DATABASE
The Journal of Biological Databases and Curation

Database, 2016, 1–6
doi: 10.1093/database/baw107
Original article

Original article

Ricebase: a breeding and genetics platform for rice, integrating individual molecular markers, pedigrees and whole-genome-based data

J. D. Edwards[1,*], A. M. Baldo[1] and L. A. Mueller[2]

[1]Dale Bumpers National Rice Research Center, USDA-ARS, Stuttgart, AR, USA, [2]Boyce Thompson Institute, Ithaca, New York, USA

***Corresponding author**: Tel: +870 672 9300; Email: jeremy.edwards@ars.usda.gov

Citation details: Edwards,J.D., Baldo,A.M., and Mueller,L.A. Ricebase: a breeding and genetics platform for rice, integrating individual molecular markers, pedigrees and whole-genome-based data. *Database* (2016) Vol. 2016: article ID baw107; doi:10.1093/database/baw107

Received 10 February 2016; Revised 30 May 2016; Accepted 23 June 2016

Ricebase (http://ricebase.org) is an integrative genomic database for rice (*Oryza sativa*) with an emphasis on combining datasets in a way that maintains the key links between past and current genetic studies. Ricebase includes DNA sequence data, gene annotations, nucleotide variation data and molecular marker fragment size data. Rice research has benefited from early adoption and extensive use of simple sequence repeat (SSR) markers; however, the majority of rice SSR markers were developed prior to the latest rice pseudomolecule assembly. Interpretation of new research using SNPs in the context of literature citing SSRs requires a common coordinate system. A new pipeline, using a stepwise relaxation of stringency, was used to map SSR primers onto the latest rice pseudomolecule assembly. The SSR markers and experimentally assayed amplicon sizes are presented in a relational database with a web-based front end, and are available as a track loaded in a genome browser with links connecting the browser and database. The combined capabilities of Ricebase link genetic markers, genome context, allele states across rice germplasm and potentially user curated phenotypic interpretations as a community resource for genetic discovery and breeding in rice.

Introduction

Rice (*Oryza sativa*) was the first crop species selected for whole genome sequencing because of its relatively small genome size (1) and its global importance in food production (2, 3). Draft genome sequences for rice were first produced in 2002 (4, 5), followed by a map-based, nearly complete genome sequence (6), and culminating in the release of a high-quality genome assembly with gene annotation (7). The high-quality reference genome has enabled high-density genotyping (8, 9) and resequencing of over 3000 additional rice varieties (10, 11).

The availability of genomic and high-density genotyping data will be of tremendous value for rice genetics and breeding. However, due to the size and structure of the datasets, working with these data is challenging for many field and bench scientists. Approaches such as genome-wide

association mapping can reveal significant phenotype associations with particular SNPs. From those significantly associated SNPs, some of the next investigative steps include: browsing the surrounding genomic region to look for likely candidate genes, examining potential functional consequences of an SNP within in a gene, and determining if the SNP is near a molecular marker that has been previously reported to be associated with the same or a related phenotype. Ricebase provides user friendly tools to address all of those needs.

Most of the current rice genetics literature has only older molecular marker technologies as genomic position reference points. These are primarily simple sequence repeat (SSR) markers (12, 13). SSRs are still in use in many laboratories because of their high levels of polymorphism because certain SSRs are known to be linked to traits of interest, such as marker RM190 at the *Waxy* gene (14), which controls starch content, and because they can provide fast results and are more flexible in comparison to large SNP sets that are pre-defined.

Most SSR markers in rice were developed before the complete genome sequence became available, and their positions on the current pseudomolecules are not reported. Determining the positions of the SSR markers can be done by BLAST (15) but primer sequences do not always match uniquely or perfectly. To address this, we have developed a pipeline to map SSR primer sequences to the pseudomolecule sequence that uses a stepwise relaxation in match stringency, and incorporates a series of rules for declaring a match for primer pairs that uses distance and orientation.

A number of other genomic databases exist for rice including Gramene (http://gramene.org) (16), Oryzabase (17) and SNP-Seek (10). These databases focus on comparative genomics, diversity data, gene annotation or mutant collection resources. Ricebase is unique in that it specifically includes SSRs as a track and has a focus on breeding and genetics and closes some critical gaps linking the latest resources with historical genetic knowledge.

