RiceVarMap: a comprehensive database of rice genomic variations

Hu Zhao, Wen Yao, Yidan Ouyang, Wanneng Yang, Gongwei Wang, Xingming Lian, Yongzhong Xing, Lingling Chen and Weibo Xie*

National Key Laboratory of Crop Genetic Improvement, National Center of Plant Gene Research (Wuhan), Huazhong Agricultural University, Wuhan 430070, China

Received August 11, 2014; Revised September 15, 2014; Accepted September 16, 2014

ABSTRACT

Rice Variation Map (RiceVarMap, http://ricevarmap.ncpgr.cn) is a database of rice genomic variations. The database provides comprehensive information of 6,551,358 single nucleotide polymorphisms (SNPs) and 1,214,627 insertions/deletions (INDELs) identified from sequencing data of 1,479 rice accessions. The SNP genotypes of all accessions were imputed and evaluated, resulting in an overall missing data rate of 0.42% and an estimated accuracy greater than 99%. The SNP/INDEL genotypes of all accessions are available for online query and download. Users can search SNPs/INDELs by identifiers of the SNPs/INDELs, genomic regions, gene identifiers and keywords of gene annotation. Allele frequencies within various subpopulations and the effects of the variation that may alter the protein sequence of a gene are also listed for each SNP/INDEL. The database also provides geographical details and phenotype images for various rice accessions. In particular, the database provides tools to construct haplotype networks and design PCR-primers by taking into account surrounding known genomic variations. These data and tools are highly useful for exploring genetic variations and evolution studies of rice and other species.

INTRODUCTION

Single nucleotide polymorphisms (SNPs) and insertions/deletions (INDELs) are two major types of genomic variations in living beings and have been widely utilized in basic research and industry. Researchers explored genetic variations as markers to study the genetic basis of complex traits via linkage and association mapping (1–3). High-density markers increase the probability that somarkers are just located in or nearby the target genes, which would provide great advantages for marker-assisted selection in breeding (4). And in some cases, sufficient information of genetic variations may result in identifying the causal genes directly. In addition, when genotype information of individuals in a population is available, we can carry out various population genetic studies, such as estimating linkage disequilibrium blocks, constructing haplotype networks (5), inferring population history (6) and identifying natural or artificial selection signatures (7).

Rice (Oryza sativa L.) is both a major staple crop which feeds nearly half of the population in the world and a model species in the plant research community. Although the draft genome sequences of both indica and japonica subspecies of rice were released in 2002 (8,9) and the genome sequences of japonica subspecies was finished in 2005 (10), the available databases of genomic variations in rice are still very limited for bench researchers. Most of the SNP data in rice deposited in dbSNP (11) and other databases (12–14) were identified based on comparisons of the draft genome sequences between the two rice subspecies. Recently, OryzaSNP consortium interrogated 20 diverse rice varieties using resequencing microarrays and obtained 160,000 non-redundant SNPs (15). The data can be queried at a convenient database OryzaSNP (http://oryzasnplantbiology.msu.edu). Huang et al. sequenced 950 rice accessions and identified 4,109,366 non-singleton SNPs (3). They further carried out genome-wide association study (GWAS) for a lot of important agronomic traits and provided a database (http://202.127.18.221/RiceHap2/) as supporting information of the corresponding publication. The HapRice database published this year provides allele frequencies of about 3300 SNPs in 253 rice accessions (16). These newly emerging databases provide progresses in some ways. However, a comprehensive database, like HapMap (17) in human research community, is highly required in rice. Such a database should include information of abundant high-quality genomic variations and detailed genotypes of a mass of rice accessions, frequency of different subpopulations and comprehensive annotation. Meanwhile, a user-friendly query interface and useful visualization tools should be available.
To achieve these objectives, we collected and analysed sequence data of 1479 rice accessions, identified 6 551 358 SNPs and 1 214 627 INDELs, and constructed a comprehensive database of rice genomic variations, RiceVarMap. Compared with extant relevant databases, RiceVarMap provides not only the largest set of genomic variations and genotype data at present, but also characteristic phenotype images for various rice accessions and a set of intuitional query interface and useful tools, such as constructing haplotype networks and designing polymerase chain reaction (PCR) primers by taking into account surrounding known genomic variations. These data and tools are highly useful for exploring genetic variations and evolution studies of rice and other species. Based on these data, we also investigated natural variations in rice metabolism using GWAS approach (18) and designed two genotyping arrays for rice breeding (19,20). These data will be added to the database in the future.

DATABASE FEATURES

RiceVarMap is free and open to the public with comprehensive functions (Figure 1b). More detailed information is described as follows. In the database, each SNP or INDEL is labelled with a unique identifier (ID, e.g. sf0100000131, vP0136465397). The first letter of the ID indicates the polymorphic type, 's' for SNP and 'v' for INDEL. The second letter represents the version of the reference genome, 'f' for the version 6.1 of Nipponbare. The number is the chromosome coordinate of a variation, e.g. sf0100000131 means a SNP at chromosome 1, 131 bp.

Search for SNPs/INDELs by region

Information of SNPs/INDELs can be queried by limiting genomic coordinates of the rice genome. Since the reference genome used by RiceVarMap is the version 6.1 of Nipponbare from MSU, it should be ensured that all coordinates correspond to this version before query. Basic Local Alignment Search Tool search in the ‘Tools’ menu can be used to obtain corrected genomic coordinates from sequence. Furthermore, SNPs can be filtered by limiting ranges of allele frequencies in different populations (maximum three simultaneous combinations).

