Published online 28 November 2016 *Nucleic Acids Research, 2017, Vol. 45, Database issue* ***D1075–D1081***
doi: 10.1093/nar/gkw1135

Rice SNP-seek database update: new SNPs, indels, and queries

Locedie Mansueto[1], Roven Rommel Fuentes[1], Frances Nikki Borja[1], Jeffery Detras[1], Juan Miguel Abriol-Santos[1], Dmytro Chebotarov[1], Millicent Sanciangco[1], Kevin Palis[1,2], Dario Copetti[3], Alexandre Poliakov[4,5], Inna Dubchak[4,5], Victor Solovyev[6], Rod A. Wing[1,3], Ruaraidh Sackville Hamilton[1], Ramil Mauleon[1], Kenneth L. McNally[1] and Nickolai Alexandrov[1,*]

[1]International Rice Research Institute, College, Los Baños, Laguna 4031, Philippines, [2]Boyce Thompson Institute, Ithaca, NY 14853, USA, [3]Arizona Genomics Institute, School of Plant Sciences, University of Arizona, Tucson, AZ 85750, USA, [4]Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, [5]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA and [6]Softberry, Inc., Mount Kisco, NY 10549, USA

Received September 15, 2016; Revised October 14, 2016; Editorial Decision October 15, 2016; Accepted November 04, 2016

ABSTRACT

We describe updates to the Rice SNP-Seek Database since its first release. We ran a new SNP-calling pipeline followed by filtering that resulted in complete, base, filtered and core SNP datasets. Besides the Nipponbare reference genome, the pipeline was run on genome assemblies of IR 64, 93-11, DJ 123 and Kasalath. New genotype query and display features are added for reference assemblies, SNP datasets and indels. JBrowse now displays BAM, VCF and other annotation tracks, the additional genome assemblies and an embedded VISTA genome comparison viewer. Middleware is redesigned for improved performance by using a hybrid of HDF5 and RDMS for genotype storage. Query modules for genotypes, varieties and genes are improved to handle various constraints. An integrated list manager allows the user to pass query parameters for further analysis. The SNP Annotator adds traits, ontology terms, effects and interactions to markers in a list. Webservice calls were implemented to access most data. These features enable seamless querying of SNP-Seek across various biological entities, a step toward semi-automated gene-trait association discovery. URL: http://snp-seek.irri.org.

INTRODUCTION

Genomic data play increasingly important roles in plant breeding by helping to discover new gene-trait associations and to understand how nucleotide variations are translated into phenotypic diversity of plants. While several other databases curate rice nucleotide variants, e.g. dbSNP at NCBI (1), Gramene (2), RiceVarMap (3), IC4R (4), RM-Breeding (5), the SNP-Seek database features interactive real-time visualization of millions of SNPs in thousands of rice varieties, making SNP-Seek a unique tool for allele mining (6).

Since the first release of Rice SNP-Seek Database (7), we have undertaken considerable development to incorporate new analysis results, datasets, viewers and query interfaces for multiple reference genomes and assemblies. We have also included features requested by users that will be useful to the broader rice research community.

NEW SNP SETS

We envision SNP-Seek to host data from not only our projects, but also as a repository for rice variant data in the public domain. Consequently, we redesigned the software architecture to handle multiple datasets and reference genomes. Further enhancement has occurred allowing the display and analysis of other variant types like indels and genetic markers and tools for analysis.

SNP and variety sets

The middleware of the SNP-Seek application was redesigned to handle multiple datasets and data formats. This enables use of the same user interface to display various SNP datasets and formats. Additional and updated analyses of the 3k varieties from the 3k Rice Genome Project (8) resulted in five SNP sets: All (32M positions), All biallelic (29M), Base (18M), Filtered (4.8M) and Core (404k). The details of these are described at the Download page (http://snp-seek.irri.org/_download.zul). The Filtered SNP

[*]To whom correspondence should be addressed. Tel: +63 2 580 55600; Fax: +63 2 580 5699; Email: n.alexandrov@irri.org

set is the default dataset. We also imported the HDRA (9) data into SNP-Seek. With various SNP datasets available in a single interface, we can now efficiently query combined genotypes from these data. The current options are to query either/or both of the 3k or HDRA varieties; if both are selected, the SNP positions may be the union or the intersection of the two datasets.

We also have access to phenotype data from the International Rice Genebank Collection database (Genetic Resource Information Management System, GRIMS) (10) linked to SNP-Seek. When a new genotyping dataset is added to our database and the varieties can be traced to genetic stocks or source accessions in GRIMS, phenotype data for the genetic stocks or legacy data for the source accessions are immediately available, adding value to the dataset.

SNPs from multiple *O. sativa* assemblies

In addition to Nipponbare (japonica), the SNP-calling pipeline (11) was run on four sequenced rice genomes representing indica and aus, two of the major rice subpopulations: IR 64 (indica) and DJ 123 (aus) (12), Kasalath (aus) (13) and 93-11 (indica) (14). To avoid including redundant SNP positions, a custom pipeline was run on a path alignment of the five genomes computed from the pair-wise alignment between the five reference genomes. New SNPs for regions unique to each of the other genomes were sequentially added, with the union of SNP calls loaded into SNP-Seek. We used VISTA (15) for the genome comparison. Results of the pairwise genome alignments are viewable in the VISTA browser accessible through the SNP-Seek menu.