Materials and methods

The Ricebase database inherits its model, view and controller architecture from the sol genomics network (SGN) (18). The model consists of DBIx::Class-based Perl modules to manipulate data in a relational database schema. The schema includes the community developed Chado natural diversity module (19) as well as several local schemas. The front end web interface is created using Javascript and Mason components. All code is publicly available on GitHub (http://github.com/solgenomics/sgn) with Ricebase specific modifications available at (http://github.com/solgenomics/ricebase).

Ricebase uses JBrowse (http://jbrowse.org) (20) as its genome browser. JBrowse is a Javascript-based genome browser that is scrollable, zoomable and supports the simultaneous display of multiple data types along the genome as tracks, such as gene annotations, sequence variants and other features. Users may overlay their own data as tracks onto the browser as well. Ricebase has pre-loaded gene annotations (7), a 700 000 SNP dataset across 413 diverse accessions (8), a 20 million SNP dataset across 3000 accessions (10), and over 17 000 SSR markers.

To determine the positions of the SSR markers, a pipeline was developed using BLAST and Perl scripts (Supplementary materials). To allow for imperfect matches, the pipeline includes a stepwise relaxation of stringency. Primer pairs are concatenated for a single BLAST search. For each SSR primer pair, the SSR search begins with a BLAST expect (*E*) value of 0.1, and if no primer is found, the search is repeated at a lower stringency with *E* values of 0.1, 1, 10, 100 and 1000. Hits found at the specified threshold are rejected if multiple loci are found, the primer pairs align in the incorrect orientation, or if the predicted amplicon exceeds a size threshold (500 bp). In addition, the genomic locations of the primer pairs are checked for correct orientation and expected distance apart. Collections of publically available SSR primer sequences were obtained from the Gramene database (http://archive.gramene.org/markers/microsat/). The datasets include the SSRs from McCouch *et al.* (13) and the complete Gramene collection. The reference genomes included the IRGSP 1.0 assembly (7) of Nipponbare, a temperate *japonica Oryza sativa* variety, the *indica Oryza sativa* variety 93-11 (5) and the related wild species *O. rufipogon, O. nivara, O. glaberrima, O. barthii, O. glumaepatula, O. longistaminata, O. meridionalis, O. punctate* and *O. brachyantha* obtained from ftp://ftp.gramene.org/pub/gramene/CURRENT_RELEASE/data/fasta/ (21).

The SSR names, primer sequences and genomic positions were used to generate Generic Feature Format Version 3 (GFF3) files for display as tracks in JBrowse. Each SSR name and position was loaded as a marker in the relational database schema. From a published capillary electrophoresis dataset of 421 diverse rice accessions (22), estimated SSR amplicon sizes for 36 of the markers also were loaded in the relational database. Connecting hyperlinks were added between each marker detail page and its corresponding JBrowse feature.

Pedigrees of rice accessions included in the Rice Diversity Panel 1 (RDP1) SSR dataset (22) were obtained from the Genetic Stocks Oryza (GSOR) collection (http://www.ars.usda.gov/gsor). Relationships between accessions are represented in the database as stock properties with controlled vocabulary terms to indicate the female and male parent.

Table 1. Number of SSR markers from the complete Gramene collection and published McCouch 2002 collection mapped to the *temperate japonica* Nipponbare rice reference sequence, the *indica* rice cultivar 93-11, and 9 *Oryza* wild relatives.

Reference	Gramene		McCouch *et al.* (13)	
	Matched	Unmatched	Matched	Unmatched
Nipponbare	17774	1706	1808	139
93-11	15080	4400	1604	343
O. rufipogon	16288	3192	1683	264
O. nivara	15983	3497	1654	293
O. glaberrima	12878	6602	1305	642
O. barthii	14425	5055	1527	420
O. glumaepatula	14960	4520	1577	370
O. longistaminata	11314	8166	1195	752
O. meridionalis	11273	8207	1203	744
O. punctata	8148	11332	825	1122
O. brachyantha	5024	14456	458	1489

