Current Plant Biology 7–8 (2016) 16–25

Contents lists available at ScienceDirect

Current Plant Biology

journal homepage: www.elsevier.com/locate/cpb

SNP-Seek II: A resource for allele mining and analysis of big genomic data in *Oryza sativa*☆

Locedie Mansueto[a], Roven Rommel Fuentes[a], Dmytro Chebotarov[a], Frances Nikki Borja[a], Jeffrey Detras[a], Juan Miguel Abriol-Santos[a], Kevin Palis[a,b], Alexandre Poliakov[c,d], Inna Dubchak[c,d], Victor Solovyev[e], Ruaraidh Sackville Hamilton[a], Kenneth L. McNally[a], Nickolai Alexandrov[a], Ramil Mauleon[a,*]

[a] *International Rice Research Institute, College, Los Baños, Laguna, 4031, Philippines*
[b] *Boyce Thompson Institute, Ithaca, NY 14853, USA*
[c] *Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[d] *DOE Joint Genome Institute, Walnut Creek, CA 94598, USA*
[e] *Softberry, Inc., Mount Kisco, NY 10549, USA*

ARTICLE INFO

Keywords:
Allele mining
Oryza
SNP
Indel
Genotype database
Genetic diversity

ABSTRACT

The 3000 Rice Genomes Project generated a large dataset of genomic variation to the world's most important crop, *Oryza sativa* L. Using the Burrows-Wheeler Aligner (BWA) and the Genome Analysis Toolkit (GATK) variant calling on this dataset, we identified ~40 M single-nucleotide polymorphisms (SNPs). Five reference genomes of rice representing the major variety groups were used: Nipponbare (temperate *japonica*), IR 64 (*indica*), 93–11 (*indica*), DJ 123 (*aus*), and Kasalath (*aus*).

The results are accessible through the Rice SNP-Seek Database (http://snp-seek.irri.org) and through web services of the application programming interface (API). We incorporated legacy phenotypic and passport data for the sequenced varieties originating from the International Rice Genebank Collection Information System (IRGCIS) and gene models from several rice annotation projects. The massive genotypic data in SNP-Seek are stored using hierarchical data format 5 (HDF5) files for quick retrieval. Germplasm, phenotypic, and genomic data are stored in a relational database management system (RDBMS) using the Chado schema, allowing the use of controlled vocabularies from biological ontologies as query constraints in SNP-Seek.

In this paper, we discuss the datasets stored in SNP-Seek, architecture of the database and web application, interoperability methodologies in place, and discuss a few use cases demonstrating the utility of SNP-Seek for diversity analysis and molecular breeding.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

One of the biggest challenges facing rice farming is increasing worldwide production by at least 25% to meet the demands imposed by the projected increase in global population by 2030, side by side with the constraints brought by reduction of arable land, less available water, and more severe environmental stresses due to climate change [1]. Genetic gains from current breeding methods are insufficient to achieve the target yield increase, but solutions such as molecular breeding technologies and use of allelic diversity for important rice traits could potentially increase genetic gains of ongoing rice breeding programs. Molecular breeding technologies have been utilized to improve disease resistance, drought tolerance, and agronomically important traits [2–4]. The rice gene bank collections serve as a potential source of allelic diversity for important genes. In 2014, the 3k RG Project [5] completed sequencing of 3024 rice genomes from the International Rice Genebank Collection at the International Rice Research Institute and The China National Crop Gene Bank (CNCGB), generating over 17 terabytes of raw sequence data. Bioinformatic analyses for discovery

Abbreviations: 3k RG, 3000 Rice Genomes; API, Application Programming Interface; DAO, Data Access Object; HDF5, Hierarchical Data Format 5 (file format); HDRA, High Density Rice Array; indel, insertion or deletion in genomic region; IRGCIS, International Rice Genebank Collection Information System (http://www.irgcis.irri.org:81/grc/irgcishome.html); IRRI, International Rice Research Institute; RDBMS, Relational Database Management System; SNP, Single Nucleotide Polymorphism.

☆ This article is part of a special issue entitled "Genomic resources and databases", published in the journal Current Plant Biology 7–8, 2016.

* Corresponding author at: International Rice Research Institute, College, Los Baños, Laguna, 4031, Philippines.
E-mail address: r.mauleon@irri.org (R. Mauleon).

http://dx.doi.org/10.1016/j.cpb.2016.12.003
2214-6628/© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

of sequence variants using GATK Unified Genotyper (GATK-UG) [6] has identified over 40 million SNP variants and over 2.4 million short indels (<=50 bases long). Information about the accessions sequenced, and the variants discovered have been made available through Rice SNP-Seek database [7]. The database aims to provide easy access, through a user-friendly web interface, to the SNPs and indels from the 3k RG. Data and tools built into SNP-Seek allow for exploratory discoveries of genomic variant – trait associations and examine allelic diversity at genome regions of interest (e.g. known genes, QTLs). One of the features that sets SNP-Seek apart from other publicly accessible databases such as dbSNP at NCBI [8], Gramene [9], RiceVarMap [10], IC4R [11], and RMBreeding [12], is the interactive real-time visualization of millions of SNPs in thousands of rice varieties. This makes SNP-Seek a unique tool for allele mining [6]. We are committed to provide continuous development support to SNP-Seek to incorporate new analyses results, datasets, viewers and query interfaces for multiple reference genomes and assemblies, as well as include features that will be useful to the broader rice research community.

2. Materials and methods

2.1. Data and code and availability

The SNP discovery pipeline scripts are available at https://github.com/IRRI-Bioinformatics/snp-discovery-pipeline with information on running the pipeline at https://github.com/IRRI-Bioinformatics/snp-discovery-pipeline/wiki/How-to-run-the-pipeline. The scripts are configured to work on an high-performance computing (HPC) cluster with a SLURM Workload Manager [13].

The SNP-Seek web application is implemented in Java Spring and ZK frameworks. The codes and instructions for development or installation are at our repository https://bitbucket.org/irridev/iric_portal. The application requires access to the SNP-Seek PostgreSQL or Oracle database server. Raw and analyzed files, including genotypic data are available for bulk download at the Download page.

2.2. Variety (Passport) and phenotype data

Information for each variety including country of origin, genetic stock account number, and variety group is documented in the IRIC information site (http://iric.irri.org/resources/3000-genomes-project). Further description of the selected germplasm is provided by the 3k RG Project [5]. The phenotypic and passport data are from IRGCIS [14] at IRRI.

2.3. Variant data generation

2.3.1. Alignment and variant calling

The SNP discovery pipeline was used to generate SNPs for 3024 rice accessions. The rice accessions were aligned with five reference genomes: Nipponbare [15], 93-11 [4], Kasalath [16], DJ 123 and IR 64 [17], details of which are in Table 1. The reference genomes were indexed with BWA version 0.7.10[18] and SAMtools version 1.0[19]. The sequence dictionary was created using Picard Tools version 1.119 [20]. Alignment was done using BWA-MEM with the parameters '-M' for Picard compatibility and '-t 8' for 8 threads. Each SAM alignment file was compressed to a sorted BAM file using Picard Tools then processed for marking duplicates, fixing mate-pair information and adding or replacing read groups. The processed BAM file was realigned for local indels using the GATK Realigner Target Creator and Indel Realigner. After realignment, the BAM files for each read-pair were merged for each accession using SAMtools. SNP calling was then performed using the GATK-UG version 3.2-2 [21] with parameters 'glm BOTH', '-mbq 20', '-genotyping_mode DISCOVERY' and '-out_mode EMIT_ALL_SITES'. One variety, PUTTIGE:IRGC 5258801 (IRIS 313-8921), with low sequence coverage had 98% missing calls so it was excluded from further analyses.

2.3.2. Whole-genome DNA alignment

We used the VISTA pipeline infrastructure [22,23] for the construction of genome-wide pairwise DNA alignments between the five reference genomes. To align genomes, we used an efficient combination of global and local alignment methods. First, we obtained a map of large blocks of conserved synteny between genome-pairs by applying Shuffle-LAGAN glocal chaining algorithm [24] to local alignments produced by translated BLAT [25]. After that we used Supermap, the fully symmetric whole-genome extension to the Shuffle-LAGAN. Then, for each syntenic block, we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions.

