Nucleic Acids Research, 2004, Vol. 32, Database issue ***D377–D382***
DOI: 10.1093/nar/gkh085

BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics

Wenming Zhao[1], Jing Wang[2], Ximiao He[1], Xiaobing Huang[1], Yongzhi Jiao[1], Mingtao Dai[1], Shulin Wei[1], Jian Fu[1], Ye Chen[1], Xiaoyu Ren[1], Yong Zhang[1,2], Peixiang Ni[1], Jianguo Zhang[1], Songgang Li[1,2], Jian Wang[1], Gane Ka-Shu Wong[1,3], Hongyu Zhao[4], Jun Yu[1], Huanming Yang[1] and Jun Wang[1,*]

[1]Beijing Genomics Institute (BGI), Chinese Academy of Sciences (CAS), Beijing Airport Industrial Zone-B6, Beijing 101300, China, [2]College of Life Sciences, Peking University, Beijing 100871, China, [3]University of Washington Genome Center, Department of Medicine, Seattle, WA 98195, USA and [4]Yale University School of Medicine, Department of Epidemiology and Public Health, New Haven, CT 06520-8034, USA

Received August 15, 2003; Revised and Accepted October 3, 2003

ABSTRACT

Rice is a major food staple for the world's population and serves as a model species in cereal genome research. The Beijing Genomics Institute (BGI) has long been devoting itself to sequencing, information analysis and biological research of the rice and other crop genomes. In order to facilitate the application of the rice genomic information and to provide a foundation for functional and evolutionary studies of other important cereal crops, we implemented our Rice Information System (BGI-RIS), the most up-to-date integrated information resource as well as a workbench for comparative genomic analysis. In addition to comprehensive data from *Oryza sativa* L. ssp. *indica* sequenced by BGI, BGI-RIS also hosts carefully curated genome information from *Oryza sativa* L. ssp. *japonica* and EST sequences available from other cereal crops. In this resource, sequence contigs of *indica (93-11)* have been further assembled into Mbp-sized scaffolds and anchored onto the rice chromosomes referenced to physical/genetic markers, cDNAs and BAC-end sequences. We have annotated the rice genomes for gene content, repetitive elements, gene duplications (tandem and segmental) and single nucleotide polymorphisms between rice subspecies. Designed as a basic platform, BGI-RIS presents the sequenced genomes and related information in systematic and graphical ways for the convenience of in-depth comparative studies (http://rise.genomics.org.cn/).

INTRODUCTION

Rice is not only the most important crop species providing staple food for more than half the world's population, but also a model organism for the biological study of the grass family of crops and other plants. The genome sequences of domesticated rice (1,2) provide a solid foundation for integrating biological information, including genetics, gene expression, development, physiology and evolution, which will be extended to cereal crops, monocots and even plants in general. Although rice, together with other cereal crops, diverged from a common ancestor some 60 million years ago (3), their genes and gene orders (synteny) are highly conserved (4). It becomes feasible and highly desirable for agricultural genomics that a robust, versatile workbench and specific tools are developed in time to facilitate biological research, starting from a single species and extending to other cereal crops (5).

At the Beijing Genomics Institute (BGI), the major genome sequencing center in China, we have been carrying out the Superhybrid Rice Genome Project (SRGP) with full efforts to understand the genome biology of rice. Using a whole-genome-shotgun approach, we have produced a draft genome sequence for rice, from *93-11*, which is a cultivar of *Oryza sativa* L. ssp. *indica*, one of the three major rice subspecies grown most widely in China and Southeast Asia. In order to maximize usage of the most up-to-date knowledge about the rice genome, we develop the Rice Information System (BGI-RIS) as a highly integrated resource database for rice data storage, retrieval, visualization and analysis. The current version of BGI-RIS is focused on assembling contigs and anchoring contigs/scaffolds onto the rice chromosomes based on mapped genetic markers and BAC-based physical maps. In the process, we have developed software packages for identification and annotation of genes and polymorphisms, as well as efficient visualization systems. We utilize rice as a framework genome to organize information for other cereal crops such as wheat,

*To whom correspondence should be addressed. Tel: +86 10 80481662; Fax: +86 10 80498676; Email: wangj@genomics.org.cn
Correspondence may also be addressed to Huanming Yang. Tel: +86 10 80494969; Fax: +86 10 80491181; Email: yanghm@genomics.org.cn

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

maize and barley, so as to bridge the model plant and its family members. Special emphasis is placed on the comparative analysis of different subspecies of rice and in the future, between rice, *Arabidopsis thaliana* and other cereal crops. BGI-RIS, together with its most up-to-date database, search engine, species-specific map viewer and comparative genomics viewer, provides both an information resource and a comparative analysis workbench for genomic research of rice, other cereal crops and plants.