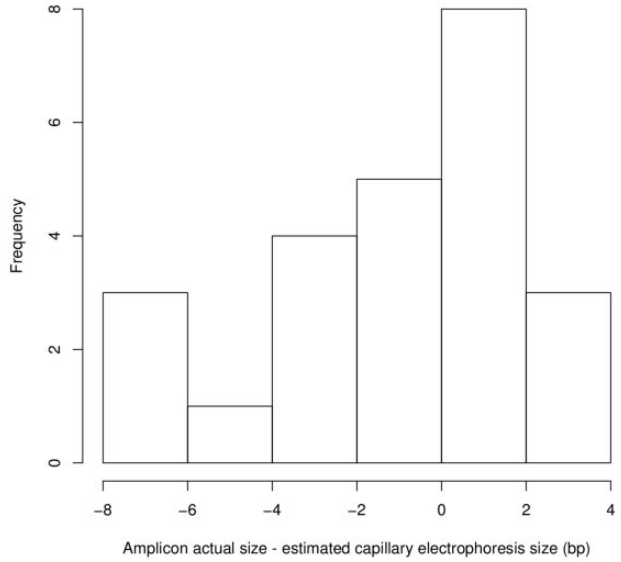

Figure 1. Distribution of differences between the estimated amplicon size using capillary electrophoresis and the sequence-determined amplicon sizes for cv Nipponbare.

Results

SSR primers from a commonly used set of 1947 (13) and the exhaustive Gramene set of 19 480 were mapped to two *O. sativa* assembled reference genomes and nine genomes of wild relatives. On the (current standard) cv. Nipponbare IRGSP 1.0 assembly 92.9 and 91.2% of the SSRs could be unambiguously located for the McCouch *et al.*'s (13) and Gramene sets, respectively (Table 1). The primary reason for inability to determine genome positions for SSRs was predicted amplicons at multiple locations. It is common for multiple amplicons to be observed for SSR markers and bands outside of expected size ranges are often ignored in allele calling. As expected the number of located SSRs decreased within the wild species as the genetic distance from *O. sativa* increased (23).

The predicted amplicon sizes, in base pairs, for each SSR primer pair were calculated based on the Nipponbare reference and compared with published estimate band sizes in Nipponbare with capillary electrophoresis. Out of 39 markers, slight deviations from the expected band sizes were seen for all but three of the SSRs with a maximum of an 8-bp difference (Figure 1). There was no strong correlation between fragment length and deviation of predicted and capillary electrophoresis estimated band sizes ($R^2 = 0.15$) (Supplementary Figure S1).

Genome browser tracks, generated using the SSR marker positions on the rice pseudomolecules, allow users to view the SSR markers in the surrounding context of annotated genes and SNPs (Figure 2). The gene annotation track and the SNP track can be used to find SNPs residing in coding regions, and clicking on the SNP displays the predicted effect (synonymous, non-synonymous, stop-codon, frameshift, etc.). Clicking on an SSR marker in the genome browser brings up information on primer sequences and amplicon size in the reference genome.

