PCP
PLANT & CELL PHYSIOLOGY

Special Online Collection – Database Paper

OryzaGenome: Genome Diversity Database of Wild *Oryza* Species

Hajime Ohyanagi[1,2,3,9], Toshinobu Ebata[4], Xuehui Huang[5], Hao Gong[5], Masahiro Fujita[1], Takako Mochizuki[6], Atsushi Toyoda[7], Asao Fujiyama[7,8], Eli Kaminuma[6,8], Yasukazu Nakamura[6,8], Qi Feng[5], Zi-Xuan Wang[1,5], Bin Han[5] and Nori Kurata[1,8,*]

[1]Plant Genetics Laboratory, National Institute of Genetics, Mishima, Japan
[2]Bioinformatics Laboratory, Meiji University, Kawasaki, Japan
[3]Tsukuba Division, Mitsubishi Space Software Co., Ltd., Tsukuba, Japan
[4]DYNACOM Co., Ltd., Chiba, Japan
[5]National Center for Gene Research, Chinese Academy of Sciences, Shanghai, PR China
[6]Genome Informatics Laboratory, National Institute of Genetics, Mishima, Japan
[7]Comparative Genomics Laboratory, National Institute of Genetics, Mishima, Japan
[8]Department of Genetics, School of Life Science, Graduate University for Advanced Studies, Mishima, Japan
[9]Present address: Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia.
*Corresponding author: E-mail, nkurata@nig.ac.jp; Fax, +81-55-981-6879.
(Received September 1, 2015; Accepted October 26, 2015)

The species in the genus *Oryza*, encompassing nine genome types and 23 species, are a rich genetic resource and may have applications in deeper genomic analyses aiming to understand the evolution of plant genomes. With the advancement of next-generation sequencing (NGS) technology, a flood of *Oryza* species reference genomes and genomic variation information has become available in recent years. This genomic information, combined with the comprehensive phenotypic information that we are accumulating in our Oryzabase, can serve as an excellent genotype–phenotype association resource for analyzing rice functional and structural evolution, and the associated diversity of the *Oryza* genus. Here we integrate our previous and future phenotypic/habitat information and newly determined genotype information into a united repository, named OryzaGenome, providing the variant information with hyperlinks to Oryzabase. The current version of OryzaGenome includes genotype information of 446 *O. rufipogon* accessions derived by imputation and of 17 accessions derived by imputation-free deep sequencing. Two variant viewers are implemented: SNP Viewer as a conventional genome browser interface and Variant Table as a text-based browser for precise inspection of each variant one by one. Portable VCF (variant call format) file or tab-delimited file download is also available. Following these SNP (single nucleotide polymorphism) data, reference pseudomolecules/scaffolds/contigs and genome-wide variation information for almost all of the closely and distantly related wild *Oryza* species from the NIG Wild Rice Collection will be available in future releases. All of the resources can be accessed through http://viewer.shigen.info/oryzagenome/.

Keywords: Database • Genome diversity • Genus *Oryza* • NIG Wild Rice Collection • NGS • Oryzabase • *Oryza rufipogon* • SNP.

Abbreviations: DRA, DDBJ Sequence Read Archive; ENA, European Nucleotide Archive; MAF, minor allele frequency; NGS, next-generation sequencing; SNP, single nucleotide polymorphism; VCF, variant call format.

Introduction

Rice is one of the three major staple crops and is widely grown around the world particularly in Asian countries, providing about 20% of the world's daily food supply as measured by calorie intake, similar to that provided by wheat (Berkman et al. 2012). In addition, historically it has been the most prominent monocot plant model species in academic biology. In particular, the whole-genome DNA sequence of a cultivated rice, *Oryza sativa* ssp. *japonica* cv. Nipponbare, was revealed in 2004 by the activity of an international research consortium, the IRGSP (International Rice Genome Sequencing Project 2005, Kawahara et al. 2013). It was followed by the whole-genome annotation project, RAP (Rice Annotation Project), also conducted in the framework of an international collaboration (Ohyanagi et al. 2006, Itoh et al. 2007, Tanaka et al. 2008, Sakai et al. 2013). With the achievements of these epic international projects, rice remains a standard model plant for the present century. Recent activities in rice research cover not only genomics itself, but transcriptomics, proteomics, phenomics and other multiple-omics research projects (Komatsu 2005, Miyao et al. 2007, Fujita et al. 2010, Hamada et al. 2011, Helmy et al. 2011, Nagamura et al. 2011, Sato et al. 2013, Ohyanagi et al. 2015).

