A draft sequence of the rice (*Oryza sativa* ssp. *indica*) genome

YU Jun[1,2,*], HU Songnian[1,*], WANG Jun[1,2,6,*], LI Songgang[1,6], WONG Ka-Shu Gane[1,2], LIU Bin[1], DENG Yajun[1,10], DAI Li[1], ZHOU Yan[2], ZHANG Xiuqing[1,3], CAO Mengliang[4], LIU Jing[2], SUN Jiandong[1], TANG Jiabin[1,3], CHEN Yanjiong[1,10], HUANG Xiaobing[1], LIN Wei[2], YE Chen[1], TONG Wei[1], CONG Lijuan[1], GENG Jianing[1], HAN Yujun[1], LI Lin[1], LI Wei[1,5], HU Guangqiang[1], HUANG Xiangang[1], LI Wenjie[1], LI Jian[1], LIU Zhanwei[1], LI Long[1], LIU Jianping[1], QI Qiuhui[1], LIU Jinsong[1], LI Li[1], WANG Xuegang[1], LU Hong[1], WU Tingting[1], ZHU Miao[1], NI Peixiang[1], HAN Hua[1], DONG Wei[1,3], REN Xiaoyu[1], FENG Xiaoli[1,3], CUI Peng[1], LI Xianran[1], WANG Hao[1], XU Xin[1], ZHAI Wenxue[3], XU Zhao[1], ZHANG Jinsong[3], HE Sijie[3], ZHANG Jianguo[1], XU Jichen[3], ZHANG Kunlin[1,6], ZHENG Xianwu[3], DONG Jianhai[2], ZENG Wanyong[3], TAO Lin[2], CHEN Xuewei[3], HE Jun[2], LIU Daofeng[3], TIAN Wei[2,10], TIAN Chaoguang[1], XIA Hongai[1], LI Gang[1], GAO Hui[1], LI Ping[3], CHEN Wei[1], WANG Xudong[3], ZHANG Yong[1,6], HU Jianfei[1,6], WANG Jing[1,6], LIU Song[1], YANG Jian[1], ZHANG Guangyu[1], XIONG Yuqing[1], LI Zhijie, MAO Long[3], ZHOU Chengshu[4], ZHU Zhen[3], CHEN Runsheng[1,5], HAO Bailin[2,7], ZHENG Weimou[1,7], CHEN Shouyi[3], GUO Wei[8], LI Guojie[9], LIU Siqi[1,2], HUANG Guyang[1,2], TAO Ming[1,2], WANG Jian[1,2], ZHU Lihuang[3,#], YUAN Longping[4,#] & YANG Huanming[1,2,3,#]

1. Beijing Genomics Institute/Center of Genomics & Bioinformatics, Chinese Academy of Sciences, Beijing 101300, China;
2. Hangzhou Genomics Institute/Institute of Bioinformatics of Zhejiang University/Key Laboratory of Bioinformatics of Zhejiang Province, Hangzhou 310007, China;
3. Institute of Genetics, Chinese Academy of Sciences, Beijing 100101, China;
4. National Hybrid Rice R & D Center, Changsha 410125, China;
5. Laboratory of Bioinformatics, Institute of Biophysics, Chinese Academy of Sciences, Beijing 100101, China;
6. College of Life Sciences, Peking University, Beijing 100871, China;
7. Institute of Theoretical Physics, Chinese Academy of Sciences, Beijing 100080, China;
8. Digital China Ltd., Beijing 100080, China;
9. Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;
10. Medical College, Xi'an Jiaotong University, Xi'an 710061, China

* These authors contributed equally to this work.

\# Corresponding author.