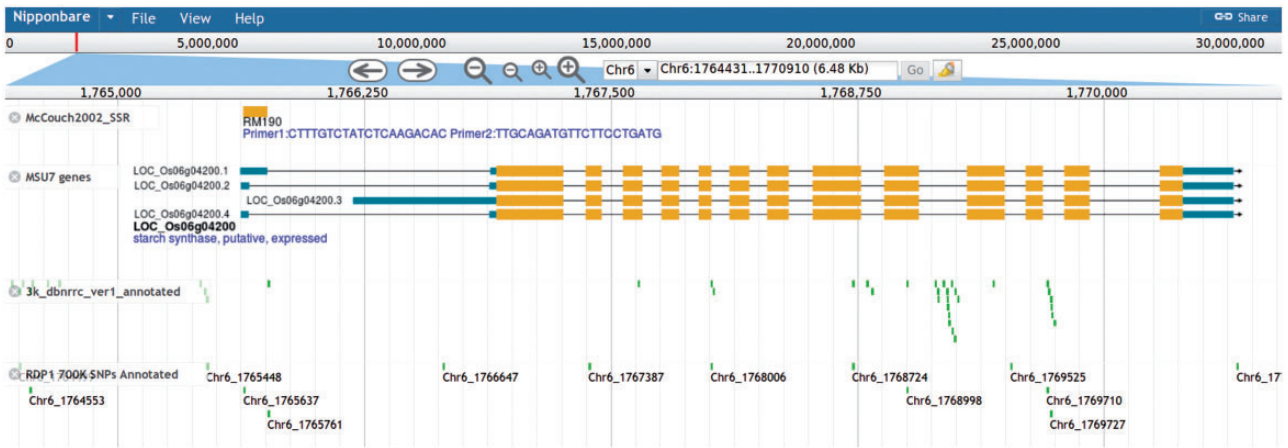

Figure 2. Screenshot showing the genome browser displaying the SSR marker RM190 and surrounding context of genes and SNPs.

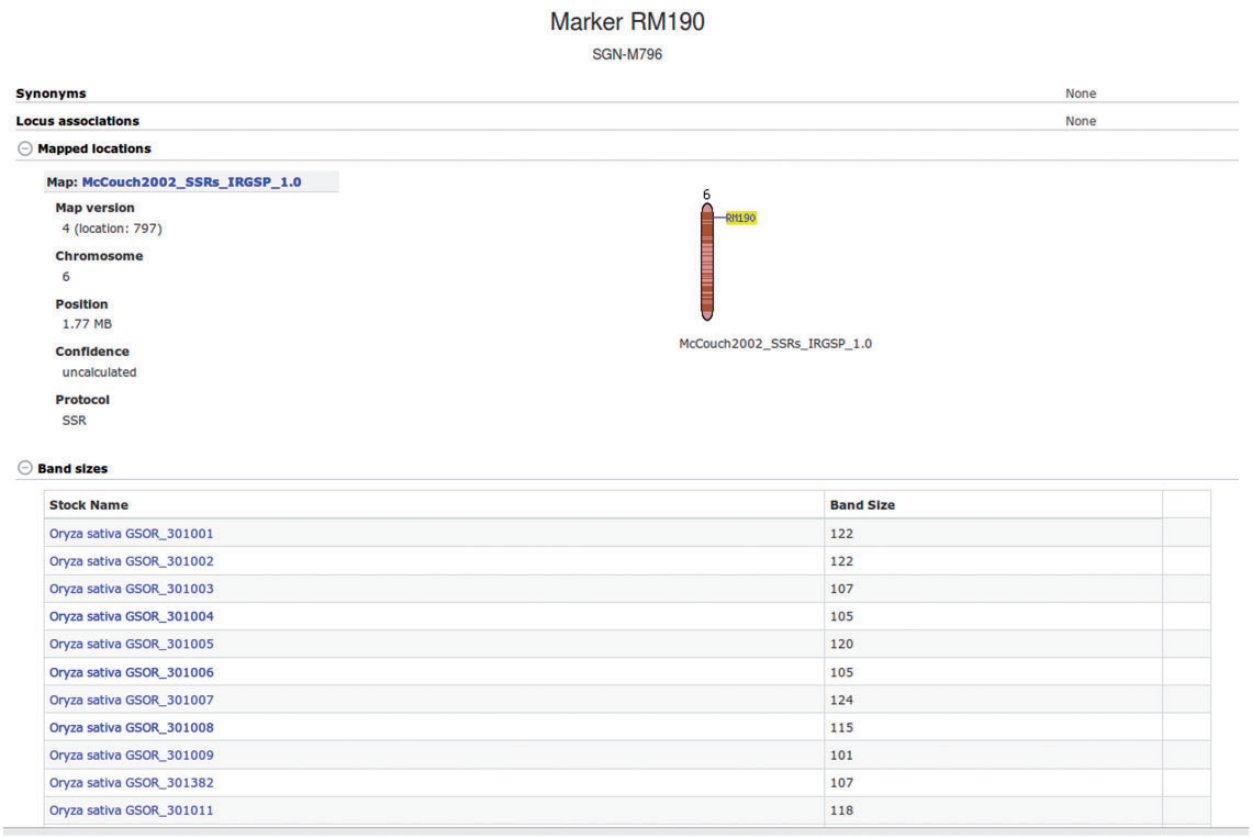

Figure 3. Allele (band size) data by accession for an SSR marker.

