Published online 6 December 2016

Nucleic Acids Research, 2017, Vol. 45, No. 2 ***597–605***
doi: 10.1093/nar/gkw958

RPAN: rice pan-genome browser for ∼3000 rice genomes

Chen Sun[1,4,†], Zhiqiang Hu[1,4,†], Tianqing Zheng[2,†], Kuangchen Lu[1,†], Yue Zhao[1], Wensheng Wang[2], Jianxin Shi[3], Chunchao Wang[2], Jinyuan Lu[1], Dabing Zhang[3,5], Zhikang Li[2,*] and Chaochun Wei[1,4,*]

[1]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China, [2]Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China, [3]Joint International Research Laboratory of Metabolic & Developmental Sciences, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China, [4]Shanghai Center for Bioinformation Technology, Shanghai 201203, China and [5]School of Agriculture, Food and Wine, University of Adelaide, Waite Campus, Urrbrae, SA 5064, Australia

Received May 05, 2016; Revised October 02, 2016; Editorial Decision October 06, 2016; Accepted October 24, 2016

ABSTRACT

A pan-genome is the union of the gene sets of all the individuals of a clade or a species and it provides a new dimension of genome complexity with the presence/absence variations (PAVs) of genes among these genomes. With the progress of sequencing technologies, pan-genome study is becoming affordable for eukaryotes with large-sized genomes. The Asian cultivated rice, *Oryza sativa* L., is one of the major food sources for the world and a model organism in plant biology. Recently, the 3000 Rice Genome Project (3K RGP) sequenced more than 3000 rice genomes with a mean sequencing depth of 14.3×, which provided a tremendous resource for rice research. In this paper, we present a genome browser, Rice Pan-genome Browser (RPAN), as a tool to search and visualize the rice pan-genome derived from 3K RGP. RPAN contains a database of the basic information of 3010 rice accessions, including genomic sequences, gene annotations, PAV information and gene expression data of the rice pan-genome. At least 12 000 novel genes absent in the reference genome were included. RPAN also provides multiple search and visualization functions. RPAN can be a rich resource for rice biology and rice breeding. It is available at http://cgm.sjtu.edu.cn/3kricedb/ or http://www.rmbreeding.cn/pan3k.

INTRODUCTION

Next-generation sequencing technologies have opened the possibility of sequencing a large number of individuals from one species, such as the 1001 Genomes Project for *Arabidopsis thaliana* (1), the 1000 Genomes Project for Human (2) and the 3000 Rice Genomes Project (3K RGP) for *Oryza sativa* L. (rice) (3). Pan-genome, a concept first introduced in the study of many genomes of a bacterial species in 2005 (4), is becoming prevalent in studies of bacteria and archaea which have small genome sizes (5). The pan-genome of a species consists of a 'core genome' that contains genes present in all individuals and a 'distributed genome' that comprises genes not shared by all individuals. In recent years, pan-genome analysis was also successfully applied to eukaryotes with large genome sizes, such as human (6), rice (7,8), soybean (9) and maize (10). The 3K RGP generated sequencing data for >3000 rice accessions with a mean coverage of 14.3×. The organization and visualization as well as the consequent analyses of >3000 genomes are big challenges; however, they are of extreme significance especially for the rice breeding. Therefore, development of a user-friendly visualization tool for rice pan-genome is in great demand.

One of the most widely used genome visualization tools, the University of California Santa Cruz Genome Browser (UCSC Genome Browser) (11,12), takes one single genome as the reference genome and its visualization is based on this individual genome. Several visualization tools for pan-genome or comparative genomics have been developed recently, including stand-alone programs, such as Pan-Tetris (13), GenPlay Multi-Genome (14) and JContextExplorer (15), and web-based browsers, such as PopGeV (16), the

*To whom correspondence should be addressed. Tel: +86 21 34204237; Fax: +86 21 20283718; Email: ccwei@sjtu.edu.cn
Correspondence may also be addressed to Zhikang Li. Email: zhkli1953@126.com
†These authors contributed equally to this work as first authors.

Table 1. Statistics about genomes in RPAN

Accession group	Count
Indica	1764
Japonica	801
AUS	221
ARO	101
ADM	123
Total	3010

Table 2. Statistics about rice gene categorization

Gene category	Count
Total genes	50 995
Core genes	23 914
Candidate core genes	4986
Distributed genes	22 095
Subspecies-unbalanced genes	13 617
Indica-dominant genes	5579
Japonica-dominant genes	6038
Subspecies-specific genes	853
Indica-specific genes	587
Japonica-specific genes	147
AUS-specific genes	67
ARO-specific genes	52
Subgroup-unbalanced genes	11 581
Indica-subgroup-unbalanced genes	9816
Japonica-subgroup-unbalanced genes	3418
Random genes	5316

browser for UK10K-cohorts project (17) and GPAC (18). These visualization tools were developed to meet specific needs. Pan-Tetris visualizes the gene occurrences for bacteria and represents each gene as a unicoloured glyph. GenPlay Multi-Genome visualizes the data in a resolution of a single base and compares allele-specific expression and functional genomic data for multiple closely related genomes. However, with multiple thousands of genomes, creating the meta-reference genome via multiple sequence alignment is prohibitively expensive. Therefore, GenPlay Multi-Genome does not fit for rice pan-genome. JContextExplorer is a tree-based approach comparing cross-species bacterial genomes. PopGeV is a web-based large-scale population genome browser mainly displaying SNP and InDel data. The browser for the UK10K-cohorts project retrieves genotype-phenotype association from the data and GPAC visualizes multiple genome-level changes. Nevertheless, none of these genome browsers visualizes genomes in a pan-genome approach (i.e. organizing a large number of genomes from a species and visualizing them accordingly as a genome of a species), nor displays the presence and absence variation (PAV) of genes, let alone the browser specific for rice pan-genome.

