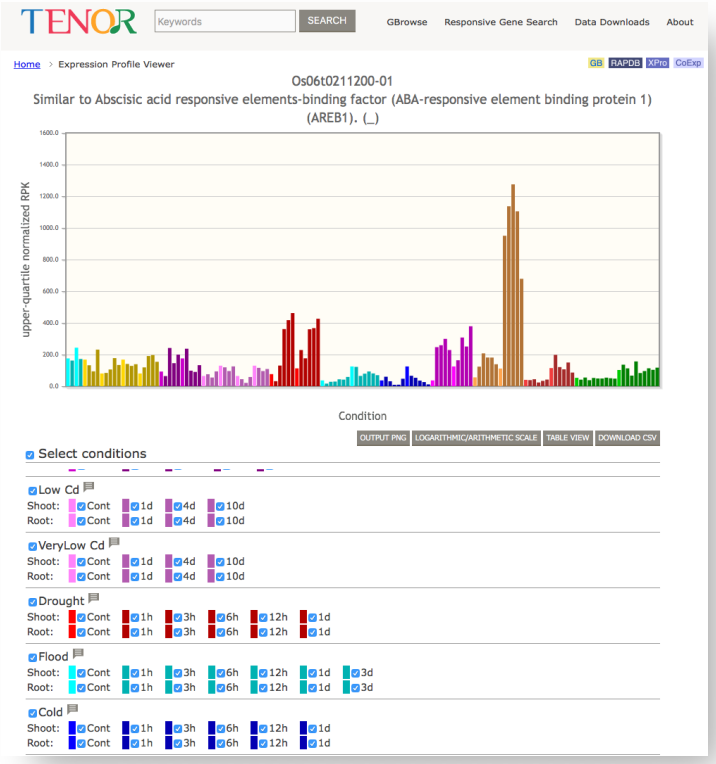
- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

Rice gene expression databases

Those expression data has been integrated into RAP-DB.

TENOR

RNA-Seq data under 140 environmental stresses and plant hormone treated conditions.



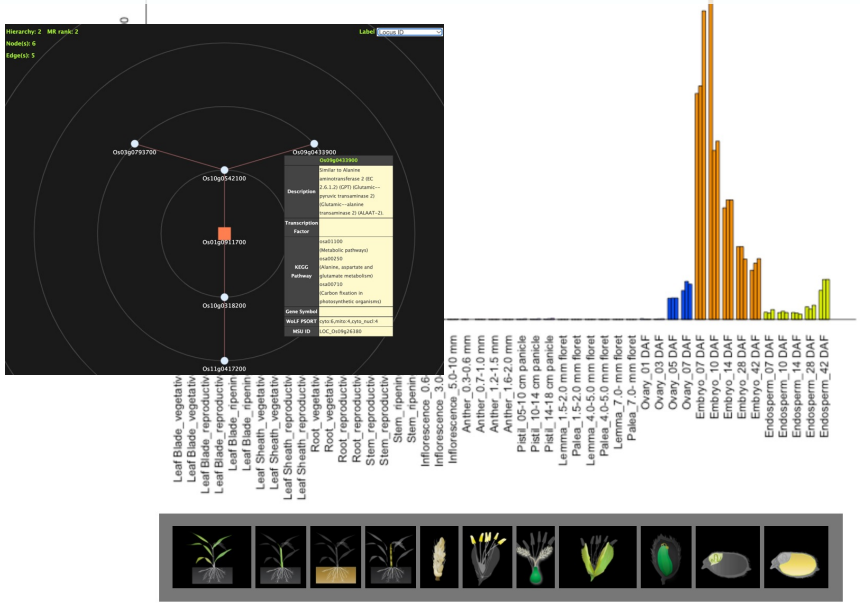
Kawahara Y. et al. 2016 Plant Cell Physiol.

RiceXPro

Microarray data of tissues/organs encompassing the entire growth of the rice plant under natural field conditions, etc.

RiceFREND

A platform for retrieving coexpressed gene networks in rice

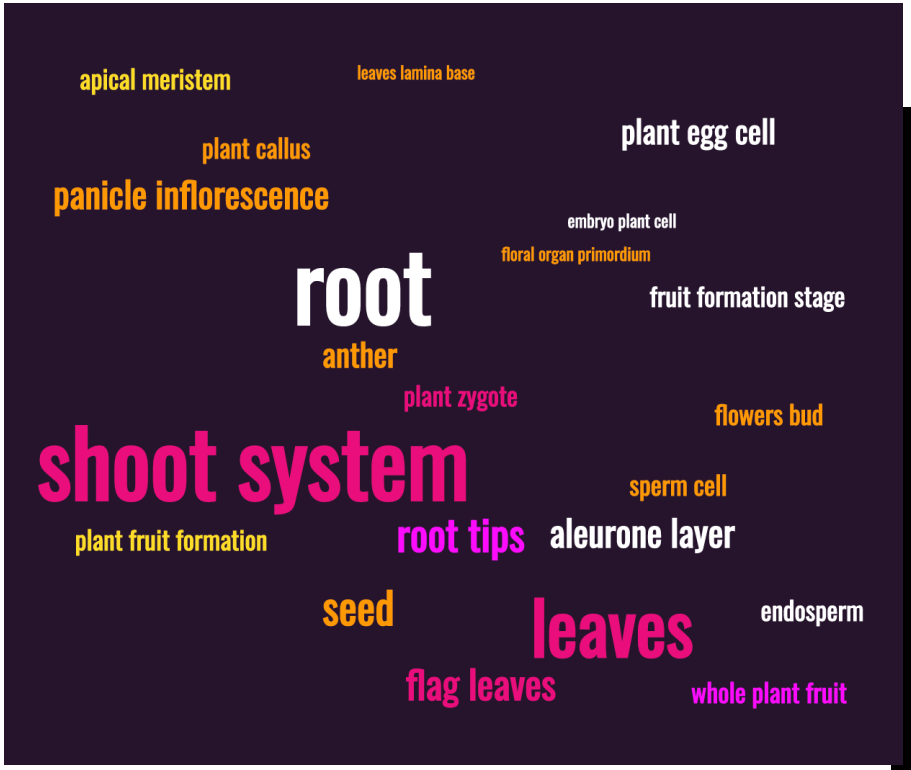


Sato Y. et al. 2013 Nucleic Acids Res.

Update of TENOR

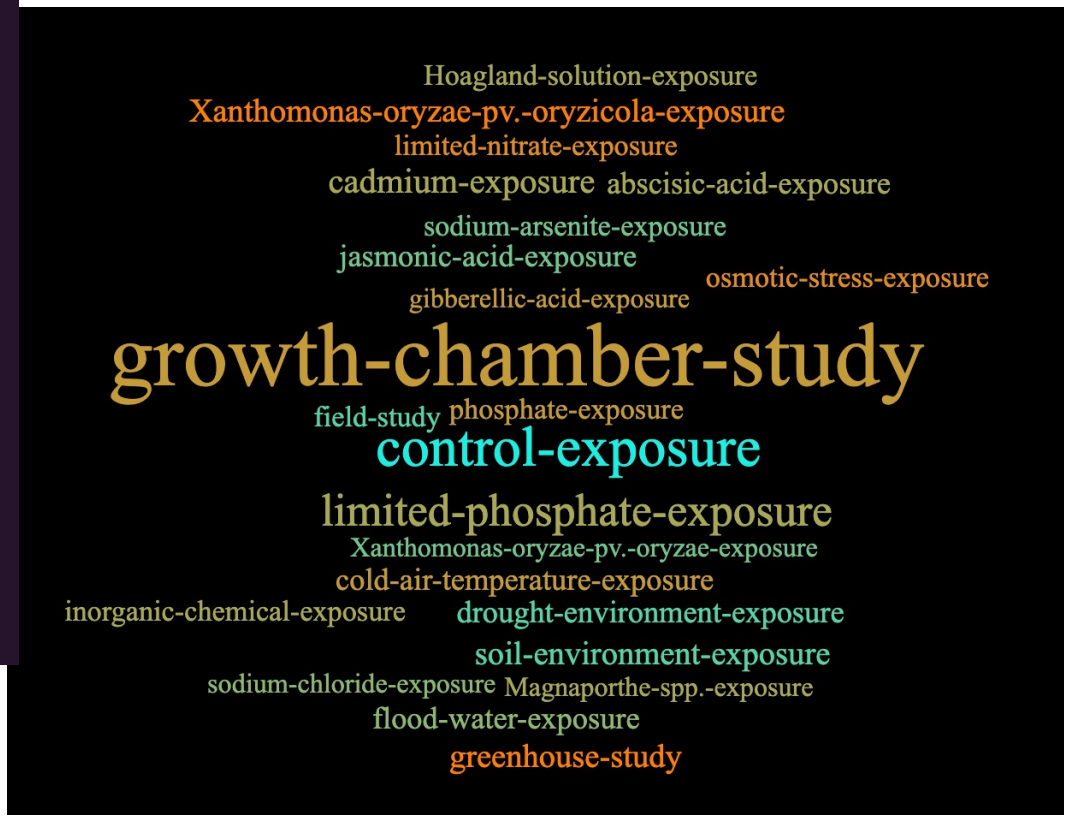
Published RNA-Seq data of 565 experiments was collected and meta-data was curated.

The expression profiles and read alignments are provided in RAP-DB.



Analyzed tissues by Trait Ontology (TO)

Analyzed conditions by Plant Experimental Conditions Ontology (PECO)



Demo:

Viewing gene annotation in RAP-DB

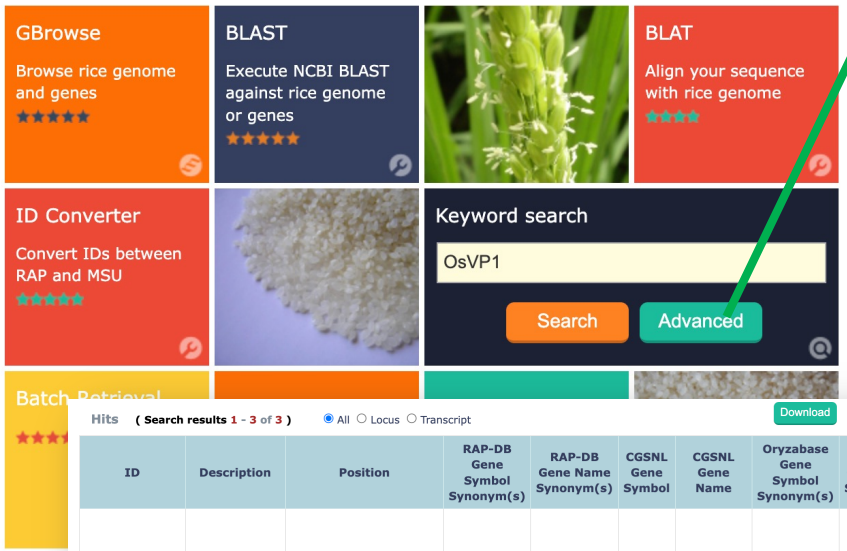
- Search by Gene ID, symbol, name, etc.
- locus/transcript annotation page (beta version)

Search by Gene ID, symbol, name, etc.