Correspondence should be addressed to Yang Huanming (e-mail: hmyang@genetics.ac.cn)

Abstract **The sequence of the rice genome holds fundamental information for its biology, including physiology, genetics, development, and evolution, as well as information on many beneficial phenotypes of economic significance. Using a "whole genome shotgun" approach, we have produced a draft rice genome sequence of *Oryza sativa* ssp. *indica*, the major crop rice subspecies in China and many other regions of Asia. The draft genome sequence is constructed from over 4.3 million successful sequencing traces with an accumulative total length of 2214.9 Mb. The initial assembly of the non-redundant sequences reached 409.76 Mb in length, based on 3.30 million successful sequencing traces with a total length of 1797.4 Mb from an *indica* variant cultivar *93-11*, giving an estimated coverage of 95.29% of the rice genome with an average base accuracy of higher than 99%. The coverage of the draft sequence, the randomness of the sequence distribution, and the consistency of BIG-ASSEMBLER, a custom-designed software package used for the initial assembly, were verified rigorously by comparisons against finished BAC clone sequences from both *indica* and *japanica* strains, available from the public databases. Over all, 96.3% of full-length cDNAs, 96.4% of STS, STR, RFLP markers, 94.0% of ESTs and 94.9% unigene clusters were identified from the draft sequence. Our preliminary analysis on the data set shows that our rice draft sequence is consistent with the comman standard accepted by the genome sequencing community. The unconditional release of the draft to the public also undoubtedly provides a fundamental resource to the international scientific communities to facilitate genomic and genetic studies on rice biology.**

Keywords: rice, genome, draft sequences.

Rice is one of the most important crops in the world and it provides the main resource of energy for more than half of the world population[1]. The estimated physical size of the rice genome is about 430 Mb[2] , the smallest among all the cereal crops. It corresponds to one seventh of the human genome whose working draft has been established[3, 4]. It is also 3.5 times the size of *Arabidopsis*[5, 6]. The well-established protocols for relatively high-efficiency genetic transformation, the genetic and physical maps of high density[7—9], and the high degree of synteny among genes in cereal genomes[10], all make rice an ideal model organism for studies on physiology, developmental biology, molecular genetics, evolution, and genomics of plants, especially of the grass family. Essential biological information from the rice genome will undoubtedly improve our understanding of the basic genomics and genetics of other related and economically significant crops, not only wheat, corn, sorghum, and members of the grass family, but also dicot crops such as soybean and cotton.

The initiation of the Human Genome Project (HGP) at the beginning of the l990s and the completion of the human genome working drafts at the beginning of this century have not only laid the ground work for genomics and opened a new era for the life science research, but also have set up an unprecedented example for genomics studies on many other organisms. HGP has developed strategies, technologies, definitions and standards for different stages of sequence assembly and analysis such as "working draft", "draft sequences" and "complete map", which are broadly applicable to other organisms[11,12].

Inspired by the Human Genome Project, the International Rice Genome Project Consortium, headed by Japan, has released 174.4 Mb of BAC/PAC-based non-redundant sequences since 1997, including the complete sequence of a single chromosome (Chr. 1)[13]. Monsanto and Syngenta, two private companies, have announced the establishment of a "working draft"[14, 15], independently, in April of 2000 and February of 2001, respectively, but neither has made their sequence data completely available to the public. All of the three projects mentioned above have used subspecies *japanica* (Nipponbare) as target materials, in spite of the fact that another subspecies, *indica*, is dominantly planted in Asia and other regions in the world, and has provided the unique template for the unique hybrid rice strain that has greatly contributed to solving the food supply problem in China[16].

Here we report a draft assembly of the genome of *93-11*, a cultivar of *Oryza sativa* ssp. *indica*, the major food crop in China. The contigs and draft sequences are being made freely available to the public, in order to provide important information for the understanding of the rice genome and its genes at molecular levels. These data will lay the foundation for a complete map of the rice genome, which is our ultimate goal.

1 Materials and methods

(ⅰ) Genomic DNA isolation and shotgun DNA library construction. A well-cultivated and well-documented strain of *indica*, *93-11*, which was bred in Jiangsu Academy of Agricultural Sciences[17] (Yangzhou, Jiangsu, China), was used for the construction of the reference genome sequence map. The protocol for DNA isolation was modified from Sambrook and Russell[18]. Briefly, fresh leaves at the seeding stage were ground in liquid nitrogen before complete lysis[19]. The purified genomic DNA in high molecular weight was sonicated and then sized on agarose gel for the fractions of 1.5—3.0 kb[20]. QIAEX Gel Extraction Kit (QIAGEN Inc., USA) was used to extract DNA from the gel slices. The genomic fragments were ligated to Sma I-linearized pUC18 plasmid and the ligation mixture was transformed into the DH10B competent cells through electroporation.