Additional information about each SSR marker is stored in the relational database and markers are searchable through the website menu. The marker detail page shows the chromosome and location of the marker on the sequence map (numerically and using a clickable diagram) and measured band (amplicon) sizes in particular accessions when available (Figure 3). Clicking on the map diagram brings up a comparative map viewer highlighting the selected marker and providing an interface to compare the current map with other sequence-based or genetic maps (Figure 4). When band size information is available for a marker, the assayed accessions are clickable to be directed to the corresponding stock (accession) detail page. The stock detail page may contain synonyms for the accession, images, phenotypic data, related accessions, pedigree and descendants. The pedigree and descendants are displayed as an interactive Scalable Vector Graphics where clicking on any displayed accession will take the user to the corresponding accession's stock detail page (Figure 5). The combination of a pedigree search and genetic marker assay data enable users to search for progenitors or descendants of a line and determine if they do or do not share an allele state at a particular genetic marker.

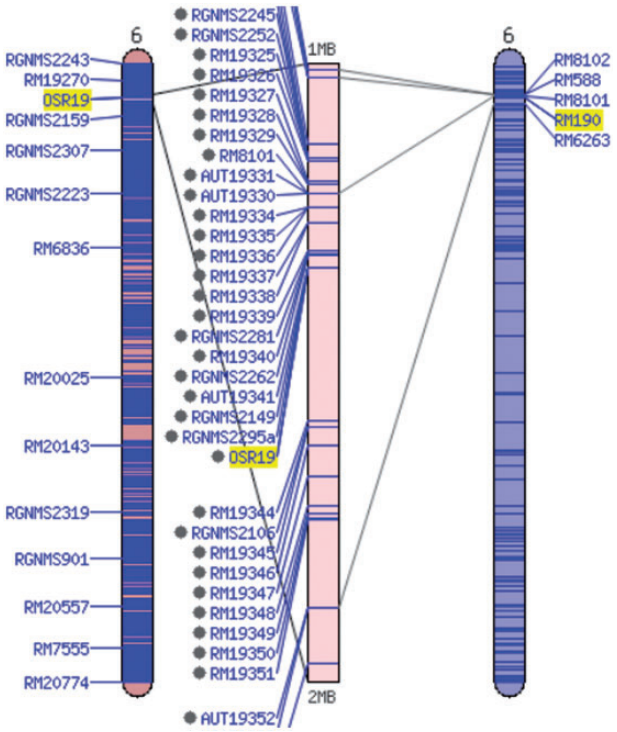

Figure 4. Comparative map viewer aligning the sequence-based maps of the Gramene SSR set (left) and the McCouch 2002 set (right).