DATA CONTENT AND SOURCING

BGI-RIS integrates our own genomic data on *O.sativa* L. ssp. *indica* (*93-11*), genome sequences of *O.sativa* L. ssp. *japonica* from other institutions, as well as EST sequences of rice from our own production and public data for rice and other cereal crops, such as maize, wheat and barley (http://www.ncbi.nlm.nih.gov/dbEST/). Additional related information from rice genomics (such as BACs; ftp://ftp.genome.arizona.edu/pub/

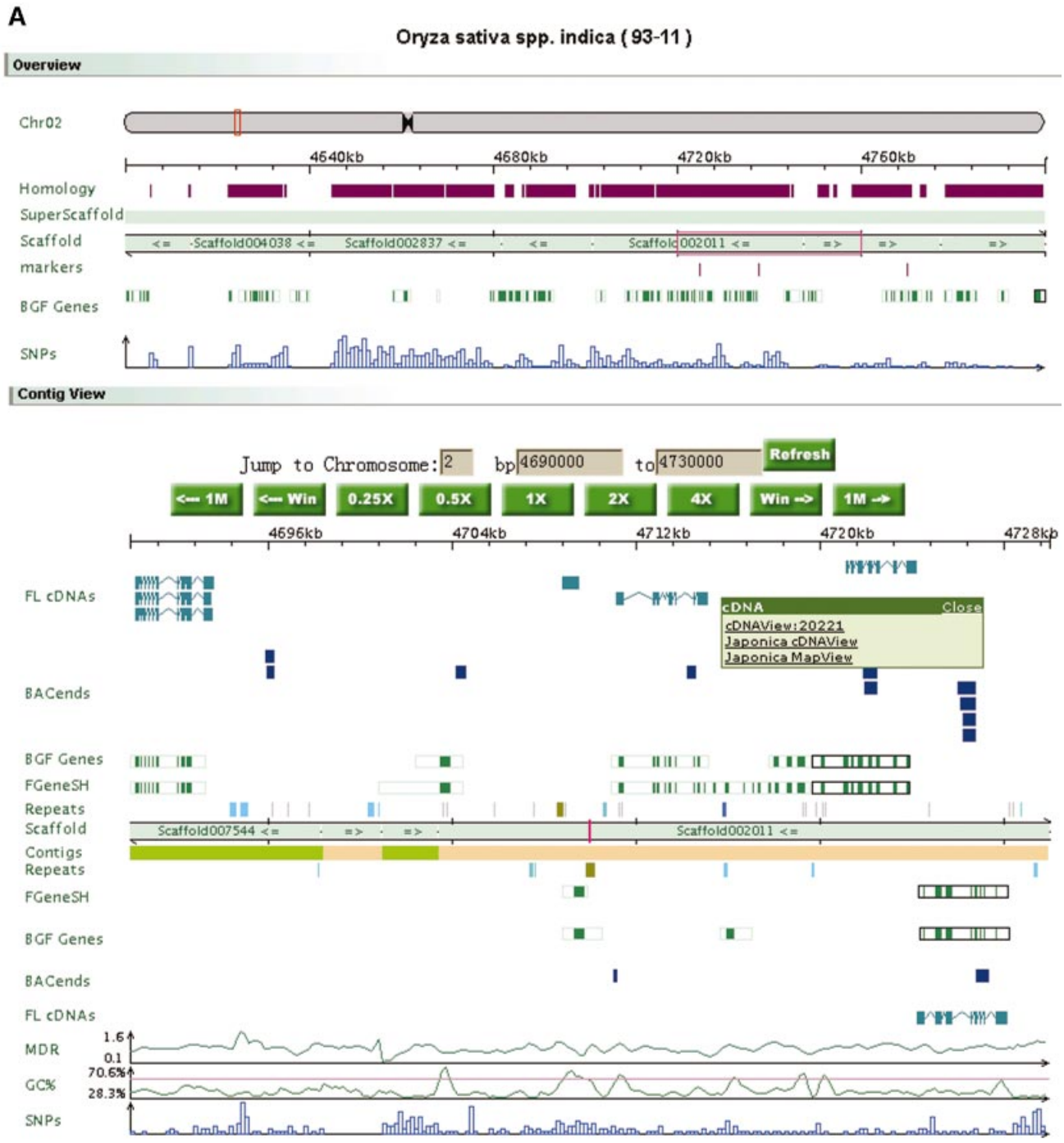

Figure 1. Examples of the of BGI-RIS viewer for *O.sativa* L. ssp. *indica* (*93-11*). Screenshots of (**A**) the ContigView, (**B**) the GeneView, which offers zoomed-in information for the predicted genes chosen in the ContigView and (**C**) the cDNAView, with two highlighted SNPs identified between the genome sequences of *indica* (*93-11*) and *japonica* (Nipponbare).