The Rice Annotation Project (RAP)

was conceptualized in 2004 upon the completion of the *Oryza sativa* ssp. *japonica* cv. Nipponbare genome sequencing by the **International Rice Genome Sequencing Project** with the aim of providing the scientific community with an accurate and timely annotation of the rice genome sequence. One of the major objectives of this project is to facilitate a comprehensive analysis of the genome structure and function of rice on the basis of the annotation.

More



Advanced keyword search

Search string

Search all words (AND search): germination
Search exact phrase (Phrase search):
Search any words (OR search):
Exclude words (NOT search):

Search attributes

- [ ] Gene identifier (e.g. Os01g0100600, Os01t0100600-01)
- [x] Gene description
- [ ] RAP-DB Gene Symbol Synonym(s)
- [ ] RAP-DB Gene Name Synonym(s)
- [ ] CGSNL Gene Symbol
- [ ] CGSNL Gene Name
- [ ] Oryzabase Gene Symbol Synonym(s)
- [ ] Oryzabase Gene Name Synonym(s)
- [ ] Gene ontology (e.g. nucleic acid binding, GO:0003676)
- [ ] Functional domain (InterPro) (e.g. Single-stranded nucleic acid binding R3H, IPR001374)
- [ ] Transcript evidence (e.g. AK121339, EU948781)
- [ ] Literature PMID (e.g. 22889013)
- [ ] Oryzabase ID (e.g. 11455)

Output options

Results

Search

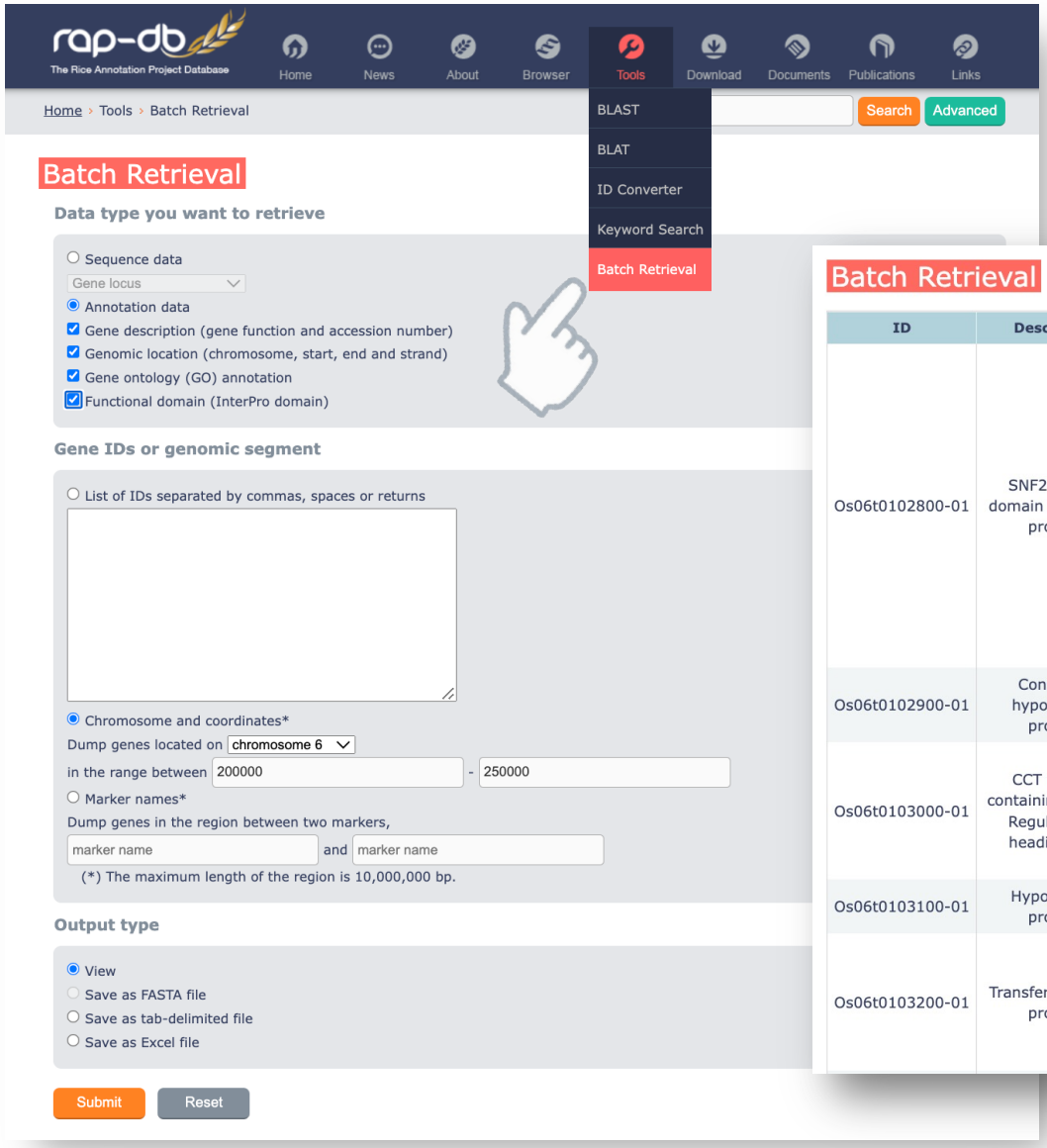
Hits (Search results 1 - 3 of 3) All Locus Transcript Download

ID	Description	Position	RAP-DB Gene Symbol Synonym(s)	RAP-DB Gene Name Synonym(s)	CGSNL Gene Symbol	CGSNL Gene Name	Oryzabase Gene Symbol Synonym(s)	Oryzabase Gene Name Synonym(s)
Os01t0911700-02	Transcription activator VP1-rice.	chr01:39723171..39726874			VP1	VIVIPAROUS 1	Vp1* (OSVP1), OsVP1, OsVp1, Vp1, ABI3, OsABI3, OsLFL4, OsVp-1, OSVP1	Viviparous-1, Protein viviparous homolog, B3 domain-containing protein VP1, Transcription activator VP1-rice, LEAFY COTYLEDON 2 and FUSCA 3-LIKE 4
Os01g0911700	B3 domain-containing transcriptional activator, Control of seed dormancy (Os01t0911700-01);Transcription activator VP1-rice. (Os01t0911700-02)	chr01:39723171..39726984	OsVP1	Viviparous1	VP1	VIVIPAROUS 1	Vp1* (OSVP1), OsVP1, OsVp1, Vp1, ABI3, OsABI3, OsLFL4, OsVp-1, OSVP1	Viviparous-1, Protein viviparous homolog, B3 domain-containing protein VP1, Transcription activator VP1-rice, LEAFY COTYLEDON 2 and FUSCA 3-LIKE 4
Os01t0911700-01	B3 domain-containing transcriptional activator, Control of seed dormancy	chr01:39723186..39726984	OsVP1	Viviparous1	VP1	VIVIPAROUS 1	Vp1* (OSVP1), OsVP1, OsVp1, Vp1, ABI3, OsABI3, OsLFL4, OsVp-1, OSVP1	Viviparous-1, Protein viviparous homolog, B3 domain-containing protein VP1, Transcription activator VP1-rice, LEAFY COTYLEDON 2 and FUSCA 3-LIKE 4

Hits (Search results 1 - 10 of 126) All Locus Transcript Download

ID	Description	Position	RAP-DB Gene Symbol Synonym(s)	RAP-DB Gene Name Synonym(s)	CGSNL Gene Symbol	CGSNL Gene Name	Oryzabase Gene Symbol Synonym(s)	Oryzabase Gene Name Synonym(s)
Os01t0111600-01	PEBP (phosphatidylethanolamine-binding protein) family protein, Regulation of ABA signaling-mediated seed germination	chr01:612564..615631	OsMFT2	MOTHER OF FT AND TFL 2	MFT2	MOTHER OF FT AND TFL 2	OsMFT2, OsMFT	MFT-like gene 2
Os01g0111600	PEBP (phosphatidylethanolamine-binding protein) family protein, Regulation of ABA signaling-mediated seed germination (Os01t0111600-01)	chr01:612564..615631	OsMFT2	MOTHER OF FT AND TFL 2	MFT2	MOTHER OF FT AND TFL 2	OsMFT2, OsMFT	MFT-like gene 2
Os01g0159800	Similar to DNA binding protein. (Os01t0159800-01);Basic helix-loop-helix (bHLH) transcription factor, Mediation of seed germination, Seedling recovery from salt stress (Os01t0159800-02)	chr01:3129056..3130722	OsbHLH035	basic helix-loop-helix protein 035			OsbHLH035, bHLH035, OsbHLH35, bHLH35, bHLH107	basic helix-loop-helix protein 035
Os01t0159800-02	Basic helix-loop-helix (bHLH) transcription factor, Mediation of seed germination, Seedling recovery from salt stress	chr01:3129106..3130722	OsbHLH035	basic helix-loop-helix protein 035			OsbHLH035, bHLH035, OsbHLH35, bHLH35, bHLH107	basic helix-loop-helix protein 035
Os01g0258600	Plant-specific small protein, Hybrid incompatibility, Pollen germination (Os01t0258600-01)	chr01:8657066..8658466	DPL1	DOPPELGANGER1	DPL1	DOPPELGANGER 1		DOPPELGANGER1
Os01t0258600-01	Plant-specific small protein, Hybrid incompatibility, Pollen germination	chr01:8657066..8658466	DPL1	DOPPELGANGER1	DPL1	DOPPELGANGER 1		DOPPELGANGER1