The constructed genome-wide pairwise alignments can be downloaded from http://pipeline.lbl.gov/downloads.shtml .The alignments are accessible for browsing and performing various types of analysis through the VISTA browser at http://pipeline.lbl.gov/or through the SNP-Seek menu.

In each of the 10 resulting alignments, we calculated overall coverage, coverage of each reference genome in the alignments, coverage of different annotated sequences [26], the fraction of unique sequence for each genome, and mapping rates of the 3023 genomes to the reference genomes (Table 1). The reference genomes demonstrate high levels of similarity among them, with the total genome coverage among alignments at 70–92% levels (Supplementary Table 1). For each reference genome, there are unique regions (from 12.3 Mbp to 79.6 Mbp) that may harbor genes found only in these variety-group-specific genome segments. Consequently, allele variants cannot be discovered in accessions that are aligned to a reference genome that belong to a different variety group. This is shown by the reduced read mapping rate of accessions aligned to a reference genome that is not of the same variety group, and the highest read mapping occurring when accessions and reference genome belong to the same variety group (Table 1). These unique regions were used in the discovery of SNPs unique to a particular reference genome.

2.3.3. SNP and indel universe creation

The 3023 emit-all-sites Variant Call Format (VCF) files generated by GATK-UG [21] were analyzed to make the union of variants (SNP and indel "universe"). Due to the size of the data and the independent calling of SNPs and indels, we developed a joint genotyping tool to merge the variants instead of constructing a large VCF file that would be difficult to query. A variant position is reported and added to the list if at least one sample supports it and if the phred-scaled quality score is at least 30. The program ran in parallel by grouping the samples and assigning each group to a processor. The resulting position lists were merged before retrieving the respective calls from each sample before finally exporting the corresponding SNP and indel universe matrices to Hierarchical Data Format v5 (HDF5) [27] files. Two versions of HDF5 files were created for each dataset for efficient retrieval: one version has varieties set in rows and variant position in columns, optimized for queries made on a given set of varieties, returning variants in a long genome region. The other is a transposed version (variant positions in rows, varieties in columns) used when querying for a given set of variant positions/loci across all varieties.

We incrementally discovered new variants from unique genome regions of the other reference genomes from the pairwise genome

Table 1
Summary statistics of five published rice genomes used in SNP discovery and sequencing read mapping rates of 3k RG accessions (binned by 3 major variety groups) to the genomes.

Variety	Genome Publication	Assembly length (bps)	Number of contigs/scaffolds/ chromosomes	Number of annotated genes	Length of unique genome region from pairwise genome alignment (bps)	% reads mapping rate of 3k RG binned to 3 major variety groups		
						Aus	Indica	Temperate Japonica
IR 64 (Indica)	[17]	345,209,449	2919 scaffolds	37,758 (MAKER genes)	14,171,198	95.6	96.4	94.6
93-11 (Indica)	[4]	423,026,874	12 chromosomes 12,718 scaffolds	40,464 (GLEAN genes mRNA)	22,642,740	96.8	97.7	96.1
DJ 123 (Aus)	[17]	345,981,746	2819 scaffolds	37,812 (MAKER genes)	12,327,785	96.9	96.4	95.4
Kasalath (Aus)	[16]	401,141,708	12 chromosomes, 1 unmapped contig (14,822 contigs concatenated by 1000 Ns)	20,869 RAP-aligned genes; 53,662 genes predicted by tophat-cufflinks-cuffmerge	79,626,875	96.7	96.5	95.2
Nippon-bare (Japonica)	[15]	373,245,519	12 chromosomes + chloroplast + mitochondria + unmapped contig	37,869 representative, 8118 predicted genes (IRGSP1.0/RAP); 16,979 TE & 39,102 non TE genes/loci (RGAP 7)	32,056,098	96.1	96.5	97.2

alignment results (Supplementary Table 1). Variants were progressively reduced following the schema in Supplementary Fig. 1.

2.3.4. Generation of SNP subsets

In addition to the full dataset of 32 million Nipponbare-based SNP calls, we provide the following PLINK-format subsets: All Biallelic SNPs (29 M), Base (18 M), Filtered (4.8 M), and Core (404k). Each subset resulted from the successive application of filtering criteria fit for different analyses. The details of these are described in the SNP-Seek Download page (http://snp-seek.irri.org/_download.zul). In brief, starting from biallelic SNP set, three rounds of filtering were applied since despite following best practices in SNP calling, there can still be false positive SNPs ("pseudoSNPs") due to many factors such as alignment ambiguity and undetected paralogs. A common signature of such SNPs is an elevated proportion of heterozygous calls. The Base SNP set was obtained from Full Biallelic SNP set by removing SNPs that have excessive heterozygosity for a given allele frequency and level of inbreeding. These SNPs have a higher chance of being false positives, although some of them may be true SNPs having high heterozygosity due to selection.

The Filtered SNP set is the subset of Base SNP set comprising the SNPs that have minor allele frequency > 1% and less than 20% of missing genotypes per SNP. The Filtered SNP set is the default dataset in SNP-Seek. It is recommended for most analyses where rare variation is not prioritized. Researchers interested in rare SNPs should use either the Base SNP set or the complete set (if sensitivity is prioritized over specificity).

Additionally, a smaller Core SNP set was created by LD-pruning the filtered set with r2 > 0.8. This dataset still shares the diversity present in Filtered SNP set with much lower number of SNPs. It is useful for getting a first-look genome-wide overview of features of 3k RG in a less data intensive setting.

2.4. Genomic data integration

2.4.1. Gene models

We merged gene models from four annotations for Nipponbare – Rice Genome Annotation Project 7 (RGAP7) [15], Rice Annotation Project Database (RAP-DB) representative, RAP-DB predicted [28], and FGenesh++ [29] as one set of loci and loaded these into SNP-Seek, simplifying query and analysis of genome and variant features. Gene loci from different annotations were merged as one common locus (assigned with a new unifying locus ID, described in Supplementary Materials II) if (1) there is at least 50% overlap from the first and last coding regions (CDS) and (2) the gene loci are in the same strand. Gene function annotation from RGAP7 is used for the merged locus, unless it is determined as 'hypothetical', 'unknown', or 'expressed'. For these cases, the functional annotation from RAP-DB is used. The merging created 72,520 gene loci. The contribution of each source is shown in the Venn diagram in Fig. 1. The values in parenthesis are the number of loci from the source, while the number directly below it is the number of merged loci they generated. This number is lower because some loci from the same source are merged into one after the gene models from all sources are compared for overlaps. The numbers inside the Venn diagram refer to the number of merged loci at every possible intersection.

We also loaded the gene models of the other four reference genomes as generated by their respective projects. However, the loci designation used the name auto-generated by their corresponding annotation pipelines. So, we introduce a uniform gene model naming convention applicable to any sequenced variety of rice as described in Supplementary Materials II.

2.4.2. Marker and gene annotation data

To facilitate the annotation of SNP markers, we also incorporated data from additional analyses we made. This data includes the promoter regions from the FGenesh++ gene prediction pipeline that generated the gene models mentioned above. We also identified the effects of SNP variant on the RGAP7 gene models using the SNPEff software[30]. Data from public rice genomic databases relevant to gene-trait association discovery were also incorporated including,

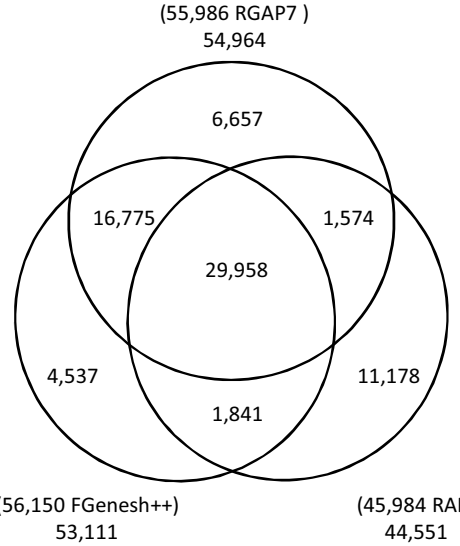

Fig. 1. Number of merged loci from the three sources of rice gene annotations.