stc/rice/) and genetics (such as genetic markers; http://www.gramene.org/resources/) wherever publicly available is also carefully curated and integrated into BGI-RIS. Due to the complexity and the large-scale nature of the genomic data, the strategy of comprehensive organization and effective management are of essence for successive analyses. In BGI-RIS, we organize the genomic data at three vertical levels as different modules: chromosome level, contig/scaffold level and genetic element level, which are in accordance with the main tables in the database schema, and link the data of the three different levels through genome-oriented MapView and CompView for comparative analysis.

The chromosome module contains basic information on the scale of chromosomes, such as length, centromere location, number of genes, markers and single nucleotide polymorphisms (SNPs). The major module in BGI-RIS is the contig/scaffold module, which contains detailed information for each assembled contig/scaffold, including BAC-end sequences, cDNAs/ESTs, physical and genetic markers, repetitive elements and SNPs. Up to the present, 24 890 full-length cDNAs have been mapped onto the 12 chromosomes of *indica* (*93-11*) in reference to cDNAs from *japonica* rice (6). The longest assembled contig of *indica* (*93-11*) (superscaffold or Mbp-sized scaffold) is nearly 18 Mb by the time of submission. We performed robust *in silico* alignments to anchor assembled contigs onto the 12 rice chromosomes. A total of 231 superscaffolds are aligned along the rice chromosomes of *indica* (*93-11*) with high stringency cut-off criteria. These alignments provide a reliable resource for positional cloning of rice genes. The contig sequences were annotated for gene content by using automated processes that involve *ab initio* gene finders, such as FgeneSH (http://www.softberry.com) and Beijing Gene Finding (BGF), and database searches against public nucleic acid and protein data. BGF is developed at BGI and is a program for gene identification in eukaryotic genomic DNA sequences (7–9). It is based on Dynamic Programming (10,11) and Hidden Markov Model (HMM) algorithm (12–14) with a special emphasis on rice genomes. We also carried out SNP analyses on both a chromosome scale and a cDNA scale. About 5 million SNPs between *indica* (*93-11*) and *japonica* (Nipponbare) have been identified. Quality values (15) for each nucleic acid are also included for empirical SNP verification. The differences between cereal genomes of distant taxa are characterized by large variations in their DNA content (4) and the expansion of genome sizes primarily due to insertions of repetitive elements (16). We applied RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) for identifying repetitive sequences and tagging repeat classes. Detailed information concerning repetitive elements, genetic markers, genes, exons and SNPs is stored in the third module, the genetic element module.

Cross-species comparisons within members of the cereal family are proven keys to understanding the evolutionary changes in genes and their structures, which are related to their functions and agronomic traits underlined by them in cereal crops (5). Based on the information organized in the three modules, we provide sequence-based marker and full-length cDNA alignments between the genome sequences of *indica* (*93-11*) and *japonica*. Synteny alignments and comparisons between rice, *A.thaliana* and other cereal crops will be added in future updates. All the data described above are available for download through our FTP site.

Besides timely updated rice genome sequence information, BGI-RIS has been constantly incorporating more data, when they become available, from other plant genomes and different types of biological data, such as tRNA, mRNA, SAGE and microarray data. To assist the users, we have introduced into BGI-RIS a versioning system and a frame of reference around different versions of the rice data. In the near future, it will be possible for users to retrieve the data of different versions, to trace up and locate changes of a given entity between different versions.

ACCESS AND WEB QUERY INTERFACE

There are two ways for users to access data stored in the genome database: browse and query. As an efficient visualization tool, MapView for each rice genome (i.e. species-specific map) is composed of three types of subviewer in hierarchical architecture, namely ChroView/OverView (17), ContigView and GeneView/cDNAView, which are in accordance with the data organization (Fig. 1). ChroView is based on a low-resolution physical map with sequence contigs/scaffolds aligned to, and relevant statistics of the aligned sequences and genomic elements. We also mark out homologous regions between *indica* (*93-11*) and *japonica*. ContigView allows users to highlight an area in ChroView to browse the annotated information for the chosen contigs/scaffolds. Anchored BAC ends, predicted genes, classes of repeats and SNPs are positioned along the targeted area with distinct color coding. Detailed information about each specific gene/cDNA is presented in the features of GeneView/cDNAView, such as exon–intron structures and protein sequences. A factual report for each element contained in the visualization system is displayed automatically by clicking. In BaseView, users can focus on a region of DNA sequence and investigate on Q-value for each nucleotide. CompView is another interactive visualization tool, which is being developed for identifying and visualizing conserved syntenic blocks (homologous chromosome segments and gene homologs) across multiple related genomes simultaneously (Fig. 2). It is designed to allow users to switch between CompView and MapView and to start with a gene or region of interest in searching for related information and will provide users with timely genomic information across species beyond genera and families.