Here we have a new mode of rice genomic research utilizing state-of-the-art DNA short read sequencing technology [next-generation sequencing (NGS)], namely population genomics with NGS. In terms of bioinformatics, it is relatively easy to resequence and analyze multiple genome sequences, and

Plant Cell Physiol. 57(1): e1(1–7) (2016) doi:10.1093/pcp/pcv171, Advance Access publication on 16 November 2015, available online at www.pcp.oxfordjournals.org

detect genomic variants from cultivated rice or species closely related to cultivated rice. This takes advantage of the highly accurate reference genome sequence of *O. sativa* ssp. *japonica* cv. Nipponbare (Os-Nipponbare-Reference-IRGSP-1.0) (Kawahara et al. 2013). Using these methods, studies with genome-wide variant information and their accumulated geographical and historical origin information are facilitating further exploration into the origin of rice cultivation and genomic regions associated with the domestication processes (Huang et al. 2012). In addition, the challenges of deciphering the whole-genome DNA sequences of distantly related wild rice species de novo have been met and the outcomes have been accumulated (Chen et al. 2013, Wang et al. 2014, Zhang et al. 2014).

In this decade, we have constructed and been maintaining an integrated biological and genome information resource, Oryz abase (http://www.shigen.nig.ac.jp/rice/oryzabase/) (Kurata and Yamazaki 2006), as one of the resources of the National BioResource Project (NBRP) in Japan hosted by the National Institute of Genetics (NIG), Japan. The genus *Oryza* is thought to be a highly genetically diverged lineage including nine genome types and 23 species, and information about plant morphology and anatomy, geographical origins, mutants and genetic resources (especially for wild accessions within the genus) has been deposited in Oryzabase. It also provides access to rice germplasm resources for those who need to obtain further biological information by conducting experiments. Oryzabase covers approximately 1,700 accessions ranging from closely to distantly related wild accessions concealing as yet unrevealed tetraploidy issues. The biological significance revealed so far indicates a high potential to make it the best repository of the genus *Oryza* in the next decade (Huang et al. 2010, Huang et al. 2012).

Here we release a database allied to Oryzabase, namely OryzaGenome, which is designed to serve and visualize a massive amount of genomic variation and genome sequence determined using NGS technology. The current version of OryzaGenome consists of genomic variants from 446 *O. rufipogon* accessions derived by an imputation method and variants from 17 accessions by imputation-free deeper (up to approximately 90×) sequencing along with the Os-Nipponbare-Reference-IRGSP-1.0 reference genome of *O. sativa* ssp. *japonica* cv. Nipponbare. Our goal is to establish a pan-*Oryza* genomic repository that covers both reference genome sequences and genomic variant information. In this article, we introduce the current status of the OryzaGenome and discuss its future perspectives.

Database Contents and Web Interface

Database contents

OryzaGenome principally provides genome sequence/variant information for wild *Oryza* species together with that of several cultivated strains, in close collaboration with Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/), ensuring easy access to information about geographical origins, phenotypic traits, mutants and genetic resources. As of the OryzaGenome release 1.0 (on June 19, 2015), SNP information in two categories (Imputation-free SNPs and Imputed SNPs) is available (see below and **Table 1**). The third category, wild genome reference sequences yet to be uncovered (and their variants in each species), will be available in a forthcoming release (see 'Conclusion and Future Directions').

Imputation-free SNPs

Currently this category comprises 11 *O. rufipogon* accessions and two *O. longistaminata* accessions (W1413 and W1508) as an outgroup. Each of them was resequenced at around 3.7–32 × average depth and covering about 51–95% of the reference genome in mapping analyses (see the **Materials** and **Methods** and **Table 1**). In addition, three cultivated *O. sativa* strains, namely ssp. *japonica* cv. Nipponbare (NIG stock, short reads have not been published), ssp. *japonica* cv. Nongken-58 and ssp. *indica* cv. Guangluai-4 sequenced in our previous work (Huang et al. 2010) were retrieved, and one more cultivar, the aus-type Kasalath sequenced by Sakai et al. (2014) was downloaded from a publicly available databank. All of the genome short reads from each of 17 wild accessions or *sativa* cultivars were mapped/aligned onto the reference genome of *O. sativa* (Os-Nipponbare-Reference-IRGSP-1.0) with various mapping rates and genome coverage (see the Materials and Methods and **Table 1**). Then according to our standard filtering policy (see the Materials and Methods), the highly accurate SNPs were extracted from the mapping analysis outcomes and presented in OryzaGenome. Currently the total number of collected SNPs is 23,458,338 in 17 accessions/cultivars. The statistics of the deeper NGS genome SNP information are summarized in **Table 1**. The raw data of our illumina Genome-Seq reads for each accession are also available in public archives with the indicated DRA or ENA numbers in the accession list (see 'Raw data availability').