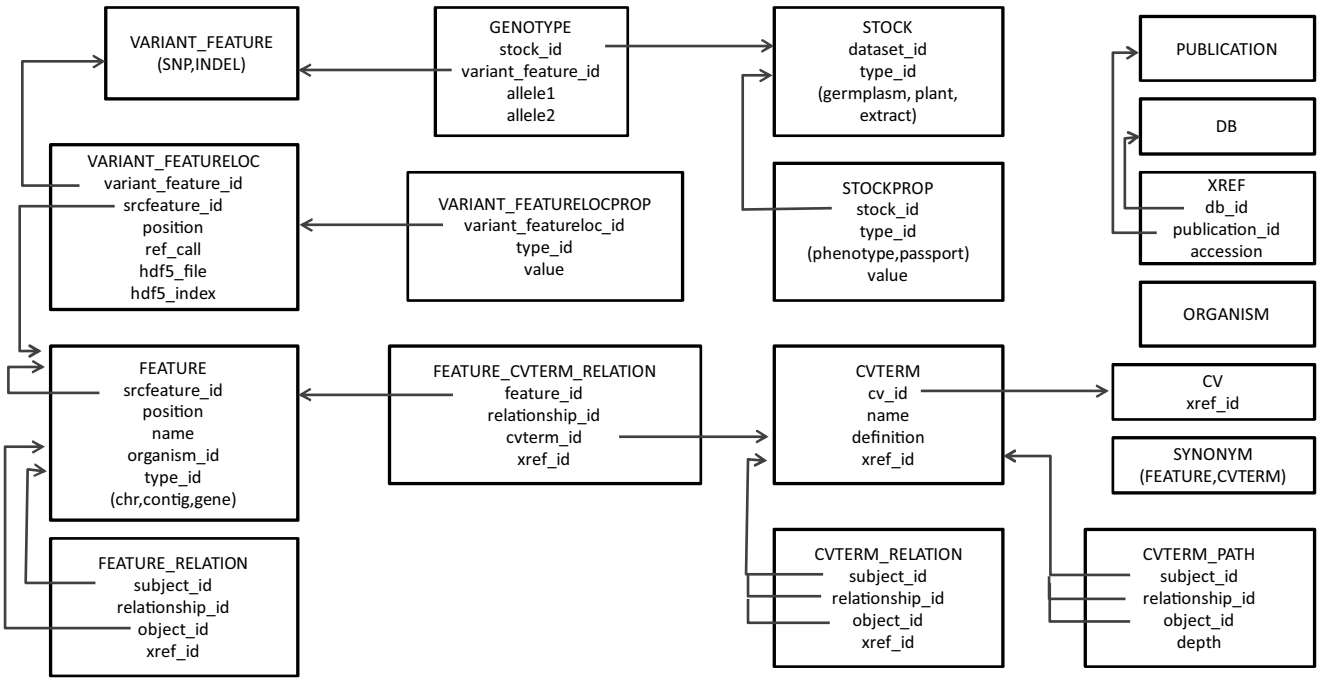

Fig. 2. Overview of SNP-Seek schema based on Chado.
Some relationships are not shown for clarity including: all relationship_id, type_id, dataset_id are referenced to CVTERM.cvterm_id; xref_id to XREF.xref_id; organism_id to ORGANISM.organism_id. Tables can be an abstract to more specific class of entities: VARIANT_FEATURE can be SNP_FEATURE or INDEL_FEATURE; SYNONYM can be CVTERM_SYNONYM or FEATURE_SYNONYM. The OGANISM table holds the different reference varieties. CVTERM_PATH is the pre-computed transitive closure relationships

but not limited to: gene ontology terms from RGAP7, Oryzabase [31], and OGRO [32]; Plant Trait (TO) and Plant (PO) ontology terms from Oryzabase; trait genes and QTL from Q-TARO [33]; rice promoter regions from PlantProm DB [34]; interacting genes from RiceNetv2 [35], and rice protein interactions from PRIN[36].

2.4.3. Biological ontologies

We designed SNP-Seek to be biologist-friendly especially in querying the diverse sets of biological data available. A simple but powerful approach is to use the biological concepts defined by the several ontology projects. This has simplified the querying of molecular markers using controlled biological terms and their transitive closure. Using the annotation data described in the Section 2.4.2 [*Marker and gene annotation data*], terms and relationships from Gene [37], Plant [38], Plant Trait [39], Sequence [40] Ontologies were loaded and referenced to gene models. We also mapped our legacy phenotypic data to Plant Trait and Crop Ontology [41] terms so that varieties can be queried using controlled trait terms.

2.4.4. Public rice genotypic data

We foresee hosting more publicly available data from other rice genotyping projects. A test case is the loading of the genotypic and variety data from the High Density Rice Array (HDRA) project [42]. When legacy passport and phenotypic data for the varieties are found in the IRGCIS, they are also loaded into SNP-Seek.

2.5. Information system design

2.5.1. Database schema

SNP-Seek uses the Chado [43] schema for storing most of the data. Chado is designed to be flexible to handle constantly evolving biological entities and relationships. Through the use of ontologies, controlled terms and hierarchy can be used to define biological entities and relationships in the relational database. A simplified overview to the schema is in Fig. 2. The major modules used are: STOCK for germplasm, phenotype and passport, CV for controlled terms and relations, FEATURE for genomic features, GENOTYPE for genotypic data and modules for referencing like PUBLICATION, DB, and SYNONYM. The X_RELATION tables handle relationships between two FEATURES, two CVTERMS or a FEATURE and a CVTERM, where the relationships are also controlled and evolving terms defined in CVTERM. The data loaded are listed in Table 2.

We made several modifications to the original Chado schema for the main purpose of performance. Ideally, SNPs and indels can be included in the FEATURE table but we separated them into SNP_FEATURE and INDEL_FEATURE. Also, SNPs or indels from different datasets can be merged into a single table but we created separate tables for each dataset. The major modification is our use of HDF5 files to store some of the genotypic dataset. The task of using the right RDBMS table or HDF5 file to query is assigned to the web application, particularly to the Data Access Object (DAO) layer in our software design.

An additional layer of abstraction between the DAO and the relational database is the use of views. Instead of querying the relational tables, the DAOs use views defined in the RDBMS so that changes in the schema only require a change in the views and not the DAO code, simplifying use of the SNP-Seek tables in either PostregSQL or Oracle RDBMS.

2.5.2. Web-application architecture

The web application architecture in Fig. 3 follows the Model-View-Controller (MVC) design to separate the viewers from the business logic, and the logic from the data storage. This is accomplished by defining interfaces in between layers and implemented in Java Spring Framework (https://spring.io). We used the ZK Framework (https://www.zkoss.org) to implement the user interfaces. The Model follows the Data Service and Data Access Layers pattern[44]. The middleware can be grouped into four modules based on the query features of SNP-Seek. The Genotype module handles SNPs and indels. The Variety module handles variety, phenotype, and passport data. The Genomics module handles genes,

Table 2
Data loaded to the Chado tables and their utility in the three (Genotype, Variety, Gene) SNP-Seek search features. Utility code: R-result, C-constraint, A-annotation/attribute, I-internal.