You can get the list of genes between chromosomal positions or markers



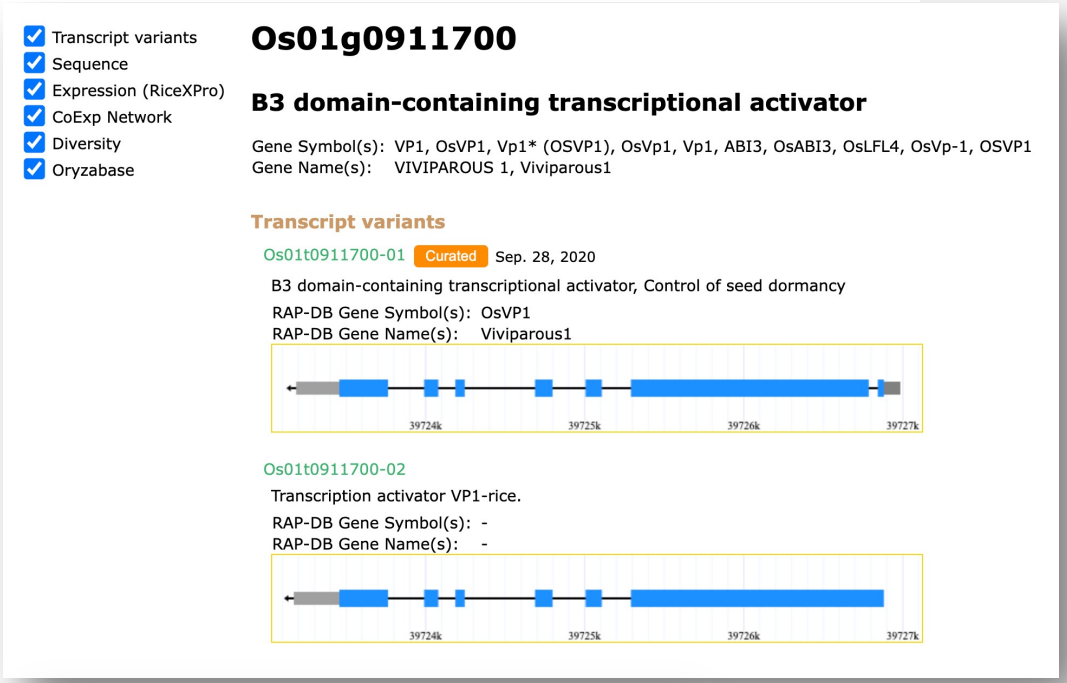
Batch Retrieval

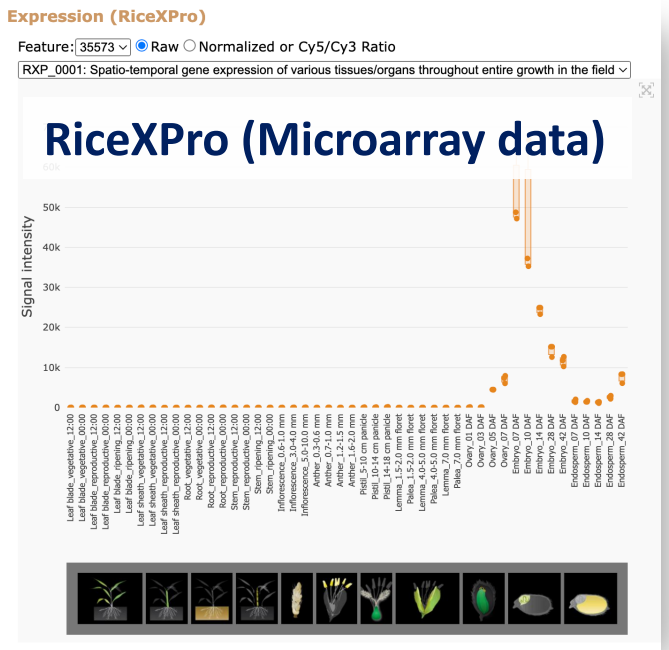
ID	Description	Chromosome	Start	End	Strand	GO	InterPro
Os06t0102800-01	SNF2-related domain containing protein.	chr06	195018	204322	+	Cellular Component: nucleus (GO:0005634), Molecular Function: DNA binding (GO:0003677), Molecular Function: protein binding (GO:0005515), Molecular Function: ATP binding (GO:0005524), Molecular Function: zinc ion binding (GO:0008270)	SNF2-related (IPR000330), Chromo domain/shadow (IPR000953), Zinc finger, PHD-type (IPR001965), Zinc finger, FYVE/PHD-type (IPR011011), Zinc finger, RING/FYVE/PHD-type (IPR013083), Chromo domain-like (IPR016197), Zinc finger, PHD-finger (IPR019787), Chromo domain (IPR023780)
Os06t0102900-01	Conserved hypothetical protein.	chr06	204487	208534	+	NONE	NONE
Os06t0103000-01	CCT domain containing protein, Regulation of heading date	chr06	209204	210107	+	Molecular Function: protein binding (GO:0005515), Molecular Function: zinc ion binding (GO:0008270)	B-box-type zinc finger (IPR000315), CCT domain (IPR010402)
Os06t0103100-01	Hypothetical protein.	chr06	212913	214438	-	NONE	NONE
Os06t0103200-01	Transferase family protein.	chr06	213020	214695	+	Molecular Function: transferase activity, transferring acyl groups other than amino-acyl groups (GO:0016747)	Transferase (IPR003480), Chloramphenicol acetyltransferase-like domain (IPR023213)

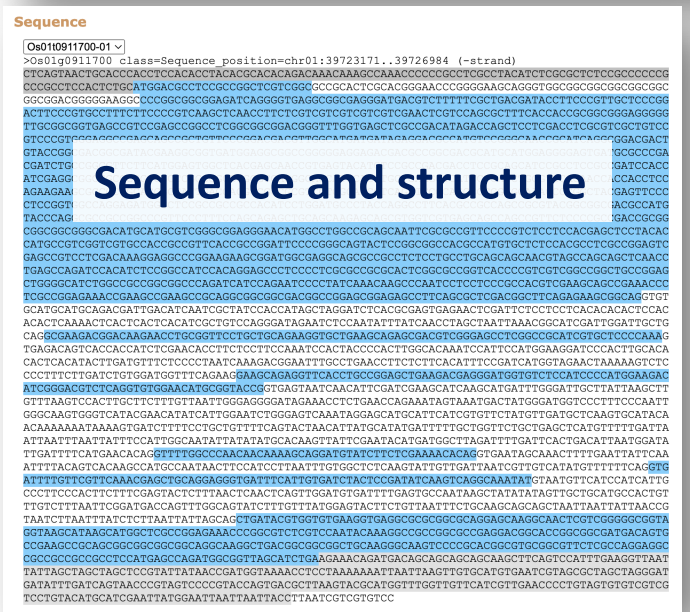
New locus/transcript annotation pages have been opened

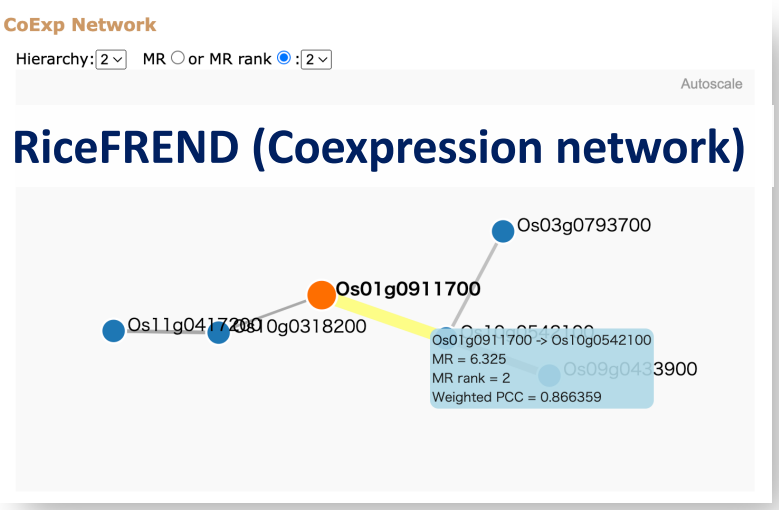
Locus annotation page (beta version)

Description, symbol, name and structures









Transcript annotation page (beta version)

農研機構 NARO

Description, symbol, name, GO, InterPro, etc.

- [x] Description
- [x] Gene Symbol(s)
- [x] Gene Name(s)
- [x] Position
- [x] Locus
- [x] Other variants
- [x] Gene ontology
- [x] Functional domain
- [x] Subcellular location
- [x] Transcript evidence
- [x] Protein evidence
- [x] Literature
- [x] Expression (TENOR)
- [x] Diversity
- [x] Oryzabase
- [x] KEGG
- [x] Sequence
- [ ] Note

Os01t0911700-01 Curated Sep. 28, 2020

Description

B3 domain-containing transcriptional activator, Control of seed dormancy

RAP-DB Gene Symbol(s)

OsVP1

RAP-DB Gene Name(s)

Viviparous1

Position

chr01:39723186.

Locus

Os01g0911700

Other variants

Os01t0911700-02

Gene ontology

Molecular Functio

Functional don

DNA-binding pseu
B3 DNA binding domain (IPR003340)

Experimentally confirmed subcellular location

nucleus (GO:0005634)

Transcript evidence

AK073805 (Oryza sativa Japonica Group, Nipponbare)
Oryza sativa Japonica Group cDNA clone:J033069I02, full insert sequence

Protein Evidence

P37398-2 (Oryza sativa subsp. japonica)
B3 domain-containing protein VP1

Literature

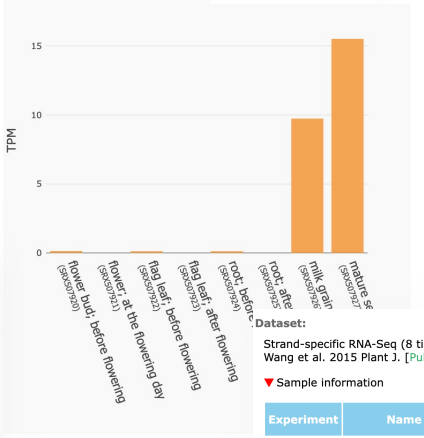
"OsVP1 activates Sdr4 expression to control rice seed dormancy via the ABA signaling pathway"
Wenqiang Chen et al. The Crop Journal, 9(1):68-78 (2021)[DOI]

TENOR (RNA-Seq data)

Expression (TENOR)

Bar Boxplot

Strand-specific RNA-Seq (8 tissues, suc



Dataset:

Strand-specific RNA-Seq (8 tissues, such as flower, leaf, root, seed, etc.)
Wang et al. 2015 Plant J. [PubMed] [DOI]

▼ Sample information

Experiment	Name	Plant Ontology (PO)	Plant Experimental Conditions Ontology (PECO)
SRX507920	flower bud; before flowering	flower bud (PO:0000056)	greenhouse study (PECO:0007248)
SRX507921	flower; at the flowering day	flower (PO:0009046)	greenhouse study (PECO:0007248)
SRX507922	flag leaf; before flowering	flag leaf (PO:0020103)	greenhouse study (PECO:0007248)
SRX507923	flag leaf; after flowering	flag leaf (PO:0020103)	greenhouse study (PECO:0007248)
SRX507924	root; before flowering	root (PO:0009005)	greenhouse study (PECO:0007248)
SRX507925	root; after flowering	root (PO:0009005)	greenhouse study (PECO:0007248)
SRX507926	milk grain	whole plant fruit formation stage (PO:0007042)	greenhouse study (PECO:0007248)
SRX507927	mature seed	seed (PO:0009010)	greenhouse study (PECO:0007248)

Genome-wide variation data
Known functional alleles & mutations

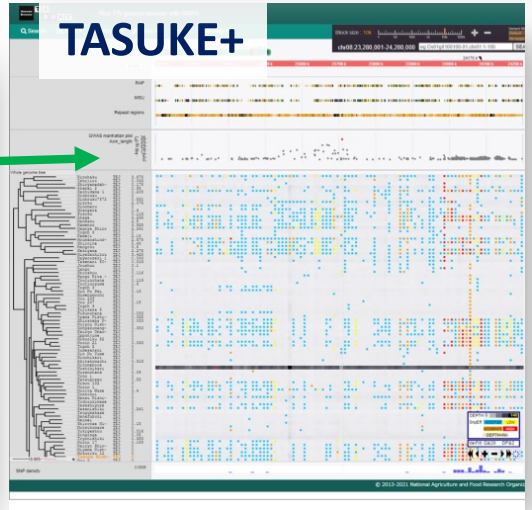
Genome-wide variation data

- TASUKE+ for RAP-DB (685 varieties)
- TASUKE+ for NARO Genebank core collection
- TASUKE+ for RAP-DB (old, 533 varieties)