(ⅱ) Preparation and sequencing of the DNA templates. Single colonies were grown in the 96-deep-well plates and plasmid DNA was prepared by conventional alkaline lysis protocol[18, 21]. Quality and insert size were determined by agarose gel electrophoresis. DNA in the aliquots of 3—5 μL, corresponding to 200 ng was used for sequencing reactions according to the manufacture's instruction (Amersham Pharmacia Biotech, Beijing, China). Sequencing was carried out on MegaBACE 1000 automatic capillary sequencers (Amersham Pharmacia Biotech, Beijing, China) with parameters adjusted to having high output, i.e. 10 runs a day on average.

(ⅲ) Base calling and initial assembly. The base calling software, Phred[22, 23], and the alignment software, CrossMatch[24], were used to remove vector sequences. A custom-designed software package named BIG-ASSEMBLER (manuscript in preparation), which explicitly detects, masks repeats, and removes mitochondria contamination and chloroplast contamination, was used to assemble the draft sequence. The flow chart of the package is illustrated in fig. 1. All draft sequences longer than 1—2 kb are being submitted to two different public databases (http://www.ncbi.nlm.nih.gov/; http://www. genomics.org.cn).

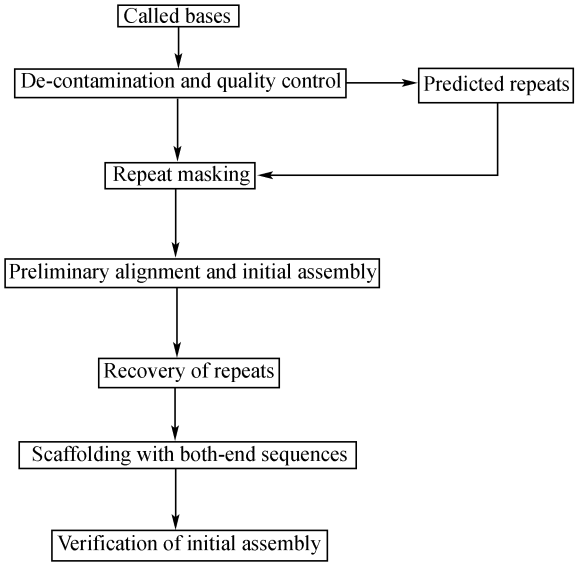

Fig. 1. A flow-chart demonstrating key procedures for initial assembly by BIG-ASSEMBLER.

2 Results and discussions

(ⅰ) Libraries and high-throughput sequencing. 55 shotgun libraries, with plasmid insert sizes ranging from 1.5 to 3.0 kb, have been used in the sequencing effort so far. 2.5 million plasmid preparations were processed and sequenced from both ends. 4.3 million successful reads have been obtained until the last data freezing on Sept. 18, 2001, giving rise to a success rate of approximately 85.30%. The average length of successful reads is 519.23 bp. The accumulated total length is 2214.9 Mb, corresponding to about 5.15 equivalents of the genome size.

(ⅱ) The initial assembly of the draft sequences. In the first assembly reported in this paper, 3.30 million reads, with high quality (Q>20) and longer than 200 bp, have been assembled. The average read-length of the sequences for the assembly is 544.34 bp. The distribution of read-length is shown in fig. 2. The cumulative sequences

with quality score greater than *Q*20 are 1797.4 Mb, or, 4.18 equivalents of the genome size. BIG-ASSEMBLER has given a total of 409.76 Mb initially assembled non-redundant sequences, indicating a coverage of 95.29% of the genome.

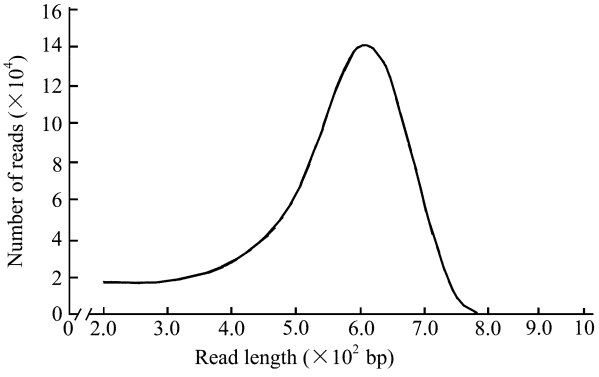

Fig. 2. Read-length distribution of the rice draft sequences.

