Published online 28 November 2012 *Nucleic Acids Research, 2013, Vol. 41, Database issue* ***D1199–D1205***
doi:10.1093/nar/gks1225

The Rice Genome Knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology

Dapeng Wang[1], Yan Xia[1,2], Xinna Li[1], Lixia Hou[1] and Jun Yu[1,*]

[1]CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029 and [2]Graduate University of Chinese Academy of Sciences, Beijing 100049, PR China

Received August 15, 2012; Revised and Accepted October 31, 2012

ABSTRACT

Over the past 10 years, genomes of cultivated rice cultivars and their wild counterparts have been sequenced although most efforts are focused on genome assembly and annotation of two major cultivated rice (*Oryza sativa* L.) subspecies, 93-11 (*indica*) and Nipponbare (*japonica*). To integrate information from genome assemblies and annotations for better analysis and application, we now introduce a comparative rice genome database, the Rice Genome Knowledgebase (RGKbase, http://rgkbase.big.ac.cn/RGKbase/). RGKbase is built to have three major components: (i) integrated data curation for rice genomics and molecular biology, which includes genome sequence assemblies, transcriptomic and epigenomic data, genetic variations, quantitative trait loci (QTLs) and the relevant literature; (ii) User-friendly viewers, such as Gbrowse, GeneBrowse and Circos, for genome annotations and evolutionary dynamics and (iii) Bioinformatic tools for compositional and synteny analyses, gene family classifications, gene ontology terms and pathways and gene co-expression networks. RGKbase current includes data from five rice cultivars and species: Nipponbare (*japonica*), 93-11 (*indica*), PA64s (*indica*), the African rice (*Oryza glaberrima*) and a wild rice species (*Oryza brachyantha*). We are also constantly introducing new datasets from variety of public efforts, such as two recent releases—sequence data from ~1000 rice varieties, which are mapped into the reference genome, yielding ample high-quality single-nucleotide polymorphisms and insertions–deletions.

INTRODUCTION

Rice is one of the economically important monocot crops in the world. Since 2002 when genome assemblies of the two major rice varieties (*Oryza sativa* L. ssp. *indica* 93-11 and *Oryza sativa* L. ssp. *japonica* Nipponbare) were published (1,2), efforts to construct better rice reference genomes continue even to this date (3,4). Comparative analyses on rice and other large plant genomes have been promoting the application of genomic research activities to agricultural practice, such as marker-assisted breeding for the improvement of biotic and abiotic stress resistances (5,6). Although there have been a number of databases or web servers constructed for rice and related plant genomes (7–9), a comprehensive database or knowledgebase for general rice genomic information is still necessary, especially when data are still being generated in a fast rate for this much treasured crop.

We started our database with sequence information from five rice varieties, including a *japonica* variety Nipponbare, an *indica* variety 93-11, a complex *indica* variety PA64s (55% *indica* + 25% *japonica* + 20% *javanica*), the African rice *Oryza glaberrima* and a wild rice species *Oryza brachyantha*, and excavated information on sequence variations, gene expression profiles and subspecific phenotypes. We have several reasons to select these five rice subspecies. First, all of them have relatively complete genome assemblies based on high-coverage sequence data, as most other assemblies are rather partial or assembled based on low-coverage sequences. Second, data from these five genomes already allow users to address many interesting biological questions and to perform certain comprehensive comparative analysis, which may include wild versus cultivated, *japonica* versus *indica*, African versus Asian and maternal parent versus paternal parent of a super-hybrid. In particular, we anticipate that RGKbase is able to provide the most up-to-date molecular information for

*To whom correspondence should be addressed. Tel: +86 10 82995357; Fax: +86 10 82995373; Email: junyu@big.ac.cn

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

deciphering mechanisms on hybrid vigor (or heterosis) and other complex agronomic traits to rice biologists (10). RGKbase adopts parallel annotation pipelines for repetitive sequences, protein-coding genes, noncoding RNAs, genome compositional changes, sequence variations (such as single-nucleotide polymorphisms or SNPs and insertions–deletions or InDels) from individual genomes and in comparative frameworks between a reference rice genome (*Oryza sativa* L. ssp. *japonica* Nipponbare) and other varieties. RGKbase not only reveals genomic characterizations among different genome assemblies but also evaluates pros and cons among various genome annotation approaches and tools used in plant genomics.