Our comprehensive all-against-all pair-wise alignments showed that all five genomes are highly similar to one another. However, a significant fraction of each genome sequence was found to be unique, namely 8.59% of Nipponbare, 5.35% of 93-11, 4.11% of IR 64, 19.85% of Kasalath and 3.56% of DJ 123 (Supplementary Table S1). Reference genome-specific regions may be informative for the discovery of novel variants since variants in these unique genomic regions (which represent additional 12 Mb to 79 Mb, Supplementary Table S1) would not be detected in accessions aligned to reference genomes that are too distant. Over ∼11 million additional SNPs and ∼0.5 million indels were discovered from the additional reference genomes (Supplementary Table S2).

In a genotype query result, a position in the genotype table is based on the selected reference genome. The chromosome/contig and locus choices also depend on the selected reference. Selecting the 'Show all reference alleles' will display the alleles for all reference genomes. A gap in the allele means the position is not found, and possibly deleted for that genome (shown in Figure 1). The location of the queried region in the other references is reported in a message at the top of the results page.

Indels

In addition to SNPs, the variant calling pipeline (11) yielded short indel data. Unlike most variant databases where indels are presented as alleles, we optionally display them in the genotype matrix along with the SNPs, in a multiple sequence alignment-like format (shown in Figure 2). To accomplish this, the longest indel at each anchor point is determined for all 3024 varieties, call this length N. Then N columns are inserted to the right of the anchor point column. For insertion of length I less than N, the inserted nucleotides are filled in the first I columns, and gap(s) are padded from columns $I + 1$ to N. If the anchor point is at position P, the insertion region columns position are P.01, P.02, to $P.N$ and the reference alleles are set to gap. For deletion D of length less than N, gap(s) are filled in the first D columns, and the reference is copied for columns $D + 1$ to N. If the anchor point is at position P, the deletion region column positions are $P + 1$, $P + 2$, to $P + N$ and the reference alleles are set from the reference genome.

Genomic features

We also developed features to query genomic data. Although most of the genomic data were incorporated as provided by external data sources, we introduced some conventions to uniformly name and merge various gene models. First, we merged the Nipponbare MSUv7 (16) and RAP (17) gene models, plus in-house FGenesh++ (18) annotation into a single set of gene models and named them using our convention, of the form OsNippo{YY}g{NNNNNN}. Details about the merging and naming procedure are described in Supplementary Information Sections I and II, respectively. The advantage with this approach is that we were able to map and interconvert all Nipponbare genes between gene models based on location and overlaps instead of by using names. We also introduced a naming convention for gene loci for the other reference genomes, described in Supplementary Information Section II and summarized in Table 1.

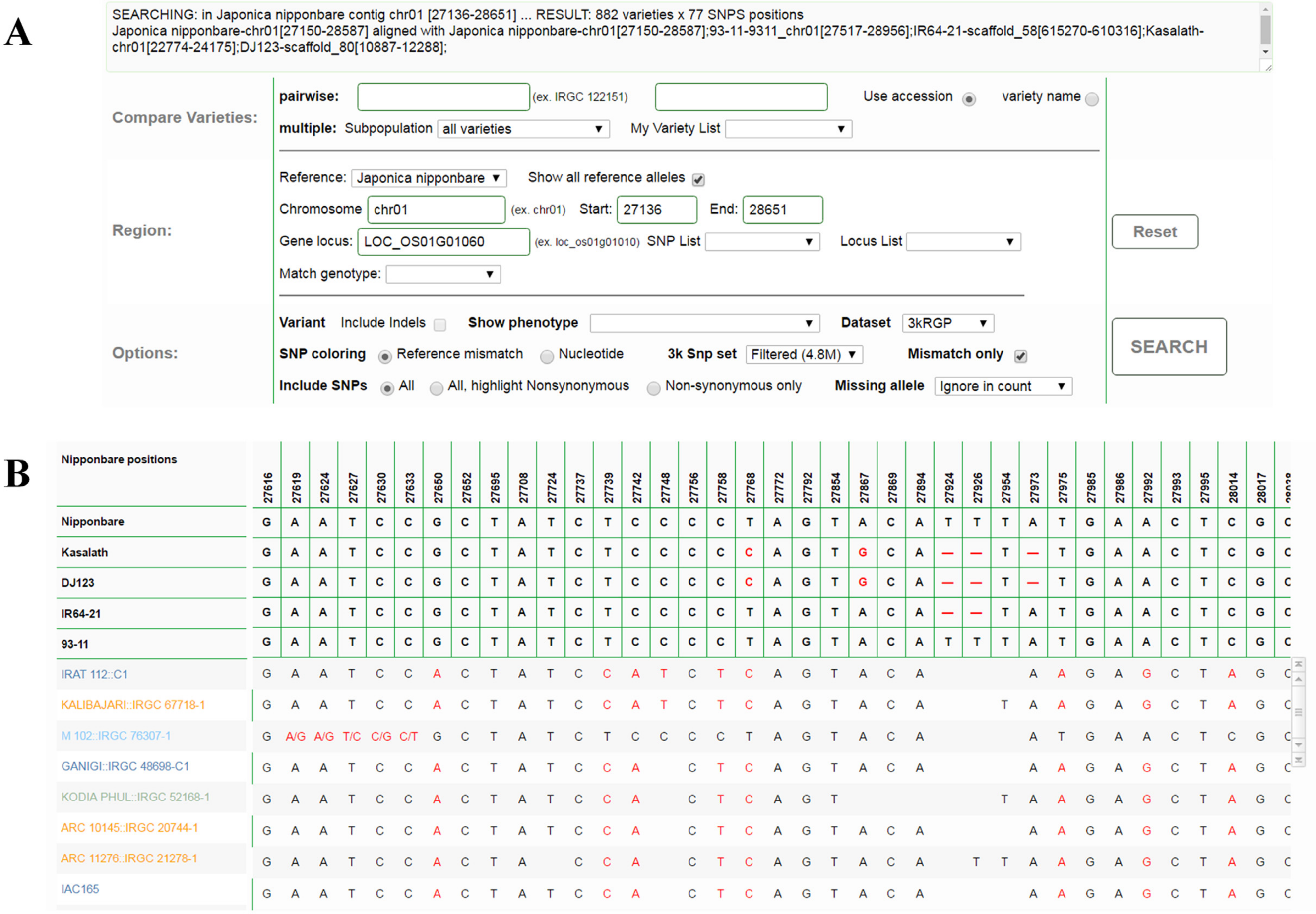
Where data are available, genes for any of the five reference assemblies can be queried using any of these constraints: functional annotation, gene name/symbols, accession number, Gene Ontology terms, traits or Trait Ontology terms, sequence or lists of SNP positions. The data sources are listed in Table 2 and were imported into SNP-Seek using the CHADO schema (19). Storing them with genotype and phenotype data allows complex queries to be performed for various analyses with display in SNP-Seek interface.

