nature genetics

Genome-wide transcription analyses in rice using tiling microarrays

Lei Li[1,2,9], Xiangfeng Wang[1,3,4,9], Viktor Stolc[2,5,9], Xueyong Li[2,6], Dongfen Zhang[7], Ning Su[2], Waraporn Tongprasit[8], Songgang Li[3], Zhukuan Cheng[7], Jun Wang[4] & Xing Wang Deng[2,3]

Sequencing and computational annotation revealed several features, including high gene numbers[1–6], unusual composition of the predicted genes[1,7] and a large number of genes lacking homology to known genes[8,9], that distinguish the rice (*Oryza sativa*) genome from that of other fully sequenced model species. We report here a full-genome transcription analysis of the *indica* rice subspecies using high-density oligonucleotide tiling microarrays. Our results provided expression data support for the existence of 35,970 (81.9%) annotated gene models and identified 5,464 unique transcribed intergenic regions that share similar compositional properties with the annotated exons and have significant homology to other plant proteins. Elucidating and mapping of all transcribed regions revealed an association between global transcription and cytological chromosome features, and an overall similarity of transcriptional activity between duplicated segments of the genome. Collectively, our results provide the first whole-genome transcription map useful for further understanding the rice genome.

Rice is an important crop and the model for grass species. To provide a comprehensive genome level transcription analysis in rice, we used custom tiling microarrays that contain 13,078,888 individual 36-mer oligonucleotide probes tiled throughout the nonrepetitive sequence of the genome. The probes were selected on the basis of an improved whole-genome shotgun (WGS) sequence of the *indica* subspecies[1,3]. The probes were synthesized in a set of 34 maskless array synthesizer–produced microarrays[10–13] and hybridized to a mixture of cDNA targets derived from four major tissues to maximize transcript detection (Methods).

Correlation of the probe fluorescence intensity with the annotated genome features showed strong agreement between hybridization signals and the predicted exons (**Fig. 1a**). To systemically analyze the hybridization data, we identified signal probes that essentially correspond to the top 10th–12th percentile specified for individual microarrays[13]. We determined an empirical cutoff value to separate probes that represent transcription from those that reflect background noise, based on the median hybridization rate value of a specific set of introns (**Supplementary Fig. 1**). We identified 43,914 non–transposable element protein-coding gene models from the improved *indica* WGS sequence[3]. (A gene model is here defined as the mRNA transcript of a putative transcription unit. Due to alternative splicing, more than one gene model can derive from a transcription unit.) Using the same cutoff value as before, we detected 84,736 of the 188,034 annotated exons (45.1%; as compared to 14.8% of introns). This result provides the first genome-wide experimental support for many of the predicted exons, for which transcription has not yet been detected by other methods.

At the gene level, we detected transcription of 35,970 (81.9%) gene models (**Fig. 1b**). Detection rate of gene models for individual chromosomes varied from 75.7% (Chr. 8) to 89.2% (Chr. 4). There was a strong linear relationship ($r^2 = 0.91$) between gene model detection rates and the percentage of signal probes among the 12 chromosomes (**Supplementary Fig. 2**), suggesting that tiling array detection was consistent with transcriptional activity reflected by the signal probes. Additionally, 10,452 (23.8%) gene models showed significant antisense transcription (**Fig. 1b**). The proportion of rice gene models showing antisense transcription was slightly lower than that reported from tiling microarray analysis in *Arabidopsis thaliana* ($\sim$30% of all annotated genes)[14], adding to an increasing body of evidence indicating that antisense transcription is an inherent property of plant genomes.

We further assessed the expression of the gene models by alignment against rice full-length cDNAs[15] and expressed sequence tags (ESTs). This resulted in the identification of 14,910 full-length cDNA supported (CG) and an additional 5,934 EST-supported (EG) gene models. The remaining 23,070 predicted genes were classified as unsupported gene (UG) models (**Fig. 2a**, **Supplementary Fig. 3**). A descending order of array detection rate for CG, EG and UG models was observed: that is, we detected 13,727 CG models (92.1%) and

[1] National Institute of Biological Sciences, Zhongguancun Life Science Park, Beijing 102206, China. [2] Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. [3] Peking-Yale Joint Research Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences, Peking University, Beijing 100871, China. [4] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China. [5] Genome Research Facility, NASA Ames Research Center, MS 239-11, Moffett Field, California 94035, USA. [6] National Center for Crop Design, China Bioway Biotech Group Co., LTD, Beijing 100085, China. [7] Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, Beijing 100101, China. [8] Eloret Corporation, Sunnyvale, California 94087, USA. [9] These authors contributed equally to this work. Correspondence should be addressed to X.W.D. (xingwang.deng@yale.edu).