The initial assembly yields 222263 contigs. The size distribution of the contigs is shown in fig. 3. The total number of contigs longer than 1.0 kb is 101733, with a total length of 236.389 Mb and a coverage of 54.97% of the genome. The total number of contigs longer than 3.0 kb is 23061 with a total length of 108.879 Mb which cover 25.3% of the whole genome.

BIG-ASSEMBLER is capable of making scaffolding automatically with the two-end sequences of templates. The total length of scaffolds with an accuracy of higher than 99.9% is about 202.4 Mb, giving a coverage of 47% of the whole genome. Scaffolds in size of 3.0 kb or longer are counted as 24576, with a coverage of 40.3% (173.187 Mb) of the whole genome. The total number of scaffolds longer than 5.0 kb is 14457 with a collective length of 133.589 Mb that cover 31.1% of the genome. There are 4282 scaffolds with size longer than 10 kb, which are cumulated as 62.8 Mb in length, or 14.6% of the genome. The length distribution of the scaffolds is shown in fig. 4.

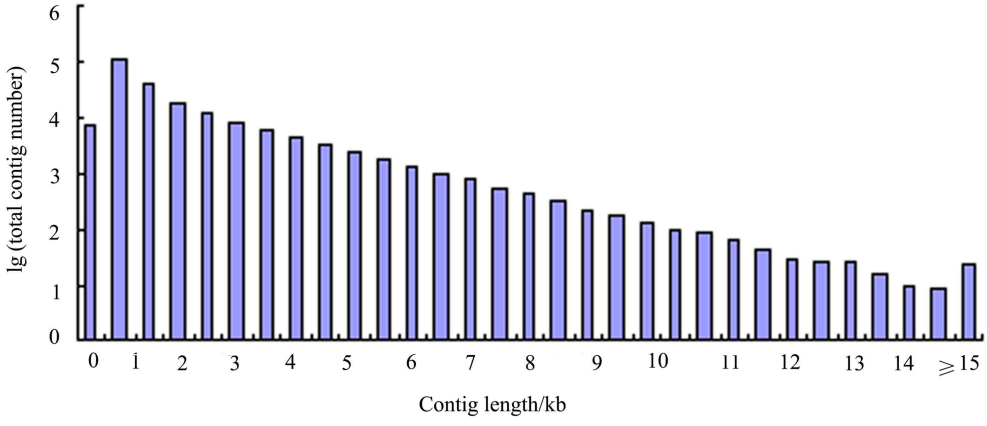

Fig. 3. Contig-length distribution of the initially assembled rice draft sequences.

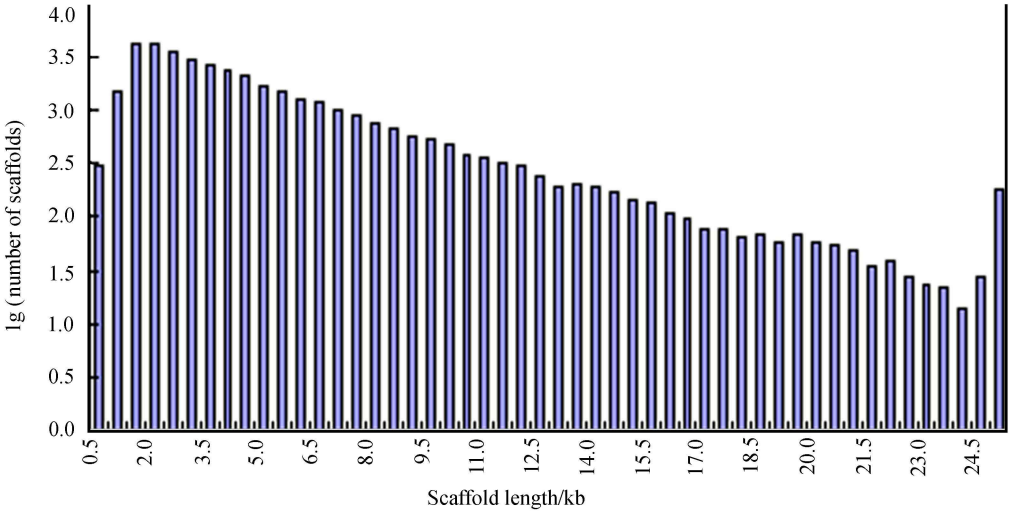

Fig. 4. Scaffold-length distribution of the initially assembled rice draft sequences.