Here we present Rice Pan-genome Browser (RPAN), an interactive web-based pan-genome browser (Supplementary Figure S1) for the rice pan-genome created from the 3K RGP. The browser contains information of the 3010 sequenced rice accessions, genomic sequences, gene annotations as well as gene expressions. It also provides several search functions, such as the search function for the PAV information with a list of rice accessions or gene IDs. Furthermore, it provides the search for specific DNA sequences against the rice pan-genome as well. The data resources and tools provided by RPAN will accelerate both the basic research in rice biology and the applied efforts in rice breeding.

MATERIALS AND METHODS

The sequencing data of 3010 rice accessions were acquired from the 3K RGP (3) and the pan-genome sequences were constructed based on IRGSP-1.0 genome, a widely used rice reference genome with high-quality annotations. The total size of the compressed raw sequencing data was about 15 TB.

Pan-genome construction

The raw sequencing data of each accession were first assembled with SOAPdenovo version r240 (19), and then the assembled contigs with lengths >500 bp were aligned to the IRGSP genome by the nucmer tool in Mummer package version 3.23 (20). For those unaligned contigs, the redundant sequences (identity threshold: 90%) were removed by CD-HIT version 4.6.1 (21). Next, various contaminants were removed by NCBI-blast (version 2.2.28+) (22) against the NT database. Then all-vs-all alignments with NCBI-blast was carried out to ensure no redundancy. The remained contigs formed the non-redundant novel sequence dataset (identity < 0.9 in comparison to IRGSP version 1.0, and identity <0.9 for all novel sequences). All the resulted non-redundant unaligned sequences were then categorized into 12 groups according to the classification of their corresponding rice accessions, which are predefined by population structure from SNP analysis (23,24). These groups include five subgroups (I_{G1}, I_{G2}, I_{G3}, I_{G4} and I_{G5}) of subspecies *Indica*, AUS_{G6}, four subgroups (J_{G7}, J_{G8}, J_{G9} and J_{G10}) of subspecies *Japonica*, ARO_{G11} and admixtures (Adm). Then, all these contigs from the same varietal subgroup were concatenated with 100 consecutive Ns as the delimiters. Finally, the IRGSP genome (373 Mbps) and these novel sequences (268 Mbps) were merged as the reference pan-genome of rice (Supplementary Figure S2).

Pan-genome annotation

The protein coding genes on unaligned sequences were predicted by MAKER-P and the gene/transcript annotation of the IRGSP-1.0 genome was downloaded from the Rice Annotation Project (RAP) (25).

PAV determination

The coverage of a gene was used to determine its presence/absence. First, all raw reads of a rice accession were mapped to the pan-genome sequences by 'bwa mem' of BWA version 0.7.10 (26). Using samtools version 0.1.19 (27), a gene with coding region coverage >0.95 and gene region coverage >0.85 was considered as a presence in the accession.

Gene categorization

Genes were categorized according to their presence/absence

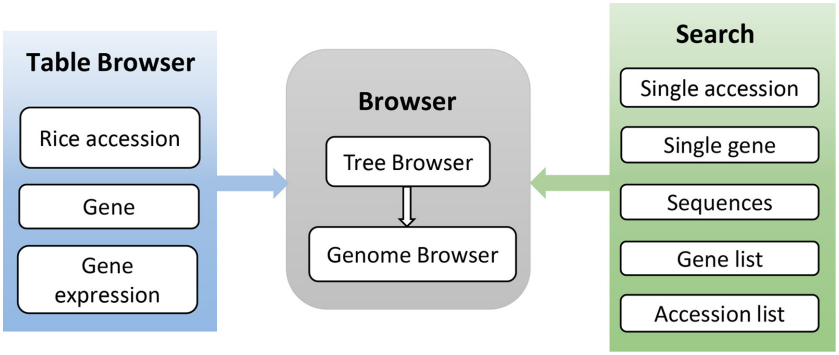

Figure 1. The architecture of RPAN. This rice pan-genome browser contains table browser, genome browser and multiple search functions. Table browser allows users to summarize, display and download the contents of tracks. Tree browser can be used to select target genomes which can be displayed as tracks in genome browser. Genome browser contains a reference individual genome as well as those novel sequences not included in the reference genome. Users can also search the pan-genome with a list of genes or accessions, for the presence/absence of the genes in the list of selected rice accessions. Searched results can be displayed in the genome browser as well as in tables and figures.