Received 25 July; accepted 14 October; published online 20 December 2005; doi:10.1038/ng1704

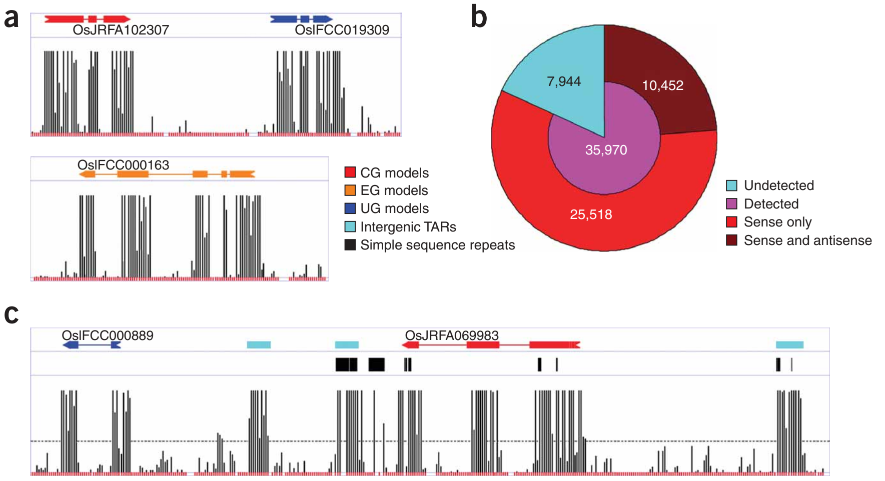


Figure 1 Tiling microarray analysis of *indica* rice gene models and intergenic regions. (**a**) Representative gene models from chromosome 1 are aligned to the chromosomal coordinates. Arrow points indicate direction of transcription. The interrogating probes are also aligned to the chromosomal coordinates, with the fluorescence intensity value depicted as a vertical bar. The red blocks at the bottom indicate the presence of an interrogating probe in the microarray at the specific genomic location. (**b**) Classification of rice gene models based on array detection. (**c**) A sample region of chromosome 1 illustrating transcription in the intergenic regions. TARs represented by at least four consecutive signal probes were identified (green bars). The cutoff for signal probe is depicted as a dashed line corresponding to a fluorescence intensity of 250. SSRs are indicated as black bars and are aligned to the chromosomal coordinates.

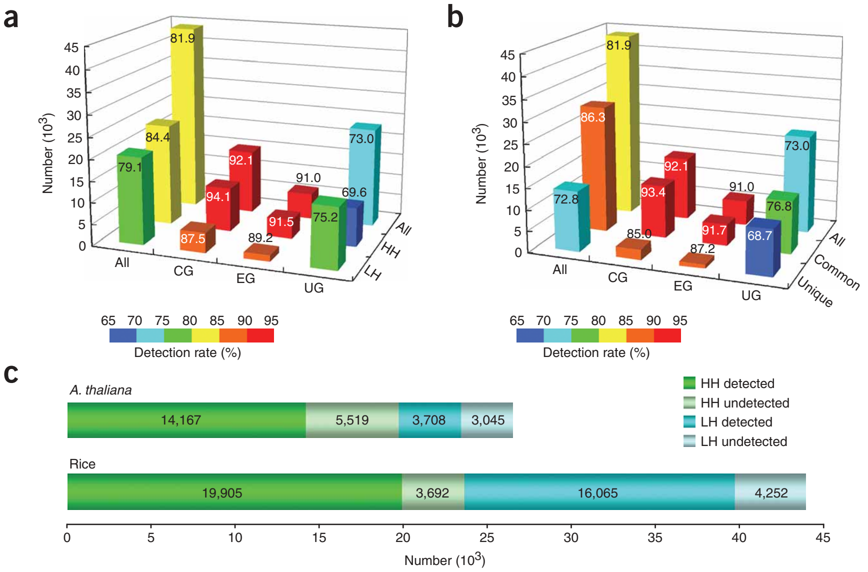
5,397 EG models (91.0%) but only 16,848 UG models (73.0%), revealing an overall agreement between tiling array detection and other experimental evidence.