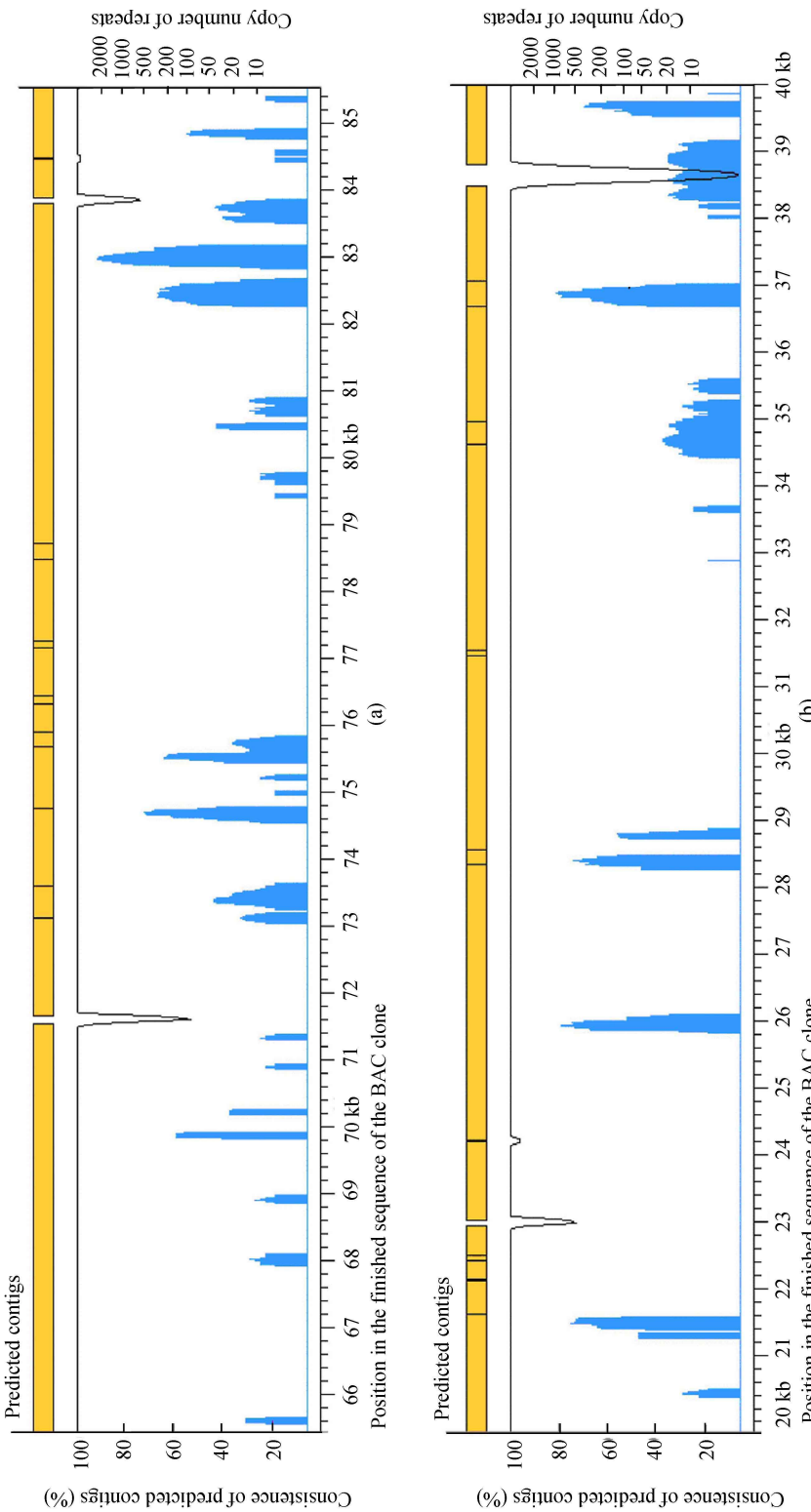

Fig. 5. Alignment of the predicted contigs with the finished sequences of the published *indica* (a) and *japanica* (b) BAC clones. Only segments of the BAC sequences, 66—85 kb of *indica* (Acc. No. AL117264) and the 20—40 kb of *japanica* (Acc. No. AP003339), are shown here as examples. The gaps between predicted contigs demonstrated in the figures could be interpreted as incompletely sequenced regions except the second gap in (a) which is caused by a large cluster of repeats.