IMPLEMENTATION

We implemented the Linux-Apache-Mysql-Php infrastructure to realize dynamic invocation and optimization and HTML/javascript to write the static and simple user-responsive webpage. To accelerate the performance of Gbrowse/Gbrowse-syn displays, we imported a large amount of trace data in the gff3 format into a mysql database. We fully configured compatible Java, Bioperl, C++ and Python in the background of Linux environment. The major programs used as tools are written in Perl codes and graphs are plotted by using R codes. Larger downloadable data stored in this database are compressed in *.bz or *.gz to significantly reduce file sizes.

LITERATURES

We collected >300 published articles related to rice genomics from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and Google Scholar/Google in eight categories: species, genomics, transcriptomics, proteomics, epigenetics, QTLs, intraspecific polymorphisms and comparative genomics. We also listed the quintessence and raw references on the RGKbase pages. Primarily, we organized the literatures chronologically, covering publications in physical mapping, sequence assemblies and functional genomics. We also introduced basic biology, agronomic trait information and following-up research activities on the five rice varieties. Users are able to review the history of rice biology from molecular genetics to genomics.

DATA CONTENT

Genome sequences

We obtained genomic data for four of the five rice varieties from Ensembl Plant Database (Version 14, http://plants.ensembl.org/, ftp.ensemblgenomes.org) and assembled the genome sequence of PA64s based on data produced in our own laboratory and acquired by using the Roche 454 and Life Technologies SOLiD platforms (D. Wang et al., in preparation).

Repetitive elements

Rice genome is composed of diversified repetitive elements (11). We adopted Repeatmasker 3.3.0 and its rice library (http://www.repeatmasker.org/) to scan sequences of long terminal repeat (LTR), long interspersed element (LINE), short interspersed element (SINE), DNA transposon, simple repeat and low complexity and RepeatModeler 1.0.5 (http://www.repeatmasker.org/RepeatModeler.html) for the identification of novel repeats predicted just based on copy numbers of selected repeat sequences. We integrated many complementary tools to identify specific repetitive elements and to scan the genome assemblies using manually created data, such as RetrOryza database (containing the low copy number elements) (12), LTR (LTRharvest1.4.1, LTR_finder1.0.5 and LTR_STRUC1.1) (13–15), MITE/miniature inverted-repeat transposable elements (MITE_Hunter 2010version) (16), tandem duplications (TRF/Tandem repeats finder 4.04) (17) and microsatellites (SciRoKo 3.3) (18). We used results derived from each method instead of combining all predictions to scan the genome assemblies, since a combined approach may lead to disruptions of transposon structures.

Genome compositional dynamics

The nucleotide frequency of G (guanine) + C (cytosine) or the alteration of the two nucleotides is associated with gene density, regulatory elements and transcription activity (19). We introduced IsoFinder (20) and GC_profile2.0 (21) to identify putative isochores (large ones of high-GC or low-GC chromosomal segments) in the rice genome; the former defines long homogeneous genomic regions and the latter uses the *Z*-curve methodology. In addition, we identified CpG islands (CGI) using CpGcluster2.0 (22) that considers the distribution of information distance between two neighboring CpG sequences. We used EP3 (23) to further predict promoters and mapped 576 experimental promoters from PlantProm DB (24) onto the genome sequences in order to search for transcription start sites and complex relationship between promoters and CGI sequences.

Gene annotation

Currently, the genome assemblies in RGKbase were annotated with 69 456 plant RefSeq mRNAs (*Arabidopsis thaliana*, *Glycine max*, *Oryza sativa japonica*, *Solanum lycopersicum* and *Zea mays*) from NCBI ftp database (ftp.ncbi.nih.gov/refseq/) (25) and 536 029 Swiss-Prot proteins from UniProt database (ftp.uniprot.org/) (26). The genes were also predicted based on FGENESH after the assemblies were repeat-masked (but not filtered for simple repeats or low-complexity sequences) and subjected to de novo prediction and evidence-based mapping. In addition, Snap (27) and Nscan/Twinscan (28) were also used for protein-coding gene annotation. The annotated protein sequences were submitted to Interproscan version 5RC2 (29) that integrated many popular tools such as ProDom, PANTHER, PROSITE, Pfam and SMART. This system can be used for gene ontology (GO) term and protein domain annotations. We also integrated pathway annotations by using KEGG API Server (http://www.genome.jp/kegg/) (30) with a bi-directional best hit approach. Regarding transcription factors and three-dimensional