in all accessions. In order to reduce the influence of the low sequencing coverages, the presence/absence of a gene was checked only for 453 high-quality accessions whose sequencing depths were $>20\times$ and mapping depths were $>15\times$. Core genes are those present in all 453 rice accessions, and other genes are distributed. In order to reduce the false discovery rate of distributed genes, those genes absent in $>1\%$ of accessions (binomial tests, P-value < 0.05, null hypothesis is 'A non-distributed gene is absent in $<1\%$ of accessions') were kept as distributed while the other (we call them candidate core genes) were removed from distributed gene set. A distributed gene can be categorized further as subspecies unbalanced gene (whose frequency in one or more subspecies is significantly higher than that in other subspecies (Fisher tests, FDR < 0.05)) and subspecies specific gene (that is present in a subspecies but absent in all other subspecies). For subspecies unbalanced genes, if a gene is significantly (5%) more frequent in *Indica* (*Japonica*) accessions than in *Japonica* (*Indica*) subspecies, it is called *Indica* (*Japonica*) dominant gene. Similarly, in each subspecies, a gene can be further categorized as sub-group unbalanced gene, whose frequency in one or more sub-groups of a subspecies is significantly higher than the frequencies in other sub-groups in this subspecies. At the end, a distributed gene is called a random gene if it is neither subspecies unbalanced nor sub-group unbalanced. A random gene shows no significant difference in frequency among subspecies or sub-groups inside a subspecies.

Expression data

In total, 226 runs of RNA-seq data from diverse rice tissues, including seedling shoots, booting panicles, callus and four-leaf stage shoots were collected from all available public databases (See detailed information in Supplementary Table S1). The obtained RNA-seq data were first trimmed with Trimmomatic version 0.32 (28) with parameters 'ILLUMINACLIP:2:30:10 LEADING:20 TRAILING:20 SLIDINGWINDOW:4:20 MINLEN:36', which yielded a clean RNA-seq data with a size of 3.4 TB. Then the RNA-seq data were aligned to the rice pan-genome with HISAT2 version 2.0.1-beta (29) with default parameters. The alignment results were converted, sorted and stored in .bam file format with samtools version 1.2 (27). The coverage of each gene was calculated with 'bedtools coverage' in bedtools suite version 2.17.0 (30).

Visualization

The visualization page consists of two parts. A tree browser in the left panel and a genome browser in the right panel. The genome browser was constructed based on the JBrowse framework (31), and the tree browser was implemented in-house with HTML5, SVG and JavaScript. The tree browser not only generates an interactive tree view, but also supports intuitive track selection with the tree browser by scrolling, searching and node selecting in addition to the traditional track selection for the 3010 accessions. Another characteristic of the tree browser is that the selected accessions from the tree browser can be submitted to the genome browser directly. RPAN uses the newly developed rice pan-genome as the reference genome, and provides gene annotation track as a default track for the whole rice pan-genome with an additional 226 tracks of RNA-seq data.

RESULTS

A rice pan-genome browser, RPAN, was developed with the rice pan-genome derived from 3010 sequenced rice accessions (Table 1). The system diagram of RPAN is shown in Figure 1. RPAN contains a database for the rice pan-genome together with search and visualization tools. All the parts in RPAN are dynamic and interactive.

The rice pan-genome

A rice pan-genome was constructed with the IRGSP-1.0 genome (25) and all novel sequences not included in the reference genome. The novel sequences were grouped according to their source rice accessions based on the phylogenic tree derived from SNPs (see Methods for more details). All novel sequences in a subgroup of rice accessions were concatenated as one pseudo-chromosome. The sequencing data of all 3010 rice accessions were mapped to the pan-genome and visualized through the JBrowse framework.