(iii) Coverage verification of the rice draft sequence. The major benefits of using BIG-ASSEMBLER is that sequences can be assembled without reference to biologically characterized repeats, which make it powerful for genomes with high repeat-content (WANG Jun and LI Songgang, manuscript in preparation). The success of its application to this project were confirmed through simulation studies and comparisons of the initially assembled draft sequences to other types of publicly available sequences, including human, rice and corn (data not shown).

(1) Comparison to finished rice sequences from BAC clones. 650 kb finished sequences from 8 BAC clones (GenBank accession numbers: AL442007, AL442112, AL442114, AL42115, AL512542, AL512545, AL512546, AL512547) of *indica* (Strain Guang Lu Ai) were compared with the initially assembled draft sequences, after repeats longer than 5 kb (41 kb) were removed. The results indicate that about 590 kb finished sequences are aligned to the draft sequence, given a coverage of 96.82% (fig. 5(a)) with a contig-length error of 0.8%. The comparative results also suggest a random distribution and high representation of the shotgun sequences.

Another confirmation of the coverage with the finished rice sequences was performed on 788 kb sequences (BAC sequences from GenBank with accession numbers: AP002538, AP003197, AP003201, AP003339, AP003537, AP003707, AP003853) from *japanica* (cultivar Nipponbare). A coverage of 95.75% were obtained in the repeat-free region of 518 kb (devoid of repeats longer than 2 kb), despite numerous sequence variations were detected in the repeat-rich regions (fig. 5(b)).

(2) Comparison to STS, STR, RFLP markers of the rice genome. A physical map with dense STS, STR, RFLP markers of *japanica* has been established, which has relatively high representation in the rice genome. Total of 4854 STS, STR, RFLP sequences, available from all public domains of genome databases were compared to our initially assembly[25]. Among them, 4679 STS, STR, RFLP sequences were unambiguously aligned and the coverage of the working draft was estimated at 96.4%.

(3) Comparison to rice full-length cDNA, EST, and unigene sequences. 896 published full-length cDNAs from *japanica*[26], retrieved from the databases, were compared to our assembly. 863 were aligned with high confidence to our assembly and the result yielded an overall coverage of 96.3%.

Furthermore, 3155 unigene clusters were compared to our assembly in a similar way. 2956 out of the total were aligned, and the alignment yielded the coverage of 94.9%.

Finally, 10831 ESTs, taken randomly from our own EST resources, were compared to the assembled draft, after repeats were masked. 10533 ESTs out of the total were aligned (97.2%). As far as the sequence length of the ESTs used for the analysis was concerned, 4.7 Mb out of 5.0 Mb were fully aligned (94.0%).

3 Conclusions

Genome coverage is the most important parameter for a working draft proposed by the International Human Genome Sequencing Consortium. The coverage is complicated by the repeat contents that are abundant in higher eukaryotic genomes, such as human and rice. The overall coverage estimate, 95.29% of the rice genome being represented in our draft sequences, resulted from the application of our BIG-ASSEMBLER. The results were confirmed vigorously by sequence comparisons with finished sequences from rice BAC clones, STS/STR/RFLP markers, full-length cDNAs, ESTs, and unigene clusters to our rice assembly. The coverage, in addition to the total accumulated length of redundant sequences and high sequence quality, has reached the standard for a genome working draft. This initially assembled working draft will be further improved through deeper sequencing, gap-filling, map integration, gene identification and data analysis. The ultimate goal of the project is to provide a highly accurate and well-annotated rice genome to the research communities worldwide in genetic and biologic studies on rice and other cereal crops.