Imputed SNPs

As of August 2015, this category contains 339,007,070 SNPs from 446 accessions of *O. rufipogon*, the direct ancestor of cultivated rice *O. sativa*, incorporated from our previous publication (Huang et al. 2012). The raw data of illumina Genome-Seq reads for the accessions can be retrieved from the archives using the indicated ENA accession number (see 'Raw data availability'). The genetic variant information was accumulated with respect to the reference genome sequences of *O. sativa*. While the genome co-ordinates were based on IRGSP-build4.0 in our previous work (Huang et al. 2012), the co-ordinates were converted to the latest Os-Nipponbare-Reference-IRGSP-1.0 in this study (see the Materials and Methods). Although most of the 446 wild accessions were sequenced at only about 1 × genome depth covering <10% of the genome, based on the mapped sequences of all 446 accessions, SNPs were imputed (method reported in Huang et al. 2010) and reconstructed for about 30% of the genome of each accession (SNPs located at 200 bp intervals on average). These Imputed SNPs were proven to show high accuracy, with a probability of >98% (Huang et al. 2010).

Table 1 Statistics of mapping analysis and detection of genome variants (SNPs) for deeper NGS genome sequences

				mapping analysis					SNP detection			
Species/ Ecotype	**Cultivar name/NIG Wid Rice accession**	**Number of read pairs (original)**	**Number of read pairs (after preprocessing)**	**Mapping rate (%, average)**	**Mapping rate (%, read1)**	**Mapping rate (%, read2)**	**Average depth (x times of reference genome)**	**Genome coverage (%)**	**Number of SNPs(all)**	**Number of SNPs (homogeneous)**	**Number of SNPs (heterogeneous)**	**heterogeneous ratio (%, hetero./all)**
Os-japonica	Nipponbare (NIG stock)	27,163,585	21,977,537	85.28	94.17	76.39	9.131	89.75	12,376	4,937	7,439	60.11
	Nongken-58	68,213,581	60,824,534	98.50	98.65	98.36	12.69	96.54	46,422	38,261	8,161	17.58
Os-indica	Guangluai-4	54,309,982	45,105,225	92.86	93.02	92.70	15.13	90.21	1,231,216	1,145,128	86,088	6.992
Os-aus	Kasalath	206,469,104	199,240,404	91.53	91.54	91.51	92.39	92.84	2,020,481	1,865,985	154,496	7.646
Or-I	W0106	25,090,954	21,291,026	69.83	76.76	62.90	7.481	78.26	269,386	225,060	44,326	16.45
	W0630	79,734,147	70,417,961	90.13	90.31	89.94	32.37	89.10	2,201,911	2,053,089	148,822	6.759
	W1230	68,999,113	61,736,419	89.72	89.89	89.55	28.36	89.27	2,143,756	1,968,156	175,600	8.191
	W1921	24,756,280	21,323,134	70.03	77.36	62.69	7.086	77.87	191,542	155,355	36,187	18.89
Or-II	W0120	69,331,945	61,953,915	89.88	90.03	89.72	28.47	90.55	2,082,384	1,569,524	512,860	24.63
	W0180	63,097,081	56,313,044	88.89	89.11	88.68	25.61	89.44	1,984,170	1,629,437	354,733	17.88
	W1236	71,082,413	63,974,789	89.31	89.45	89.17	29.21	91.44	2,507,539	1,583,395	924,144	36.85
	W1715	59,688,661	53,023,455	89.84	90.03	89.64	24.36	95.20	2,776,086	974,774	1,801,312	64.89
	W1981	68,868,645	61,843,301	89.13	89.27	89.00	28.17	89.95	2,374,591	1,687,582	687,009	28.93
Or-III	W0593	72,006,505	63,695,769	88.68	88.86	88.50	28.94	88.44	2,117,072	1,963,157	153,915	7.270
	W1943	32,091,193	30,369,210	94.38	94.34	94.41	15.06	92.10	769,640	703,427	66,213	8.603
O. longistaminata	W1413	36,304,086	21,237,761	61.89	67.79	56.00	6.571	54.95	424,135	280,952	143,183	33.76
	W1508	15,768,341	11,679,929	66.15	68.95	63.34	3.766	51.07	305,631	198,973	106,658	34.90