The BGI-RIS query interface is the entry point for searching the major data types housed in BGI-RIS. It provides two kinds of searches for users: a keyword-based subject search and a BLAST-based homology search, including identification numbers of contigs/scaffolds, physical/genetic markers, genes, repeats, cDNAs and BAC ends as well as their DNA sequences.

SYSTEM DESIGN AND IMPLEMENTATION

BGI-RIS has three hardware components: a world wide web server, a database server and a sequence analysis/homology search server. Its back end is an Oracle9i relational database, and the front end consists of a set of JSP scripts running on a

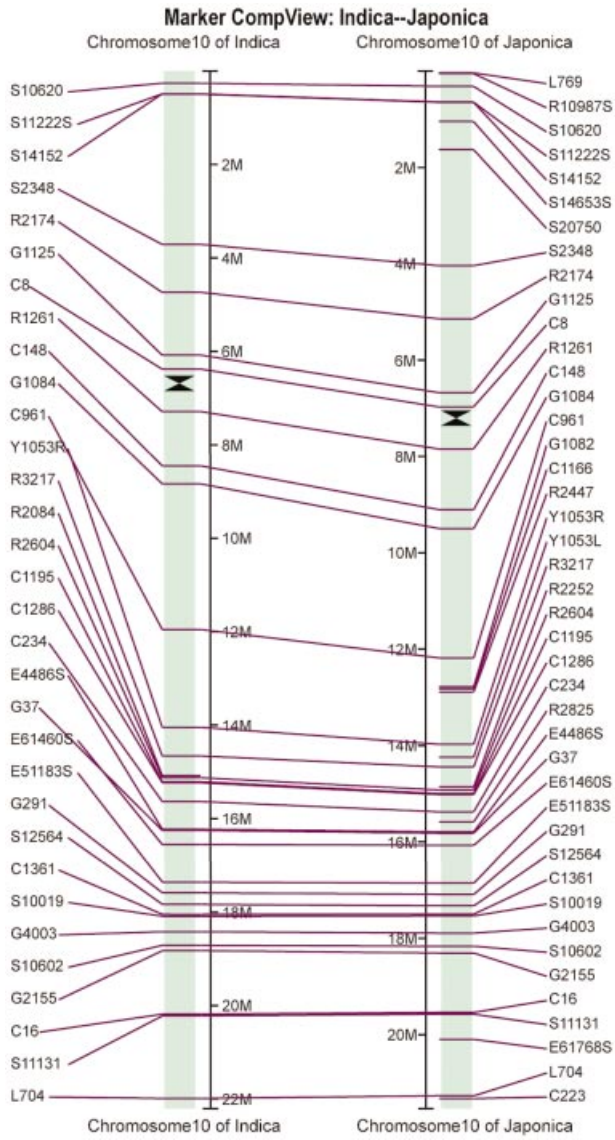

Figure 2. An example of the CompView features. Rice genetic markers are identified on rice chromosome 10 of *indica* (*93-11*) and *japonica* (Nipponbare).

TomCat web server. The search engine and visualization system are both based on the Model–View–Controller (MVC) system. This system consists of a Model where the business logic resides, a View that is generated by JSP pages and a Controller that is a servlet or a collection of servlets to provide centralized process handling. In our BGI-RIS, the Model is EJB. XML, as an effective data storage and exchange format for biological data, is gaining more attention due to its self-defining, uniform, easily parseable and transmissible format. To handle the large amount of even more complex rice genome data, we developed our own standard set of genome-based Bio-XML format that lays the foundation for our research work and allows BGI-RIS to accommodate the fast-accumulating data and to integrate new data types when encountered.

