After the initial assembly of sequence data, stretches of poor or ambiguous quality and apparent gap regions were identified for further sequencing to obtain greater than 99.99% sequence accuracy. But despite extensive efforts to improve the sequence quality and to fill the gaps, 4 of the 390 PAC/BAC clones sequenced are still at phase 1 (GenBank, http://www.ncbi.nlm.nih.gov/HTGS/) because the consensus sequence could not be ordered correctly owing to numerous repeats. The remainder comprises 16 phase 2 and 370 phase 3 clones. The nine contigs for chromosome 1 representing the non-overlapping segments of continuous sequence were conjoined by inserting into the gap regions nucleotides that were calculated on the basis of the results of FISH experiments. All of the sequence information of chromosome 1 has been submitted to the DNA Data Bank of Japan (DDBJ, http://www.ddbj.nig.ac.jp/) with the accession number BA000010 (Con Division).

Gene prediction and functional classification

We carried out gene prediction using our in-house automated gene prediction system RiceGAAS[13]. The algorithm for gene domain prediction in RiceGAAS was designed by combining several prediction programs including GENSCAN[21] for maize, GENSCAN[21] for *Arabidopsis*, RiceHMM (http://rgp.dna.affrc.go.jp/RiceHMM/index.html) and the exon-finding program MZEF (http://argon.cshl.org/genefinder/), with homology search results from BLASTN and BLASTX (http://www.ncbi.nlm.nih.gov/BLAST/). These results were merged and integrated for gene prediction. Domain search was done using InterPro (http://www.ebi.ac.uk/interpro/scan.html), and repeats were identified using RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker). The predicted proteins were used to query the nonredundant protein database using BLASTP and categorized according to functional categories defined for *Arabidopsis* by MIPS (http://mips.gsf.de/cgi-bin/proj/thal/filter_funcat.pl?all) with a threshold probability value of 10^{-20}.

Received 4 April; accepted 19 September 2002; doi:10.1038/nature01184.

1. Harushima, Y. *et al.* A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* **148,** 479–494 (1998).
2. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).
3. Yamamoto, K. & Sasaki, T. Large-scale EST sequencing in rice. *Plant Mol. Biol.* **35,** 135–144 (1997).
4. Sasaki, T. & Burr, B. International rice genome sequencing project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3,** 138–141 (2000).
5. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–91 (2002).
6. Goff, S. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
7. Cheng, Z. *et al.* Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* **14,** 1691–1704 (2002).
8. Dong, F. *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl Acad. Sci. USA* **95,** 8135–8140 (1998).
9. Richards, E. J. & Ausubel, F. M. Isolation of a higher eukaryotic telomere from *Arabidopsis thaliana*. *Cell* **53,** 127–136 (1988).
10. The Arabidopsis Genome Initiative Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–820 (2000).
11. Cheng, Z. *et al.* Toward a cytological characterization of the rice genome. *Genome Res.* **11,** 2133–2141 (2001).
12. Barakat, A., Carels, N. & Bernardi, G. The distribution of genes in the genomes of Gramineae. *Proc. Natl Acad. Sci. USA* **94,** 6857–6861 (1997).
13. Sakata, K. *et al.* RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res.* **30,** 98–102 (2002).
14. Hiratsuka, J. *et al.* The complete sequence of rice (*Oryza sativa*) chloroplast genome: Intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. *Mol. Gen. Genet.* **217,** 185–194 (1989).
15. Lowe, T. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).
16. Wessler, S. R., Bureau, T. E. & White, S. E. LTR-retrotransposons and MITEs: important players in the evolution of plant genomics. *Curr. Opin. Genet. Dev.* **5,** 814–821 (1995).
17. Sasaki, A. *et al.* A mutant gibberellin-synthesis gene in rice. *Nature* **416,** 701–702 (2002).
18. Feng, Q. *et al.* Sequence and analysis of rice chromosome 4. *Nature* **420,** 316–320 (2002).
19. Baba, T. *et al.* Construction and characterization of rice genome libraries: PAC library of *japonica* variety, Nipponbare, and BAC library of *indica* variety, Kasalath. *Bull. Natl. Inst. Agrobiol. Resour. (Japan)* **14,** 41–52 (2000).
20. Barry, G. The use of the Monsanto draft rice genome sequence in research. *Plant Physiol.* **125,** 1164–1165 (2001).
21. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com/nature).

Acknowledgements We thank Monsanto for the BAC contig information, BAC clones and their sequence data; R. Wing of Clemson University Genomics Institute and Novartis for the rice Nipponbare BAC library and its fingerprint data, respectively; M. Hattori for technical assistance; B. Burr and F. Burr for critically reading the manuscript; T. Slezak for comments; and K. Eguchi for encouragement.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to T. Sasaki (e-mail: tsasaki@nias.affrc.go.jp).