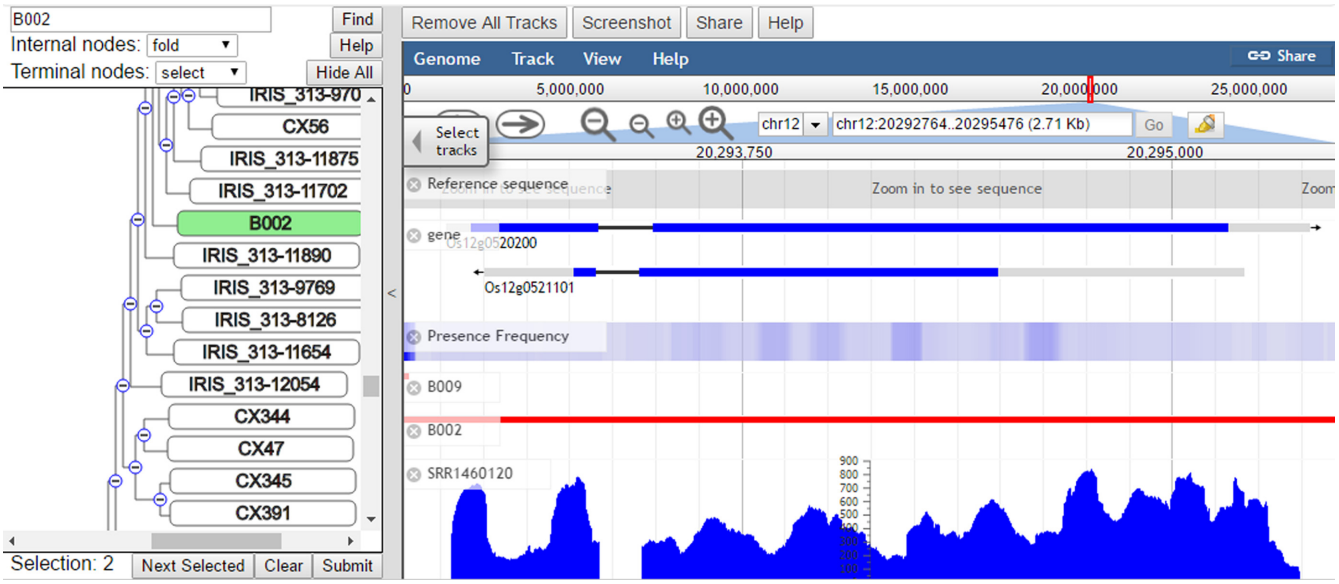

Figure 2. The tree browser and genome browser of RPAN. The left panel is the tree browser representing the clustering of ∼3000 individual rice genomes. The tree browser can be used to select genomic tracks for visualization in the genome browser. The right panel is the genome browser. The tracks in it from top to bottom are reference, gene annotation, presence frequency, accessions (red) and RNA-seq (blue). Its genomic sequences contain a reference individual rice genome as well as those novel sequences not included in the reference genome.

RPAN database also includes basic information of >3000 rice accessions, genome-wide expression profile data, gene annotations, and the presence-absence variations.

1. Basic information of 3010 rice accessions derived from the 3K RGP, including accession names, sequencing depths, mapping depths on the IRGSP-1.0 genome and meta-information such as geological locations, subspecies (or subgroup) categorization, etc.
2. Coding sequences, protein sequences and annotations for 50 995 full-length coding genes (Table 2) in the rice pan-genome.
3. Gene presence-absence variations (PAVs). The presence/absence of genes in the rice pan-genome were determined by 453 high-quality accessions. All genes were then categorized as core, candidate core or different types of distributed genes. In total, there are 23 914 core genes, 4986 candidate core genes and 22 095 distributed genes. Of the distributed genes, 853 genes are subspecies or varietal group specific, including 587, 147, 67 and 52 genes for *Indica* and *Japonica* subspecies, *Aus* and *Aro* groups, respectively (Table 2).
4. Genome-wide expression profiles for the rice pan-genome, including 226 publicly available RNA-seq runs of pan-genome expression profiles (Supplementary Table S1).

Basic search functions

Users can search with a gene ID, a rice accession code, or genomic sequences. When a gene ID is searched, RPAN provides the results from eight aspects: basic gene information, gene categorization information, gene family, gene presence frequency and distribution, gene ontology, protein coding sequence and protein sequence. The basic gene information such as location on the pan-genome, the gene categorization information (including whether the gene is a core gene or distributed one, an unbalanced gene among subspecies or varietal subgroup, its gene age, etc.) and gene family information are displayed in the first three tables. The frequency of this gene present in five subspecies and 12 subgroups are shown in two heat maps. In addition, the relevant GO terms, protein coding sequence and protein sequence are also provided.

When a rice accession code is searched, the basic information about the rice accession and the statistics of the genes in this rice accession are displayed in two tables. Meanwhile, three pie charts show the categorization of these genes in this rice accession. Users can also visualize the alignment of sequencing data of the queried rice accession against the rice pan-genome in the visualization page after clicking the 'Genome Browser' button.

Users can also search with genomic sequences against the rice pan-genome directly with BLAT (32). One or more sequences in the FASTA format can be searched. All alignments can be further checked in a detailed page by clicking the 'Genome Browser' button in the record line and visualized in the pan-genome browser.

Advanced search functions

Searching multiple rice accession codes or a list of gene IDs is also supported in RPAN. Users can input multiple accession codes in the search box or upload a file containing accession codes. The least number of rice accessions sharing a specific gene can be an optional parameter. If this number is set to 1, the search result will be all genes existing in all the input accessions; similarly, if the number is set to the number of all input accessions, the core genes of all input accessions would be acquired. Then, the basic information of