Known functional alleles & mutations

Ref. CVs	Target CVs	Var. Type	HGVS (DNA)	HGVS (CDS/Protein)	Traits of Target CVs	References
Nipponbare	cr-osvp1-1 mutant	CRISPR/Cas9 mutant	g.39726701_39726702insT	p.Asp35Glufs*35	high germination rate	Wenqiang Chen et al. The Crop Journal, 9(1):68-78 (2021)[DOI]
Zhonghua 11	Osvp1 mutant	EMS mutant	-	-	increased pre-harvest sprouting; low ABA sensitivity	Wang J et al. J Agric Food Chem, 68(50):14748-14757 (2020)[PubMed]

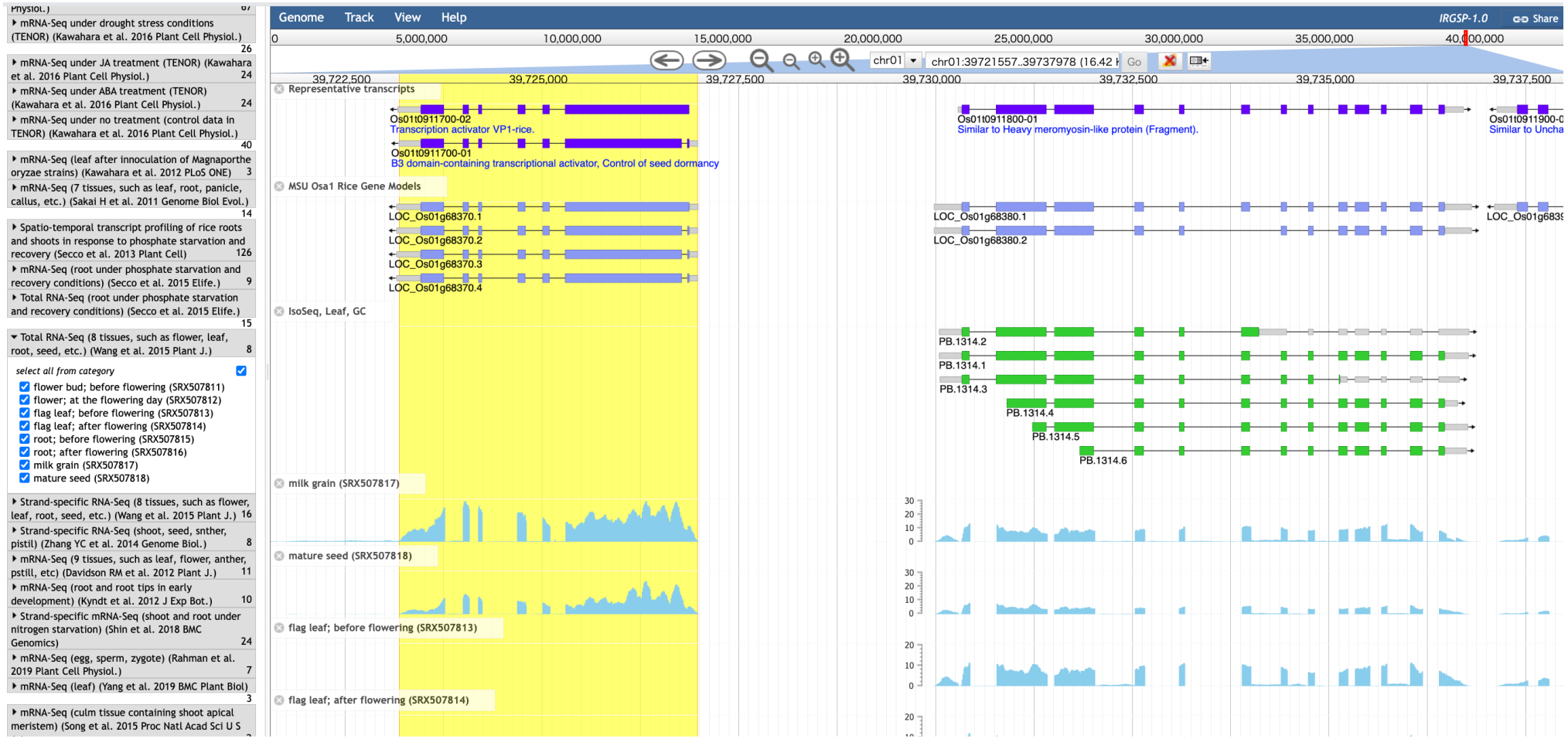
TASUKE+



JBrowse

- RAP-DB, MSU, RefSeq, IsoSeq transcripts structures
- Alignment of RNA-Seq reads

etc.



Take-home messages

RAP-DB …

- is the only primary database for rice that continues to update gene annotation data by manual curation.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integrated into RAP-DB.

- provides genome-wide variations among >600 rice varieties through TASUKE+.

- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

Towards population-level genomics

農研機構 NARO

Most resequencing data is used only once.
It could be used more effectively.

Xu et al. 2011 Nature Biotech.

Resequencing 50 accessions of cultivated and w rice yields markers for identifying agronomically important genes

50

Xun Xu[1–3,12], Xin Liu[2,12], Song Ge[4,12], Jeffrey D Jensen[5,12], Fengyi Hu[6,12], Xin Li[1,12], Yang Dong[1,12], Ryan N Gutenkunst[7], Lin Fang[2], Lei Huang[3,4], Jingxiang Li[2], Weiming He[2,8], Guojie Zhang[1,2,4], Xiaoming Zheng[3,4], Fumin Zhang[3], Yingrui Li[2], Chang Yu[2], Karsten Kristiansen[2,9], Xiuqing Zhang[2], Jian Wang[2], Mark Wright[10], Susan McCouch[10], Rasmus Nielsen[1,9,11], Jun Wang[2,9] & Wen Wang[1]

Rice is a staple crop that has undergone substantial phenotypic and physiological changes during domestication. Here we resequenced the genomes of 40 cultivated accessions selected from the major groups of rice and 10 accessions of their wild progenitors (*Oryza rufipogon* and *Oryza nivara*) to >15 × raw data coverage. We investigated genome-wide variation patterns in rice and obtained 6.5 million high-quality single nucleotide polymorphisms (SNPs) after excluding sites with missing data in any accession. Using these population SNP data, we identified thousands of genes with significantly lower diversity in cultivated but not wild rice, which represent candidate regions selected during domestication. Some of these variants are associated with important biological features, whereas others have yet to be functionally characterized. The molecular markers we have identified should be valuable for breeding and for identifying agronomically important genes in rice.

Zhao et al. 2018 Nature Genetics

Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice

63

Qiang Zhao[1], Qi Feng[1], Hengyun Lu[1], Yan Li[1], Ahong Wang[1], Qilin Tian[1], Qilin Zhan[1], Yiq
Lei Zhang[1], Tao Huang[1], Yongchun Wang[1], Danlin Fan[1], Yan Zhao[1], Ziqun Wang[1], Congcon
Jiaying Chen[1], Chuanrang Zhu[1], Wenjun Li[1], Qijun Weng[1], Qun Xu[2], Zi-Xuan Wang[1], Xinghua Wei[2],
Bin Han[1] and Xuehui Huang[1,3]*

The rich genetic diversity in *Oryza sativa* and *Oryza rufipogon* serves as the main sources in rice breeding. Large-scale resequencing has been undertaken to discover allelic variants in rice, but much of the information for genetic variation is often lost by direct mapping of short sequence reads onto the *O. sativa japonica* Nipponbare reference genome. Here we constructed a pan-genome dataset of the *O. sativa–O. rufipogon* species complex through deep sequencing and de novo assembly of 66 divergent accessions. Intergenomic comparisons identified 23 million sequence variants in the rice genome. This catalog of sequence variations includes many known quantitative trait nucleotides and will be helpful in pinpointing new causal variants that underlie complex traits. In particular, we systemically investigated the whole set of coding genes using this pan-genome data, which revealed extensive presence and absence of variation among rice accessions. This pan-genome resource will further promote evolutionary and functional studies in rice.

Tanaka et al. 2020 Nature Biotech.

Whole-Genome Sequencing of the NARO World Rice Core Collection (WRC) as the Basis for Diversity and Association Studies

119

N. Tanaka[1], M. Shenton[1], Y. Kawahara[1,2], M. Kumagai[2], H. Sakai[2], H. Kanamori[1], J. Yone
K. Sugimoto[1], M. Ishimoto[1], J. Wu[1] and K. Ebana[3,*]

[1]Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
[2]Advanced Analysis Center, National Agriculture and Food Research Organization, Tsukuba Ibaraki, 305-8517, Japan
[3]Genetic Resources Center, National Agriculture and Food Research Organization, Plant Genetic Diversity Laboratory, Tsukuba, Ibaraki 305-8502, Japan

Investigation of the Genetic Diversity of a Rice Core Collection of Japanese Landraces using Whole-Genome Sequencing

Nobuhiro Tanaka[1], Matthew Shenton[1], Yoshihiro Kawahara[1,2], Masahiko Kumagai[2], Hiroaki Sakai[2], Hiroyuki Kanamori[1], Jun-ichi Yonemaru[1], Shinichi Fukuoka[1], Kazuhiko Sugimoto[1], Masao Ishimoto[1], Jianzhong Wu[1] and Kaworu Ebana[3]*

[1]Institute of Crop Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
[2]Advanced Analysis Center, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
[3]Genetic Resources Center, National Agriculture and Food Research Organization, Tsukuba, Ibaraki, 305-8518 Japan
*Corresponding author: E-mail, ebana@affrc.go.jp; Fax, +81-29-838-7408.
(Received 14 August 2020; Accepted 22 September 2020)