Os and *Or* stand for *O. sativa* and *O. rufipogon*, respectively.
The ecotype categories for *O. rufipogon* are according to our previous work (Huang et al. 2012).

Plant Cell Physiol. 57(1): e1(1–7) (2016) doi:10.1093/pcp/pcv171

PCP PLANT & CELL PHYSIOLOGY

Overview of the OryzaGenome Web Interface

The accumulated genomic information in OryzaGenome is available via the internet through HTTP access either as visual images or by batch download. The portal of OryzaGenome (http://viewer.shigen.info/oryzagenome/) provides two major features: SNP Viewer (**Fig. 1**) and Downloads. It also offers a few more basic hyperlinks (About OryzaGenome, Contact Information and Links).

SNP Viewer and Variant Table

The SNP Viewer has two modes, namely SNP Viewer (default, **Fig. 1A, B**) and Variant Table (**Fig. 1C**). The mode can be switched using the blue button on the right-hand side of the map window ('Variant Table' or 'Return to SNP Viewer').

As we have already described, all of the genomic resources stored in the current version of OryzaGenome (release 1.0) are based on the genome co-ordinates of the latest Os-Nipponbare-Reference-IRGSP-1.0 (*O. sativa* ssp. *japonica* cv. Nipponbare). The main genome map window (**Fig. 1B**) is shown in the bottom part of the page (below the control functionalities, **Fig. 1A**). In order to navigate to a particular physical position on the genome, keyword search, chromosome/position jump and clickable karyotype are available in the top part of SNP Viewer (**Fig. 1A**). By clicking on the long vertical buttons beside the map window (**Fig. 1B**), the window will be relocated to the flanking regions. In addition, by using the sliding scale, clicking on the +/− buttons or selecting between any two positions on the map window, the corresponding region will be redrawn in a zoomed-in or -out state (**Fig. 1B**). Just below the position control section, check boxes for toggling on/off the tracks for common annotations (MSU gene models, Rice-FL-cDNA, RAP gene models, BAC/PAC and GC content) are provided (**Fig. 1A**) to redraw the annotations in the map window. Those annotations were collected from the MSU Rice Genome Annotation Project Database (http://rice.plantbiology.msu.edu/) (Ouyang et al. 2007) and the Rice Annotation Project Database (http://rapdb.dna.affrc.go.jp/) (Sakai et al. 2013). In the next section, any accessions to be drawn in the map window can be selected and reordered by geographical and other properties (**Fig. 1A**). In the case of each of the 17 deeply sequenced accessions, the distribution of SNPs is shown in a pink heat map (**Fig. 1B**) and the reads themselves are shown in a histogram (zoomed-out view) or a piled graph (zoomed-in view) (**Fig. 1B**). In the case of the 446 shallowly sequenced accessions, the distribution of SNPs is shown as blue dots on the map (**Fig. 1B**), but the reads themselves are not shown. Here the accessions can be reordered vertically by particular information so that the rough associations between particular phenotypes and SNPs can be browsed using the pink and blue maps. Further inspection can be made utilizing the Variant Table (see below). For more detail on SNP Viewer functionalities, please refer to the tutorial (**Fig. 1A**, available by clicking the yellow link button 'Tutorial').