structures, we mapped protein sequences to plant transcriptional factors in reference to PlnTFDB 3.0 (http://plntfdb.bio.uni-potsdam.de/v3.0/) (31) and PlantTFDB 2.0 (http://planttfdb.cbi.edu.cn/) (32) as well as proteins in PDB (www.rcsb.org/) (33). In fact, we only selected the best hit as the annotation of each protein. We subsequently used tRNAscan1.3.1 (34) to generate information on location, sequence and secondary structure of tRNAs. We annotated rRNAs by adopting RNAmmer (35) and utilized snoReport for snoRNAs (small nucleolar RNAs) (36). We used 1 331 936 rice expressed sequence tags (ESTs) from dbEST database (http://www.ncbi.nlm.nih.gov/dbEST/) (37) as the first gene annotation dataset, including those from *Oryza sativa* (1 252 989), *Oryza longistaminata* (71 367), *Oryza minuta* (5309), *Oryza officinalis* (1468) and *Oryza punctata* (803). Both UniProt proteins and RefSeq plant mRNAs were also used independently for the validation of de novo predictions and EST-based gene annotations. miRNA sequences from miRBase (Version 19, http://www.mirbase.org/) (38) covering species with diverse evolutionary history and PMRD (http://bioinformatics.cau.edu.cn/PMRD/) focusing on plants (39) were also mapped onto the genome assemblies by using MapMi (40).

Gene transcription and translation

Both microarray-based and high-throughput RNA-seq data from the same rice variety are used for RGKbase. For the microarray data, we mapped probe sequences onto the predicted mRNAs or coding sequences (CDSs) and only selected the best or the only match as derived position. As to the RNA-seq data from three subspecies (41,42), we combined Tophat and Cufflink to map transcripts for estimating gene expression with FPKM values (43). We used the gff guiding for gene expression estimation but did not adopt this parameter for building RNA-seq gene models in order to find novel splicing events. In addition, we downloaded eight epigenetic datasets from four types of libraries (DNA methylation McrBC-Seq, H3K4me3 ChIP-Seq, H3K9ac ChIP-Seq and H3K27me3 ChIP-Seq) and the two major rice subspecies (Nipponbare and 93-11) (44). We subsequently mapped the data onto the two genome assemblies and integrated them as tracks in Gbrowse. We extracted proteome data from OryzaPG-BD (http://oryzapg.iab.keio.ac.jp/) (45) and used results from our previous study on 93-11 and PA64s (46), employing blastp for choosing the best hit as a standard for protein identification.

Intraspecific polymorphisms

The polymorphism data are essentially from two sequencing projects; one is a comparative analysis of the five model rice genome assemblies and the other is a collaborative project to sequence 1000 rice varieties for polymorphism studies (6,47). Since the original data were mapped onto the IRGSP's v4 reference (http://rgp.dna.affrc.go.jp/IRGSP/), RGKbase uses it as the intermediate genome to locate sequence variations.