Wang et al. 2018 Nature

Genomic variation in 3,010 di accessions of Asian cultivated

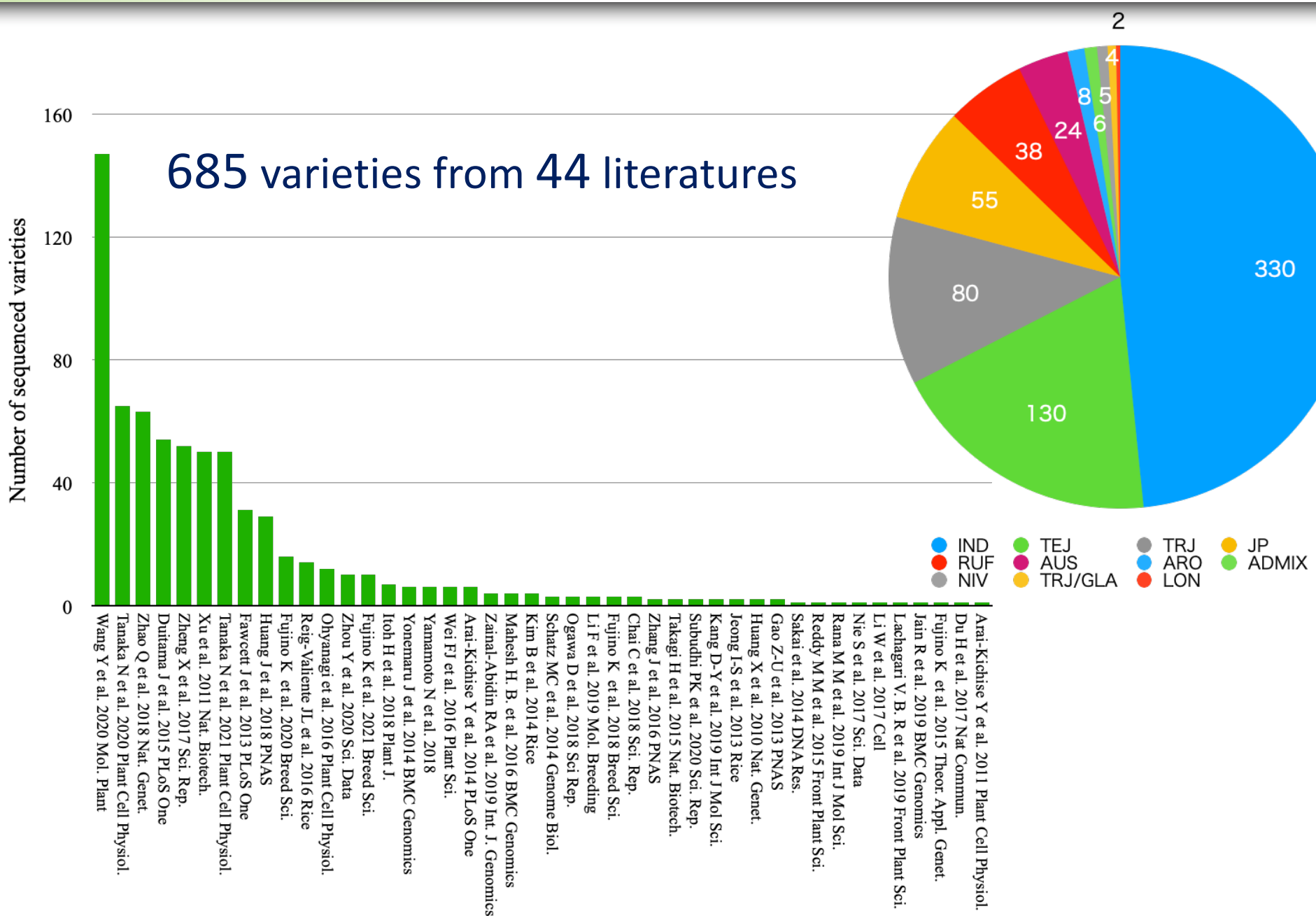
3,010

Wensheng Wang[1,17], Ramil Mauleon[2,17], Zhiqiang Hu[1,3,17], Dmytro Chebotarov[2,17], Shuaishua
Tianqing Zheng[1,17], Roven Rommel Fuentes[2,17], Fan Zhang[1,17], Locedie Mansueto[2,17], Dario C
Kevin Christian Palis[2], Jianlong Xu[1,5,6], Chen Sun[3], Binying Fu[1,6], Hongliang Zhang[9], Yongming Gao
Xiao Cui[3], Hong Yu[10], Zichao Li[9], Miaolin Chen[3], Jeffrey Detras[2], Yongli Zhou[1,6], Xinyuan Zhang[5], Yue Zhao[4], Dave Kudrna[8],
Chunchao Wang[1], Rui Li[3], Ben Jia[3], Jinyuan Lu[3], Xianchang He[3], Zhaotong Dong[3], Jiabao Xu[4], Yanhong Li[4], Miao Wang[4],
Jianxin Shi[3], Jing Li[3], Dabing Zhang[3], Seunghee Lee[8], Wushu Hu[4], Alexander Poliakov[11], Inna Dubchak[11,12], Victor Jun Ulat[2],
Frances Nikki Borja[2], John Robert Mendoza[13], Jauhar Ali[2], Jing Li[3], Qiang Gao[4], Yongchao Niu[4], Zhen Yue[4],
Ma. Elizabeth B. Naredo[2], Jayson Talag[8], Xueqiang Wang[9], Jinjie Li[9], Xiaodong Fang[4], Ye Yin[4], Jean-Christophe Glaszmann[14,15],
Jianwei Zhang[8], Jiayang Li[1,10], Ruaraidh Sackville Hamilton[2], Rod A. Wing[2,8]*, Jue Ruan[5]*, Gengyun Zhang[4,6]*,
Chaochun Wei[3,16]*, Nickolai Alexandrov[2]*, Kenneth L. McNally[2]*, Zhikang Li[1,6]* & Hei Leung[2]

Here we analyse genetic variation, population structure and diversity among 3,010 diverse Asian cultivated rice (*Oryza sativa* L.) genomes from the 3,000 Rice Genomes Project. Our results are consistent with the five major groups previously recognized, but also suggest several unreported subpopulations that correlate with geographic location. We identified 29 million single nucleotide polymorphisms, 2.4 million small indels and over 90,000 structural variations that contribute to within- and between-population variation. Using pan-genome analyses, we identified more than 10,000 novel full-length protein-coding genes and a high number of presence–absence variations. The complex patterns of introgression observed in domestication genes are consistent with multiple independent rice domestication events. The public availability of data from the 3,000 Rice Genomes Project provides a resource for rice genomics research and breeding.

Number of varieties sequenced in the paper

Collection & curation of resequence data



Standardized analysis pipeline

Our analysis pipeline for SNP & InDel detection is publicly available

Analysis workflow for genome-wide variations in TASUKE+ for RAP-DB (version 2.0)

Genome-wide variations were detected using publicly available genome resequencing data for various rice varieties and provided in the multiple genome browser TASUKE+. Here we provide the analysis workflow for the detection of variations.

History

- 11/Mar/2022 Analysis workflow (version 2.0) and TASUKE+ with 685 rice varieties were released.
- 29/Aug/2019 Analysis workflow (version 1.0) and TASUKE+ with 533 rice varieties were released.
- 15/Nov/2017 TASUKE with 333 rice varieties were released.

Reference data

- Genome sequences [FASTA]
  - IRGSP-1.0 genome (including organel
- Gene annotation for snpEff [Annotation
  - RAP-DB (both representative and pre
  - MSU (all genes in RGAP 7)
- Illumina adapter sequences attached wit
- Chromosome information (chromosome_

```
$ cat chromosome_list.csv
chr01,43270923,16610866,17243770
chr02,35937250,13541821,13872411
chr03,36413819,19431743,19745569
chr04,35502694,9744480,9973218
```

Analysis tools

- Java (JDK 1.8.0_191)
- FastQC v0.11.8
- BWA (bwa v0.7.17)
- SamTools (v1.9)
- BamTools (as of 4 Dec 2018)
- GATK (v4.0.11.0)/GATK (v4.2.4.0)
- Trimmomatic (v0.38)
- Picard (v2.18.17)
- SnpEff (v5.0e)
- TASUKE+ (version 20210831)

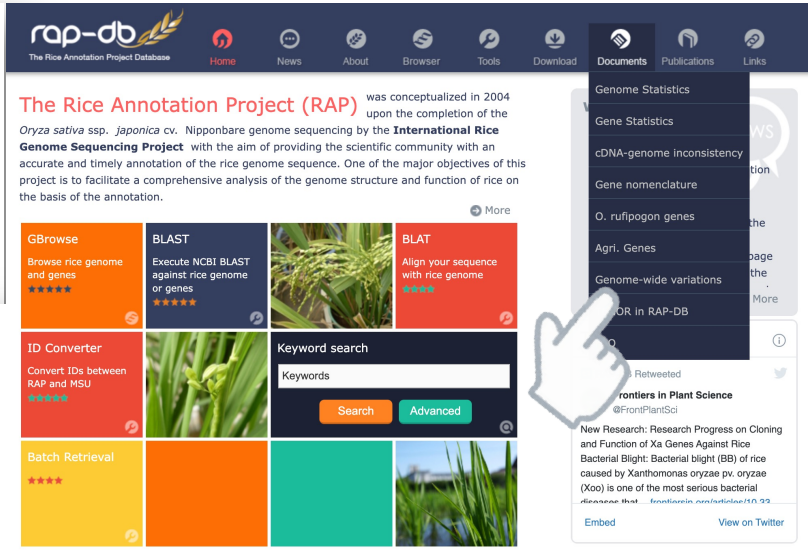
Commands and parameters used in the workflow

1. **Preprocessing of Illumina paired-end reads**

```
$ java -jar trimmomatic-0.38.jar PE \
    -phred33 read.r1.fastq.gz read.r2.fastq.gz \
    read.pe.r1.fastq.gz read.se.r1.fastq.gz read.pe.r2.fastq.gz read.se.r2.fastq.gz \
    ILLUMINACLIP:adapters.fa:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:10:20 MINLEN:30
```

2. **Making index of the genome**

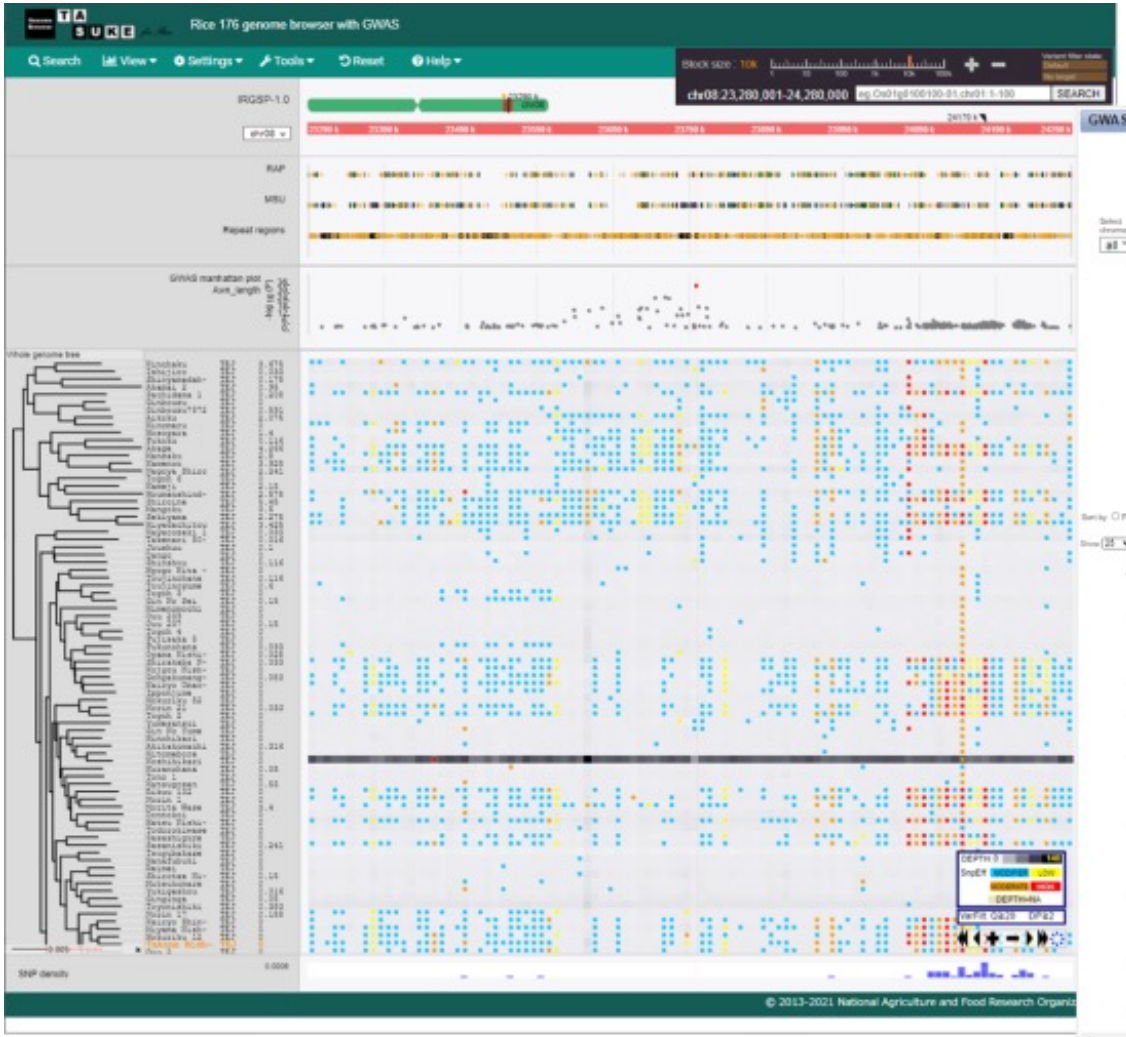
```
$ bwa index genome.fa
$ samtools faidx genome.fa
$ java -jar picard.jar CreateSequenceDictionary \
    REFERENCE=genome.fa \
    OUTPUT=genome.dict
```