Sequence and analysis of rice chromosome 4

Qi Feng*†, Yujun Zhang*†, Pei Hao*†, Shengyue Wang†‡, Gang Fu‡, Yucheng Huang*, Ying Li*, Jingjie Zhu*, Yilei Liu*, Xin Hu*, Peixin Jia*, Yu Zhang*, Qiang Zhao*, Kai Ying*, Shuliang Yu*, Yesheng Tang*, Qijun Weng*, Lei Zhang*, Ying Lu*, Jie Mu*, Yiqi Lu*, Lei S. Zhang*, Zhen Yu*, Danlin Fan*, Xiaohui Liu*, Tingting Lu*, Can Li*, Yongrui Wu*, Tongguo Sun*, Haiyan Lei*, Tao Li*, Hao Hu*, Jianping Guan*, Mei Wu*, Runquan Zhang*, Bo Zhou*, Zehua Chen*, Ling Chen*, Zhaoqing Jin*, Rong Wang*, Haifeng Yin‡, Zhen Cai‡, Shuangxi Ren‡, Gang Lv‡, Wenyi Gu‡, Genfeng Zhu‡, Yuefeng Tu‡, Jia Jia‡, Yi Zhang‡, Jie Chen‡, Hui Kang‡, Xiaoyun Chen‡, Chunyan Shao‡, Yun Sun‡, Qiuping Hu‡, Xianglin Zhang‡, Wei Zhang‡, Lijun Wang‡, Chunwei Ding‡, Haihui Sheng‡, Jingli Gu‡, Shuting Chen‡, Lin Ni‡, Fenghua Zhu‡, Wei Chen§, Lefu Lan§, Ying Lai§, Zhukuan Cheng‖¶, Minghong Gu‖, Jiming Jiang¶, Jiayang Li§, Guofan Hong*, Yongbiao Xue§ & Bin Han*

** National Center for Gene Research, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 500 Caobao Road, Shanghai 200233, China*
‡ Chinese National Human Genome Center at Shanghai, 351 Guo Shoujing Road, Shanghai 201203, China
§ Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Datun Road, Andingmenwai, Beijing 100101, China
‖ Yangzhou University, 27 Wenhua Road, Yangzhou, Jiangsu 225009, China
¶ Department of Horticulture, University of Wisconsin, Madison, Wisconsin 53706, USA
† These authors contributed equally to this work

Rice is the principal food for over half of the population of the world. With its genome size of 430 megabase pairs (Mb), the cultivated rice species *Oryza sativa* is a model plant for genome research[1]. Here we report the sequence analysis of chromosome 4 of *O. sativa*, one of the first two rice chromosomes to be sequenced completely[2]. The finished sequence spans 34.6 Mb and represents 97.3% of the chromosome. In addition, we report the longest known sequence for a plant centromere, a completely sequenced contig of 1.16 Mb corresponding to the centromeric region of chromosome 4. We predict 4,658 protein coding genes and 70 transfer RNA genes. A total of 1,681 predicted genes match available unique rice expressed sequence tags. Transposable elements have a pronounced bias towards the euchromatic regions, indicating a close correlation of their distributions to genes along the chromosome. Comparative genome analysis between cultivated rice subspecies shows that there is an overall syntenic relationship between the chromosomes and divergence at the level of single-nucleotide polymorphisms and insertions and deletions. By contrast, there is little conservation in gene order between rice and *Arabidopsis*.

The rice genome has been well mapped both genetically and physically[3–5] and has a syntenic relationship with other cereals[6]. *Arabidopsis thaliana* (Arabidopsis), a member of the Brassica family of dicotyledonous (dicot) plants, has become an important model flowering plant for studying many aspects of plant biology[7]. The completion of the Arabidopsis genome[8–10] has afforded an unprecedented opportunity for systematic studies of plant gene function. Equally, the complete rice genome sequence will provide a catalogue of genes that may be important for improving not only rice but also other cereals, as functionally important sequences are conserved and may be identified by their similarity[11]. The International Rice Genome Sequencing Project (IRGSP) has adopted the clone-by-clone approach for obtaining a finished rice genome sequence, because it is modular, allows efficient gap filling, avoids problems arising from distant repetitive sequences and results in the early completion of larger contiguous segments of a genome. Here we describe the completed sequence of rice *O. sativa* ssp. *japonica* cv.

Nipponbare chromosome 4 to an accuracy of greater than 99.99%, together with its analysis.

We constructed a comprehensive clone-based physical map of the Nipponbare chromosome 4 through an integrated approach (ref. 12 and Fig. 1). We identified a tiling path that consisted of 287 bacterial artificial chromosomes (BACs) and two phage (P1)-derived artificial chromosomes (PACs; Supplementary Fig. 1). Each clone was sequenced by a random shotgun approach on both strands of small-insert subclones to achieve a tenfold coverage. Sequence assemblies were verified by comparing *Not*I and *Hin*dIII restriction profiles predicted from the sequence against experimentally determined restriction profiles.

The collinearity of assembled sequences with the chromosome was confirmed further by integrating 176 sequenced genetic and 402 expressed sequence tag (EST) markers (Supplementary Table 1). Sequence overlaps of adjacent clones were used to assess potential sequence differences between them, which resolved the sequence with less than 1 mismatch per 14,000 base pairs (bp); these mismatches were corrected further by manual checking or by additional sequencing reactions. We estimate the overall accuracy of our finished sequence to exceed 99.99%.