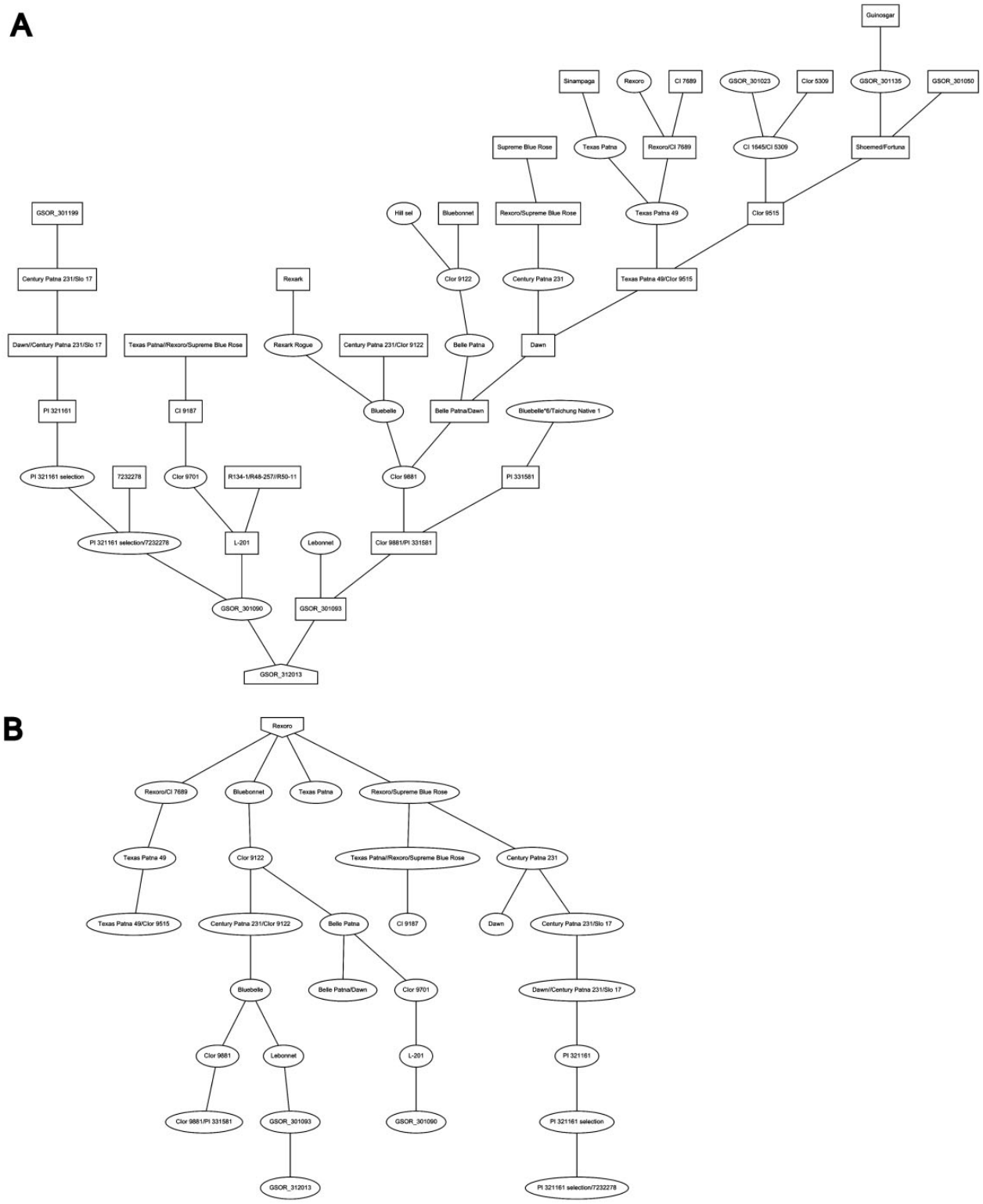

Figure 5. Pedigree and descendent display: **A,** Pedigree of cultivar "Cypress" and **B**, descendants of the cultivar "Rexoro" as displayed on the accession detail view.

Discussion

The development of a fully automated pipeline for determining pseudomolecule assembly positions of SSRs will be useful to overlay marker information on additional *de novo* assemblies of other rice accessions or other species as they become available. This will help maintain a connection between the latest genomic discoveries and the extensive body of (largely SSR-based) rice genetics literature.

With the results of this pipeline, Ricebase distinguishes itself from other genomic databases for rice such as Gramene, Rice Genome Annotation Project and the 3000 Genome Project by providing continuity from past and current marker technologies to whole genome resequencing data.

Fine mapping/positional cloning research often requires new markers to be designed near or between pairs of existing markers. Using the browser, new markers may be developed near an existing marker using SNP information and even gene annotation to potentially capture functional polymorphisms. Having a common coordinate system based on pseudomolecule position combined with the ability to browse the genomic context will ease the transition of SSRs to SNPs. Additional SNP or other marker data may be added as tracks to the database in the future, or users may overlay their own tracks on the browser. This will ease the transition for users when working across different genotyping platforms.

The inclusion of pedigree information, along with molecular marker assay data, presents the possibility of tracing the transmission of particular allele states through a series of crosses. This information may be also used for quality control to detect events where the observed allele state is not possible or highly unlikely given the allele states of the parents or accessions in the pedigree. The integration of pedigrees and breeding records is a unique feature of Ricebase among the existing rice genomics databases. With a relational database of genetic markers in rice, there is now an opportunity to include additional user curated information for each marker. Any type of metadata may be attached to a genetic marker, such as experimental validation, inclusion in a genotyping project, or multiplexing protocols. Users may wish to record phenotypes associated with a particular genetic marker (or allele states of that marker) and supporting literature. Establishment of a user curated resource that contains a collection of markers tightly linked to genes of interest with known phenotypic effects may help researchers pool their collective knowledge to facilitate gene discovery and accelerate rice breeding through marker assisted selection.