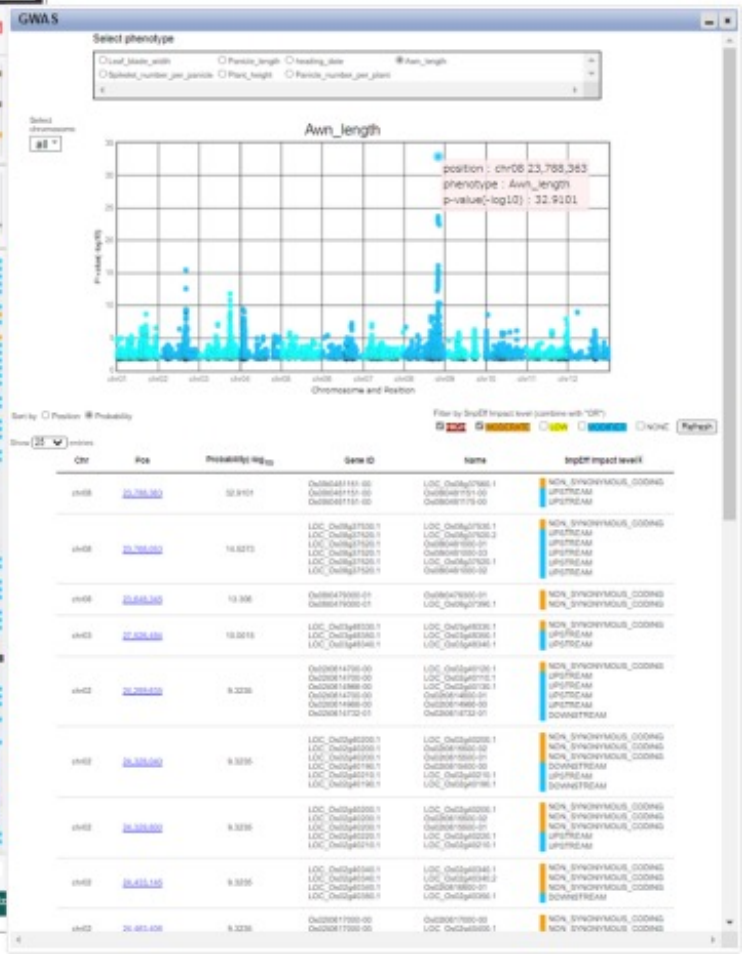
Genome-wide variation among 685 rice varieties on TASUKE+

農研機構 NARO

TASUKE+ is a web-based genome browser that visualizes variation and read depth data of genome resequencing data analysis



https://tasuke.dna.affrc.go.jp/



Genome-wide variation among 685 rice varieties on TASUKE+

Variations within the Hd16 locus. Koshihikari and some TEJ varieties have a nonsynonymous mutations (Ala331Thr).

Curation of known alleles and mutations

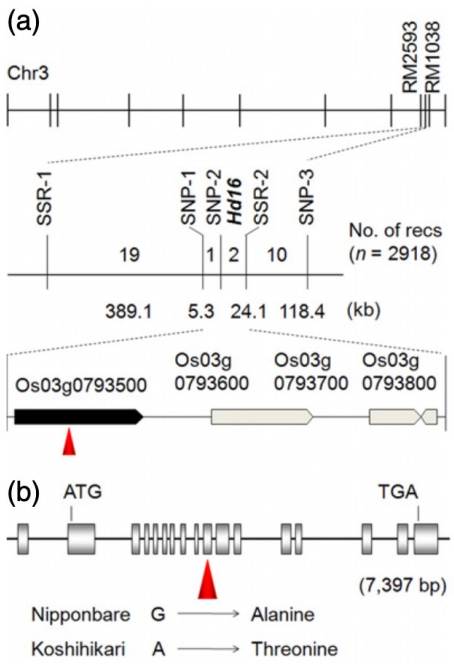
the plant journal
SEB Society for Experimental Biology
The Plant Journal (2013) **76**, 36–46 doi: 10.1111/tpj.12268

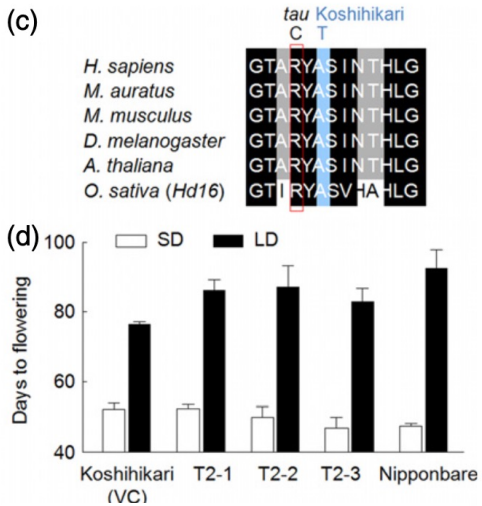
Hd16, a gene for casein kinase I, is involved in the control of rice flowering time by modulating the day-length response

Kiyosumi Hori†, Eri Ogiso-Tanaka†, Kazuki Matsubara‡, Utako Yamanouchi, Kaworu Ebana and Masahiro Yano*
National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, Ibaraki 305-8602, Japan

Received 8 March 2013; revised 29 May 2013; accepted 17 June 2013; published online 21 June 2013.
*For correspondence (e-mail myano@affrc.go.jp).
†K.H. and E.O.T. contributed equally to this work.
Data deposition: The sequences reported in this paper have been deposited in the DDBJ/EMBL/GenBank database [accession nos. AB753041 (Nipponbare) and AB753042 (Koshihikari)].
‡Present address: Institute of Crop Sciences, National Agriculture and Food Organization, 2-1-18 Kannondai, Tsukuba, Ibaraki 305-8518, Japan.

an alanine amino acid in Nipponbare changed to a threonine amino acid in Koshihikari.
…
One non-synonymous substitution in Hd16 resulted in decreased photoperiod sensitivity in rice.





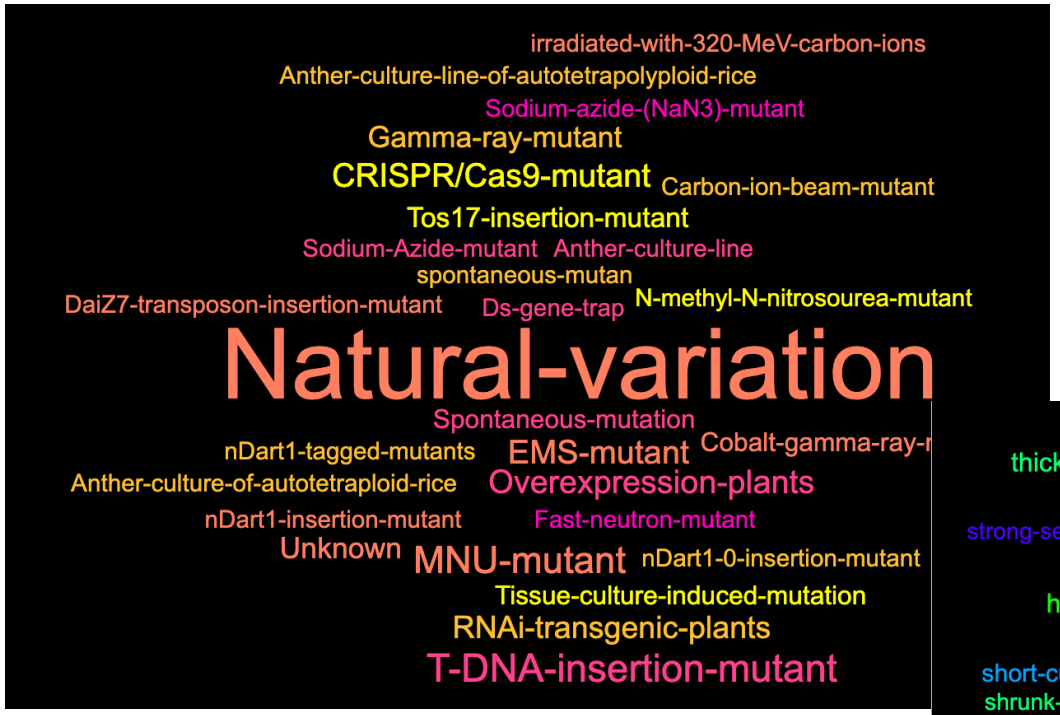
Known functional alleles & mutations

Ref. CVs	Target CVs	Var. Type	HGVS (DNA)	HGVS (CDS/Protein)	Traits of Target CVs	References
Nipponbare	Koshihikari	Natural variation	g.33002789G >A	p.Ala331Thr	low photoperiod sensitivity	Hori K et al. Plant J, 76(1):36-46 (2013)[PubMed]