DB Table	Data	Source	Pub.	Genotype	Variety	Gene
STOCK	3024 varieties	3k RGP	[5]	C	R	
	1529 varieties	HDRA project	[42]			
STOCKPROP	Passport data	IRGCIS	[14]		C,A	
	Phenotype data	IRGCIS	[14]	A	C,A	
SNP_FEATURE, SNP_FEATURELOC	42,553,103 positions	this project		C		C
	700k positions	HDRA project	[42]			
INDEL_FEATURE, INDEL_FEATURELOC	3,218,491 indel positions	this project		C		C
SNP_FEATUREPROP	SNP effects on RGAP7gene models using SNPEff	this project		A		
SNP_GENOTYPE	HDRA alleles	HDRA project	[42]	R		
INDEL_GENOTYPE	3k indel alleles	this project		R		
FEATURE, FEATURELOC	chromosomes for Nipponbare	IRGSP v1	[15]	C		R
	RGAP7 gene models	MSU RGAP7	[15]			
	RAP gene models	RAP-DB	[28]			
	FGenesh++ gene models	this project	[29]			
	Merged RGAP7, RAP, FGenesh++ gene models	this project				
	chromosomes, contigs and gene models for 93-11		[4]			
	Chromosomes and gene models for Kasalath		[16]			
	Contigs and gene models for DJ 123		[17]			
	Contigs and gene models for IR 64		[17]			
	QTL	http://qtaro.abr.affrc.go.jp (July 2015)	[33]			
CV, CVTERM, CVTERM_RELATION, CVTERM_PATH, CVTERM_SYNONYM	GO: gene ontology	http://obofoundry.org (updated July 2015)	[37]	A		C
	SO: sequence ontology		[40]	I		I
	PO: plant ontology		[38]	A		C
	TO: plant trait ontology		[39]	A	C	C
	RO: relationship ontology			I	I	I
	CO: crops ontology (rice)	http://cropontology.org	[41]	A	C	C
	Q-TARO traits	http://qtaro.abr.affrc.go.jp (July 2015)	[33]	A	C	C
	SNP-Seek controlled terms			I	I	I
FEATURE_CVTERM_RELATION	Trait genes	http://shigen.nig.ac.jp/rice/oryzabase, http://qtaro.abr.affrc.go.jp (July 2015)	[31,32]	A		C
	Plant anatomy, development genes	http://shigen.nig.ac.jp/rice/oryzabase, (July 2015)	[31]	A		C
	Gene ontology genes	RGAP7, Oryzabase		A		C
FEATURE_RELATION	Ricenet interations	RiceNetv2	[35]	A		C
	PRIN interactions	PRIN	[36]	A		C
	Merged RGAP7, RAP, FGenesh++ gene models mapping	this project		A		C
FEATURE_SYNONYM	gene symbols, accessions	http://shigen.nig.ac.jp/rice/oryzabase, Uniprot	[31]			C

marker annotations, and other sequence features. The Workspace module handles user lists and bulk downloads processing. The Genotype, Variety, and Genomics modules use ontology services for ontology-related queries.

3. Results and discussion

3.1. SNPs and indels

The number of SNPs and indels generated from the 3023 accessions on all five reference genomes are summarized in Table 3. Bulk of the primary SNP discovery was made on the gold standard reference genome for rice, Nipponbare, since there is high degree of similarity of the other four genomes relative to Nipponbare (77–87%, Supplementary Table 1). Reference genome-specific SNPs are discovered mainly in unique genome region (13–20%, Supplementary Table 1), hence a much lower number was observed (Table 3). The quality of genome assembly also affects the discovery rate of SNPs. Kasalath and 93-11 utilized longer read technology [16,4], allowing sequencing across repetitive genome regions. On the other hand, IR 64 and DJ 123 are from short-reads technology [17] which is problematic for highly repetitive genomes such as rice.

The distribution of SNP counts across the 3k RG accessions partitioned into the six major subpopulations is shown in Fig. 4. As expected, SNP density is lowest in accessions belonging to the same group (temperate japonica, subtropical) or closely related to Nipponbare (which is temperate japonica), and it increases as the accessions become more genetically distant from the reference genome (indica, aus).

Majority of the insertions discovered are from 1 to 5 bases long, followed by those of 6–17 bases. Insertions of 18–25 bases are rare (Fig. 4). For deletions, majority was also of size 1–5 bases, followed by sizes of 6–28 bases, and the rarest ones were those between sizes 29–36 bases (Fig. 4). We did not see any significant difference in the discovery rate of indels in these size ranges between the six *Oryza* subgroups (data not shown). The overall discovery rate may be biological in nature, but detection of longer indels may be incomplete, likely caused by short sequencing read length (83 bases), depth of coverage, and the GATK-UG detection method, which is best for detecting small indels only (<=50 bases).

3.2. The SNP-Seek web interface

The rice variant and related data can be accessed by the rice research community through the Rice SNP-Seek website (http://snp-seek.irri.org). We describe here the major features of the site.

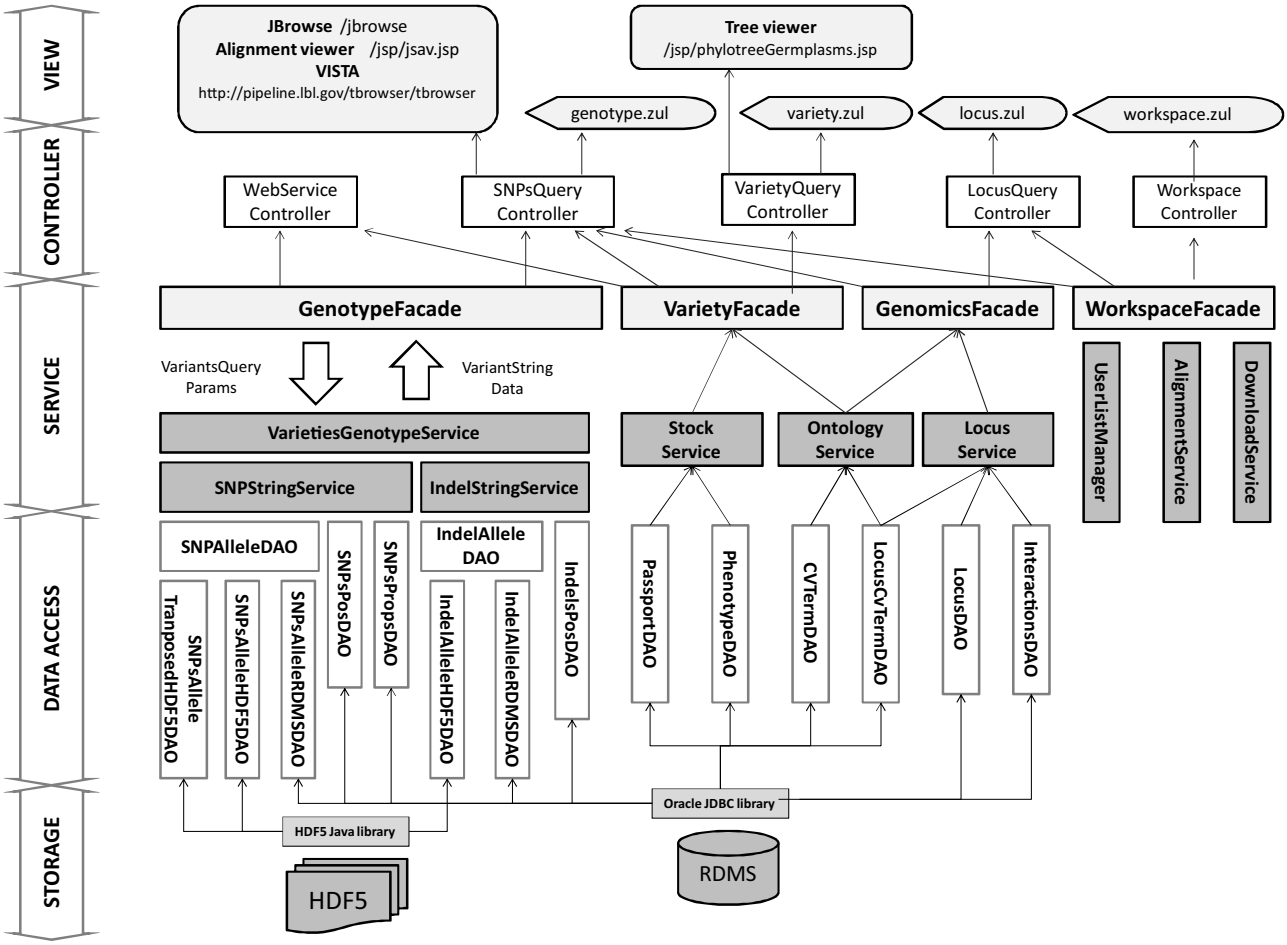

Fig. 3. Web-application software architecture.
The web-application follows the Model-View-Controller design. The VIEW layer are jsp or ZK termplates with UI components data and events handled by the CONTROLLER Java classes. The MODEL is based on Data Service and Data Access Objects layers from core J2EE patterns [44]. The storage uses HDF5 for the large 3kRGP genotype matrix, and RDBMS for the rest.