Comparative genomics

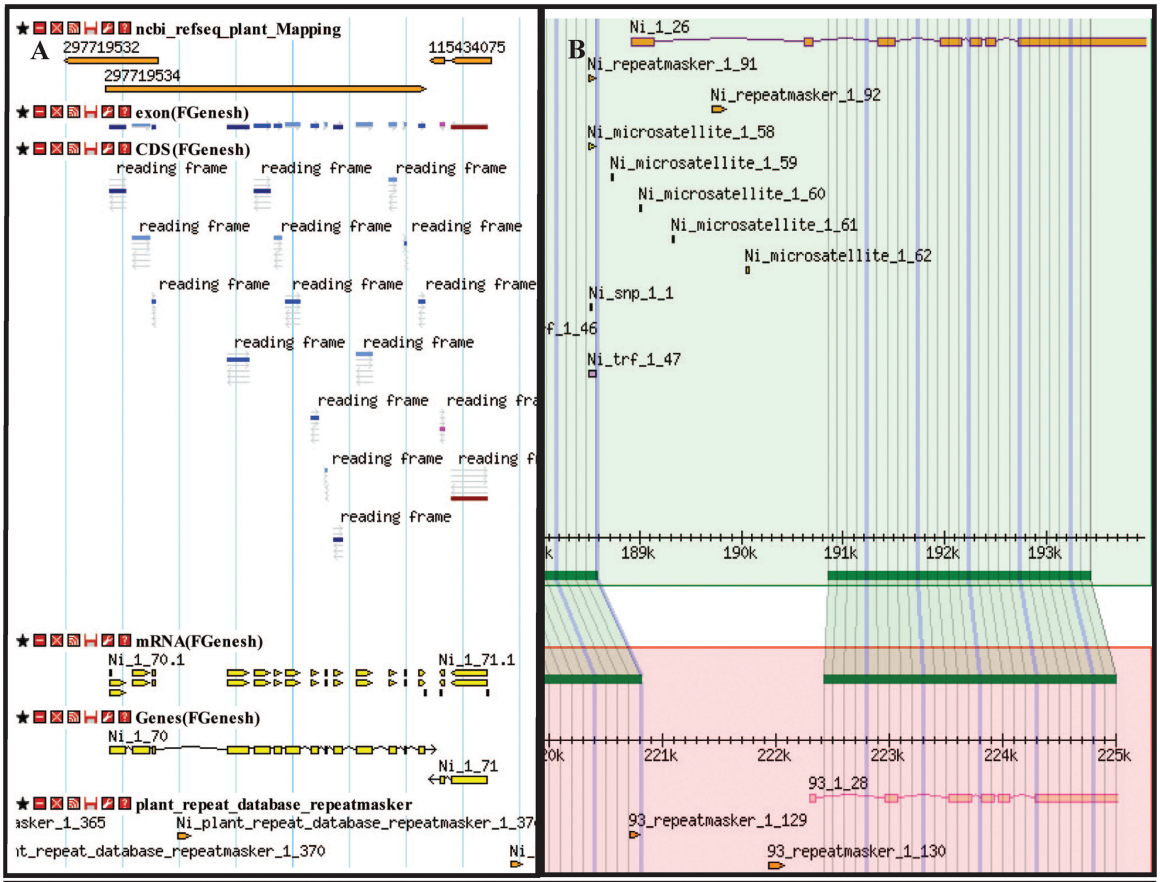
Since our data are mainly from a single species and its assumed wild ancestors, most of the variations we collected are intraspecific (specific to rice) and subspecific (specific to subspecies or varieties). Our data were structured at three levels. First, we chose the Nipponbare genome as the reference and compared it with the other four genome assemblies in a pairwise fashion. The sequences were aligned based on Lastz (48). Briefly, the subject sequence was mapped in both forward and reverse directions onto the reference, and gaps were allowed since the intergenic sequence has more repeats so that it is not usually mapped well (49). In addition, we also provided comparative analysis for each chromosome to itself and to other chromosomes within each rice genome assembly with gap penalty since plant segmental duplication events happen not only within internal regions of the same chromosomes but also between chromosomes (3,50). Pecan was used for alignment in a unit of chromosome and a Perl script was written to identify consensus sequences that are assumed as ancestral sequences (51). Second, we identified homologous genes by carrying out an all-blast (to)-all protein matching using E-value = 1E−5 and identity $\geq$30% as cut-offs. Protein families were determined based on Tribe-MCL with $I = 2$ (52). In particular, we combined all curated proteins into a library, rice proteins-UniProt. The median length and the number of proteins in each family were calculated. Third, we used ParaAT1.0 to align the coding regions of genes (53) and various methods built in KaKs_Calculator2.0 (54) to compute nonsynonymous (K_a) and synonymous (K_s) substitution rates and selection strength (K_a/K_s). K_a and K_s analyses are widely demonstrated to be effective in making a distinction between fast- and slow-evolving protein-coding genes, or variable and conservative protein-coding genes (55).

User entry

We provide three approaches for users to access information in this database. First, five different types of BLAST families are used to compare DNA/protein full-length sequences or fragments with gene and protein fasta files (56). Second, BLAT is used to map DNA/protein (especially mRNA/EST) sequences onto genome assemblies, leading to fine mapping of gene structures (57). Last, the search text box supports interactive keyword searching, including Ensembl gene ID, RGKbase gene ID and keywords in title or abstract of the related literature involved in RGKbase.

Browser

We utilize three data browser systems to display the core data. First, we convert analyzed data into to gff3 files and guide them into a popular Gbrowse package (58). The tracks are divided into the default and the optional. Users can selectively display different tracks (Figure 1A). The visualization of different results derived from different methods helps users to identify high-confidence regions and to compare characteristics of different algorithms



C

Species Name	Gene Name	Chromosome Name	Strand	Transcript Start	Transcript End	Length	Exon Number
Oryza sativa L. ssp. Indica 93-11	93_2_1337	2	-	13620737	13625183	4447	9
Oryza sativa L. ssp. Indica 93-11	93_2_1338	2	-	13648621	13653814	5194	6
Oryza sativa L. ssp. Indica 93-11	93_2_1339	2	+	13660399	13664676	4278	8
Oryza sativa L. ssp. Indica 93-11	93_2_1340	2	+	13684335	13685707	1373	3
Oryza sativa L. ssp. Indica 93-11	93_2_1341	2	+	13687193	13688853	1661	4
Oryza sativa L. ssp. Indica 93-11	93_2_1342	2	-	13697745	13698918	1174	2
Oryza sativa L. ssp. Indica 93-11	93_2_1343	2	+	13712217	13716923	4707	10
Oryza sativa L. ssp. Indica 93-11	93_2_1344	2	-	13723609	13727349	3741	1
Oryza sativa L. ssp. Indica 93-11	93_2_1345	2	+	13729278	13734509	5232	4
Oryza sativa L. ssp. Indica 93-11	93_2_1346	2	+	13748503	13750618	2116	4
Oryza sativa L. ssp. Indica 93-11	93_2_1347	2	-	13766250	13768325	2076	1