The rice gene models were divided into 23,597 (53.7%) high-homology (HH) and 20,317 (46.3%) low- or no-homology (LH) models on the basis of predicted protein homology between rice and *A. thaliana* ($E \leq 10^{-7}$; **Fig. 2a**). Higher proportions of the CG (68.6%) and EG (74.7%) groups than of the UG group (38.7%) were HH models. The CG/HH models, considered the most reliable, were detected at the highest rate (94.1%; **Fig. 2a**). The UG/HH models were detected at a much lower rate (69.6%), suggesting that these models might be expressed in specialized conditions. LH models in general, and UG/LH models in particular, are shorter, contain fewer exons and have different GC composition as compared to the HH models (**Supplementary Figs. 3** and **4**). In contrast to CG and EG models, LH models in the UG group were detected at a slightly higher rate (75.2%) than the UG/HH models (**Fig. 2a**). This difference might be explained by the unusual GC composition[1,7] of the UG/LH models, which could artificially enhance their hybridization strength (**Supplementary Fig. 4**).

We used the sequence conservation between *indica* and *japonica* rice to identify 29,761 (67.8%) common and 14,153 (32.2%) unique *indica* models (**Fig. 2b**; Methods). The common models showed a greater abundance of supported models (CG and EG), and a higher array detection rate (86.3%), than the unique models (72.8%; **Fig. 2b**). Barring biological polymorphism, the unique *indica* models could represent false annotations or transposable element–related genes, which have been proposed to be the cause of the superabundance of predicted genes in rice[8,9]. Nevertheless, array detection of many of the unique *indica* gene models suggests that they could represent functional genes.

Rice and *A. thaliana* are models for monocotyledonous and eudicotyledonous plants, respectively, and are the first two plant species whose genome sequences were determined[1–3,16]. Thus, comparison of gene homology and expression between rice and *A. thaliana* constitutes an important step toward understanding the functional genes that are common and unique to different higher plants and that underlie their great physiological and chemical diversity. To this end, we divided the 26,439 *A. thaliana* gene models into 19,686 HH and 6,753 LH models[16]. In a previous *A. thaliana* genome tiling analysis involving tissues similar to those used in this study[14], 14,167 (72.0%) HH and

Figure 2 Tiling array detection of annotated *indica* gene models. (**a**) The gene models are categorized on the basis of previous expression support (CG, EG and UG) and protein homology to *A. thaliana* genes (HH and LH) using $E \leq 10^{-7}$. Array detection rate for gene models in each category is shown. (**b**) The gene models are categorized on the basis of previous expression support and alignment against *japonica* gene models (common and unique). Array detection rate for gene models in each group is shown. (**c**) Comparison of tiling array detection of rice and *Arabidopsis* gene models. Rice and *A. thaliana* gene models are categorized on the basis of protein homology in the reciprocal genomes. Expression data on the *A. thaliana* gene models was obtained from a previous whole-genome tiling array study[14].