Table 3
Count of small indels and SNPs discovered from Nipponbare and unique regions of four additional genomes.

Genome assembly	Variety group	Count of indels	Count of SNPs
Nipponbare (gold standard)	Temperate japonica	Biallelic: 2,354,934 Multiallelic: 344,545 Total: 2,699,479	Biallelic: 29,635,225 Multiallelic: 2,428,992 Total: 32,064,217
93-11	Indica	308,834	5,540,366
IR 64	Indica	24,041	163,879
Kasalath	Aus	173,892	4,744,760
DJ 123	Aus	12,245	39,881
Total		3,218,491	42,553,103

3.2.1. Data browsers

Several visualization tools are accessible from the Browse menu to display genotypic data. The JBrowse genome browser [45] displays genomic features for each of the five reference genomes used. The gene models and annotation data described in Section 2.4 [*Genomic data integration*] and the BAM and VCF files from the variant calling pipeline described in Section 2.3.1 [*Alignment and variant calling*] are added as tracks to the Nipponbare genome browser. We also generated the phylogenetic tree for all 3024 varieties and display it using the jsPhyloSVG [46] library. Another way to visualize the evolutionary relationships between the varieties is through the display of multidimensional scaling (MDS) plots. The results of the pairwise genome alignments between the five reference genomes described in Section 2.3.2 [*Whole-genome DNA alignment*] can be viewed in VISTA browser [22].

3.2.2. Search features

The major queries available in SNP-Seek are genotype, varieties, and gene loci. Genotype queries return the allele matrix for a given region or list of SNP positions. The varieties can be set to all or constraint by subpopulation or a list. It is also possible to rank the varieties by genotype similarity with a list of positions and alleles, or with any of the varieties in a particular region, or SNP positions. More options are available as shown by the Genotype Query interface in Fig. 5.

The Variety Query interface in Fig. 6 shows the constraints to query varieties including name, genebank accession, and country of origin or subpopulation. It is also possible to query a set of varieties to satisfy phenotypic or passport value. When several constraints are set, the results satisfy all these values. The Gene Locus Query interface returns gene loci from the selected gene model name,

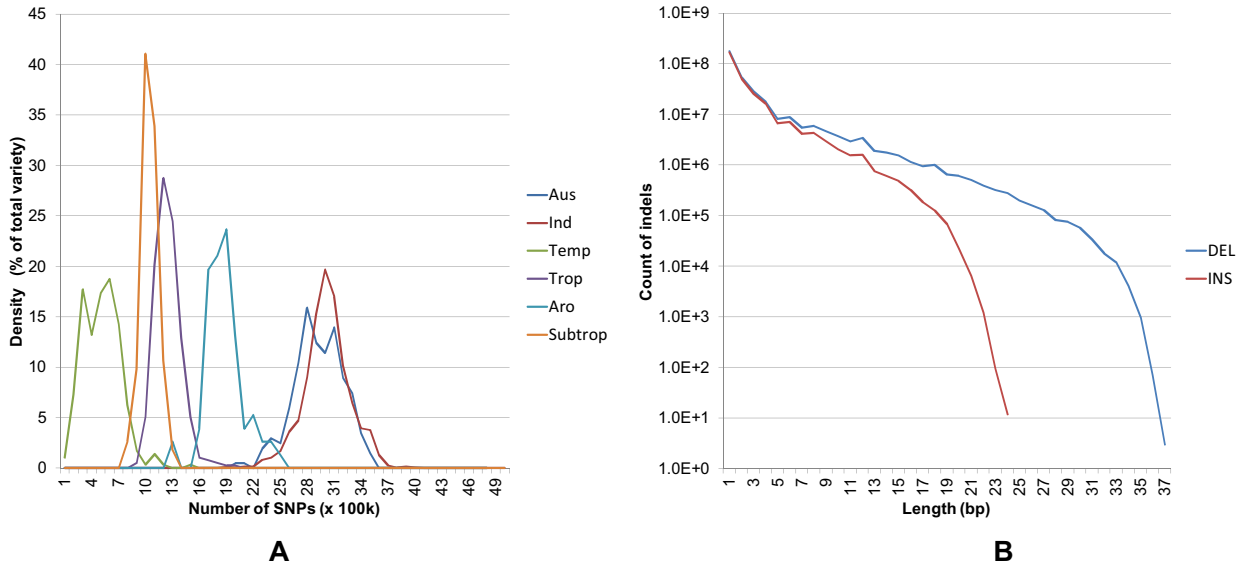

Fig. 4. Distribution of SNP and indels in the 3k RG accessions.
Panel A shows the distribution of SNPs in the 3k RG accessions partitioned to the six major variety groups. Panel B plots the count of indels against indel length (in basepairs) for all 3K RG accessions

function, ontology term, chromosome region, or sequence as constraint. When the constraint is a list of SNP positions, this feature annotates the positions with the available data described in Section 2.4.2 [*Marker and gene annotation data*].

The utility of data in any of the queries described is summarized in Table 2. The use of the data is marked as follows: Result (denoted as R) if it is the returned entity; Constraint (C) when it is used as query constraint; Annotation/Attribute (A) if it is added as an attribute or annotation to the returned entities; and Internal (I) if it is used internally in the database.

Another useful SNP-Seek feature is the possibility to create a list of varieties, SNP positions, or gene loci from query results. The list can also be defined manually by the user through the My List page. Once defined, the list can be used as constraint in succeeding queries. The list can be downloaded for reuse in future sessions.

3.2.3. Data download

All raw and analyses files generated for each variety and the genotype matrices are available through the Download page. There is also a feature to generate the alternate sequence from the VCF file for any variety and specific regions using GATK's FastaAlternateReferenceMaker option.

3.3. SNP-Seek use cases

We present some use cases of SNP-Seek for rice research. The instructional details, with screenshots, are in Supplementary Material IV. SNP-Seek has also been highly utilized by scientists at IRRI for rice allele mining exercises [6].

3.3.1. Use case 1: given a region or gene of interest, get diversity of varieties

One of the important traits in domesticated rice is seed shattering. This is encoded by the *sh-4* gene (LOC_Os04g57530). Examining the variation of the region for this gene across the 3k RG may elucidate more information about this trait.

Utilizing the Genotype search function, accessions that showed variation in the region of interest were identified. In the Genotype search page, the subpopulation was chosen and the gene locus was supplied. After supplying this information, the start and end position of the gene will be automatically filled in. The table matrix showing the accession with variation in the region will be returned.