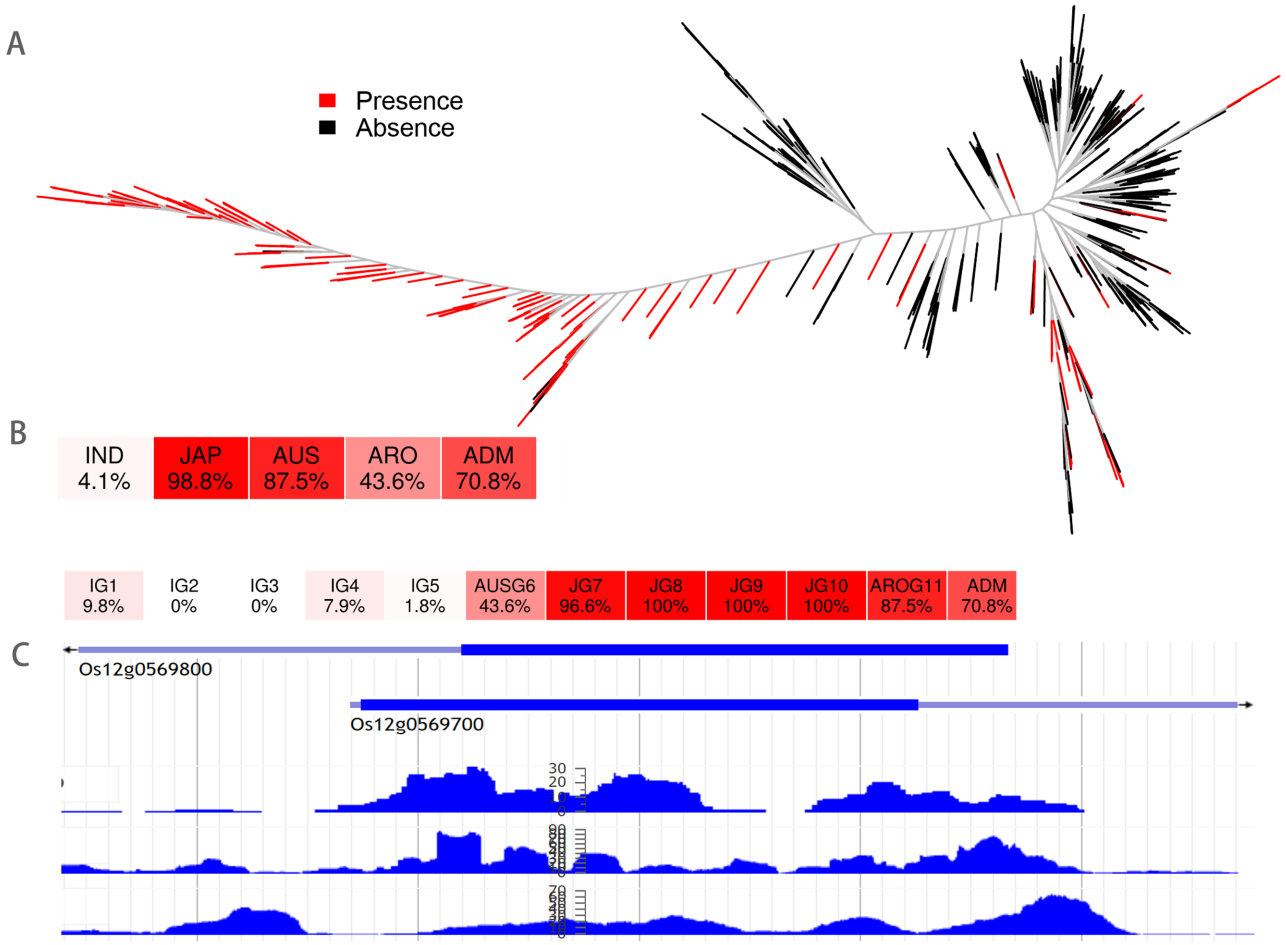

Figure 3. Examples of search and visualization functions of RPAN. Os12g0569700, a gene related to rice acclimation to salt and drought stresses, was searched. Search results include (**A**) distribution of the gene in high-quality accessions and (**B**) heat maps of the gene presence frequency in different subspecies and subgroups. (**C**) The visualization of this gene with three RNA-seq tracks.

these accessions and the resulted genes can be downloaded and the statistics tables and charts for these genes are also provided. Likewise, users can search with a gene ID list to obtain rice accessions in which all genes present and acquire relevant information of accessions and genes.

Visualization

The visualization page contains two parts, a dynamic tree browser on the left panel and a genome browser on the right panel (Figure 2). The tree was constructed from the SNP data. Users can select multiple nodes (including leaf nodes and internal nodes) and click the 'Submit' button to visualize these rice accessions in the genome browser. The tree browser also supports search function to accelerate target genome selection. The pan-genome reference sequence, gene annotation and overall presence frequency of high quality accession are three basic tracks. There are 3010 rice genome tracks and 226 RNA-seq tracks. Users can select any number of accessions or expression data through the hidden 'Select tracks' panel or the tree browser as well. For the performance concern, we recommend to select less than 300 tracks each time. In addition, multiple useful tools such as screenshot and share link are listed in the toolbar.

Table browser

All information in the pan-genome browser was stored in tables that can be downloaded. These tables include the rice accession information table, the gene information table and gene expression profile table.

In the rice accession information table, users can filter the results by selecting browse options such as categories, geological regions and sequencing depth status (high/low). A summary table can be generated for filtered results.

In the gene information table, there are 50 995 full-length genes. The basic gene information including chromosome positions on the reference pan-genome, strand, CDS length and exon number, are contained in the table. Visualization and detailed gene information, such as gene categorization (core/distributed), gene presence frequency, gene ontology, coding sequence and protein sequence could be acquired by clicking the related links. The location of a genomic region can also be searched in a format of 'chromosome ID: start coordinate-end coordinate'.

A total of 226 runs of RNA-seq data from diverse rice tissues were collected in the gene expression profile table. The detailed information of gene expression profiles could be acquired and visualized in the genome browser.

Figure 4. An example of searching the shortlist of candidate donor for early-*japonica* breeding by RPAN. With the visualization function for the distributed key gene Os08g0174500 (**A**–**C**) controlling the day-length sensitivity of rice, the donors were shortlisted by visualization of the PAVs of Os08g0174500 (**D**).

Use cases

Here we give two examples on using RPAN for rice breeding study. One case is that researchers often search candidate genes for abiotic stress tolerances. Os12g0569700 is a gene with potentially important roles in rice acclimation to salt and drought stresses (33). Users can input the gene ID, Os12g0569700, to RPAN, and RPAN will show that this gene is present in 1107 accessions and 795 of them are *Japonica*. The phylogenetic tree of frequency distribution and heat map of this gene (Figure 3) show this gene is *Japonica*-dominant with very low frequencies in the other varietal groups. Further screening of donors for salt/drought tolerances should be focused on the accessions with full length (presence) of this gene.