While the SNP Viewer shows the global aspect of SNP distribution, the Variant Table provides precise positional information on individual SNPs (**Fig. 1C**). The nucleotide characters

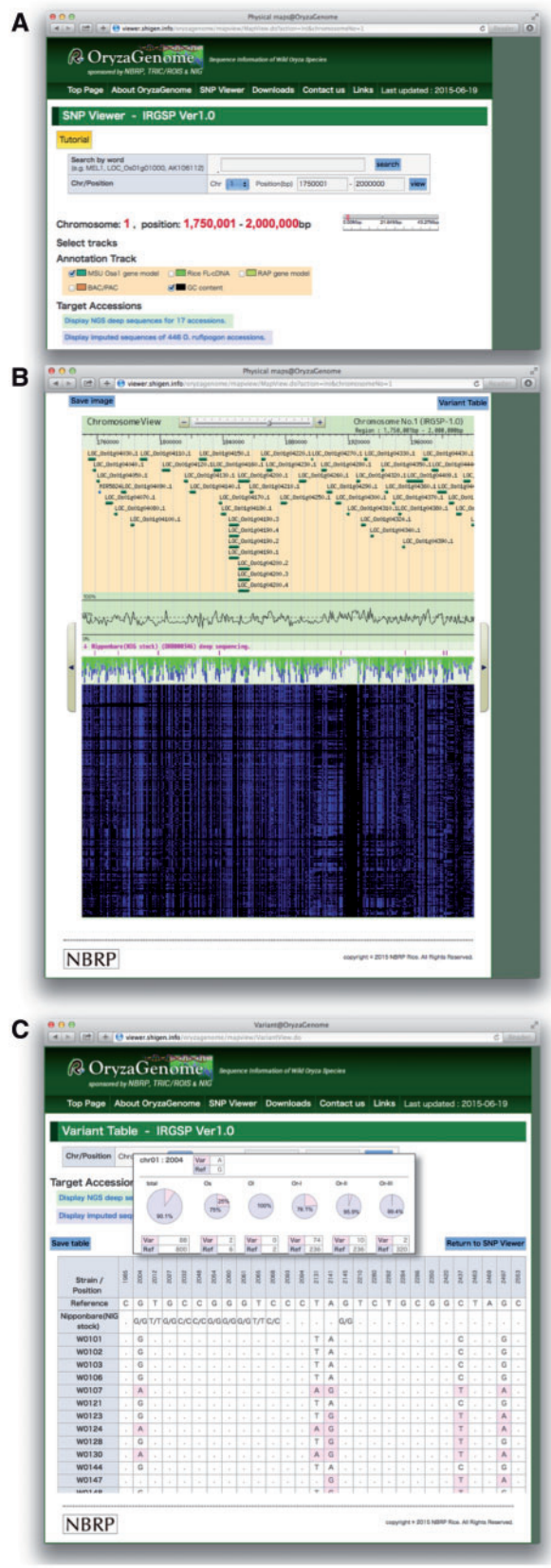

Fig. 1 OryzaGenome Web Interface. The SNP Viewer includes the control function (A) and the map window (B). It is possible to switch between the Variant Table (C) and the SNP Viewer, using the blue button on the right-hand side of the map window ('Variant Table' or 'Return to SNP Viewer').

of both the 17 deeply sequenced accessions and the 446 shallowly sequenced accessions that are selected and ordered in the SNP Viewer will be shown as they are, along with the reference nucleotides (**Fig. 1C**). If the mouse is hovered over a chromosome position, pie charts for allele frequencies in all species/ecotype categories will be visualized summarizing the SNP characteristics (**Fig. 1C**). The physical position is relocatable and SNP accessions to be drawn in the map window can be selected and reordered with the control functions in the top portion of the page (**Fig. 1C**).

The information in the SNP Viewer and Variant Table can be saved as PNG files or CSV files, using the 'Save image' button or 'Save table' button, respectively (**Fig. 1B, C**).

Batch download

The genome variant (SNP) information is also available as text. The download page offers hyperlinks to SNP information of each of the deep 17 accessions in VCF (variant call format) files and all of the shallow 446 accessions in a custom-defined tab-delimited file. The lists of Sequence Read Archive accession numbers and NIG Wild Rice Collection accession numbers (Wxxxx, xxxx is a four digit number) are also available (**Supplementary data**).

Conclusions and Future Directions

Here we introduce OryzaGenome, a united repository of genotype–phenotype information for the evolutionarily diverged species of *Oryza* genus. The currently available data set in OryzaGenome covers in particular cultivated rice (*O. sativa*) and closely related species (*O. rufipogon* and *O. longistaminata*). This is valuable genomic information, particularly for rice research scientists interested in associating the genotypic entity with phenotypic aspect, in order to elucidate hidden molecular mechanisms behind biological phenomena by means of population genomics, i.e. GWAS (genome-wide association studies). Their biological information (e.g. geographical origins, phenotypic/physiological traits and mutant phenotypes) can be accessed via hyperlinks from OryzaGenome to our allied database, Oryzabase. If further biological information is needed, the rice germplasm stock is available upon request. For more details, please refer to the Distribution section in Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/request/help). In future releases, we are planning to provide a more streamlined connection between OryzaGenome and Oryzabase to facilitate association analyses for a broader range of rice germplasm.