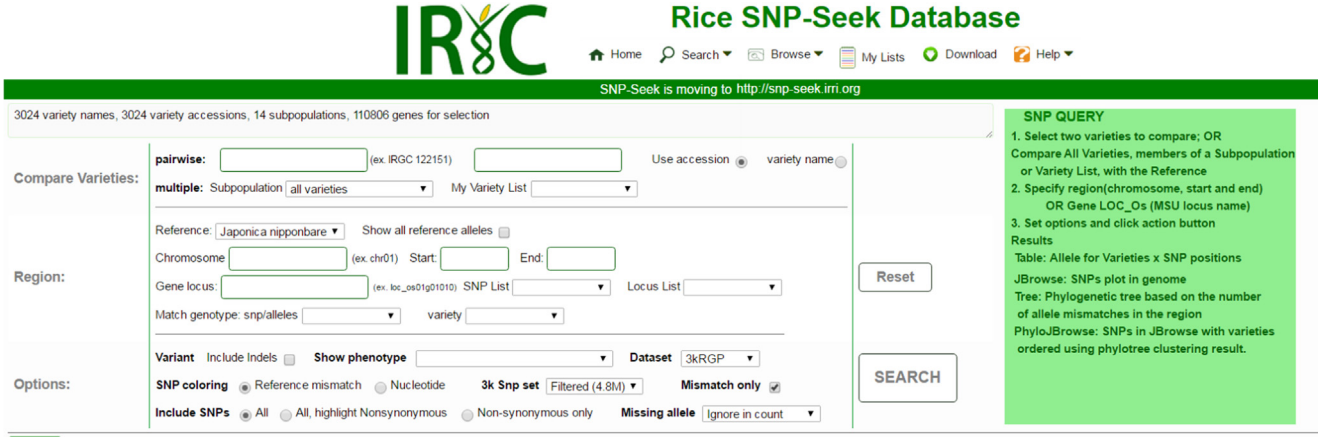

Fig. 5. Genotype query interface.

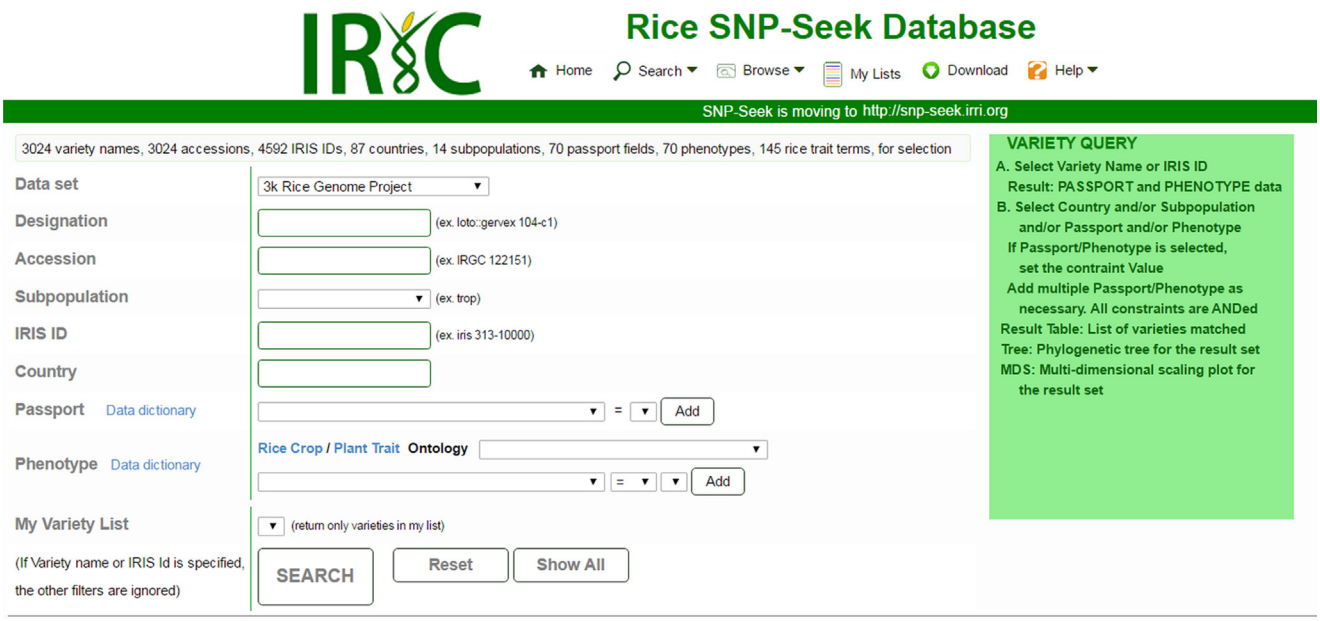

Fig. 6. Variety query interface.

There are 78 varieties that show variation in this region. The table matrix is downloaded in comma-separated value format to further examine the region.

3.3.2. *Use case 2: find varieties from the 3k RG panel that is similar to a particular variety*

Grain yield is an important trait in rice breeding. One of the components in determining grain yield is grain weight. In this case study, we will find similar accessions to an accession with the heaviest 100-grain weight.

The Genotype search function was used to search across all varieties in the region of interest. Additional information was added to specifically identify varieties with specific phenotypic data. The table matrix for 3024 accessions will be returned. The accession of interest is identified by sorting the table matrix according to the phenotypic data. This accession is used as a reference for the rest of the 3k RG panel. A new table matrix is returned and can be downloaded for further examination.

3.3.3. *Use case 3: SNP discovery from a region not found in the nipponbare reference genome*

There are genes for important traits in rice that can only be found in genome regions of the donor variety but not in the Nipponbare genome. An example is *Pstol1*, a gene encoding a protein kinase responsible for the phosphorus-uptake efficiency phenotype [47]. It is located within a major rice QTL, Pup1, which is associated with tolerance to P deficiency in soils [48,49]. Pup1 QTL was first identified to be 90 kb region in chromosome 12 in the P-deficiency tolerant variety Kasalath, and it is absent (deleted) in P-deficiency intolerant variety Nipponbare. Variant discovery in this region should be done in the Kasalath reference. Using the Genotype search function, over a thousand SNP positions were identified in the 3k RG accessions. This was done by selecting Kasalath as reference genome and inputting the published region of *Pstol1*. Using Nipponbare as reference resulted in no SNPs being discovered. Marker assays could now be designed around these SNP positions for molecular breeding purposes.

3.4. *Web service APIs*

It is possible to query data from SNP-Seek programmatically using the RESTful web service APIs we defined. One example is our in-house Galaxy instance (http://galaxy.irri.org) which queries genotypic data. We also implemented the Breeding API (http://docs.brapi.apiary.io) to adhere to the standards in phenotype/genotype databases. This allows other data providers and consumers to interact with SNP-Seek programmatically in the future. An intuitive and functional documentation page is available in SNP-Seek's Help page.

4. Conclusion

We described the development of the Rice SNP-Seek database and data portal for the 3k RG and HDRA Projects. The massive genotypic data required the adoption of modern storage and retrieval technologies. With the integration of diverse biological data, from legacy genebank information system to results from recent high-throughput genomic and computational analyses projects, the need for flexible database and middleware software design is ineviTable SNP-Seek is designed to be adaptive and responsive to the data from the two aforementioned projects as well as that from future sequencing and high-density genotyping projects. Trait data from germplasm panels with curated genotype data can also be accommodated, and legacy passport and phenotypic data for germplasm from the IRGCIS can also be accessed by SNP-Seek. We also demonstrated the utility of ontologies to simplify biological database design and for users to perform effective queries.

Based on user-community feedback and access statistics, SNP-Seek is being heavily utilized as a tool for allele discovery across the diverse 3k RG panel. It is also used for the discovery of SNPs and indels for QTL mapping experiments, and finding candidate molecular markers associated to important agronomic traits of interest for molecular breeding use. SNP-Seek promises to be an indispensable resource and tool for rice genomics and allele discovery. In the future, SNP-Seek will (1) integrate more publicly available rice genotypic and phenotypic data and (2) develop and integrate analysis and visualization tools for genome-wide association study and genomic selection studies.

Acknowledgements

We would like to thank Rolando Santos, Jr. for the administration of the database, Rogelio Alvarez, Denis Diaz and the IRRI ITS team for the operation of the web application servers. Variant calling was supported by the Philippine Genome Center Core Facility for Bioinformatics and DNAnexus, Inc. Computing and data/results hosting were graciously provided by the Extreme Science and Engineering Discovery Environment (XSEDE, which is supported by National Science Foundation grant number ACI-1053575), the Computing and Archiving Research Environment project of the Advanced Science and Technology Institute, Department of Science and Technology of the Philippines, and Amazon Public Data. The work (software development, operation of the database and web application) is being continuously funded by IRRI, the CGIAR Research Program CRP 3.3 (Global Rice Science Partnership) and the International Rice Informatics Consortium (IRIC). Special thanks are in order to CIRAD, Bayer Crop Sciences and Syngenta for providing IRIC funding, and to the Taiwan government for the grant to IRRI.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cpb.2016.12.003.