In total, we finished 34,689,786 bp of sequence in eight sequence contigs (Fig. 1). The pachytene spread fluorescence *in situ* hybridization (FISH) and fibre FISH analyses indicted that the sizes of gaps 1–7 are about 125.91 kb, 112.84 kb, 47.82 kb, 110.89 kb, 73.57 kb, 101.06 kb and 301.76 kb, respectively, which leaves 885.49 kb of unsequenced regions (Supplementary Fig. 2 and Supplementary Table 2). The total length of chromosome 4 was therefore calculated to be 35.6 Mb, which is 1.3 Mb shorter than previous estimates of 36.8 Mb (refs 6, 13). Thus, 97.3% (34.6/35.6 Mb) of the chromosome including the centromeric region has been sequenced and represents a practically finished chromosome 4.

The distribution of various repetitive DNA elements along chromosome 4 is shown in Fig. 2. In addition, details of the content of repetitive sequences in rice chromosome 4 are given in Supplementary Table 3. Overall, the repeats are mainly located in the heterochromatic regions including the centromere. But the miniature inverted-repeats transposable elements (MITEs) have a pronounced bias towards the euchromatic regions along the long arm south (Fig. 2); these are the most frequently occurring class of repetitive elements and, with a total of 7,666, account for 49% of the total number of repetitive DNA elements. No MITEs have been found in exons and they are rare in introns. Other types of repeat seem to be dispersed in the rest of the chromosome, with no obvious clustering. This is similar to the Arabidopsis genome[14], but in contrast to what has been observed in maize, where nested retroelements are the main component of the intergenic regions. Overall, according to an estimation based on current information in the rice RepeatDatabase, the total length of the repetitive sequence is 6.3 Mb, accounting for 18.2% of chromosome 4 excluding the unsequenced regions. The G + C content of chromosome 4 is 44.16%. The heterochromatic region shows a very low recombination frequency with a ratio of physical distance to genetic distance of 1.57 cM Mb^{-1}; however, the euchromatic region has a value of 4.79 cM Mb^{-1}.

We used gene prediction programs and homology searches against all available DNA and protein sequences, in particular rice

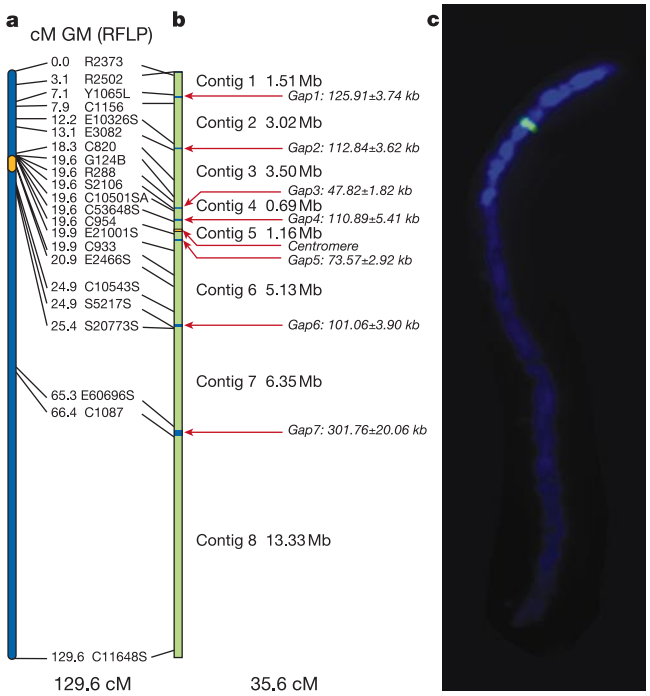

Figure 1 Maps of rice chromosome 4. **a**, Genetic map showing the positions of the centromere (yellow) and genetic markers. **b**, Physical map showing the locations of eight sequenced contigs (green) numbered from the top, seven remaining gaps (blue) indicated by red arrows, and the centromere (yellow) on the chromosome. The sizes of the unsequenced gaps measured by fibre FISH analysis are given in kilobases. The whole chromosome is about 35.53 Mb. **c**, A 4′,6-diamidino-2-phenylindole dihydrochloride (DAPI)-stained pachytene spread of rice chromosome 4 showing heterochromatic (bright blue) and euchromatic (light blue) regions. The centromere (green) was detected by FISH with a centromere-specific probe (pRCS2 clone).