D

Figure 1. Snapshots of selected browser functions. (**A**) Gbrowse; an example displays RefSeq mapping, gene structure and repetitive elements annotations in a selected window. (**B**) Gbrowse-syn; a window displays syntenic regions and selected tracks between two genomes, Nipponbare and 93-11. (**C**) GeneBrowse; it lists details of gene and genome annotations (93-11 genome). (**D**) Circos; it displays a link map of homologous genic regions of rice chromosomes (93-11).

and software packages. We also divided all tracks into two major categories: nested and discontinuous features, which are denoted with a combination of basic shapes (rectangle, arrow and line). Other quantitative data are shown in barplot, curve and heatmap. In addition, Gbrowse-syn is used to compare the co-linear conserved regions between reference genome and other genomes identified by Lastz (48). We also added selected valuable tracks from

Gbrowse to Gbrowse-syn and a new track to show similarity of compared segments (Figure 1B). Second, gene browse directs users to genes in a genome and their detailed structural and functional annotations, including two tables 'Basic and structural information' and 'Function and domain annotation', such as gene, protein, chromosome, start and end positions, exon and intron boundaries, GO terms, pathways, PDB ID and classification of transcription factors. This function allows users to look into genes of interest in the database based on variety of initial information (Figure 1C). Last, aside from the first two dynamic or interactive browsers, we also provided general static figures generated beforehand by using Circos (59). We mapped gene and repetitive elements onto chromosome and make them comparable in both interchromosomal and intrachromosomal modes. In addition, we drew 'link map of homologous genic regions' and added hotspot and hyperlink on the static figure and linked them to compared regions in Gbrowse-syn (Figure 1D). In details, we only selectively showed the anchors in the duplicated regions best identified both at the protein and nucleotide levels. In addition, 'Syntenic block' displays collinear blocks in the same chromosome or different chromosomes for the same species or different species.

Tools

First, we built in a tool for compositional analysis on transcript-centric GC-gradient that is related to the mechanism of transcription-coupled DNA repair and is found in the rice genome and other grass family plants but appears absent in other families of monocotyledons and all dicotyledons, such as onion and *Arabidopsis* (60,61). This tool helps users calculate sequence content variations, including the gradient or alteration of mononucleotide, dinucleotide, GC content, codon usage and codon bias along transcripts. Second, for better sequence alignment, RGKbase provides multiple sequence alignment function for gene families, taking the advantage of two aligning algorithms: Clustalw (62) and Muscle (63). Users are able to select gene family of their interest and align all protein members of the family to check for change and gain/loss of amino acids. Third, GO annotation and pathway annotation are not only done in general for each gene but also done for multiple functional annotations so that each functional annotation involves multiple genes. GO (64) and KEGG pathways (30) have been widely used in the functional identification of genes and their interaction networks. For simple statistics, RGKbase provides Fisher's exact test and multiple testing correction method with different thresholds for users to investigate the five rice genome assemblies. Fourth, we implemented ID convertor to use chromosome position information and to concatenate genes and their products from different annotation systems, such as Ensembl. In fact, the Ensembl genome annotation system adopts the two existing rice genome annotations from RGAP (9) and BGI-RIS (7) databases, where they maintain data from Nipponbare and 93-11, respectively. Therefore, users can easily look up genes of interest in between RGKbase and other related databases. Fifth, for visualization of syntenic genes and chromosomal segments, we previously constructed a genome co-linearity database that focuses on animal gene position and distance (65), and we now introduced its framework into RGKbase. This tool helps users to observe and understand the conservation and dynamics of homologous genes and gene clusters in different evolutionary lineages of rice genomes. Sixth, we also built gene co-expression networks by mapping selected gene expression datasets onto annotated genomes to build the expression matrix for multiple tissues or developmental stages, adopting Pearson coefficient parameters and calculating correlations between any of two genes (66,67). We used two types of criteria, rank and absolute value, to determine the edges of gene expression network. Users can enter the name of certain genes, and this tool generates a figure to show the network that contains these genes as key nodes. Seventh, we introduced two datasets of QTLs from Q-TARO (68) and Gramene database (69) and find their locations by medium of Nipponbare genome. Users can search their traits of interest and check sequence variations in the corresponding regions by means of hyperlink to Gbrowse.