NEW QUERY METHODS AND TOOLS

Data alone will not have an impact unless it's made available to those who need it the most and who have the best understanding of its biological significance. Large data sets require above average data handling and programming skills that may not be available to the average biologist. With this motivation, we implemented several query, data management and visualization features that use large variant and genomic data sets, but are relatively quick and intuitive for the general researcher.

Ontology-driven queries

Taking advantage of the CHADO data model, we use ontology terms to constrain gene queries, exploiting transitive closure. That is, selecting a term in the ontology will



SEARCHING: in Japonica nipponbare contig chr01 [27136-28651] ... RESULT: 882 varieties x 77 SNPS positions
Japonica nipponbare-chr01[27150-28587] aligned with Japonica nipponbare-chr01[27150-28587];93-11-9311_chr01[27517-28956];IR64-21-scaffold_58[615270-610316];Kasalath-chr01[22774-24175];DJ123-scaffold_80[10887-12288];

Compare Varieties:
pairwise: (ex. IRGC 122151) Use accession ● variety name ○
multiple: Subpopulation all varieties ▼ My Variety List ▼

Region:
Reference: Japonica nipponbare ▼ Show all reference alleles ☑
Chromosome chr01 (ex. chr01) Start: 27136 End: 28651
Gene locus: LOC_OS01G01060 (ex. loc_os01g01010) SNP List ▼ Locus List ▼
Match genotype: ▼

Options:
Variant Include Indels □ **Show phenotype** ▼ **Dataset** 3kRGP ▼
SNP coloring ● Reference mismatch ○ Nucleotide **3k Snp set** Filtered (4.8M) ▼ **Mismatch only** ☑
Include SNPs ● All ○ All, highlight Nonsynonymous ○ Non-synonymous only **Missing allele** Ignore in count ▼