Supplementary data

Supplementary data are available at *Database* online.

Conflict of interest. None declared.

References

1. Arumuganathan,K. and Earle,E. (1991) Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep*., 9, 208–218.
2. Khush,G.S. (2005) What it will take to feed 5.0 billion rice consumers in 2030. *Plant Mol. Biol*., 59, 1–6.
3. FAO. (2009) *Increasing Crop Production Sustainably. The Perspective of Biological Processes*. Food and Agriculture Organization of the United Nations, Rome.
4. \Goff,S.A., Ricke,D., Lan,T.H., *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science*, 296, 92–100.
5. Yu,J., Hu,S., Wang,J., *et al*. (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science*, 296, 79–92.
6. International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. *Nature*, 436, 793–800.
7. Kawahara,Y., de la Bastide,M., Hamilton,J.P., *et al*. (2013) Improvement of the *Oryza sativa* nipponbare reference genome using next generation sequence and optical map data. *Rice*, 6, 4.
8. McCouch,S.R., Wright,M.H., Tung,C., *et al*. (2016) Open access resources for genome-wide association mapping in rice. *Nat. Commun*, 7, 10532.
9. Zhao,K., Tung,C., Eizenga,G.C., *et al*. (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun*., 2, 467.
10. Alexandrov,N., Tai,S., Wang,W., *et al*. (2015) SNP-seek database of SNPs derived from 3000 rice genomes. *Nucleic Acids Res*., 43, D1023–D1027.
11. Duitama,J., Silva,A., Sanabria,Y., *et al*. (2015) Whole genome sequencing of elite rice cultivars as a comprehensive information resource for marker assisted selection. *PloS One*, 10, e0124617.
12. Temnykh,S., DeClerck,G., Lukashova,A., *et al*. (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativ*a L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res*., 11, 1441–1452.
13. McCouch,S.R., Teytelman,L., Xu,Y., *et al*. (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res*, 9, 199–207.
14. Bligh,H., Till,R., and Jones,C. (1995) A microsatellite sequence closely linked to the waxy gene of *Oryza sativa*. *Euphytica*, 86, 83–85.
15. Altschul,S.F., Gish,W., Miller,W., *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol*., 215, 403–410.
16. Monaco,M.K., Stein,J., Naithani,S., *et al*. (2014) Gramene 2013: comparative plant genomics resources. *Nucleic Acids Res*., 42, D1193–D1199.
17. Kurata,N. and Yamazaki,Y. (2006) Oryzabase. an integrated biological and genome information database for rice. *Plant Physiol*., 140, 12–17.
18. Fernandez-Pozo,N., Menda,N., Edwards,J.D., *et al*. (2015) The sol genomics network (SGN)—from genotype to phenotype to breeding. *Nucleic Acids Res*., 43, D1036–D1041.
19. Jung,S., Menda,N., Redmond,S., *et al*. (2011) The chado natural diversity module: a new generic database schema for large-scale phenotyping and genotyping data. *Database*, 2011, bar051.
20. Skinner,M.E., Uzilov,A.V., Stein,L.D., *et al*. (2009) JBrowse: a next-generation genome browser. *Genome Res*., 19, 1630–1638.
21. Jacquemin,J., Bhatia,D., Singh,K., *et al*. (2013) The international *Oryza* map alignment project: development of a genus-wide comparative genomics platform to help solve the 9 billion-people question. *Curr. Opin. Plant Biol*., 16, 147–156.
22. Ali,M.L., McClung,A.M., Jia,M.H., *et al*. (2011) A rice diversity panel evaluated for genetic and agro-morphological diversity between subpopulations and its geographic distribution. *Crop Sci*., 51, 2021–2035.
23. Ammiraju,J.S., Lu,F., Sanyal,A., *et al*. (2008) Dynamic evolution of *Oryza* genomes is revealed by comparative genomic analysis of a genus-wide vertical data set. *Plant Cell*, 20, 3191–3209.