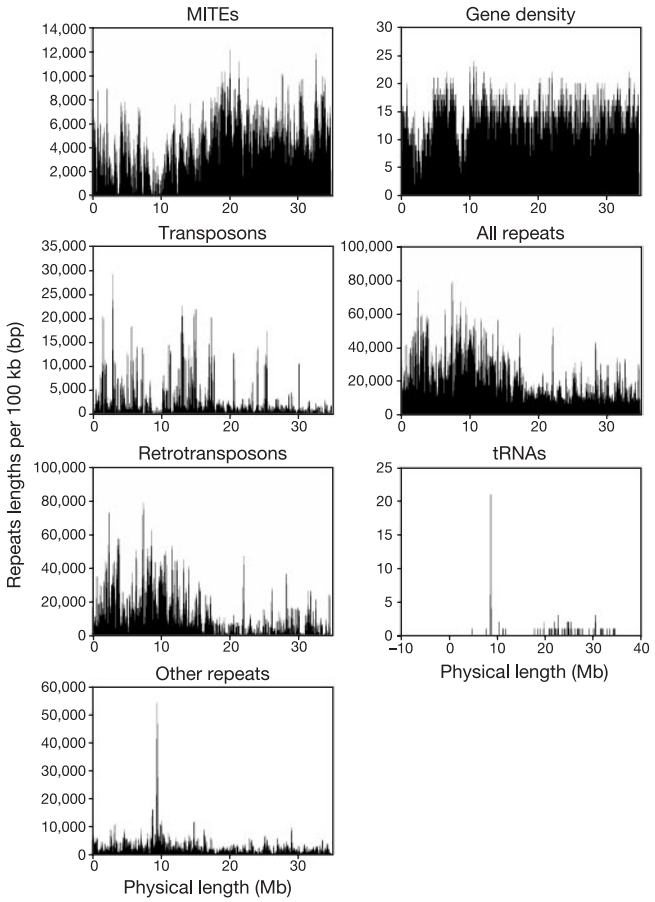

Figure 2 Distribution of various repeats and features along chromosome 4. The chromosome is represented on the *x* axis and is numbered from base pair 1 of the north telomere to base pair 34,689,786 bp of the south telomere by joining up the physical gaps. The *y* axis shows the various repeats plotted relative to the nucleotide position on the chromosome. The densities of the repeats and tRNA were obtained by analysing the sequence every 100 kb using a 10-kb sliding window.

Table 1 **General structural features of rice chromosome 4**

Statistics of annotated chromosome 4	
Total genes	4,658
Average gene size (bp)	2,779
Gene density (bp per gene)	7,441
Average exon per gene	4.4
Average exon size (bp)	340
Average intron size (bp)	376
Average BAC GC (%)	44.16
Average gene GC (%)	47.22
Average intergenic GC (%)	42.30
Average exon GC (%)	53.0
Average intron GC (%)	39.0

EST sequences including 115,000 reads from *indica* Guangluai 4 generated by ourselves, to annotate the rice genomic sequences by *ab initio* methods. This analysis relied much more heavily on the results of a manual curation. There are a total of 4,658 coding genes predicted on chromosome 4. The average gene density is 7,441 bp per gene, which is lower than in Arabidopsis (4.5 kb per gene). We estimate that the total number of rice genes is about 57,000, with the assumption that the rest of the genome has a similar gene density to that of chromosome 4. On average, each gene contains 4.4 exons and is about 2,779 bp. The largest gene contains 32 exons and encodes a protein of 1,483 amino acids with 69% similarity to the polyadenylation protein of Arabidopsis[15]. Seventy tRNA and four small nucleolar RNA genes were found. Twenty-four of the tRNA genes are chloroplast-derived tRNA genes located in a cluster near centromere (Supplementary Table 4). We identified 1,681 genes to have unique rice EST matches. As only a limited number of ESTs were available, we classified 2,073 of the predicted genes as hypothetical genes. The remaining 2,585 predicted genes have been classified as 'known', 'novel' or 'putative' because they have EST matches, protein matches or any detectable domain signatures. The gene map of chromosome 4 is shown in Supplementary Fig. 1 together with its other features. The statistics of the annotated chromosome 4 and its general features are given in Table 1.

To classify the function of the predicted genes, we used BLASTP (basic local alignment search tool) to search for homologous proteins in protein databases, as well as InterPro to search for domain signatures. It was difficult to assign a function to most of the predicted coding proteins, and these are therefore classified as 'others'. Clearly, the proteins involved in cellular metabolism (555 proteins, 11.92%), transcription (259, 5.56%) and signal transduction (256, 5.51%) typically represent a high proportion of the protein set (Table 2). Notably, 161 predicted proteins containing a disease resistance protein signature and other associated domains were identified to be involved in plant defense. Among these, a bacterial blight-resistance gene *Xa1* cluster consisting of six members was identified. As it has been reported that *Xa1* gene is a single-copy gene in the rice genome[16], it will be interesting to test whether each member of the *Xa1* gene cluster has a role in recognizing different races of the rice bacterial pathogen *Xanthomonas oryzae* pv. *oryzae*. In total, 1,004 genes are in gene families with more than two members on the chromosome, indicating a high proportion (at least 21.5%) of gene families in the rice genome (Supplementary Fig. 3).