Nipponbare positions	27616	27619	27624	27627	27630	27633	27650	27652	27695	27708	27724	27737	27739	27742	27748	27756	27758	27768	27772	27792	27854	27867	27869	27894	27924	27926	27954	27973	27975	27985	27986	27992	27993	27995	28014	28017	
Nipponbare	G	A	A	T	C	C	G	C	T	A	T	C	T	C	C	C	C	T	A	G	T	A	C	A	T	T	T	A	T	G	A	A	C	T	C	G	C
Kasalath	G	A	A	T	C	C	G	C	T	A	T	C	T	C	C	C	C	C	A	G	T	G	C	A	—	—	T	—	T	G	A	A	C	T	C	G	C
DJ123	G	A	A	T	C	C	G	C	T	A	T	C	T	C	C	C	C	C	A	G	T	G	C	A	—	—	T	—	T	G	A	A	C	T	C	G	C
IR64-21	G	A	A	T	C	C	G	C	T	A	T	C	T	C	C	C	C	T	A	G	T	A	C	A	—	—	T	A	T	G	A	A	C	T	C	G	C
93-11	G	A	A	T	C	C	G	C	T	A	T	C	T	C	C	C	C	T	A	G	T	A	C	A	T	T	T	A	T	G	A	A	C	T	C	G	C
IRAT 112::C1	G	A	A	T	C	C	A	C	T	A	T	C	C	A	T	C	T	C	A	G	T	A	C	A				A	A	G	A	G	C	T	A	G	C
KALIBAJARI::IRGC 67718-1	G	A	A	T	C	C	A	C	T	A	T	C	C	A	T	C	T	C	A	G	T	A	C	A			T	A	A	G	A	G	C	T	A	G	C
M 102::IRGC 76307-1	G	A/G	A/G	T/C	C/G	C/T	G	C	T	A	T	C	T	C	C	C	C	T	A	G	T	A	C	A				A	T	G	A	A	C	T	C	G	C
GANIGI::IRGC 48698-C1	G	A	A	T	C	C	A	C	T	A	T	C	C	A		C	T	C	A	G	T	A	C	A				A	A	G	A	G	C	T	A	G	C
KODIA PHUL::IRGC 52168-1	G	A	A	T	C	C	A	C	T	A	T	C	C	A		C	T	C	A	G	T						T	A	A	G	A	G	C	T	A	G	C
ARC 10145::IRGC 20744-1	G	A	A	T	C	C	A	C	T	A	T	C	C	A		C	T	C	A	G	T	A	C	A				A	A	G	A	G	C	T	A	G	C
ARC 11276::IRGC 21278-1	G	A	A	T	C	C	A	C	T	A		C	C	A		C	T	C	A	G	T	A	C	A		T	T	A	A	G	A	G	C	T	A	G	C
IAC165	G	A	A	T	C	C	A	C	T	A	T	C	C	A		C	T	C	A	G	T	A	C	A				A	A	G	A	G	C	T	A	G	C

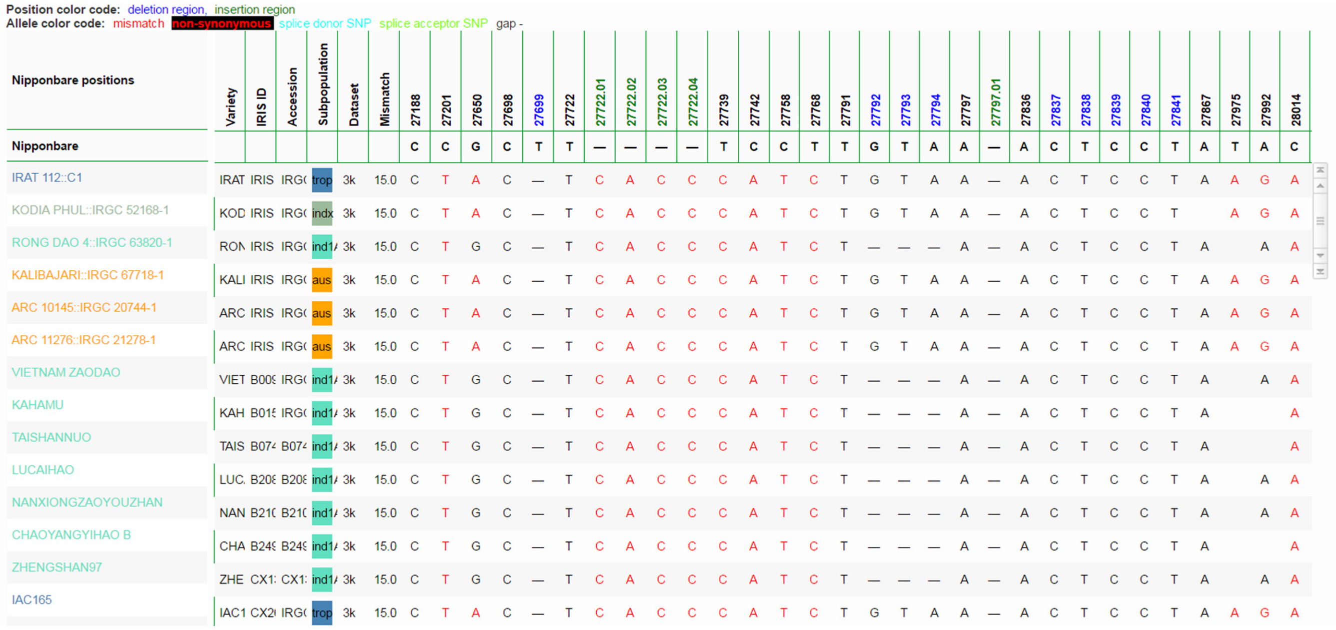
Figure 1. Genotype query options (**A**) and results table (**B**) with multiple reference genomes alleles. The selected reference genome (Nipponbare) is displayed at the top header row just below the SNP positions. The alleles for the other genomes (Kasalath, DJ 123, IR 64, 93-11) are shown below in the table header. The corresponding positions in the other genomes are displayed in the message box.