Another case is selecting parental lines from the 3010 accessions for developing early maturing *Japonica* cultivars. In this work, donors with insensitivity to day length would be desirable. Instead of phenotyping of the whole set of 3010 accessions, breeders can shortlist the candidate accessions with genes controlling day length insensitivity. Gene Os08g0174500 is known to be able to suppress flowering time under the long-day condition and regulate the plant height and grain yield in rice (34–36). Users can firstly search the gene ID, and get the results shown in Figure 4A–C. RPAN shows that Os08g0174500 is a distributed gene, and is *Japonica*-dominant with the highest frequency of 96.8% in varietal subgroup J_{G9} (Figure 4B and C). To further figure out the possible donors insensitive to the day-length carrying the non-functional alleles of Os08g0174500, users can pick multiple tracks with relatively significant differences in the Os08g0174500 regions (Figure 4D). After clicking the 'submit' button, as shown by the genome browser panel, some accessions (CX106 and B026 as examples here) can be found to carry the complete sequences of Os08g0174500, while the others have different sizes of deletions. Then, users are able to shortlist the candidates to smaller number of accessions (B024, IRIS_313-11275, B060, B067, B112 and IRIS_313-11859 in this example) which do not have this gene for further phenotyping. Finally, based on their flowering times under the long-day (LD) condition at Northern China, users may find that B024 could be a desirable donor for their breeding purposes (Supplementary Table S2).

Here, we show an example we did with RPAN for rice biology. Because the distributed genes of the rice pan-genome, particularly those group specific ones, are expected to have contributed significantly to the adaptations of specific rice varietal populations to their environments. Gene information in our rice pan-genome browser can be very useful for rice scientists to determine functions of those 'domestication' genes of rice (*O. sativa* L.) (37). We extracted the 132 domestication protein-coding genes reported previously (Supplementary Table S3). Through searching (Fig-

A Gene Statistics

#Total genes	132
#Core genes	103
#Candidate core genes	9
#Distributed genes	20
#Subspecies-unbalanced genes	6
#*Indica*-dominant genes	3
#*Japonica*-dominant genes	3
#Subspecies-specific genes	0
#*Indica*-specific genes	0
#*Japonica*-specific genes	0
#AUS-specific genes	0
#ARO-specific genes	0
#Subgroup-unbalanced genes	15
#*Indica*-subgroup-unbalanced genes	14
#*Japonica*-subgroup-unbalanced genes	3
#Random genes	5