The sequences of ten minimally tiled BACs of the 1.16-Mb contig derived from the centromeric region have been determined completely (Supplementary Fig. 4). Sequences in a BAC clone OSJNBb0062N22 located 0.4 Mb from the top end of the contig have a significant similarity to a satellite repeat (pRCS2) associated with rice centromeres[17]. The physical structure of the core centromeric sequence consists of 341 copies of pRCS2-like repeats with 80–100% similarity. There are 15 genes predicted in this region. Among them, ten seem to be related to retroelements and five are hypothetical. Thirteen have unique rice EST matches. Because the centromere is apparently not subject to recombination, these genes will be useful to investigate chromosome evolution at the molecular level. The dispersed repeats consist predominantly of long terminal repeats and simple sequence repeats, which are all found frequently in the heterochromatin (Supplementary Fig. 5). The boundaries of the centromere could be defined further by cytogenetic and functional analysis. Nevertheless, the centromere-related sequence of chromosome 4 provides both a structural basis for studying centromere function in plants and an opportunity to construct rice or plant artificial chromosomes that can be potentially used in the cloning and transformation of large DNA fragments.

The cultivated rice species *O. sativa* is classified into two main subspecies, *indica* and *japonica*[18], that represent most of the rice crops grown in the world[19]. We carried out sequence alignments between 2.3 Mb of three contiguous segments of chromosome 4 from *indica* Guangluai 4 and its collinear sequences from *japonica* Nipponbare, which revealed extensive sequence collinearity including gene

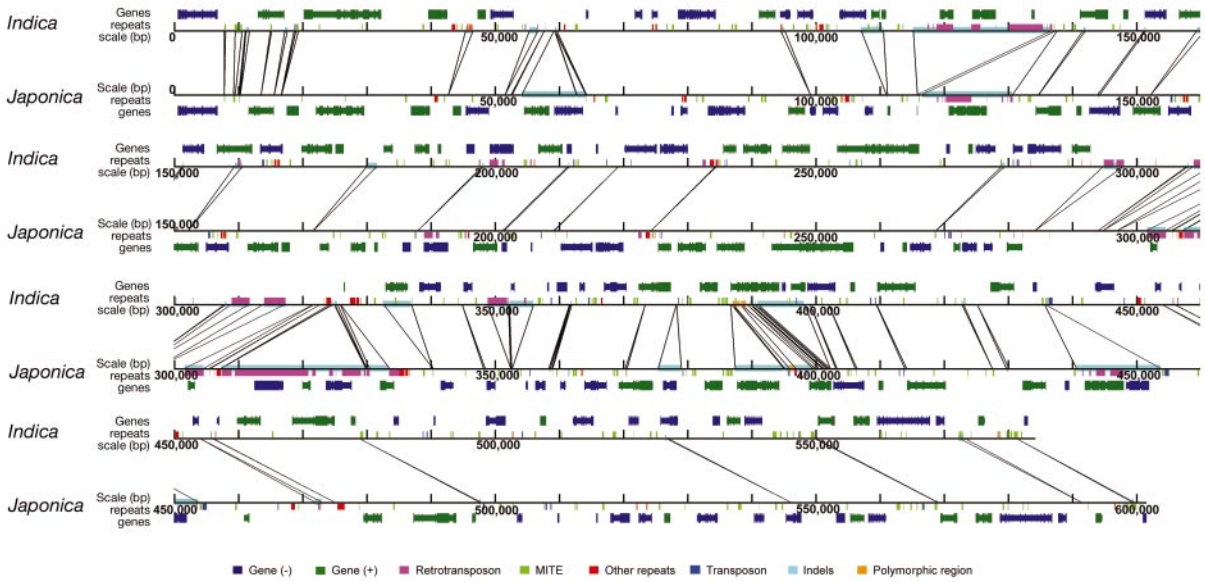

Figure 3 Comparative sequence analyses between rice subspecies *indica* GLA4 and *japonica* Nipponbare. The *indica* (top) and *japonica* (bottom) sequences were aligned over contiguous segments of 583 kb (*indica*) and 602 kb (*japonica*). Insertions and deletions are indicated in light blue. Genes on the two strands are indicated in green and blue, respectively. Polymorphic regions are indicated in orange, and other types of repeats are indicated in different colours.

Table 2 **Functional classification of rice chromosome 4**

Cellular metabolism	555	11.92%
Transcription	259	5.56%
Signalling	256	5.51%
Transport	198	4.14%
Plant defence	161	3.47%
Protein fate	85	1.84%
Growth	81	1.76%
Other	3,063	65.76%
Total	4,658	

contents and orders between them. But deviations from the collinearity are frequent with insertions or deletions (Indels; Fig. 3). Chromosome 4 of *japonica* is probably larger than that of *indica* because of an expansion by insertions of transposable elements. If this were also true for other chromosomes, it would mean that the *indica* genome is smaller than that of the *japonica*. The insertions occur not only in the intergenic regions, but also in some of the coding regions. We identified 9,056 single-nucleotide polymorphisms (SNPs) in 2.3 Mb of sequence, indicating an average frequency of 1 SNP per 268 bp. There are 63 Indels for the *indica* sequence and 138 Indels for the *japonica* sequence. The results are summarized in Supplementary Table 5. These findings may reflect an overall conservation of the two genomes. Because the *indica* and *japonica* rice species had a polyphyletic origin from a common wild annual rice species *O. nivara*[19], some of the differences identified from the two chromosomes may simply have resulted from the differences of the two varieties used. We need more genomic data from other rice varieties of *indica* and *japonica* to identify the evolutionary relationships of the cultivated rice species.