Known functional alleles & mutations

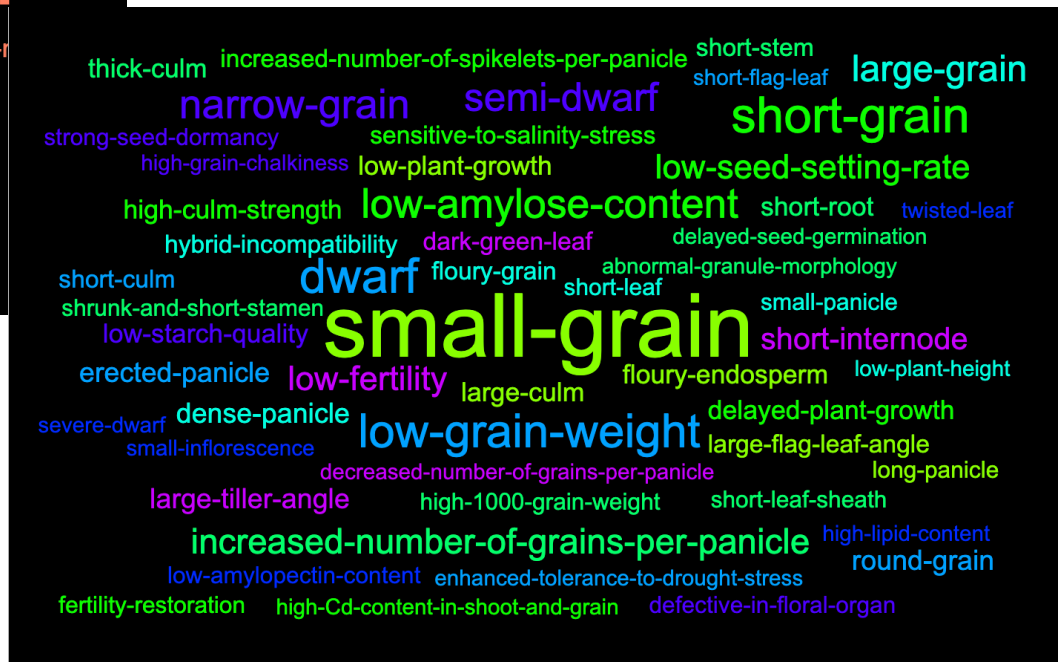
農研機構 NARO

We're collecting known functional alleles from literatures



Types of variations

Related traits

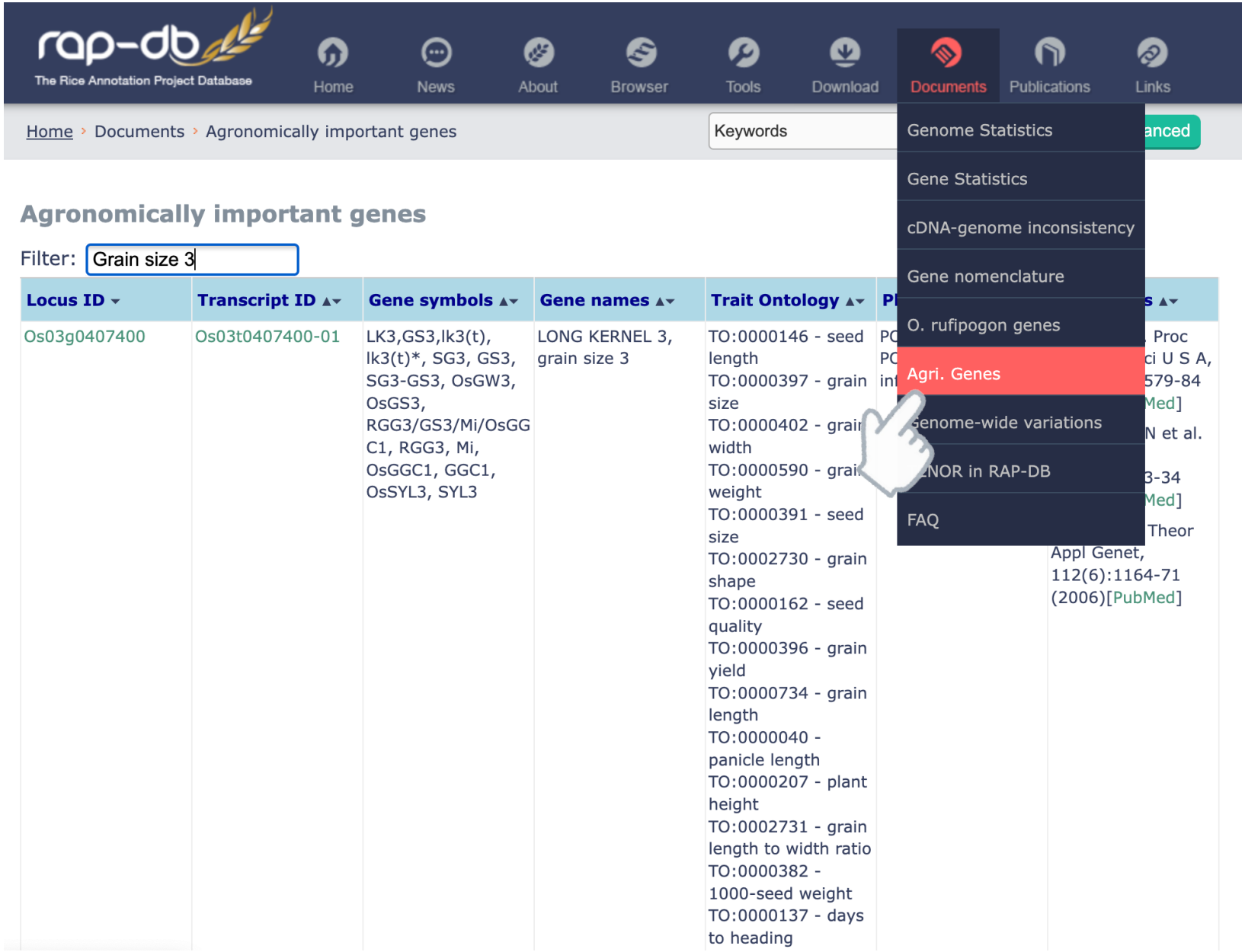


Demo:

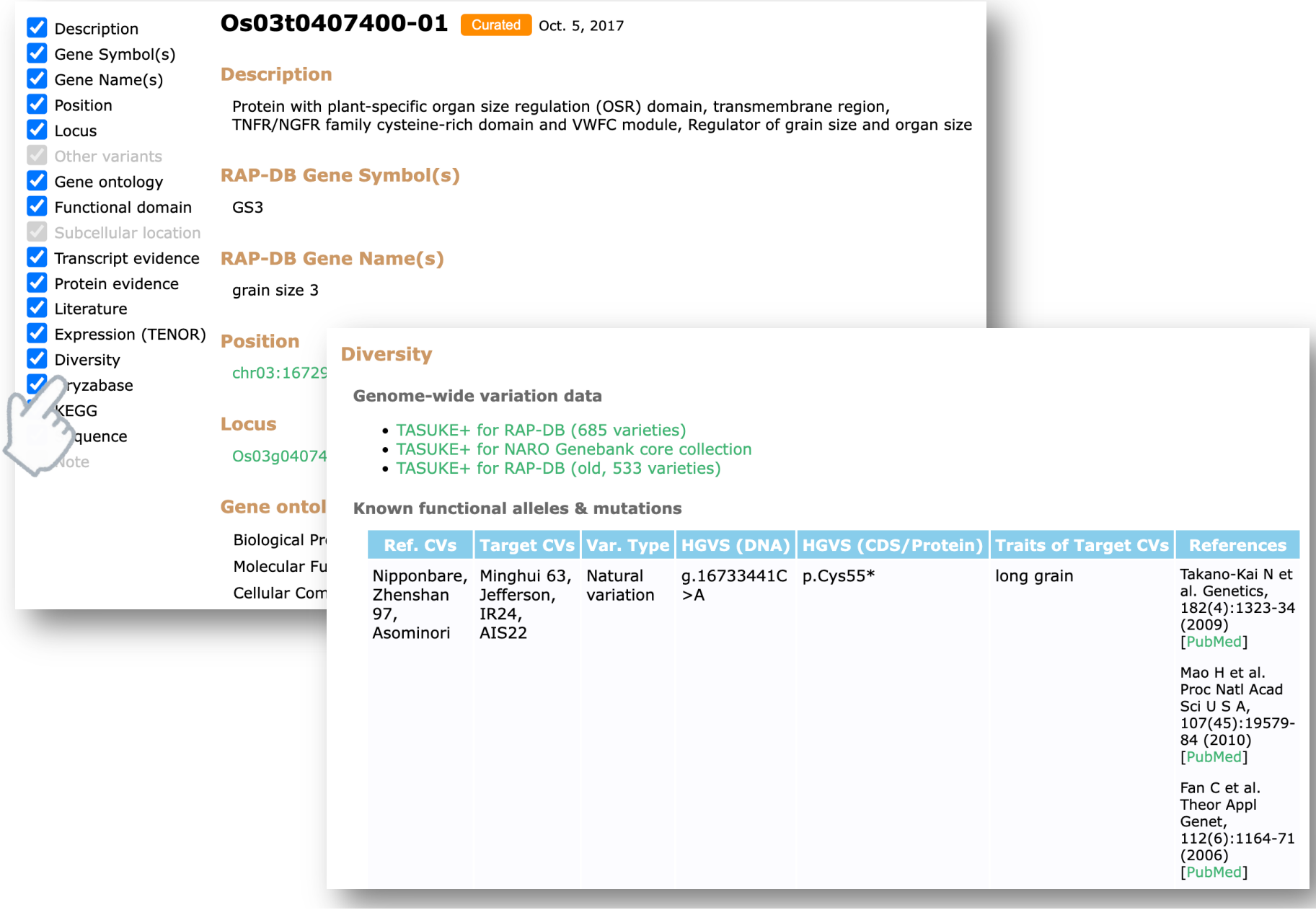
Viewing genome-wide variation on TASUKE+
and known functional alleles/mutations in RAP-DB

Agronomically important genes

365 agronomically important rice genes are listed.