FUTURE DEVELOPMENTS

We are aiming to build a genomic information resource and comparative analysis workbench for rice with the intention to expand to other cereal crops and model plants. Refinement of the system and addition of new applications are continuous efforts for the BGI-RIS project. A key enhancement to comparative study (CompView) will be the replacement of static homology relationships with a dynamic mapping tool, allowing users to evaluate the alignment of conserved regions with alternative views of genome evolution (18). With the accomplishment of high-quality sequence assembly for both *indica* (*93-11*) and *japonica* (Nipponbare), comparative analysis between rice and *A.thaliana* becomes feasible. We also have an up-to-date collection of EST sequences from maize, wheat, barley and many other cereal crops, which enables us to align as many genes as possible to our rice genomic sequences for biologists to study the function and evolution of genes and gene families among the cereal crops (19). Functional annotations for the cereal crops based on the gene ontology (GO) tree, InterPro and other protein domain-based analysis tools, are other planned additions.

To enrich and extend BGI-RIS, we will develop several secondary databases, such as the repeat databases for the grass family of plants (http://www.girinst.org/index.html, http://www.tigr.org/tdb/e2k1/osa1/blastsearch.shtml). Repetitive DNA is a major component of eukaryotic genomes and contributes to increased genome size, gene regulation and genome evolution (19). Identifying active transposons, reconstructing transposon-like elements and understanding their origin, evolution and impact on the host genome and gene functions are of fundamental importance to genome studies (20).

ACKNOWLEDGEMENTS

Most of this work was jointly sponsored by Chinese Academy of Sciences, Commission for Economy Planning, Ministry of Science and Technology and National Natural Science Foundation of China.

REFERENCES

1. Yu,J., Hu,S.N., Wang,J., Gane,K.W., Li,S.G., Liu,B., Deng,Y.J., Li,D., Zhou,Y., Zhang X. Q. *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. Chen,M., SanMiguel,P., De Oliveira,A.C., Woo,S.-S., Zhang,H., Wing,R.A. and Bennetzen,J.L. (1997) Microcolinearity in sh2-homologous regions of the maize, rice and sorghum genomes. *Proc. Natl Acad. Sci. USA*, **94**, 3431–3435.
4. Bevan,M. and Murphy,G. (1999) The small, the large and the wild—the value of comparison in plant genomics. *Plant Genomics*, **15**, 211–215.
5. McCouch,S. (1998) Toward a plant genomics initiative: thought on the value of cross-species and cross-genera comparisons in the grasses. *Proc. Natl Acad. Sci. USA*, **95**, 1983–1985.
6. Kikuchi,S., Satoh,K., Nagata,T., Kawagashira,N., Doi,K., Kishimoto,N., Yazaki,J., Ishikawa,M., Yamada,H., Ooka,H. *et al.* (2003) Collection, mapping and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376–379.
7. Burge,Ch. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

8. Burset,M. and Guigo,R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34**, 353–367.
9. Fickett,J.W. (1996) Finding genes by computer: the state of the art. *Trends Genet.*, **12**, 316–320.
10. Bellman,R. (1957) *Dynamic Programming*. Princeton University Press, Princeton, NJ.
11. Bellman,R. and Dreyfus,S.E. (1962) *Applied Dynamic Programming*. Princeton University Press, Princeton, NJ.
12. Krogh,A., Mian,I.S. and Haussler,D. (1994) A hidden Markov model that finds genes in *E. coli* DNA. *Nucleic Acids Res.*, **22**, 4768–4778.
13. Krogh,A., Brown,M., Mian,I.S., Sjolander,K. and Haussler,D. (1994) Hidden Markov Models in computational biology applications to protein modeling. *J. Mol. Biol.*, **235**, 1501–1531.
14. Rabiner,L.R. and Juang,B.H. (1986) An introduction to Hidden Markov Models. *IEEE ASSP Mag.*, **3**, 4–16.
15. Ewing,B. and Green,P. (1998) Basecalling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.*, **8**, 186–194.
16. Ware,D., Jaiswal,P., Ni,J., Pan,X., Chang,K., Clark,K., Teytelman,L., Schmidt,S., Zhao,W., Cartinhour,S. *et al.* (2002) Gramene: a resource for comparative grass genomics. *Nucleic Acids Res.*, **30**, 103–105.
17. Hubbard,T. Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
18. Tatusov,R.L., Natale,D.A., Garkavtsev,I.V., Tatusova,T.A., Shankavaram,U.T., Rao,B.S., Kiryutin,B., Galperin,M.Y., Fedorova,N.D. and Koonin,E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.
19. Kellogg,E.A. (1999) Relationships of cereal crops and other grasses. *Proc. Natl Acad. Sci. USA*, **95**, 2005–2010.
20. Jurka,J. (1998) Repeats in genomic DNA: mining and meaning. *Curr. Opin. Struct. Biol.*, **8**, 333–337.