The working drafts of the rice *indica* variety 93-11 genome and the *japonica* Nipponbare have been completed[20,21]. We identified 7,423 contigs with a length of 26,142,459 bp from a total of 127,550 contigs of the 93-11 draft sequences on chromosome 4 by using a similarity of >90% as a match cutoff. These contigs matched 7,551 regions on the *japonica* Nipponbare chromosome 4, indicating 128 possible insertions on this chromosome, and covered 75.42% of the chromosome (Supplementary Fig. 6). We estimated that about 2,225 predicted genes (47.77%) are present in full in the 93-11 draft sequences, 1,788 genes (38.38%) are partially identified, and 645 genes (13.85%) are not found from the located draft sequences. All of these identified free contigs of the 93-11 draft sequences might serve as a scaffold for comparative genome analysis of *indica* and *japonica*.

Both rice and Arabidopsis are model angiosperm plants, but they are distantly related species. We compared the proteins predicted on rice chromosome 4 with a complete set of proteins of *Arabidopsis thaliana* using the dps program (dna against protein database search) of the AAT package with stringency value of parameter $f > 200$ (ref. 22). We identified 2,040 Arabidopsis proteins with significant homology in amino acid sequence. The distribution of rice genes that encode protein products related to the Arabidopsis proteins has a pronounced bias toward the euchromatic region of rice chromosome 4 (Supplementary Fig. 7). It is evident that the 130–200 million years that separate the two main angiosperm groups[23] have eroded conservation of gene order to the point where the comparative approach is no longer a useful tool with predictive power, consistent with previous observations that suggest that collinearity between dicot and monocot species seems to be limited to a few very small regions. The remaining 2,618 proteins that have no significant homology to Arabidopsis proteins may represent rice- or monocot-specific proteins. In this set, 14 protein-coding genes matched maize unique ESTs. Two of these proteins have known functions in maize (a phytase for hydrolysis of phytin[24]) and in coconut (a lysophosphatidic acid acyltransferase for accepting medium-chain-length substrates[25]), respectively. Of 1,794 maize marker sequences, 114 markers were mapped on the rice chromosome 4 (Supplementary Fig. 8). Macrocollinearities between rice chromosome 4 and maize chromosome 2 and 10 are observed, but microcollinearity between the two genomes was not apparent.

The sequence of rice chromosome 4 provides us with further insight into the structure and organization of eukaryotic chromosomes. Functional analysis of the 65% of genes on rice chromosome 4 that are predicted but have unknown function clearly represents a considerable challenge. The overall breakdown of genetic synteny between monocots and dicots, seen through the comparison of rice chromosome 4 and Arabidopsis, suggests that extensive genome reshuffling has occurred since their divergence. The comparison between rice subspecies, as well as varieties, shows that the principal sequence differences include SNPs and Indels, which might be useful for dissecting the molecular basis of the phenotypic differences that are often associated with different subspecific or varietal lines. □

Methods

Chromosome sequencing

Four genomic libraries made from the rice *O. sativa* ssp. *japonica* cv. Nipponbare were used to sequence chromosome 4: two BAC libraries of OSJNBa and OSJNBb were provided by the Clemson University Genomics Institute (CUGI); a PAC library was provided by the Rice Genome Project (RGP); and some BACs were provided by Monsanto. We purified the selected BACs and PACs by a caesium chloride gradient. For the shotgun approach, sheared BAC or PAC DNA (3 kb) was ligated into a pBluescript vector and transformed into *Escherichia coli* strain DH5α. The shotgun subclones with inserts of 3 kb were sequenced from both ends by the dideoxy chain termination method by using either BigDye Terminator Cycle Sequencing V2.0 Ready Reaction (Applied Biosystems) or Dyenamic ET Dye Terminator kit (MegaBACE; Amersham Pharmacia). Most of the reactions were analysed on MegaBACE 1000 Capillary Sequencing machines (Amersham Pharmacia). ABI377 and ABI3700 sequencers (Applied Biosystems) were also used in this study. The subclones were sequenced to generate between eight- and tenfold coverage of each BAC or PAC clone. Individual BACs or PACs were assembled from the shotgun sequences using the PHRED and PHRAP[26] programs. For sequence finishing, we used the GAP4 program of the Staden Packages to help edit and select the additional sequencing reactions for gap closure and problem solving. The *Consed* program[27] was also used for sequence finishing. Sequence gaps were closed by using various dye-labelled terminator chemistries or by a combination approach of primer walking, small- or big-insert library constructing, and polymerase chain reaction with oligonucleotides.

Gene prediction and sequence alignments

We analysed the DNA sequence of each BAC for protein-coding genes using gene-prediction software and alignments with the sequences from the EST and nonredundant protein databases. For initial gene and exon predictions, we used FGENESH (http://www.softberry.com/), Genscan[28] and GeneMark.HMM (http://dixie.biology.gatech.edu/Genemark/eukhmm.cgi). We aligned sequences in the EST and nonredundant protein databases to the BAC sequences using BLAST[29] and AAT (http://genome.cs.mtu.edu/aat.html) sequence database search and alignment software. We identified repeats using RepeatMasker2 (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) and used diffseq (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/diffseq.html) to find SNPs.