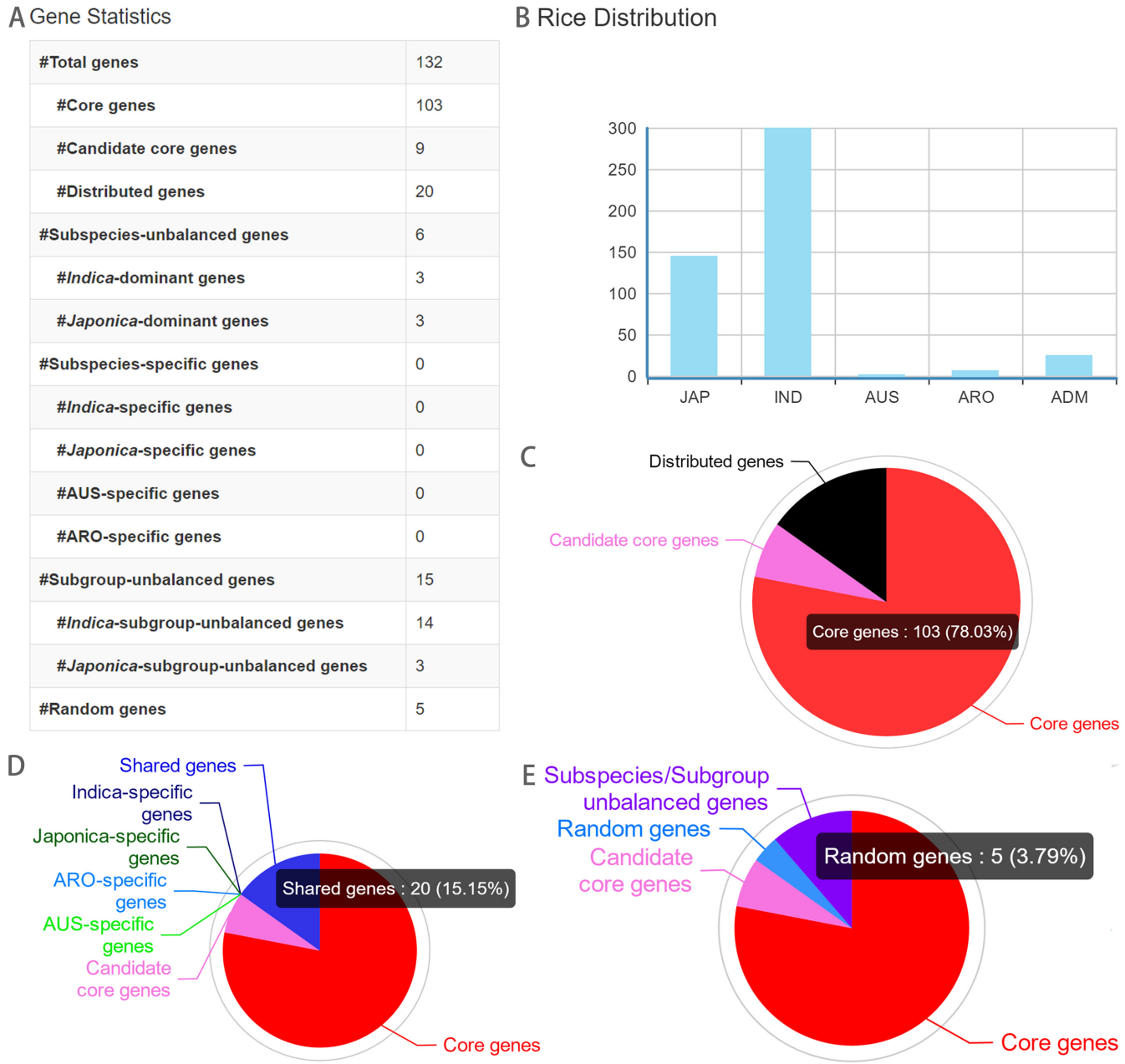

Figure 5. Examples of search and visualization functions of RPAN. A list of 132 genes associated with domestication were searched. Search results include (**A**) statistics about categorization of the genes; (**B**) the distribution of rice accessions containing all the genes and (**C**), (**D**) and (**E**), visualization of gene categorization in pie charts.

ure 5) in our pan-genome browser, we observed that 112 (84.8% of them) are the core or candidate core genes. Of the remaining 20 distributed genes, six genes are dominant in one of the subspecies, implicating their greater contributions to the population differentiation in rice.

DISCUSSION

Clearly, pan-genome analyses have expanded the genome analyses from the level of focusing on all functional genes in single or few reference genomes of a species to the comprehensive analyses of all genes in large numbers of individuals of different populations of a species. This expansion is essential to understand the whole genomic diversity of any species. Since the public availability of the sequencing data and seeds of the 3010 rice accessions from the 3K RGP, tremendous efforts have been taken to analyze this huge amounts of rice genome sequence data to understand the population genomic organization of rice and to phenotype the 3K rice accessions to identify loci associated with important traits by GWAS. Clearly, the rice pan-genome browser and the analytic tools developed in this study are expected to contribute significantly to these efforts. As the functions and associated traits of more genes in the rice pan-genome

are determined in the global rice functional genomics efforts, our browser will be updated regularly to facilitate various needs from the scientific community.

As the fast progress of sequencing technology, it is becoming increasingly tractable to generate whole-genome sequencing data for a large number of individuals of important plant and animal species, in which, visualization of the generated huge data is of particular importance in subsequent data analysis. RPAN, the rice pan-genome browser we present here, will help rice researchers to search and visualize their results in a pan-genome context and also provide a useful template and tool to facilitate searching and visualization of results from future pan-genome analyses. In addition, the strategy and visualization method in RPAN can be a good example to many other species with a large number of individual genomes available.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Center for High Performance Computing at Shanghai Jiao Tong University for computing. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

FUNDING

National Natural Science Foundation of China [61272250, 61472246, J1210047]; National Basic Research Program of China [2013CB956103]; National High-Tech R&D Program (863) [2014AA02150, 2012AA101601]; China National Transgenic Plant Special Fund [2014ZX08012-002, 2016ZX08012-002]; Programme of Introducing Talents of Discipline to Universities [111 Project, B14016]; Bill & Melinda Gates Foundation Project [OPPGD1393]; Shenzhen Peacock Plan; and the Key Special Project on Molecular Design Breeding for Rice [2016YFD0101800]. Funding for open access charge: National Natural Science Foundation of China [61272250, 61472246, J1210047]; National Basic Research Program of China [2013CB956103]; and National High-Tech R&D Program (863) [2014AA02150, 2012AA101601].

Conflict of interest statement. None declared.