Table 1. Gene loci names we used for rice reference genomes we are using

Reference genome	Reference	Gene loci names
93-11	(14)	Os9311_{YY}g{NNNNNN}, Os9311_{XXXXX}g{NNNNNN}
IR 64	(12)	OsIR64_{XXXXX}g{NNNNNN}
DJ 123	(12)	OsDJ123{XXXXX}g{NNNNNN}
Kasalath	(13)	OsKasal{YY}g{NNNNNN}
Nipponbare	(16,17)	OsNippo{YY}g{NNNNNN}

Table 2. Data sources for genomics data

Data	Source	URL	Reference
Gene model	MSU v7	http://rice.plantbiology.msu.edu	(16)
Gene model	RAP	http://rapdb.dna.affrc.go.jp	(17)
Gene names/symbols	Oryzabase	http://shigen.nig.ac.jp/rice/oryzabase	(20)
Gene ontology	MSU v7	http://rice.plantbiology.msu.edu	(16)
Trait genes	OGRO	http://qtaro.abr.affrc.go.jp/ogro/table	(21)
Trait ontology-genes	Oryzabase	http://shigen.nig.ac.jp/rice/oryzabase	(20)
Plant ontology-genes	Oryzabase	http://shigen.nig.ac.jp/rice/oryzabase	(20)
QTL	Q-TARO	http://qtaro.abr.affrc.go.jp	(26)
Sequence	MSU v7	http://rice.plantbiology.msu.edu	(16)



Position color code: deletion region, insertion region
Allele color code: mismatch non-synonymous splice donor SNP splice acceptor SNP gap -

Nipponbare positions	Variety	IRIS ID	Accession	Subpopulation	Dataset	Mismatch	27188	27201	27650	27698	27699	27722	27722.01	27722.02	27722.03	27722.04	27739	27742	27758	27768	27791	27792	27793	27794	27797	27797.01	27836	27837	27838	27839	27840	27841	27867	27975	27992	28014
Nipponbare							C	C	G	C	T	T	—	—	—	—	T	C	C	T	T	G	T	A	A	—	A	C	T	C	C	T	A	T	A	C
IRAT 112::C1	IRAT	IRIS	IRG(	trop	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T	A	A	G	A
KODIA PHUL::IRGC 52168-1	KOD	IRIS	IRG(	indx	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T		A	G	A
RONG DAO 4::IRGC 63820-1	RON	IRIS	IRG(	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A		A	A
KALIBAJARI::IRGC 67718-1	KALI	IRIS	IRG(	aus	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T	A	A	G	A
ARC 10145::IRGC 20744-1	ARC	IRIS	IRG(	aus	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T	A	A	G	A
ARC 11276::IRGC 21278-1	ARC	IRIS	IRG(	aus	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T	A	A	G	A
VIETNAM ZAODAO	VIET	B00!	IRG(	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A		A	A
KAHAMU	KAH	B01!	IRG(	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A			A
TAISHANNUO	TAIS	B07‹	B07‹	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A			A
LUCAIHAO	LUC.	B20!	B20!	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A		A	A
NANXIONGZAOYOUZHAN	NAN	B21(	B21(	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A		A	A
CHAOYANGYIHAO B	CHA	B24!	B24!	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A			A
ZHENGSHAN97	ZHE	CX1:	CX1:	ind1/	3k	15.0	C	T	G	C	—	T	C	A	C	C	C	A	T	C	T	—	—	—	A	—	A	C	T	C	C	T	A		A	A
IAC165	IAC1	CX2(	IRG(	trop	3k	15.0	C	T	A	C	—	T	C	A	C	C	C	A	T	C	T	G	T	A	A	—	A	C	T	C	C	T	A	A	G	A

Figure 2. Genotype matrix with short indels. The table displays deletions (positions in blue) at anchor positions (region): 27698 (27699), 27791 (27792–27794), 27836 (27837–27841). Insertion regions (positions in green) are at 27722.01 27722.04 and 27797.01. For deletion regions, the reference is copied from the reference genome, while for insertions the reference genome is set to gaps.

return entities related to the term and its descendants as defined in the ontology. In the gene locus query, Gene Ontology terms can be used to constrain gene locus queries. Trait Ontology terms are also used to query genes where gene–trait association data are available from Oryzabase (20) or OGRO (21). In variety queries, phenotypes are mapped to Trait Ontology (http://browser.planteome.org/amigo/term/TO:0000387) and Crop Ontology (Rice) terms (http://www.cropontology.org/terms/CO_320:ROOT/).

Allele frequency display

The allele frequency chart (shown in Figure 3) displays the frequency or count of major and minor alleles for the queried region or positions, for all varieties or by subpopulation. The chart can be useful for detecting haplotype blocks in the queried region, since adjacent SNPs in a block tend to have the same allele frequencies. If the major frequency of a subpopulation is 1, but less for the other groups, it means the subpopulation has no variant at those positions, which may be an important discriminator for the group. It can also show genotype instead of allele statistics.

New jbrowse tracks and reference genomes

This is a significant update to the JBrowse genome browser (22) since our first release. First, new sets of tracks for Nipponbare are organized by categories. We added tracks for trait associated genes and the BAM and VCF analysis results for each of the 3024 varieties stored in Amazon S3. All tracks for Nipponbare are listed in Table 3.

The second update is a separate JBrowse instance for each of the other four reference assemblies. The tracks in JBrowse are loaded with the same data as Nipponbare using gff format. Each instance has the option to display the sequence and gene models as provided by the source using original locus names and another track using our locus naming convention described in Supplement II. The alignments with the other four genomes can also be viewed as tracks. These instances are accessible through the main menu.