Future work

We will keep updating the genome assemblies of the four rice varieties by filling gaps, correcting assembly errors and annotating genes based on both software upgrades and new data since only the Nipponbare genome has been considered a 'finished map' by the rice community until now. We will also carry out large-scale sequencing and re-sequencing projects to generate more data on different rice breeds or varieties. We will build a reference genome or 'pangenome' of rice, identify variety-specific variations and establish genotype–phenotype relations for agronomically important traits. The ultimate goals of constructing RGKbase include genome annotations, comparative genomics, evolutionary studies, trait mapping and identifications of genetic markers for marker-assisted breeding of rice and other grass crops.

DISCUSSION AND CONCLUSION

We built RGKbase for three fundamental scientific questions. First, what are the common features or conserved contents within the rice genome and distinct from other monocotyledons? Second, what are discriminative patterns among various rice subspecies? And how does rice evolve during domestication processes from the wild to the cultivated in multiple geographic regions? Last, how can we link genome variations/genetic markers to phenotypes/traits? We are sure that such information provides valuable clues for molecular biologists and agriculturists to exploit potentials of rice as a major crop to feed the world. We not only contribute original data for downloading but also visualization tools for multiple datasets suitable for comparative genome studies. We are working on a series of specific methodologies for plant research, especially rice genome annotations and population

genetics analyses to address specific questions for rice biology, genetics and evolution. For example, we will build knowledge networks for the sequence-composition-to-function complexity. Ultimately, we will update the data and annotations hosted by RGKbase when any new data or novel analysis methods become available and keep this database up-to-date, becoming a curation and distribution center of the knowledge of rice genome and biology.

FUNDING

Funding for open access charge: National Basic Research Program [973 Program; 2011CB944100 and 2011CB944101]; National Programs for High Technology Research and Development [863 Program; 2012AA020409]; National Natural Science Foundation of China [90919024]; Special Foundation Work Program [2009FY120100].