Besides their use in the latest functional genomics, wild rice reference genome sequences themselves are invaluable as standard resources for all kinds of biological studies in future rice science and crop breeding. In 2014, we initiated the ORYZA200 project with the aim of promoting our genotyping activities of approximately 200 wild rice genomes (N. Kurata et al. unpublished). In this project, three reference pseudomolecules/scaffolds in the *Oryza officinalis* complex are in the finishing stages and will be released via OryzaGenome in the near future. Subsequently, several more reference pseudomolecules/scaffolds/contigs will be reconstructed and released.

Furthermore, genome resequencing reads from each of around 200 *Oryza* accessions have already been generated with at least 10 × coverage, with the aim of documenting and characterizing the great genetic diversity in the genus *Oryza* including its enigmatic tetraploidy. Precise analyses of their genomic diversity are underway. The ORYZA200 accessions are selected from the NIG Wild Rice Collection covering nine genome types (AA, BB, BBCC, CC, CCDD, EE, FF, GG and HHJJ) and 21 species in the genus *Oryza*. For further details of the biological significance of genus *Oryza* species, please refer to the Rice in the World page in Oryzabase (http://www.shigen.nig.ac.jp/rice/oryzabase/education/riceInTheWorld). Eventually, all of the results of these comprehensive pan-*Oryza* genome diversity analyses will be accumulated in OryzaGenome released in close collaboration with Oryzabase.

In our ORYZA200 project, we are releasing further genomic information not only from resequencing, but also from novel genomic reference pseudomolecules/scaffolds/contigs. The additional genomic information is planned to be incorporated into OryzaGenome and released continually as it becomes available. We believe that the accumulation and dissemination of genomic information in the genus *Oryza* will facilitate total rice functional genomics and rice breeding science, and make bold contributions to solving the imminent global food security issue.

Materials and Methods

Reference information

For the reference genome sequences and reference gene annotations, the latest reference Nipponbare genome Os-Nipponbare-Reference-IRGSP-1.0 (*O. sativa* ssp. *japonica* cv. Nipponbare) (Kawahara et al. 2013), MSU Rice annotations (Ouyang et al. 2007) and RAP-DB annotations (Sakai et al. 2013) were employed.

Imputation-free SNPs of 17 accessions/cultivars

The cultivated and wild rice accessions were from the NIG Wild Rice Collection. They were selected to cover the greatest possible genetic diversity within the lineage. Their NIG Wild Rice Collection accession numbers and Sequence Read Archive accession numbers are listed in the **Supplementary data**. Each accession was maintained by a couple of self-propagations, and genomic DNA was extracted from a single plant for sequencing. In this study, 11 accessions/cultivars were newly sequenced on the Illumina GAIIx or HiSeq2000 platforms generating paired-end reads. Short reads of six other accessions/cultivars (W0106, W1921, W1943, Nongken-58, Guangluai-4 and Kasalath) were retrieved from our previous studies or from the DRA (DDBJ Sequence Read Archive, http://trace.ddbj.nig.ac.jp/dra/) (Huang et al. 2010, Huang et al. 2012, Sakai et al. 2014, Shenton et al. 2015).

The generated genome short reads were firstly quality inspected by FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), then adaptor sequences were trimmed out using cutadapt (https://code.google.com/p/cutadapt/). Low-quality reads were trimmed or filtered out by an empirically optimized custom Perl script. After those pre-processing steps, the remaining reads were mapped onto the reference genome using the 'bwa aln' and 'bwa sampe' commands in BWA (http://bio-bwa.sourceforge.net) (Li and Durbin 2009) with default parameters except for the proper insert size limitation (sampe -a 500). Repeat sequences scattered within the reference genome were not masked in the mapping process. Next, the variants were called using the 'samtools mpileup' command in samtools (http://samtools.sourceforge.net) (Li et al. 2009) with default parameters. High-confidence variants (SNPs) were extracted by the

following steps: (i) discard multiple-mapped reads; (ii) discard indels; (iii) coverage of each SNP position should be $\geq$8, and $\leq$100; and (iv) each SNP position should be covered on both the plus and minus strands. Finally, by empirically assessing the MAF (minor allele frequency) of each high-confidence SNP, they were divided into two categories, homozygous SNPs (MAF $<$ 0.25) and heterozygous SNPs (MAF $\geq$ 0.25).