Large snp queries

In the prior version of SNP-Seek, querying SNPs for large regions or using a list of many positions was prohibited by server timeout where the client application needed to wait for a query to finish before the user could proceed with other tasks. To accommodate large queries, we implemented an asynchronous query engine, utilizing Spring (https://spring.io) @Async annotation to manage parallel processes. This allowed us to extend the genotype query limit to a 5Mb region, 500kb SNPs or 1000 gene loci; however, the results are not displayed but are available for download. The user need not wait for the task to finish since a dynamic link is given to monitor the progress and download the results when ready.

Alternate sequence download

A common task is to reconstruct the alternate sequence for a list of regions and varieties, by substituting the SNPs, and integrating the insertions and deletions into the reference genome. This feature uses tabix (23) to query the VCF files from Amazon S3 and process the VCF using the FastaAlternateReferenceMaker tool in GATK (11). Results are downloadable as compressed Fasta files in gzip format (*.fasta.gz).

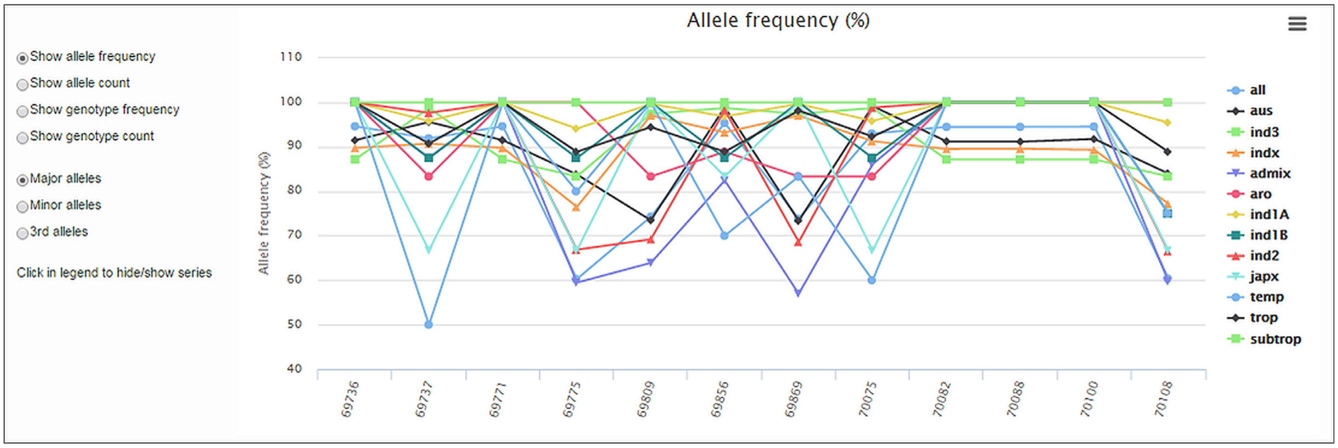


Figure 3. Allele frequency chart with major/minor allele/genotype frequency/count at each SNP position in the queried region for all or each subpopulation.

Table 3. JBrowse tracks for Nipponbare

Category	Track names (count)	Reference
Gene model	MSU7 RAP representative RAP predicted FGenesh++ Merged MSU7, RAP, FGenesh++	(16,17)
Trait Genes	28 OGRO trait track OGRO all traits genes Oryzabase all trait genes	(20,21)
QTL	28 QTARO QTL tracks QTARO all QTL	(26)
BAM	3024 varieties	https://aws.amazon.com/public-data-sets/3000-rice-genome/
BAM Coverage	3024 varieties	https://aws.amazon.com/public-data-sets/3000-rice-genome/
VCF	3024 varieties	https://aws.amazon.com/public-data-sets/3000-rice-genome/
Alignment	Nipponbare versus 9311 Nipponbare versus IR64-21 Nipponbare versus DJ123 Nipponbare versus Kasalath	This project
Variants	SNPs v2 INDELs v2 SNPs v1	This project

List management

We want SNP-Seek to be used by researchers with minimal data interface/conversion issues. The List Manager is designed for this purpose. There are currently three types of lists implemented: variety, SNP and locus lists. The user can create a list, use it to constrain a query, generate a list from the query results and then download or submit it to other analysis tools or queries. The flow of data and queries available in SNP-Seek is illustrated in Figure 4. Arrows show the possible data from an initial set of information. Along with the available set operations, the system can be used for gene-trait association discovery. We extended the SNP List functionality to perform SNP-Marker annotation and genotype matching:

SNP/Marker annotator

The user can create a list of SNP positions and this feature annotates the markers with evidence collected from various other databases and analyses. This list may be significant markers from gene expression or GWAS studies. The annotations can include gene models (RAP (17), MSUv7 (16) or FGenesh++ (18)) or promoter regions (FGenesh++, PlantPromDB (24)) if SNPs are located within these loci. The effects of SNP variants were also added using results from SNPEff (25). For SNPs within gene models, additional evidence about the gene are included using Gene Ontology terms, Plant and Trait Ontology terms and gene names collected from Oryzabase (20), trait genes from OGRO (21), and QTL from Q-TARO (26), interacting genes from RiceNet v2 (27) and rice proteins from PRIN (28). The list of annotations and references are in Supplementary Table S3.