Acknowledgements The authors are specially indebted to the other co-authors who have contributed to this work of team efforts (http://www.genomics.org.cn) and to Amersham Pharmacia Biotech (China) Ltd., Beijing, China, and SUN Microsystems (China) Inc., and Dawning Computer Corp. for their continuous support and excellent service. This work was sponsored by the Chinese Academy of Sciences, the Commission for Economy Planning, the Ministry of Science and Technology, the National Natural Science Foundation of China, Beijing Municipal Government, Zhejiang Provincial Government, and Hangzhou Municipal Government.

References

1. Sasaki, T., Burr, B., International Rice Genome Sequencing Project: the effort to completely sequence the rice genome, Curr. Opin. Plant. Biol., 2000, 3: 138.
2. Eckardt, N. A., Sequencing the Rice Genome, The Plant Cell, 2000, 12: 2011.
3. Lander, E. S., Linton, L. M., Birren, B. et al., Initial sequencing and analysis of the human genome, Nature, 2001, 409: 860.
4. Venter, J. C., Adams, M. D., Myers, E. W. et al., The sequence of the human genome, Science, 2001, 291: 1304.
5. Bevan, M., Murphy, G., The small, the large and the wild: the value of comparison in plant genomics, Trends Genet., 1999, 15: 211.
6. The Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant Arabidopsis thaliana, Nature, 2000, 408: 796.
7. Tao, Q., Zhao, H., Qiu, L. et al., Construction of a full bacterial artificial chromosome (BAC) library of Oryza sativa genome, Cell Res., 1994, 4: 127.
8. Umehara, Y., Miyazaki, A., Tanoue, H., Construction and characterization of rice YAC library for physical mapping,, Molecular Breeding, 1995, 1: 79.
9. Wang, G. L., Holsten, T. E., Song, W. Y. et al., Construction of rice bacterial artificial chromosome library and identification of clones linked to the Xa-21 disease resistance locus, Plant. J., 1995, 7: 525.
10. Gale, M. D., Devos, K. M., Comparative genetics in the grasses, Proc. Natl. Acad. Sci. USA, 1998, 95: 1971.
11. Rowen, L., Wong, G. K., Lane, R. P. et al., Publication rights in

the era of open data release policies, Science, 2000, 289:1881.
12. http://www.ornl.gov/hgmis/research/bermuda.html#3.
13. http://www.tigr.org/tdb/e2k1/osa1/BACmapping/description.shtml
14. http://www.rice-research.org/
15. http://www.syngenta.com/
16. Yuan, L. P., Breeding of super hybrid rice for super high yield production, Hybrid Rice (in Chinese), 1997, 1: 1.
17. Dai, Z. Y., Zhao, B. H., Liu, X. J., Yangdao 6 (*93-11*), a new medium *indica* variety with fine quality, high yield and multi-disease resistance (in Chinese), Jiangsu Agricultural Sciences, 1997, 4: 13.
18. Sambrook, J., Russell, J. D., Molecular Cloning, 3rd ed., New York: Cold Spring Harbor Laboratory Press, 2001.
19. Hatano, S., Yamaguchi, J., Hirai, A., The preparation of high-molecular-weight DNA from rice and its analysis by pulsed-field gel electrophoresis, Plant Sci., 1992, 83: 55.
20. Myers, E. W., Sutton, G. G., Delcher, A. L. et al., A whole-genome assembly of Drosophila, Science, 2000, 287: 2196.
21. Birnboim, H. C., A rapid alkaline extraction method for the isolation of plasmid DNA, Methods Enzymol, 1983, 100: 243.
22. Ewing, B., Hillier, L., Wendl, M. C. et al., Base-calling of automated sequencer traces using Phred, Ⅰ. Accuracy assessment, Genome Res., 1998, 8: 175.
23. Ewing, B., Green, P., Base-calling of automated sequencer traces using Phred, Ⅱ. Accuracy assessment, Genome Res., 1998, 8: 186.
24. http://www.phrap.org/
25. Sources of STS, STR, RFLP sequences: http://ars-genome.cornell.edu/rice/quickqueries.html; http://www.ncbi.nlm.nig.gov/; http://rgp.dna.affrc.go.jp/publicdata/geneticmap2000/index.html
26. Wong, G. K., Passey, D. A., Huang, Y. et al., Is "junk" DNA mostly intron DNA? Genome Res., 2000, 10: 1672.

(Received September 19, 2001; accepted September 29, 2001)