We used tRNA scan-SE[30] and manual inspection to identify cytoplasmic tRNAs and organelle-derived tRNA.

Received 4 April; accepted 16 September 2002; doi:10.1038/nature01183.

1. Sasaki, T. & Burr, B. International rice genome sequencing project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.* **3,** 138–141 (2000).
2. Sasaki, T. *et al.* The genome sequence and structure of rice chromosome 1. *Nature* **420,** 312–316 (2002).
3. Harushima, Y. *et al.* A high-density rice genetic linkage map with 2275 markers using a single F_2 population. *Genetics* **148,** 479–494 (1998).
4. Wu, J. *et al.* A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* **14,** 525–535 (2002).
5. Chen, M. *et al.* An integrated physical and genetic map of the rice genome. *Plant Cell* **14,** 537–545 (2002).
6. Gale, M. D. & Devos, K. M. Comparative genetics in the grasses. *Proc. Natl Acad. Sci. USA* **95,** 1971–1974 (1998).
7. Meinke, D. W., Cherry, J. M., Dean, C. D., Rounsley, S. & Koornneef, M. *Arabidopsis thaliana*: a model plant for genome analysis. *Science* **282,** 662–682 (1998).
8. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408,** 796–815 (2000).
9. Lin, X. *et al.* Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. *Nature* **402,** 761–768 (1999).
10. Mayer, K. *et al.* Sequence and analysis of chromosome 4 of the plant *Arabidopsis thaliana*. *Nature* **402,** 769–777 (1999).
11. Messing, J. & Llaca, V. Importance of anchor genomes for any plant genome project. *Proc. Natl Acad. Sci. USA* **95,** 2017–2020 (1998).
12. Zhao, Q. *et al.* A fine physical map of the rice chromosome 4. *Genome Res.* **12,** 817–823 (2002).
13. Saji, S. *et al.* A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. *Genome* **44,** 32–37 (2001).
14. Copenhaver, G. *et al.* Genetic definition and sequence analysis of *Arabidopsis* centromeres. *Science* **286,** 2468–2474 (1999).

15. Jenny, A. & Keller, W. Cloning of cDNAs encoding the 160 kDa subunit of the bovine cleavage and polyadenylation specificity factor. *Nucleic Acids Res.* **23,** 2629–2635 (1995).
16. Yoshimura, S. *et al.* Expression of *Xa1*, a bacterial blight-resistance gene in rice, is induced by bacterial inoculation. *Proc. Natl Acad. Sci. USA* **95,** 1663–1668 (1998).
17. Dong, F. *et al.* Rice (*Oryza sativa*) centromeric regions consist of complex DNA. *Proc. Natl Acad. Sci. USA* **95,** 8135–8140 (1998).
18. Oka, H. I. in *Rice Biotechnology* (eds Khush, G. S. & Toenniessen, G. H.) 55–80 (CAB International, Oxon, 1991).
19. Khush, G. S. Origin, dispersal, cultivation and variation of rice. *Plant Mol. Biol.* **35,** 25–34 (1997).
20. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296,** 79–92 (2002).
21. Goff, S. A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* **296,** 92–100 (2002).
22. Huang, X., Adams, M. D., Zhou, H. & Kerlavage, A. R. A tool for analyzing and annotating genomic sequences. *Genomics* **46,** 37–45 (1997).
23. Paterson, A. H. *et al.* Toward a unified genetic map of higher plants, transcending the monocot–dicot divergence. *Nature Genetics* **14,** 380–382 (1996).
24. Maugenest, S., Martinez, I., Godin, B., Perez, P. & Lescure, A. M. Structure of two maize phytase genes and their spatio-temporal expression during seedling development. *Plant Mol. Biol.* **39,** 503–514 (1999).
25. Knutzon, D. S. *et al.* Cloning of a coconut endosperm cDNA encoding a 1-acyl-sn-glycerol-3-phosphate acyltransferase that accepts medium-chain-length substrates. *Plant Physiol.* **109,** 999–1006 (1995).
26. Ewing, B. & Green, P. Base-calling of automated sequencer traces using Phred. II. Error probabilities. *Genome Res.* **8,** 186–194 (1998).
27. Gordon, D., Abajian, C. & Green, P. Consed. A graphical tool for sequence finishing. *Genome Res.* **8,** 195–202 (1998).
28. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268,** 78–94 (1997).
29. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).
30. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25,** 955–964 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (**http://www.nature.com/nature**).

Acknowledgements We thank T. Sasaki and the RGP for rice genetic and EST markers and a PAC genomic library of rice Nipponbare; R. Wing and the CUGI for providing BAC libraries of the Nipponbare variety; Monsanto for the rice working-draft sequence data; G. Barry and J. Liu for help; R. Buell and Q. Yuan for help with the annotation and analysis of chromosome 4 sequences; X. Huang and Z. Ning for help with using the *AAT* and the *ssaha* programs, respectively; members of the National Centre for Gene Research for assistance; Z. Xu, Z. Chen, G. Wang, Q. Ma and Q. Zhang for support; and X. Lin, X. Deng, Y. Li, L. Zhou, N. Zheng, X. Liu and members of the IRGSP for discussion. This work was supported by grants from the Ministry of Science and Technology of the People's Republic of China, Chinese Academy of Sciences, and the Shanghai Municipal Commission of Science and Technology.