REFERENCES

1. Weigel,D. and Mott,R. (2009) The 1001 genomes project for Arabidopsis thaliana. *Genome Biol.*, **10**, 107.
2. Siva,N. (2008) 1000 Genomes project. *Nat. Biotechnol.*, **26**, 256.
3. Rice Genome Project. (2014) The 3,000 rice genomes project. *Gigascience*, **3**, 7.
4. Tettelin,H., Masignani,V., Cieslewicz,M.J., Donati,C., Medini,D., Ward,N.L., Angiuoli,S.V., Crabtree,J., Jones,A.L., Durkin,A.S. *et al.* (2005) Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U.S.A.*, **102**, 13950–13955.
5. Vernikos,G., Medini,D., Riley,D.R. and Tettelin,H. (2015) Ten years of pan-genome analyses. *Curr. Opin. Microbiol.*, **23**, 148–154.
6. Li,R., Li,Y., Zheng,H., Luo,R., Zhu,H., Li,Q., Qian,W., Ren,Y., Tian,G., Li,J. *et al.* (2010) Building the sequence map of the human pan-genome. *Nat. Biotechnol.*, **28**, 57–63.
7. Yao,W., Li,G.W., Zhao,H., Wang,G.W., Lian,X.M. and Xie,W.B. (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. *Genome Biol.*, **16**, 187.
8. Schatz,M.C., Maron,L.G., Stein,J.C., Hernandez Wences,A., Gurtowski,J., Biggers,E., Lee,H., Kramer,M., Antoniou,E., Ghiban,E. *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.*, **15**, 506.
9. Li,Y.H., Zhou,G., Ma,J., Jiang,W., Jin,L.G., Zhang,Z., Guo,Y., Zhang,J., Sui,Y., Zheng,L. *et al.* (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. *Nat. Biotechnol.*, **32**, 1045–1052.
10. Hirsch,C.N., Foerster,J.M., Johnson,J.M., Sekhon,R.S., Muttoni,G., Vaillancourt,B., Peñagaricano,F., Lindquist,E., Pedraza,M.A. and Barry,K. (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.
11. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler,D. (2002) The human genome browser at UCSC. *Genome Res.*, **12**, 996–1006.
12. Speir,M.L., Zweig,A.S., Rosenbloom,K.R., Raney,B.J., Paten,B., Nejad,P., Lee,B.T., Learned,K., Karolchik,D., Hinrichs,A.S. *et al.* (2016) The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res.*, **44**, D717–D725.
13. Hennig,A., Bernhardt,J. and Nieselt,K. (2015) Pan-Tetris: an interactive visualisation for Pan-genomes. *BMC Bioinformatics*, **16**, S3.
14. Lajugie,J., Fourel,N. and Bouhassira,E.E. (2015) GenPlay Multi-Genome, a tool to compare and analyze multiple human genomes in a graphical interface. *Bioinformatics*, **31**, 109–111.
15. Seitzer,P., Huynh,T.A. and Facciotti,M.T. (2013) JContextExplorer: a tree-based approach to facilitate cross-species genomic context comparison. *BMC Bioinformatics*, **14**, 18.
16. Shi,X., Peng,J., Yu,X., Zhang,X., Li,D., Liu,B., Kong,F. and Yuan,X. (2015) PopGeV: a web-based large-scale population genome browser. *Bioinformatics*, **31**, 3048–3050.
17. Geihs,M., Yan,Y., Walter,K., Huang,J., Memari,Y., Min,J.L., Mead,D., Consortium,U.K., Hubbard,T.J., Timpson,N.J. *et al.* (2015) An interactive genome browser of association results from the UK10K cohorts project. *Bioinformatics*, **31**, 4029–4031.
18. Noll,A., Grundmann,N., Churakov,G., Brosius,J., Makalowski,W. and Schmitz,J. (2015) GPAC-genome presence/absence compiler: a web application to comparatively visualize multiple genome-level changes. *Mol. Biol. Evol.*, **32**, 275–286.
19. Luo,R., Liu,B., Xie,Y., Li,Z., Huang,W., Yuan,J., He,G., Chen,Y., Pan,Q., Liu,Y. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.
20. Kurtz,S., Phillippy,A., Delcher,A.L., Smoot,M., Shumway,M., Antonescu,C. and Salzberg,S.L. (2004) Versatile and open software for comparing large genomes. *Genome Biol.*, **5**, R12.
21. Fu,L., Niu,B., Zhu,Z., Wu,S. and Li,W. (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.
22. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
23. Alexandrov,N., Tai,S., Wang,W., Mansueto,L., Palis,K., Fuentes,R.R., Ulat,V.J., Chebotarov,D., Zhang,G., Li,Z. *et al.* (2015) SNP-Seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res.*, **43**, D1023–D1027.
24. Zheng,T., Yu,H., Zhang,H., Wu,Z., Wang,W., Tai,S., Chi,L., Ruan,J., Wei,C., Shi,J. *et al.* (2015) Rice functional genomics and breeding database (RFGB): 3K-rice SNP and InDel sub-database. *Chin. Sci. Bull.*, **60**, 367–371.
25. Sakai,H., Lee,S.S., Tanaka,T., Numa,H., Kim,J., Kawahara,Y., Wakimoto,H., Yang,C.C., Iwamoto,M., Abe,T. *et al.* (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.*, **54**, e6.
26. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
27. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G., Durbin,R. and Genome Project Data ProcessingSubgroup. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

28. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
29. Kim,D., Langmead,B. and Salzberg,S.L. (2015) HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods*, **12**, 357–360.
30. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
31. Skinner,M.E., Uzilov,A.V., Stein,L.D., Mungall,C.J. and Holmes,I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
32. Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
33. Zou,J., Liu,C., Liu,A., Zou,D. and Chen,X. (2012) Overexpression of OsHsp17. 0 and OsHsp23. 7 enhances drought and salt tolerance in rice. *J. Plant Physiol.*, **169**, 628–635.
34. Yan,W.H., Wang,P., Chen,H.X., Zhou,H.J., Li,Q.P., Wang,C.R., Ding,Z.H., Zhang,Y.S., Yu,S.B., Xing,Y.Z. *et al.* (2011) A major QTL, Ghd8, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Mol. Plant*, **4**, 319–330.
35. Thirumurugan,T., Ito,Y., Kubo,T., Serizawa,A. and Kurata,N. (2008) Identification, characterization and interaction of HAP family genes in rice. *Mol. Genet. Genomics: MGG*, **279**, 279–289.
36. Wei,X., Xu,J., Guo,H., Jiang,L., Chen,S., Yu,C., Zhou,Z., Hu,P., Zhai,H. and Wan,J. (2010) DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. *Plant Physiol.*, **153**, 1747–1758.
37. Huang,X., Kurata,N., Wei,X., Wang,Z.-X., Wang,A., Zhao,Q., Zhao,Y., Liu,K., Lu,H. and Li,W. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497–501.