Genotype match

A common use case is to find the most related genotypes among the 3k/HDRA varieties when given a particular genotype. We created a query where the constraint is a list of SNP positions with allele values, and the result is a genotype table where the varieties are sorted based on the number of matching alleles between the query and each variety in the selected dataset. The genotype table can also display values for a selected phenotype allowing quick evaluation of the effect of the queried genotype on the phenotype.

SOFTWARE AND DATABASE UPDATES

To provide the data and query requirements described in the previous sections, several modifications were made in the underlying software architecture. Our major objectives

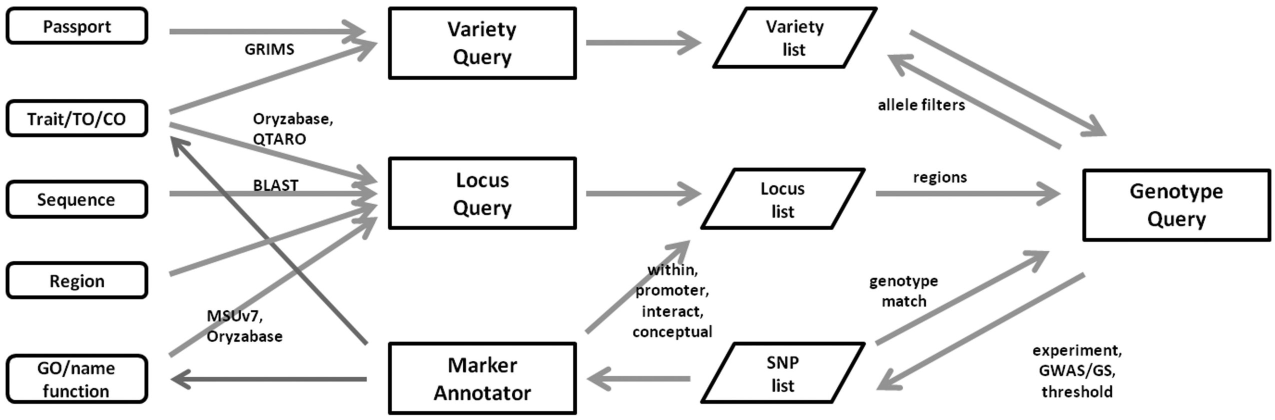


Figure 4. Query capabilities of SNP-Seek. The blocks at the left (rounded) are possible query constraints to the query modules (rectangles). The query results may be stored as a list (parallelograms), and used as constraints in further queries. Lists may also be created by the user as initial constraints. The marker annotator accepts a list of SNP positions, which may be the result from experiments or GWAS studies, to generate constraints for further queries or loop back to the initial constraints, increasing the confidence of the association.

for these updates were to improve query performance, to increase query capacity, to easily integrate new datasets and to serve data to other software systems. Query speed was improved by using hybrid storage wherein genotype matrices are located on the web server as HDF5 (https://www.hdfgroup.org/HDF5) formatted files while other data and the indices for HDF5 files are located in the RDMS. The middleware was re-designed into Service and Data Access layers using Java interfaces, further allowing multiple genotypic data sets to be accommodated. The details of these updates are described in Supplementary Information III.

Web services

We defined a set of RESTful web-service calls for internal use and shared to collaborators through our development site (http://snp-seek.irri.org/dev). The calls are focused on Germplasm, Phenotypes and Genotypes. Another set of calls were implemented for SNP-Seek for compliance to the Breeding API (BrAPI, http://docs.brapi.apiary.io) for Germplasm, Phenotypes, Maps, Studies and Genotypes. The web services documentation can be accessed from the Help menu. Most calls are open but some require login and password. Interested collaborators may contact the authors to use the protected calls.

CONCLUSION AND PERSPECTIVE

The 2016 release of SNP-Seek is designed to be adaptive and responsive to the deluge of genomic data from various sequencing and high-density genotyping projects. It can also accommodate phenotyping (trait) data from germplasm panels with curated genotype data and connect to legacy, phenotypic data for germplasm from the International Rice Genebank Collection (GRIMS) database. SNP-Seek, promises to continue to be an indispensible resource and tool for rice genomics and allele discovery. Our next efforts will focus on analysis and visualizations tools for GWA and genomic selection studies and integrating more of the public genotypic and phenotypic datasets.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank the IRRI ITS team (especially Rogelio Alvarez and Denis Diaz) and Rolando Santos Jr, for support in the operation and administration of the database and web application servers. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1053575. We are also grateful to the Advanced Science and Technology Institute of the Philippines for providing computational resources and Amazon Public Data for hosting the resulting files. The work was funded by IRRI, the CGIAR Research Program CRP 3.3 (Global Rice Science Partnership) and the International Rice Informatics Consortium (IRIC).