Competing interests statement The authors declare that they have no competing financial interests.

Correspondence and requests for materials should be addressed to B.H. (e-mail: bhan@ncgr.ac.cn).

Multiplicative computation in a visual neuron sensitive to looming

Fabrizio Gabbiani*†, Holger G. Krapp*‡, Christof Koch* & Gilles Laurent*

* *Division of Biology, California Institute of Technology, Pasadena, California 91125, USA*
† *Division of Neuroscience, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA*
‡ *Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK*

Multiplicative operations are important in sensory processing[1–5], but their biophysical implementation remains largely unknown[5–10]. We investigated an identified neuron (the lobula giant movement detector, LGMD, of locusts) whose output firing rate in response to looming visual stimuli has been described by two models, one of which involves a multiplication. In this model, the LGMD multiplies postsynaptically two inputs (one excitatory, one inhibitory) that converge onto its dendritic tree[11,12]; in the other model, inhibition is presynaptic to the LGMD[13,14]. By using selective activation and inactivation of pre- and postsynaptic inhibition, we show that postsynaptic inhibition has a predominant role, suggesting that multiplication is implemented within the neuron itself. Our pharmacological experiments and measurements of firing rate versus membrane potential also reveal that sodium channels act both to advance the response of the LGMD in time and to map membrane potential to firing rate in a nearly exponential manner. These results are consistent with an implementation of multiplication based on dendritic subtraction of two converging inputs encoded logarithmically, followed by exponentiation through active membrane conductances.

Several insect behaviours rely on tracking motion in depth: landing[15], hovering flight[16] and collision avoidance[17–19]. These behaviours probably depend on different neural computations as animals actively move towards a target or, conversely, experience the approach of a moving threat. The LGMD (Fig. 1a) is an identified neuron located in the third visual neuropile of the locust optic lobe (lobula), and is part of a circuit thought to be involved in the generation of escape behaviours (Fig. 1b). It responds vigorously to solid objects approaching on a collision course with the animal[17,18] (Fig. 1b, c). The LGMD fires throughout object approach with a rate that increases, peaks, and decreases as collision becomes imminent[11,12]. Responses are typically brisker for small or fast-moving objects, leading to higher peak firing rates (f_{peak}, Fig. 1d, top). But the timing of the peak firing rate is independent of the object's approach speed or size, and always follows with a fixed delay the time at which the object reaches a fixed threshold angular size, θ_{thres}, on the retina[12]. This is seen by plotting the timing of the peak relative to collision (t_{peak}) as a function of the ratio $l/|v|$, where l is the object's half-size and $|v|$ its approach speed (Fig. 1b, d, bottom). The trigonometric relationship between l, v and the angular size, $\theta(t)$, subtended by the object during approach causes the points of such a graph to fall on a straight line if t_{peak} occurs with a fixed delay, δ, relative to θ_{thres} (refs 11, 12). The slope, α, of this straight line is related to θ_{thres} (ref. 12) and its y intercept equals δ. Such a linear relation is indeed observed experimentally[11,12] (Fig. 1d, bottom; $\theta_{thres} = 17°$, $\delta = 25$ ms). Thus for the LGMD in this animal, the peak firing rate always occurred 25 ms after the object reached 17° in angular size, independent of v or l.

The peak firing rate of the LGMD was not the only parameter related to threshold angular size during object approach: the time, t_{thres}, at which the firing rate reached a given value, here arbitrarily set at 50 spikes s^{-1} (Fig. 1c triangle), also fell on a straight line ($n = 10$) as a function of $l/|v|$ (Fig. 1d, bottom), and therefore anticipated with a fixed delay (here 101 ms, Fig. 1d) the time at which the object's angular size reached a fixed value (10°). This implies that decoding the LGMD's output by detecting either the peak or an arbitrary threshold instantaneous firing rate (>50 spikes s^{-1}) conveys a reliable indicator of angular size during looming. The delay between angular threshold and peak firing time is independent of body temperature, arousal level, contrast, variations in shape or texture of the approaching object and direction of approach[12,20]. Thus, angular threshold might be the image-based retinal variable used to trigger escape responses in the face of an impending collision. Indeed, a leg flexion (presumably in preparation for an escape jump) has been shown to follow the peak LGMD firing rate with a fixed delay[11]. Angular object size is also closely related to obstacle avoidance behaviours in tethered flying locusts[19].

How then does the LGMD compute this angular threshold? The LGMD receives, onto a large dendritic fan, excitatory retinotopic inputs that convey, to a first approximation, the angular velocity, $\dot{\theta}(t)$ of the approaching object[21] (field A, Fig. 1a and b, green; Supplementary Information). In addition, two dendritic fields arborize in distinct regions of the lobula and receive phasic non-retinotopic, feedforward inhibition related to object size, $\theta(t)$