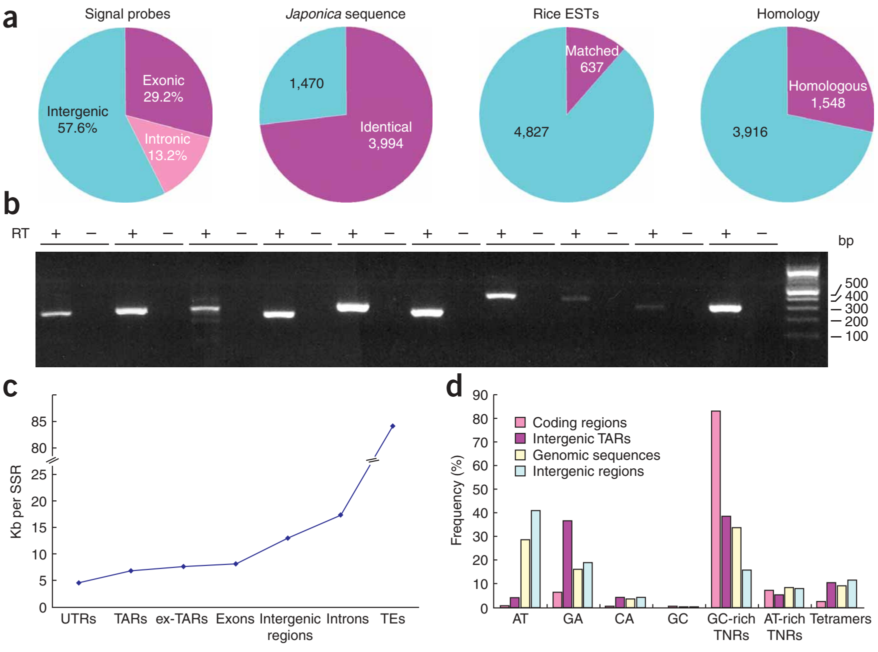


Figure 3 Identification and characterization of TARs. (**a**) Distribution of signal probes in the annotated exons, introns and intergenic regions (far left). The 5,464 identified TARs were compared against the *japonica* sequences at the corresponding genomic positions (second from left), the rice EST collections using $E \leq 10^{-20}$ (second from right) and all major plant TCs at the protein level using $E \leq 10^{-10}$ (far right). (**b**) RT-PCR analysis of selected TARs. PCR targeting each TAR was carried out on reverse transcribed cDNA (RT+) and mRNA (RT–) and resolved on gel side by side. (**c**) Average distance (in kb) between two neighboring simple sequence repeats associated with different *indica* genome components as indicated. (**d**) Relative frequency of simple sequence repeat motifs in the four *indica* genome components as indicated. TEs, transposable elements; TNRs, trinucleotide repeats.

3,708 (54.9%) LH gene models were detected (**Fig. 2c**). These results indicate that, as in rice, expression of *A. thaliana* LH gene models is more restricted than that of HH gene models. In addition, as compared to *A. thaliana*, rice contains abundant LH genes, and these are primarily responsible for its greater number of annotated genes (6,753 LH genes in *A. thaliana* versus 20,317 in rice) and the greater size of its transcriptome (3,708 expressed LH genes in *A. thaliana* versus 16,065 in rice; **Fig. 2c**). Although the LH models are prone to annotation artifacts[8,9], one explanation for the abundant LH genes in rice is lineage-specific gene divergence resulting from selection for new functions in grasses. In support of this view, homologs for many of the rice LH genes are found in the sequenced portion of the sorghum genome[17].

Consistent with results from tiling microarray analyses in other model organisms[12,14,18–21], most of the signal probes (57.6%) were located in the annotated intergenic regions (**Fig. 3a**). To systematically score transcription outside the annotated gene models, we identified 5,464 unique, previously unidentified transcriptionally active regions (TARs) (Methods), ranging in size from 179 to 1,211 nucleotides. We found identical sequences corresponding to 3,994 TARs in the *japonica* genome, suggesting that transcripts represented by these TARs are conserved (**Fig. 3a**). When aligned against the rice ESTs using a high-stringency cutoff ($E \leq 10^{-20}$), 637 (11.7%) TARs showed at least one match. This result indicates that transcripts tagged by the TARs have significant overlap with those tagged by the ESTs (**Fig. 3a**). To validate transcription of the TARs by an independent method, we carried out reverse transcriptase PCR (RT-PCR) targeting 116 randomly selected TARs. Of these, 102 (87.9%) generated PCR products of the expected size in a reverse transcription–dependent manner (**Fig. 3b**). These results confirm that the TARs represent polyadenylated transcripts and thus provide a reliable estimation of additional transcribed genomic loci beyond the predicted exons.