FUNDING

Global Rice Science Partnership (GRiSP); International Rice Informatics Consortium (http://iric.irri.org); Cirad, Bayer Crop Sciences and Syngenta (to IRIC); Taiwan Government Grant (to IRRI). Funding for open access charge: IRRI and the CGIAR Research Program CRP 3.3 (Global Rice Science Partnership).
Conflict of interest statement. None declared.

REFERENCES

1. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
2. Tello-Ruiz,M.K., Stein,J., Wei,S., Preece,J., Olson,A., Naithani,S., Amarasinghe,V., Dharmawardhana,P., Jiao,Y., Mulvaney,J. *et al.* (2016) Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Res.*, **44**, D1133–D1140.
3. Zhao,H., Yao,W., Ouyang,Y., Yang,W., Wang,G., Lian,X., Xing,Y., Chen,L. and Xie,W. (2015) RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Res.*, **43**, D1018–D1022.

4. The IC4R Consortium. (2015) Information Commons for Rice (IC4R). *Nucleic Acids Res.*, **44**, D1172–D1180.
5. 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. *Chinese Sci. Bull.*, **60**, 367.
6. Leung,H., Raghavan,C., Zhou,B., Oliva,R., Choi,I.R., Lacorte,V., Jubay,M.L., Cruz,C.V., Gregorio,G., Singh,R.K. *et al.* (2015) Allele mining and enhanced genetic recombination for rice breeding. *Rice*, **8**, 34.
7. 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.
8. 3K R.G.P. (2014) The 3, 000 rice genomes project. *Gigascience*, **3**, 7.
9. McCouch,S.R., Wright,M.H., Tung,C.-W., Maron,L.G., McNally,K.L., Fitzgerald,M., Singh,N., DeClerck,G., Agosto-Perez,F., Korniliev,P. *et al.* (2016) Open access resources for genome-wide association mapping in rice. *Nat. Commun.*, **7**, 10532.
10. Jackson,M.T. (1997) Conservation of rice genetic resources: the role of the international rice genebank at IRRI. *Plant Mol. Biol.*, **35**, 61–67.
11. McKenna,A., Hanna,M., Banks,E., Sivachenko,A., Cibulskis,K., Kernytsky,A., Garimella,K., Altshuler,D., Gabriel,S., Daly,M. *et al.* (2010) The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.*, **20**, 1297–1303.
12. 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.
13. Sakai,H., Kanamori,H., Arai-Kichise,Y., Shibata-Hatta,M., Ebana,K., Oono,Y., Kurita,K., Fujisawa,H., Katagiri,S., Mukai,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.
14. Yu,J., Hu,S., Wang,J., Wong,G.K., Li,S., Liu,B., Deng,Y., Dai,L., Zhou,Y., Zhang,X. *et al.* (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science*, **296**, 79–92.
15. Frazer,K.A., Pachter,L., Poliakov,A., Rubin,E.M. and Dubchak,I. (2004) VISTA: Computational tools for comparative genomics. *Nucleic Acids Res.*, **32**, 273–279.
16. Kawahara,Y., de la Bastide,M., Hamilton,J.P., Kanamori,H., McCombie,W.R., Ouyang,S., Schwartz,D.C., Tanaka,T., Wu,J., Zhou,S. *et al.* (2013) Improvement of the *Oryzasativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice (N. Y).*, **6**, 4.
17. Sakai,H., Lee,S.S., Tanaka,T., Numa,H., Kim,J., Kawahara,Y., Wakimoto,H., Yang,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.
18. Solovyev,V., Kosarev,P., Seledsov,I. and Vorobyev,D. (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*, **7**(Suppl. 1), S10.
19. Mungall,C.J., Emmert,D.B. and The FlyBase Consortium (2007) A Chado case study: An ontology-based modular schema for representing genome-associated biological information. *Bioinformatics*, **23**, 337–346.
20. Kurata,N. and Yamazaki,Y. (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol.*, **140**, 12–17.
21. Yamamoto,E., Yonemaru,J., Yamamoto,T. and Yano,M. (2012) OGRO: the overview of functionally characterized Genes in Rice online database. *Rice*, **5**, 26.
22. 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.
23. Li,H. (2011) Tabix: Fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
24. Shahmuradov,I.A., Gammerman,A.J., Hancock,J.M., Bramley,P.M. and Solovyev,V. V. (2003) PlantProm: a database of plant promoter sequences. *Nucleic Acids Res.*, **31**, 114–117.
25. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) 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**, 80–92.
26. Yonemaru,J., Yamamoto,T., Fukuoka,S., Uga,Y., Hori,K. and Yano,M. (2010) Q-TARO: QTL annotation rice online database. *Rice*, **3**, 194–203.
27. Lee,T., Oh,T., Yang,S., Shin,J., Hwang,S., Kim,C.Y. eong, Kim,H., Shim,H., Shim,J.E. un, Ronald,P.C. *et al.* (2015) RiceNet v2: an improved network prioritization server for rice genes. *Nucleic Acids Res.*, **43**, W122–W127.
28. Gu,H., Zhu,P., Jiao,Y., Meng,Y., Chen,M., Zhang,Y., Gao,P., Yuan,J., Plewczynski,D., Ginalski,K. *et al.* (2011) PRIN: a predicted rice interactome network. *BMC Bioinformatics*, **12**, 161.