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Goff,S.A., Ricke,D., Lan,T.H., Presting,G., Wang,R., Dunn,M., Glazebrook,J., Sessions,A., Oeller,P., Varma,H. *et al.* (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica)*. *Science*, **296**, 92–100.
3. Yu,J., Wang,J., Lin,W., Li,S., Li,H., Zhou,J., Ni,P., Dong,W., Hu,S., Zeng,C. *et al.* (2005) The Genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.*, **3**, e38.
4. International Rice Genome Sequencing Project. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.
5. Zhao,K., Tung,C.W., Eizenga,G.C., Wright,M.H., Ali,M.L., Price,A.H., Norton,G.J., Islam,M.R., Reynolds,A., Mezey,J. *et al.* (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat. Commun.*, **2**, 467.
6. Huang,X., Zhao,Y., Wei,X., Li,C., Wang,A., Zhao,Q., Li,W., Guo,Y., Deng,L., Zhu,C. *et al.* (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat. Genet.*, **44**, 32–39.
7. Zhao,W., Wang,J., He,X., Huang,X., Jiao,Y., Dai,M., Wei,S., Fu,J., Chen,Y., Ren,X. *et al.* (2004) BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics. *Nucleic Acids Res.*, **32**, D377–D382.
8. Youens-Clark,K., Buckler,E., Casstevens,T., Chen,C., Declerck,G., Derwent,P., Dharmawardhana,P., Jaiswal,P., Kersey,P., Karthikeyan,A.S. *et al.* (2011) Gramene database in 2010: updates and extensions. *Nucleic Acids Res.*, **39**, D1085–D1094.
9. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
10. Huang,Y., Zhang,L., Zhang,J., Yuan,D., Xu,C., Li,X., Zhou,D., Wang,S. and Zhang,Q. (2006) Heterosis and polymorphisms of gene expression in an elite rice hybrid as revealed by a microarray analysis of 9198 unique ESTs. *Plant Mol. Biol.*, **62**, 579–591.
11. Turcotte,K., Srinivasan,S. and Bureau,T. (2001) Survey of transposable elements from rice genomic sequences. *Plant J.*, **25**, 169–179.
12. Chaparro,C., Guyot,R., Zuccolo,A., Piegu,B. and Panaud,O. (2007) RetrOryza: a database of the rice LTR-retrotransposons. *Nucleic Acids Res.*, **35**, D66–D70.
13. Xu,Z. and Wang,H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*, **35**, W265–W268.
14. McCarthy,E.M. and McDonald,J.F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics*, **19**, 362–367.
15. Ellinghaus,D., Kurtz,S. and Willhoeft,U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.
16. Han,Y. and Wessler,S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.*, **38**, e199.
17. Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
18. Kofler,R., Schlotterer,C. and Lelley,T. (2007) SciRoKo: a new tool for whole genome microsatellite search and investigation. *Bioinformatics*, **23**, 1683–1685.
19. Zhang,R. and Zhang,C.T. (2004) Isochore structures in the genome of the plant *Arabidopsis thaliana*. *J. Mol. Evol.*, **59**, 227–238.
20. Oliver,J.L., Carpena,P., Hackenberg,M. and Bernaola-Galvan,P. (2004) IsoFinder: computational prediction of isochores in genome sequences. *Nucleic Acids Res.*, **32**, W287–W292.
21. Gao,F. and Zhang,C.T. (2006) GC-Profile: a web-based tool for visualizing and analyzing the variation of GC content in genomic sequences. *Nucleic Acids Res.*, **34**, W686–W691.
22. Hackenberg,M., Previti,C., Luque-Escamilla,P.L., Carpena,P., Martinez-Aroza,J. and Oliver,J.L. (2006) CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, **7**, 446.
23. Abeel,T., Saeys,Y., Bonnet,E., Rouze,P. and Van de Peer,Y. (2008) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res.*, **18**, 310–323.
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. Pruitt,K.D., Tatusova,T., Brown,G.R. and Maglott,D.R. (2012) NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Res.*, **40**, D130–D135.
26. The UniProt Consortium. (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **40**, D71–D75.
27. Korf,I. (2004) Gene finding in novel genomes. *BMC Bioinformatics*, **5**, 59.
28. van Baren,M.J., Koebbe,B.C. and Brent,M.R. (2007) Using N-SCAN or TWINSCAN to predict gene structures in genomic DNA sequences. *Curr. Protoc. Bioinformatics*, **Chapter 4**, Unit 4.8.
29. Hunter,S., Jones,P., Mitchell,A., Apweiler,R., Attwood,T.K., Bateman,A., Bernard,T., Binns,D., Bork,P., Burge,S. *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.*, **40**, D306–D312.
30. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
31. Perez-Rodriguez,P., Riano-Pachon,D.M., Correa,L.G., Rensing,S.A., Kersten,B. and Mueller-Roeber,B. (2010) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.*, **38**, D822–D827.
32. Zhang,H., Jin,J., Tang,L., Zhao,Y., Gu,X., Gao,G. and Luo,J. (2011) PlantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.*, **39**, D1114–D1117.
33. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
34. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
35. Lagesen,K., Hallin,P., Rodland,E.A., Staerfeldt,H.H., Rognes,T. and Ussery,D.W. (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.*, **35**, 3100–3108.

36. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2008) SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
37. Boguski,M.S., Lowe,T.M. and Tolstoshev,C.M. (1993) dbEST—database for "expressed sequence tags". *Nat. Genet.*, **4**, 332–333.
38. Kozomara,A. and Griffiths-Jones,S. (2011) miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res.*, **39**, D152–D157.
39. Zhang,Z., Yu,J., Li,D., Liu,F., Zhou,X., Wang,T., Ling,Y. and Su,Z. (2009) PMRD: plant microRNA database. *Nucleic Acids Res.*, **38**, D806–D813.
40. Guerra-Assuncao,J.A. and Enright,A.J. (2010) MapMi: automated mapping of microRNA loci. *BMC Bioinformatics*, **11**, 133.
41. Zhang,G., Guo,G., Hu,X., Zhang,Y., Li,Q., Li,R., Zhuang,R., Lu,Z., He,Z., Fang,X. *et al.* (2010) Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res.*, **20**, 646–654.
42. Kyndt,T., Denil,S., Haegeman,A., Trooskens,G., De Meyer,T., Van Criekinge,W. and Gheysen,G. (2012) Transcriptome analysis of rice mature root tissue and root tips in early development by massive parallel sequencing. *J. Exp. Bot.*, **63**, 2141–2157.
43. Trapnell,C., Roberts,A., Goff,L., Pertea,G., Kim,D., Kelley,D.R., Pimentel,H., Salzberg,S.L., Rinn,J.L. and Pachter,L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.
44. He,G., Zhu,X., Elling,A.A., Chen,L., Wang,X., Guo,L., Liang,M., He,H., Zhang,H., Chen,F. *et al.* (2010) Global epigenetic and transcriptional trends among two rice subspecies and their reciprocal hybrids. *Plant Cell*, **22**, 17–33.
45. Helmy,M., Tomita,M. and Ishihama,Y. (2011) OryzaPG-DB: rice proteome database based on shotgun proteogenomics. *BMC Plant Biol.*, **11**, 63.
46. Wang,W., Meng,B., Ge,X., Song,S., Yang,Y., Yu,X., Wang,L., Hu,S., Liu,S. and Yu,J. (2008) Proteomic profiling of rice embryos from a hybrid rice cultivar and its parental lines. *Proteomics*, **8**, 4808–4821.
47. Xu,X., Liu,X., Ge,S., Jensen,J.D., Hu,F., Li,X., Dong,Y., Gutenkunst,R.N., Fang,L., Huang,L. *et al.* (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.*, **30**, 105–111.
48. Harris,R.S. (2007) Improved pairwise alignment of genomic DNA, Ph.D. Thesis. The Pennsylvania State University.
49. Yu,J., Wong,G.K.S., Wang,J. and Yang,H. (2005) Shotgun sequencing (SGS). *Encyclopedia of Molecular Cell Biology and Molecular Medicine*, Vol. 13, 2nd edn. Wiley-VCH, Weinheim, Germany, pp. 71–114.
50. Guyot,R. and Keller,B. (2004) Ancestral genome duplication in rice. *Genome*, **47**, 610–614.
51. Paten,B., Herrero,J., Beal,K., Fitzgerald,S. and Birney,E. (2008) Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.*, **18**, 1814–1828.
52. Enright,A.J., Van Dongen,S. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.*, **30**, 1575–1584.
53. Zhang,Z., Xiao,J., Wu,J., Zhang,H., Liu,G., Wang,X. and Dai,L. (2012) ParaAT: a parallel tool for constructing multiple protein-coding DNA alignments. *Biochem. Biophys. Res. Commun.*, **419**, 779–781.
54. Wang,D., Zhang,Y., Zhang,Z., Zhu,J. and Yu,J. (2010) KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics Proteomics Bioinformatics*, **8**, 77–80.
55. Wang,D., Liu,F., Wang,L., Huang,S. and Yu,J. (2011) Nonsynonymous substitution rate (Ka) is a relatively consistent parameter for defining fast-evolving and slow-evolving protein-coding genes. *Biol. Direct*, **6**, 13.
56. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
57. Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
58. Donlin,M.J. (2007) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinformatics*, **Chapter 9**, Unit 9.9.
59. Krzywinski,M., Schein,J., Birol,I., Connors,J., Gascoyne,R., Horsman,D., Jones,S.J. and Marra,M.A. (2009) Circos: an information aesthetic for comparative genomics. *Genome Res.*, **19**, 1639–1645.
60. Wong,G.K., Wang,J., Tao,L., Tan,J., Zhang,J., Passey,D.A. and Yu,J. (2002) Compositional gradients in Gramineae genes. *Genome Res.*, **12**, 851–856.
61. Kuhl,J.C., Cheung,F., Yuan,Q., Martin,W., Zewdie,Y., McCallum,J., Catanach,A., Rutherford,P., Sink,K.C., Jenderek,M. *et al.* (2004) A unique set of 11,008 onion expressed sequence tags reveals expressed sequence and genomic differences between the monocot orders Asparagales and Poales. *Plant Cell*, **16**, 114–125.
62. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
63. Edgar,R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.
64. The Gene Ontology Consortium. (2012) The Gene Ontology: enhancements for 2011. *Nucleic Acids Res.*, **40**, D559–D564.
65. Wang,D., Zhang,Y., Fan,Z., Liu,G. and Yu,J. (2012) LCGbase: a comprehensive database for lineage-based co-regulated genes. *Evol. Bioinform. Online*, **8**, 39–46.
66. Jiao,Y., Tausta,S.L., Gandotra,N., Sun,N., Liu,T., Clay,N.K., Ceserani,T., Chen,M., Ma,L., Holford,M. *et al.* (2009) A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat. Genet.*, **41**, 258–263.
67. Wei,G., Tao,Y., Liu,G., Chen,C., Luo,R., Xia,H., Gan,Q., Zeng,H., Lu,Z., Han,Y. *et al.* (2009) A transcriptomic analysis of superhybrid rice LYP9 and its parents. *Proc. Natl Acad. Sci. USA*, **106**, 7695–7701.
68. 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.
69. Ni,J., Pujar,A., Youens-Clark,K., Yap,I., Jaiswal,P., Tecle,I., Tung,C.W., Ren,L., Spooner,W., Wei,X. *et al.* (2009) Gramene QTL database: development, content and applications. *Database*, **2009**, bap005.