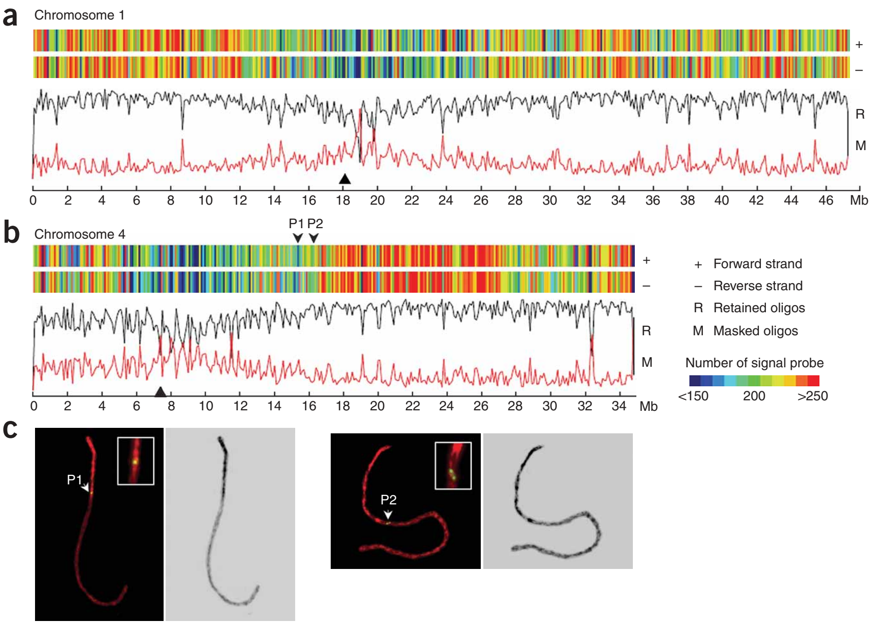


Figure 4 Tiling microarray analysis of chromosome-wide transcriptional activities. (**a**) Number of signal probes was calculated in 100-kb windows along both strands of chromosome 1 and depicted as color-coded vertical bars. Cumulative length of interrogating and masked probes in the same 100-kb windows along the length of the chromosome is shown below. Black triangle marks the starting position of the annotated centromere[3]. (**b**) Number of signal probes and accumulative length of interrogating and masked probes in 100-kb windows along the length of chromosome 4. Results for the other 10 rice chromosomes are shown in **Supplementary Figure 5**. (**c**) Euchromatin and heterochromatin of *indica* rice chromosome 4 were mapped by DAPI staining (red). **Supplementary Figure 6** shows the results for the other 11 chromosomes. Two selected probes P1 and P2, located at 15.3 Mb and 16.2 Mb, respectively (marked by arrows in **b**), were used for FISH (yellow signal). At right, the stained images converted to black and white to enhance visualization of the euchromatin and heterochromatin domains.