TASUKE+, known functional alleles & mutations in the “Diversity” section



TASUKE+

農研機構 NARO

GS3 locus related to grain size

Temperate japonica

Tropical japonica

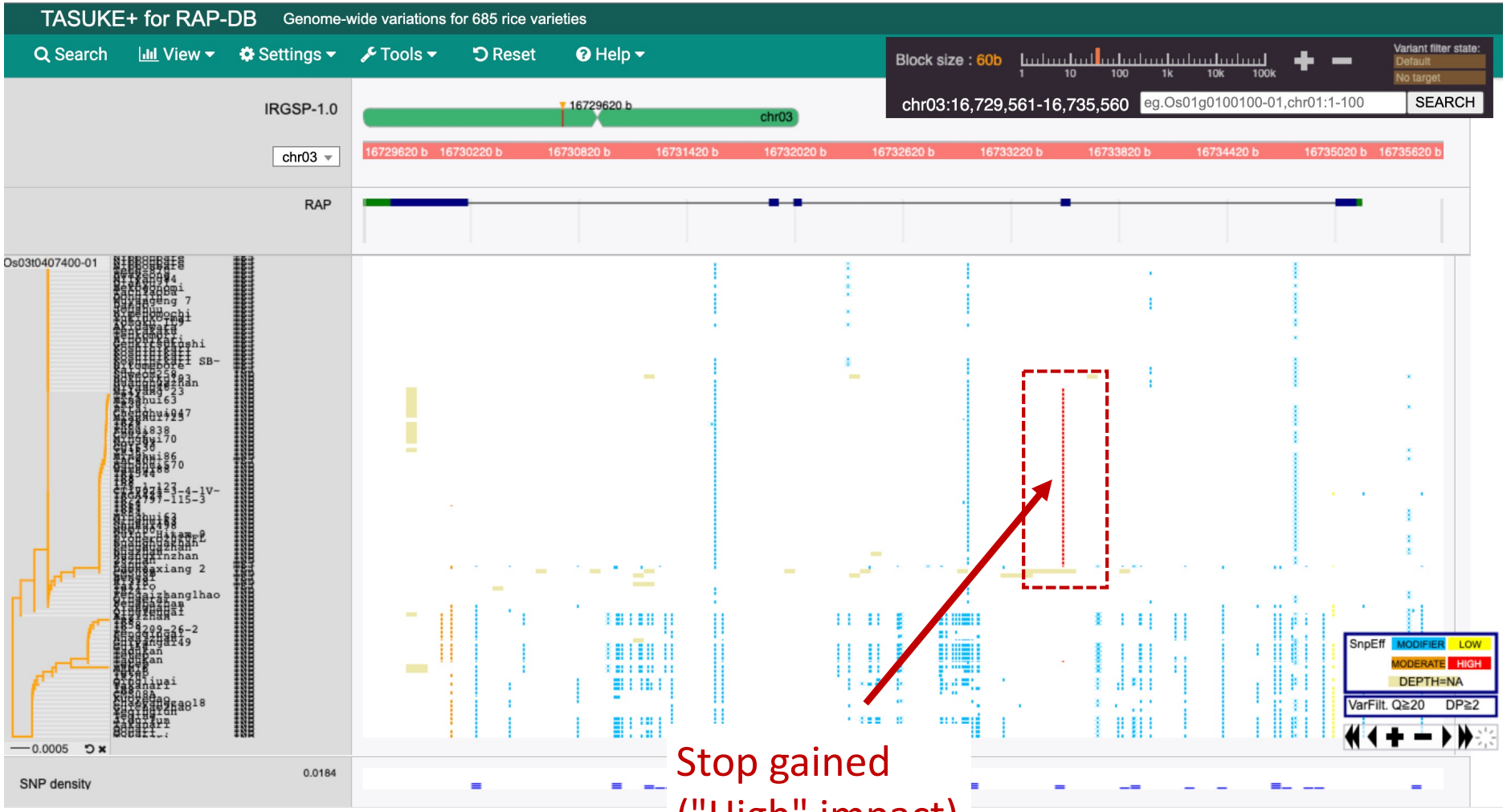
Aroma

Aus

Indica

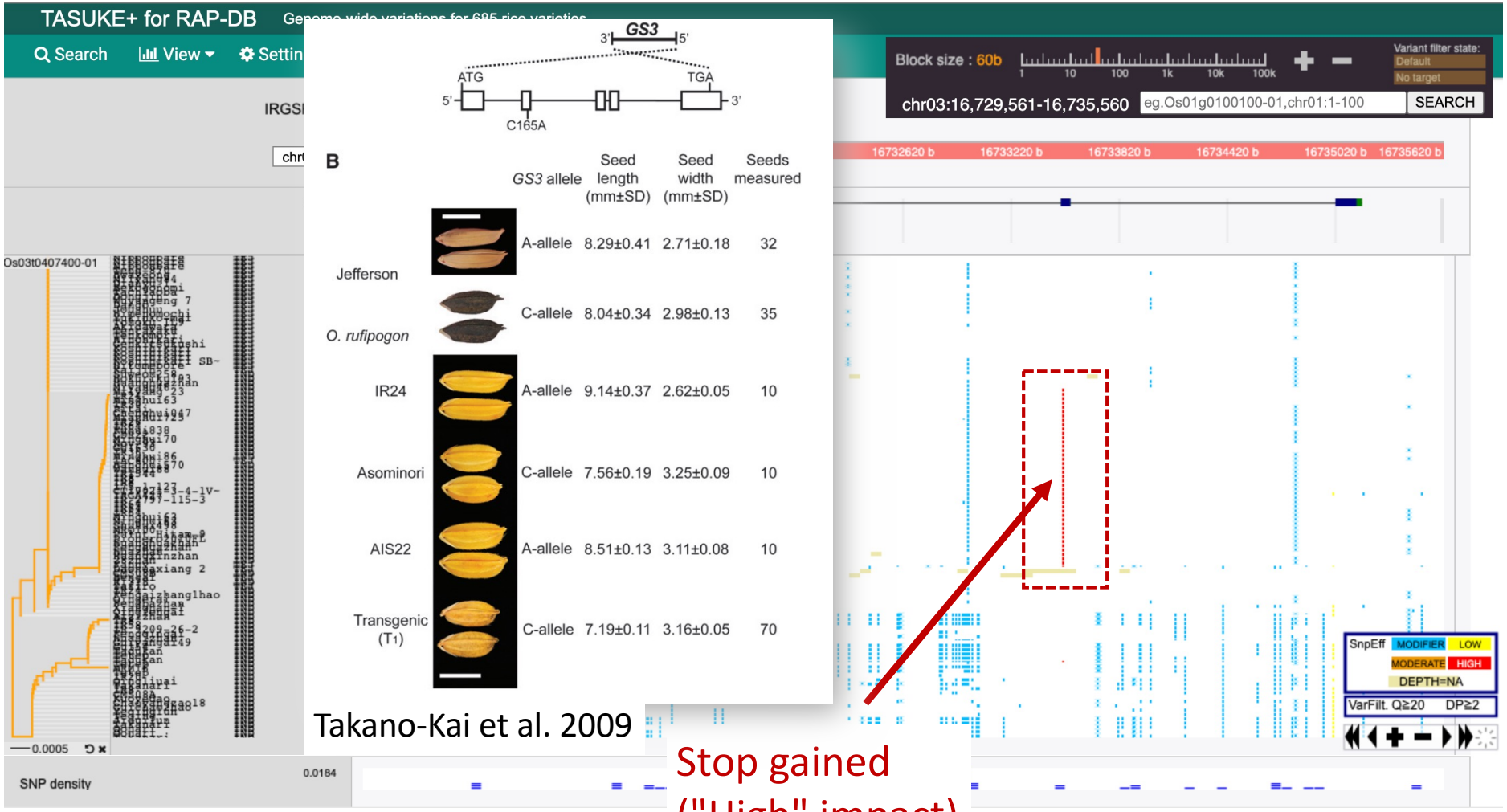
TASUKE+

Sorted by the pattern of variations at the GS3 locus



Known functional alleles & mutations

We’re collecting known functional alleles from literatures



Stop gained
("High" impact)
c.165C>A
p.Cys55*

Take-home messages

RAP-DB …

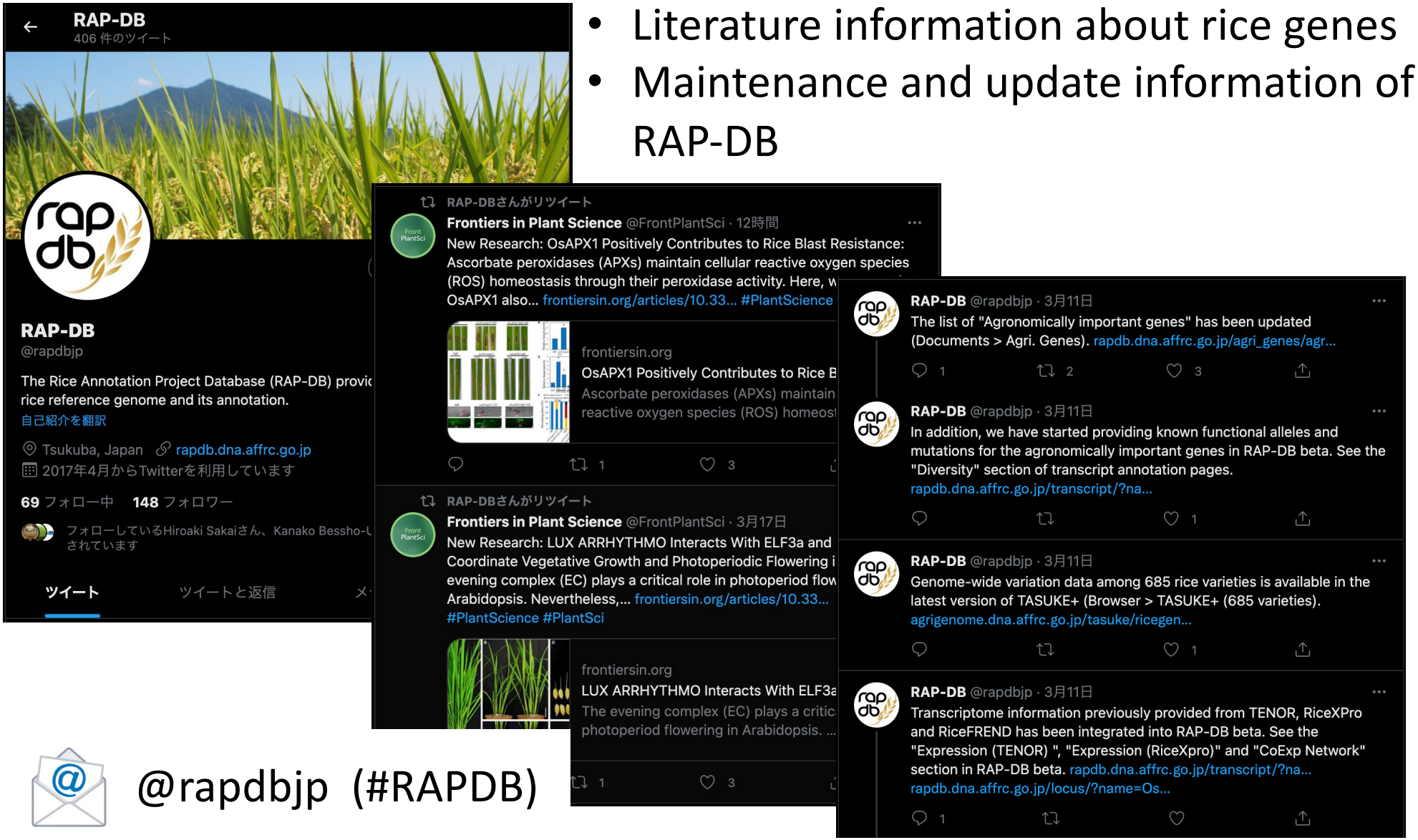
- is the only primary database for rice that continues to update gene annotation data by manual curation.

- provides gene expression profiles under >500 experimental conditions. Currently, TENOR, RiceXPro and RiceFREND are partially integrated into RAP-DB.

- provides genome-wide variations among >600 rice varieties through TASUKE+.

- provides the list of 365 agronomically important rice genes and 762 known functional alleles/mutations on the genes.

Information about RAP-DB and rice genes



- Literature information about rice genes
- Maintenance and update information of RAP-DB

@rapdbjp (#RAPDB)

rapdb@ml.affrc.go.jp