Search for SNPs/INDELs within gene

The SNPs/INDELs may have great influence on gene functions. We provide a function for users to search SNPs/INDELs by gene identifiers, gene symbols or key-words of gene annotation, and wildcard characters are accepted. In order to better explain the functional changes caused by SNPs/INDELs, we utilized SNP effector (29) to annotate SNPs/INDELs and the ones with large-effect changes would be highlighted in the result page. Moreover, user can define upstream and downstream regions of genes and retrieve SNPs/INDELs in these regions. The acquired SNPs/INDELs would be displayed in the result page with the structure of the gene in a graph.

Search for genotypes with SNP/INDEL ID

In this interface, users can fetch the genotypes of different accessions through entering SNP/INDEL ID and selecting
Figure 1. Contents and functions of RiceVarMap. (a) Population structure of the 1479 rice accessions included in RiceVarMap. Labels denote the names of subpopulations and the number of accessions in the subpopulation. (b) The functions of RiceVarMap. (c) Haplotype network of S5 constructed using the tool ‘Haplotype Network Analysis’ implemented in RiceVarMap. Each circle represents a haplotype and the size is proportionate to the number of accessions with that haplotype. Branch length represents the genetic distance between two haplotypes.

Search for SNP/INDEL information with ID

Detailed information of single SNP/INDEL is described in this page, which includes not only basic SNP/INDEL information (e.g., position, major allele, minor allele) but also allele frequencies in different populations and information of SNP/INDEL variation effects.

Search for polymorphic positions between two cultivars

It is easy to figure out the polymorphic SNPs between two cultivars in this page, which makes it convenient to generate new molecular markers for bench work in further study.

Search for information of the cultivars

The geographical information of 1479 rice accessions is provided, and each cultivar is assigned with a cultivar ID (e.g. C001, HP84 and W001). The classification information and phenotype images of these rice accessions are available from this page. At the same time, we use Google map to help users to locate the position of each accession.

Design primer by SNP/INDEL ID

RiceVarMap also provides tools to facilitate bench works and further analysis. This tool is provided for researchers to pick PCR primers to validate SNPs/INDELS or develop molecular markers. Primers will be designed to flank the target SNP/INDEL and avoid to overlap with known SNPs/INDELS.

Design primer by region

This tool is designed for researchers to pick PCR primers to amplify target genomic regions and avoid to overlap with known SNPs/INDELS.

Haplotype network analysis

Haplotype network is frequently used in population genetic analysis. RiceVarMap provides a simple tool to generate haplotype network using modified functions from R package pegas (30). Users can download PDF format graph and the detailed haplotype information for further analysis.

A CASE OF APPLICATION

Hybrid sterility is a common phenomenon between different populations; one of the best-known examples for hybrid sterility is the case between the two rice subspecies, indica and japonica. The S5 locus has been well characterized to regulate fertility in indica-japonica hybrids, which are encoded by three tightly linked genes (31,32). We used one of the genes, LOC_Os06g11010, a ‘killer’ of gametes, to demonstrate the functions of RiceVarMap. A total of 14 SNPs were found from ‘Search for SNPs within gene’ corresponding accessions. If the target genotype is a minor allele, it will be displayed in red in the search results. The genotypes can be downloaded in csv format.
function using the gene locus name as input. Five of them (sfo0605759642, sfo0605759919, sfo0605760007, sfo0605760352, sfo0605760512) with minor allele frequencies greater than 0.05 in the whole population could be identified easily from the results and we observed that all of them were non-

synonymous mutations, suggesting that this gene might undergo rapid evolution. We then copied IDs of these SNPs and generated a haplotype network using the tool ‘Haplo-
type Network Analysis’ (Figure 1c). There are five haplo-
types found with at least 10 accessions and we found that nearly all aus accessions were of haplotype IV, which was characterized as wide-compatibility accessions that could overcome the sterility of the hybrid. Thus, accessions with this haplotype would be very useful for hybrid breeding. Finally, we could use the tool ‘Design Primer by SNP/INDEL ID’ to develop molecular markers based on these SNPs as well.

FUTURE DEVELOPMENT

We will make efforts to improve and update the database from the following three aspects. First, as more rice accessions are sequenced (6,33) and publicly available, the size of the database will be enlarged with more rice accessions, SNPs/INDELs and genotypes. And the reference genome used will be updated as well. Second, we have identified thousands of significant loci regulating metabolism (18) and constructed high-throughput phenotyping platforms (34). We are planning to add these data into RiceVarMap, making it a comprehensive database of rice geno-

momic, metabolomic and phenomic variations. Third, we will take efforts to make the database more user-friendly and more efficient with the reflection and feedback of the first version of RiceVarMap.

ACKNOWLEDGEMENTS

We thank Prof. Siben Yu and Mr. Zilong Guo for helping to prepare plant photos, and Mr. Zuoxiong Liu for English editing. We also thank anonymous reviewers for the help in improving our database and the manuscript.

FUNDING

National High Technology Research and Development Program of China (863 Program) [2012AA10A304, 2014AA10A602]; National Natural Science Foundation of China [31100962, 31123009, J1103510]; Fundamental Research Funds for the Central Universities [2011PY608]. Funding for open access charge: National High Technology Research and Development Program (863 Program) [2012AA10A304], the Ministry of Science and Technology of China.

Conflict of interest statement. None declared.

REFERENCES