References

[1] P.A. Seck, A. Diagne, S. Mohanty, M.C.S. Wopereis, Crops that feed the world 7: Rice, Food Secur. 4 (2012) 7–24, http://dx.doi.org/10.1007/s12571-012-0168-1.

[2] S. Fahad, L. Nie, F.A. Khan, Y. Chen, S. Hussain, C. Wu, D. Xiong, W. Jing, S. Saud, F.A. Khan, Y. Li, W. Wu, F. Khan, S. Hassan, A. Manan, A. Jan, J. Huang, Disease resistance in rice and the role of molecular breeding in protecting rice crops against diseases, Biotechnol. Lett. 36 (2014) 1407–1420, http://dx.doi.org/10.1007/s10529-014-1510-9.

[3] H. Hu, L. Xiong, Genetic engineering and breeding of drought-resistant crops, Annu. Rev. Plant Biol. 65 (2014) 715–741, http://dx.doi.org/10.1146/annurev-arplant-050213-040000.

[4] Z.Y. Gao, S.C. Zhao, W.M. He, L.B. Guo, Y.L. Peng, J.J. Wang, X.S. Guo, X.M. Zhang, Y.C. Rao, C. Zhang, G.J. Dong, F.Y. Zheng, C.X. Lu, J. Hu, Q. Zhou, H.J. Liu, H.Y. Wu, J. Xu, P.X. Ni, D.L. Zeng, D.H. Liu, P. Tian, L.H. Gong, C. Ye, G.H. Zhang, J. Wang, F.K. Tian, D.W. Xue, Y. Liao, L. Zhu, M.S. Chen, J.Y. Li, S.H. Cheng, G.Y. Zhang, J. Wang, Q. Qian, Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 14492–14497, http://dx.doi.org/10.1073/pnas.1306579110.

[5] 3k RGP, The 3,000 rice genomes project, (2014) 1–6. http://dx.doi.org/10.1186/2047-217X-3-7.

[6] H. Leung, C. Raghavan, B. Zhou, R. Oliva, I.R. Choi, V. Lacorte, M.L. Jubay, C.V. Cruz, G. Gregorio, R.K. Singh, V.J. Ulat, F.N. Borja, R. Mauleon, N.N. Alexandrov, K.L. McNally, R. Sackville Hamilton, Allele mining and enhanced genetic recombination for rice breeding, Rice 8 (2015) 34, http://dx.doi.org/10.1186/s12284-015-0069-y.

[7] N. Alexandrov, S. Tai, W. Wang, L. Mansueto, K. Palis, R.R. Fuentes, V.J. Ulat, D. Chebotarov, G. Zhang, Z. Li, R. Mauleon, R.S. Hamilton, K.L. McNally, SNP-Seek database of SNPs derived from 3000 rice genomes, Nucleic Acids Res. 43 (2015) D1023–D1027, http://dx.doi.org/10.1093/nar/gku1039.

[8] S.T. Sherry, M.H. Ward, M. Kholodov, J. Baker, L. Phan, E.M. Smigielski, K. Sirotkin, dbSNP: the NCBI database of genetic variation, Nucleic Acids Res. 29 (2001) 308–311, http://dx.doi.org/10.1093/nar/29.1.308.

[9] M.K. Tello-Ruiz, J. Stein, S. Wei, J. Preece, A. Olson, S. Naithani, V. Amarasinghe, P. Dharmawardhana, Y. Jiao, J. Mulvaney, S. Kumari, K. Chougule, J. Elser, B. Wang, J. Thomason, D.M. Bolser, A. Kerhornou, B. Walts, N.A. Fonseca, L. Huerta, M. Keays, Y.A. Tang, H. Parkinson, A. Fabregat, S. McKay, J. Weiser, P. D'Eustachio, L. Stein, R. Petryszak, P.J. Kersey, P. Jaiswal, D. Ware, Gramene 2016: Comparative plant genomics and pathway resources, Nucleic Acids Res. 44 (2016) D1133–D1140, http://dx.doi.org/10.1093/nar/gkv1179.

[10] H. Zhao, W. Yao, Y. Ouyang, W. Yang, G. Wang, X. Lian, Y. Xing, L. Chen, W. Xie, RiceVarMap: a comprehensive database of rice genomic variations, Nucleic Acids Res. 43 (2015) D1018–D1022, http://dx.doi.org/10.1093/nar/gku894.

[11] The IC4R project consortium, information commons for rice (IC4R), Nucleic Acids Res. 44 (2015) gkv1141, http://dx.doi.org/10.1093/nar/gkv1141.

[12] T. Zheng, H. Yu, H. Zhang, Z. Wu, W. Wang, S. Tai, L. Chi, J. Ruan, C. Wei, J. Shi, Y. Gao, B. Fu, Y. Zhou, X. Zhao, F. Zhang, K.L. McNally, Z. Li, G. Zhang, J. Li, D. Zhang, J. Xu, Z. Li, Rice functional genomics and breeding database (RFGB)-3K-rice SNP and InDel sub-database, Chinese Sci. Bull. 60 (2015) 367, http://dx.doi.org/10.1360/N972014-01231 (Chinese Version).

[13] M. Jette, M. Grondona, SLURM: simple linux utility for resource management, Clust Conf. Expo CWCE (2003) 44–60, http://dx.doi.org/10.1007/10968987.

[14] M.T. Jackson, Conservation of rice genetic resources: the role of the International Rice Genebank at IRRI, Plant Mol. Biol. 35 (1997) 61–67 http://www.ncbi.nlm.nih.gov/pubmed/9291960.

[15] Y. Kawahara, M. delaBastide, J.P. Hamilton, H. Kanamori, W.R. McCombie, S. Ouyang, D.C. Schwartz, T. Tanaka, J. Wu, S. Zhou, K.L. Childs, R.M. Davidson, H. Lin, L. Quesada-Ocampo, B. Vaillancourt, H. Sakai, S.S. Lee, J. Kim, H. Numa, T. Itoh, C.R. Buell, T. Matsumoto, Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data, Rice (N. Y) 6 (4) (2013) 397–405, http://dx.doi.org/10.1186/1939-8433-6-4.

[16] S. Hiroaki, K. Hiroyuki, A.K. Yuko, S.H. Mari, E. Kaworu, O. Youko, K. Kanako, F. Hiroko, K. Satoshi, M. Yoshiyuki, H. Masao, I. Takeshi, M. Takashi, K. Yuichi, W. Kyo, Y. Masahiro, W. Jianzhong, Construction of pseudomolecule sequences of the aus rice cultivar Kasalath for comparative genomics of Asian cultivated rice, DNA Res. 21 (2014) 397–405, http://dx.doi.org/10.1093/dnares/dsu006.

[17] M.C. Schatz, L.G. Maron, J.C. Stein, A. Hernandez Wences, J. Gurtowski, E. Biggers, H. Lee, M. Kramer, E. Antoniou, E. Ghiban, M.H. Wright, J. Chia, D. Ware, S.R. McCouch, W.R. McCombie, Whole genome *de novo* assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica, Genome Biol. 15 (2014) 506, http://dx.doi.org/10.1186/PREACCEPT-2784872521277375.

[18] H. Li, R. Durbin, Fast and accurate short read alignment with Burrows-Wheeler transform, Bioinformatics 25 (2009) 1754–1760, http://dx.doi.org/10.1093/bioinformatics/btp324.

[19] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools, Bioinformatics 25 (2009) 2078–2079, http://dx.doi.org/10.1093/bioinformatics/btp352.

[20] Broad Ins,titute, Picard tools, (2016). http://broadinstitute.github.io/picard/.

[21] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M.A. DePristo, The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, Genome Res. 20 (2010) 1297–1303, http://dx.doi.org/10.1101/gr.107524.110.