The statistics concerning the imputation-free SNP information are summarized in **Table 1**.

Imputed SNPs of 446 *O. rufipogon* accessions

All of the genome SNP data for 446 *O. rufipogon* accessions were adapted from our previous study (Huang et al. 2012). Their NIG Wild Rice Collection accession numbers are listed in the **Supplementary data**. While the genome co-ordinates were based on IRGSP-build4.0 in the previous work (Huang et al. 2012), the co-ordinates were converted to the latest Os-Nipponbare-Reference-IRGSP-1.0 in this study (see below).

Reference genome co-ordinate conversion

In order to convert the co-ordinate of each SNP, the IRGSP-build4.0 and Os-Nipponbare-Reference-IRGSP-1.0 reference genome sequences were locally aligned around each SNP position by BLAST (Altschul et al. 1997), according to the following empirically defined procedure: (i) for each SNP, extract the 201 bp flanking sequence (the SNP nucleotide itself and the flanking 100 bp nucleotides on both sides) on the original genome (IRGSP-build4.0); (ii) search the counterpart of the 201 bp on the latest genome (Os-Nipponbare-Reference-IRGSP-1.0) with BLASTN homology search, allowing no indel and one mismatch at most, with 100% coverage; (iii) if there is more than one homologous region, discard the SNP; and (iv) if the SNP was converted onto a different chromosome, discard the SNP.

Raw data availability

DNA short reads data for imputation-free SNPs are deposited in the DRA (DDBJ Sequence Read Archive, http://trace.ddbj.nig.ac.jp/dra/) or the ENA (European Nucleotide Archive, http://www.ebi.ac.uk/ena). Their accession numbers are listed in the **Supplementary data**. Sequence reads for Imputed SNPs are from our previous studies (Huang et al. 2012), which have already been deposited in the ENA under accession number ERP001143.

System architecture and software

OryzaGenome was implemented on a UNIX server with CentOS version 7, Apache/Tomcat web server and PostgreSQL Database server. Java and C++ were employed as server-side application languages. JavaScript was adopted to implement client-side rich applications. The JavaScript libraries, jQuery (http://jquery.com), DataTables (https://www.datatables.net/), Magnific Popup (http://dimsemenov.com/plugins/magnific-popup/), Prototype (http://prototypejs.org/) and script.aculo.us (https://script.aculo.us/) were employed. Other conventional utilities for UNIX computing were appropriately installed on the server if necessary. All of the OryzaGenome resources are stored on the server and are available through HTTP access.

Supplementary data

Supplementary data are available at PCP online.

Funding

This work was supported by the National BioResource Project (NBRP) [the Genome Information Upgrading Program to (N.K. and A.T.)]; the Genetic Function Systems Project in the Transdisciplinary Research Integration Center at Research Organization of Information and Systems (TRIC/ROIS) [to N.K. and A.F.].

Acknowledgements

We thank Dr. Matthew R. Shenton for proofreading of this manuscript. Data sets provided in OryzaGenome are collected from the wild *Oryza* species maintained at NIG, Mishima, Japan under NBRP (the National BioResource Project in Japan) for 345 accessions and China National Rice Research Institute, Hangzhou, China, for 101 accessions. We would also like to thank the NBRP and NIG Bioresource Project for supporting the preservation of wild accessions and database construction. Data analyses were partially performed on the NIG SuperComputer Facilities hosted at NIG/ROIS (Research Organization of Information and Systems).

Disclosures

The authors have no conflicts of interest to declare.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Berkman, P.J., Lai, K., Lorenc, M.T. and Edwards, D. (2012) Next-generation sequencing applications for wheat crop improvement. *Amer. J. Bot.* 99: 365–371.

Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., et al. (2013) Whole-genome sequencing of Oryza brachyantha reveals mechanisms underlying Oryza genome evolution. *Nat. Commun.* 4: 1595.

Fujita, M., Horiuchi, Y., Ueda, Y., Mizuta, Y., Kubo, T., Yano, K., et al. (2010) Rice expression atlas in reproductive development. *Plant Cell Physiol.* 51: 2060–2081.

Hamada, K., Hongo, K., Suwabe, K., Shimizu, A., Nagayama, T., Abe, R., et al. (2011) OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. *Plant Cell Physiol.* 52: 220–229.

Helmy, M., Tomita, M. and Ishihama, Y. (2011) OryzaPG-DB: rice proteome database based on shotgun proteogenomics. *BMC Plant Biol.* 11: 63.

Huang, X., Kurata, N., Wei, X., Wang, Z.X., Wang, A., Zhao, Q., et al. (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490: 497–501.

Huang, X., Wei, X., Sang, T., Zhao, Q., Feng, Q., Zhao, Y., et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* 42: 961–967.

International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature* 436: 793–800.

Itoh, T., Tanaka, T., Barrero, R.A., Yamasaki, C., Fujii, Y., Hilton, P.B., et al. (2007) Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res.* 17: 175–183.

Kawahara, Y., de la Bastide, M., Hamilton, J.P., Kanamori, H., McCombie, W.R., Ouyang, S., et al. (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (N Y)* 6: 4.

Komatsu, S. (2005) Rice proteome database: a step toward functional analysis of the rice genome. *Plant Mol. Biol.* 59: 179–190.

Kurata, N. and Yamazaki, Y. (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol.* 140: 12–17.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25: 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25: 2078–2079.

Miyao, A., Iwasaki, Y., Kitano, H., Itoh, J., Maekawa, M., Murata, K., et al. (2007) A large-scale collection of phenotypic data describing an insertional mutant population to facilitate functional analysis of rice genes. *Plant Mol. Biol.* 63: 625–635.

Nagamura, Y., Antonio, B.A., Sato, Y., Miyao, A., Namiki, N., Yonemaru, J., et al. (2011) Rice TOGO Browser: a platform to retrieve integrated information on rice functional and applied genomics. *Plant Cell Physiol.* 52: 230–237.

Ohyanagi, H., Takano, T., Terashima, S., Kobayashi, M., Kanno, M., Morimoto, K., et al. (2015) Plant Omics Data Center: an integrated web repository for interspecies gene expression networks with NLP-based curation. *Plant Cell Physiol.* 56: e9.

Ohyanagi, H., Tanaka, T., Sakai, H., Shigemoto, Y., Yamaguchi, K., Habara, T., et al. (2006) The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. *Nucleic Acids Res.* 34: D741–D744.

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., et al. (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.* 35: D883–D887.

Sakai, H., Kanamori, H., Arai-Kichise, Y., Shibata-Hatta, M., Ebana, K., Oono, Y., et al. (2014) Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice. *DNA Res.* 21: 397–405.

Sakai, H., Lee, S.S., Tanaka, T., Numa, H., Kim, J., Kawahara, Y., et al. (2013) Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol.* 54: e6.

Sato, Y., Takehisa, H., Kamatsuki, K., Minami, H., Namiki, N., Ikawa, H., et al. (2013) RiceXPro version 3.0: expanding the informatics resource for rice transcriptome. *Nucleic Acids Res.* 41: D1206–D1213.

Shenton, M.R., Ohyanagi, H., Wang, Z.X., Toyoda, A., Fujiyama, A., Nagata, T., et al. (2015) Rapid turnover of antimicrobial-type cysteine-rich protein genes in closely related Oryza genomes. *Mol. Genet. Genomics* 290: 1753–1770.

Tanaka, T., Antonio, B.A., Kikuchi, S., Matsumoto, T., Nagamura, Y., Numa, H., et al. (2008) The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res.* 36: D1028–D1033.

Wang, M., Yu, Y., Haberer, G., Marri, P.R., Fan, C., Goicoechea, J.L., et al. (2014) The genome sequence of African rice (Oryza glaberrima) and evidence for independent domestication. *Nat. Genet.* 46: 982–988.

Zhang, Q.J., Zhu, T., Xia, E.H., Shi, C., Liu, Y.L., Zhang, Y., et al. (2014) Rapid diversification of five Oryza AA genomes associated with rice adaptation. *Proc. Natl. Acad. Sci. USA* 111: E4954–E4962.