[22] K.A. Frazer, L. Pachter, A. Poliakov, E.M. Rubin, I. Dubchak, VISTA: Computational tools for comparative genomics, Nucleic Acids Res. 32 (2004) 273–279, http://dx.doi.org/10.1093/nar/gkh458.

[23] I. Dubchak, A. Poliakov, A. Kislyuk, M. Brudno, Multiple whole-genome alignments without a reference organism, Genome Res. 19 (2009) 682–689, http://dx.doi.org/10.1101/gr.081778.108.

[24] M. Brudno, S. Malde, A. Poliakov, C.B. Do, O. Couronne, I. Dubchak, S. Batzoglou, Glocal alignment: finding rearrangements during alignment, Bioinformatics (2003), http://dx.doi.org/10.1093/bioinformatics/btg1005.

[25] W.J. Kent, BLAT – The BLAST-like alignment tool, *Genome Res.* **12**, (2002) 656–664. http://dx.doi.org/10.1101/gr.229202. Article published online before March 2002.

[26] S. Schwartz, W.J. Kent, A. Smit, Z. Zhang, R. Baertsch, R.C. Hardison, D. Haussler, W. Miller, Human-mouse alignments with BLASTZ, Genome Res. 13 (2003) 103–107, http://dx.doi.org/10.1101/gr.809403.

[27] M. Folk, G. Heber, Q. Koziol, An overview of the HDF5 technology suite and its applications, Proc. EDBT (2011) 36–47, http://dx.doi.org/10.1145/1966895.1966900.

[28] H. Sakai, S.S. Lee, T. Tanaka, H. Numa, J. Kim, Y. Kawahara, H. Wakimoto, C.C. Yang, M. Iwamoto, T. Abe, Y. Yamada, A. Muto, H. Inokuchi, T. Ikemura, T. Matsumoto, T. Sasaki, T. Itoh, Rice annotation project database (RAP-DB): An integrative and interactive database for rice genomics, Plant Cell Physiol. 54 (2013), http://dx.doi.org/10.1093/pcp/pcs183.

[29] V. Solovyev, P. Kosarev, I. Seledsov, D. Vorobyev, Automatic annotation of eukaryotic genes, pseudogenes and promoters, Genome Biol. 7 (Suppl. 1) (2006) S10.1–S10.12, http://dx.doi.org/10.1186/gb-2006-7-s1-s10.

[30] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3, Fly (Austin) 6 (2012) 80–92, http://dx.doi.org/10.4161/fly.19695.

[31] N. Kurata, Y. Yamazaki, Oryzabase. An integrated biological and genome information database for rice, Plant Physiol. 140 (2006) 12–17, http://dx.doi.org/10.1104/pp.105.063008.

[32] E. Yamamoto, J. Yonemaru, T. Yamamoto, M. Yano, OGRO: The Overview of functionally characterized Genes in Rice online database, Rice 5 (2012) 26, http://dx.doi.org/10.1186/1939-8433-5-26.

[33] J. ichi Yonemaru, T. Yamamoto, S. Fukuoka, Y. Uga, K. Hori, M. Yano, Q-TARO: QTL annotation rice online database, Rice 3 (2010) 194–203, http://dx.doi.org/10.1007/s12284-010-9041-z.

[34] I.A. Shahmuradov, A.J. Gammerman, J.M. Hancock, P.M. Bramley, V.V. Solovyev, PlantProm: a database of plant promoter sequences, Nucleic Acids Res. 31 (2003) 114–117, http://dx.doi.org/10.1093/nar/gkg041.

[35] T. Lee, T. Oh, S. Yang, J. Shin, S. Hwang, C.Y. eong Kim, H. Kim, H. Shim, J.E. unShim, P.C. Ronald, I. Lee, RiceNet v2: an improved network prioritization server for rice genes, Nucleic Acids Res 43 (2015) W12–W127, http://dx.doi.org/10.1093/nar/gkv253.

[36] H. Gu, P. Zhu, Y. Jiao, Y. Meng, M. Chen, PRIN: A predicted rice interactome network, BMC Bioinf. 12 (2011), http://dx.doi.org/10.1186/1471-2105-12-161, 13 str.

[37] The Gene Ontology Consortium, Gene Ontology Consortium: going forward, *Nucleic Acids Res.* **43**, 2015, D1049–D1056. http://dx.doi.org/ doi: 10.1093/nar/gku1179.
[38] P. Jaiswal, S. Avraham, K. Ilic, E.A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S.Y. Rhee, M.M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, D. Ware, F. Zapata, Plant Ontology (PO): A controlled vocabulary of plant structures and growth stages, Comp. Funct. Genomics 6 (2005) 388–397, http://dx.doi.org/10.1002/cfg.496.
[39] E. Arnaud, L. Cooper, R. Shrestha, Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes, Proc. Int. Conf. Knowl. Eng. Ontol. Dev. (2012) 220–225, http://dx.doi.org/10.5220/0004138302200225.
[40] C.J. Mungall, C. Batchelor, K. Eilbeck, Evolution of the Sequence Ontology terms and relationships, J. Biomed. Inform. 44 (2011) 87–93, http://dx.doi.org/10.1016/j.jbi.2010.03.002.
[41] R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G.F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, G. McLaren, Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature, AoB Plants (2010), http://dx.doi.org/10.1093/aobpla/plq008.
[42] S.R. McCouch, M.H. Wright, C.-W. Tung, L.G. Maron, K.L. McNally, M. Fitzgerald, N. Singh, G. DeClerck, F. Agosto-Perez, P. Korniliev, A.J. Greenberg, M.B. Elizabeth Naredo, S.Q. Mae Mercado, S.E. Harrington, Y. Shi, D.A. Branchini, P.R. Kuser-Falcão, H. Leung, K. Ebana, M. Yano, G. Eizenga, A. McClung, J. Mezey, Open access resources for genome-wide association mapping in rice, Nat. Commun. 7 (2016) 10532, http://dx.doi.org/10.1038/ncomms10532.
[43] C.J. Mungall, D.B. Emmert, W.M. Gelbart, A. de Grey, S. Letovsky, S.E. Lewis, G.M. Rubin, S.Q. Shu, C. Wiel, P. Zhang, P. Zhou, A Chado case study: an ontology-based modular schema for representing genome-associated biological information, Bioinformatics 23 (2007) 337–346, http://dx.doi.org/10.1093/bioinformatics/btm189.
[44] D. Alur, J. Crupi, D. Malks, Core J2EE patterns, Design (2003) 650, http://dx.doi.org/10.1017/CBO9781107415324.004.
[45] M.E. Skinner, A.V. Uzilov, L.D. Stein, C.J. Mungall, I.H. Holmes, JBrowse: a next-generation genome browser, Genome Res. 19 (2009) 1630–1638, http://dx.doi.org/10.1101/gr.094607.109.
[46] S.A. Smits, C.C. Ouverney, jsPhyloSVG: A javascript library for visualizing interactive and vector-based phylogenetic trees on the web, PLoS One 5 (2010), http://dx.doi.org/10.1371/journal.pone.0012267.
[47] M. Wissuwa, J. Wegner, N. Ae, M. Yano, Substitution mapping of Pup1: A major QTL increasing phosphorus uptake of rice from a phosphorus-deficient soil, Theor. Appl. Genet. 105 (2002) 890–897, http://dx.doi.org/10.1007/s00122-002-1051-9.
[48] M. Wissuwa, M. Yano, N. Ae, Mapping of QTLs for phosphorus-deficiency tolerance in rice (*Oryza sativa* L.), Theor. Appl. Genet. 97 (1998) 777–783, http://dx.doi.org/10.1007/s001220050955.
[49] J.H. Chin, R. Gamuyao, C. Dalid, M. Bustamam, J. Prasetiyono, S. Moeljopawiro, M. Wissuwa, S. Heuer, Developing rice with high yield under phosphorus deficiency: Pup1 sequence to application, Plant Physiol. 156 (2011) 1202–1216, http://dx.doi.org/10.1104/pp.111.175471.