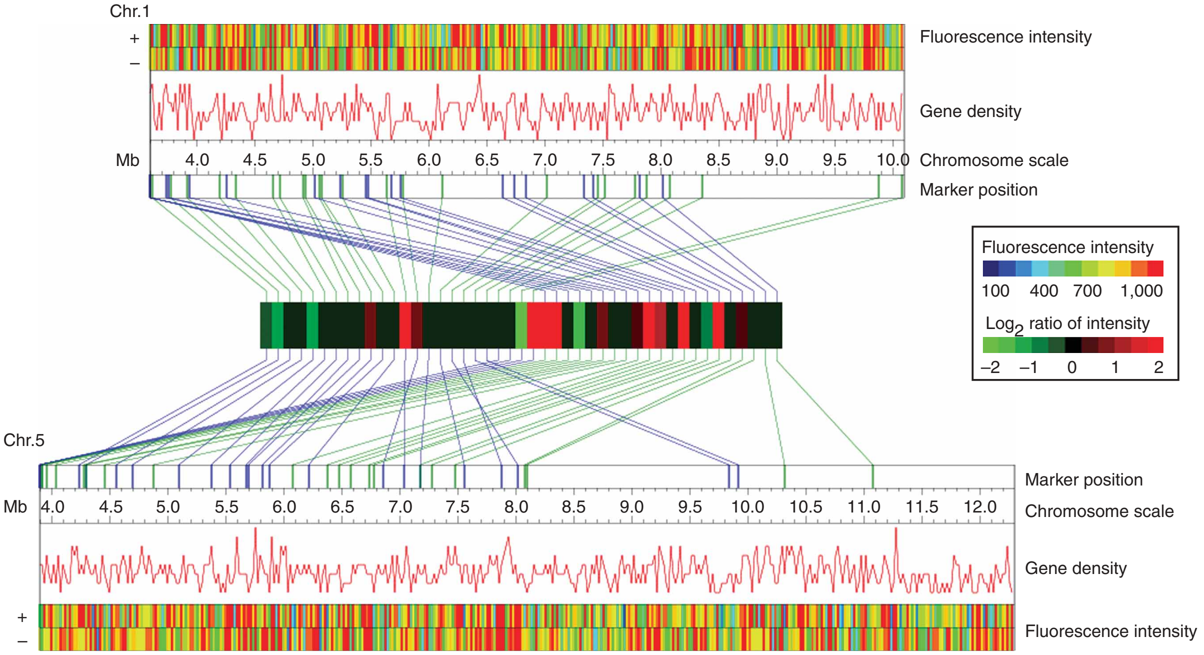


Figure 5 Transcription analysis of duplicated segments in the *indica* genome. Depicted here are the duplicated segments located in chromosome 1 and 5, which are 6.5 and 8.4 Mb in size, respectively. Average probe fluorescence intensity and number of gene models are calculated in 20-kb windows along the length of each segment and represented by the color-coded bars and the red wavy line, respectively. Positions of the markers used to define the segmental duplication are indicated. The original cDNAs are represented by green lines and their homologous counterparts by blue lines. The calculated log2 ratio of average probe intensity in the genomic regions containing the cDNAs and their counterparts is shown in the center.

To determine their protein-coding capacity, we analyzed the TARs using a homology-based method whereby the TAR sequences were used to tag the tentative consensus (TC) sequences generated from structured ESTs of plant origin[22]. When we used a stringent cutoff ($E \leq 10^{-10}$), we found at the protein level that 1,548 (28.3%) TARs had at least one hit against the 427, 428 TCs used (**Fig. 3a**). When we used a less stringent cutoff ($E \leq 10^{-5}$), we found 2,444 TARs (44.7%) to be homologous to the TCs. In a control experiment in which 5,411 introns of the CG/HH models were used in the same analysis, no significant hits (~5%) were found using both cutoffs. These results demonstrated the strong coding potentials of the TARs, which warrant further analysis to fully recognize the coding content of the rice genome.

We also assessed the TARs for their association with simple sequence repeats (SSRs; **Fig. 1c**), which are tandemly arranged repeats of short DNA motifs. Analysis of the rice SSRs revealed an association of distinct SSR motifs with different genomic components[17,23,24]. We analyzed the distribution of SSRs in seven genomic components: untranslated regions (UTRs), exonic regions, intronic regions, intergenic regions of non–transposable element models, coding sequences of transposable element models, TARs and ex-TARs (TARs extended by 100 nucleotides on both flanking sides). The prevalence of SSRs shows very distinctive distribution in these different genomic components, with transposable element models having the lowest SSR density, UTRs and exons of the non–transposable element models the highest density, and intergenic and intronic regions intermediate density (**Fig. 3c**). The SSR frequency of TARs is very similar to that of UTRs and exons, indicating that the composition of TARs is most similar to that of the transcribed regions of non–transposable element models (**Fig. 3c**). In addition, the relative frequencies of GC-rich trinucleotide and AT dinucleotide repeats vary inversely with respect to one another in the coding and noncoding regions. In this regard, too, the TARs are similar to the transcribed regions (**Fig. 3d**). Collectively, these results indicate that the TARs resemble the exonic regions compositionally and thus that they represent transcripts that have yet to be characterized.

Examination of the signal probe distribution provides an unbiased means to score genome-level transcription. Decreased transcriptional activity was found in the pericentromeric regions (**Fig. 4a**,**b**, and **Supplementary Figs. 5** and **6**), although the repression was generally not as marked as that reported for *A. thaliana*[14] and human[12]. This probably reflects the fact that for our tiling arrays we used WGS sequences, in which the highly repetitive pericentromeric sequences were 'simplified' by the WGS procedure, which has been documented in human genome sequence analysis[25–27]. Besides the pericentromeric regions, a number of chromosomal domains, including regions of chromosomes 4, 5, 7, 8, 9, 10, 11 and 12, showed relatively repressed transcription that seemed to associate with the cytologically defined heterochromatins (**Supplementary Figs. 5** and **6**).

Previously, the association of chromatin organization and transcription had been investigated for *japonica* chromosomes 4 and 10 (refs. 13,28). Here we assessed *indica* chromosome 4 for confirmation of the correlation between cytological features and transcription. Distribution of signal probes indicates that the first half of the chromosome (~16 Mb) was generally less transcriptionally active than the second half (**Fig. 4b**). As has been reported for the *japonica* chromosome[5,28,29], *indica* chromosome 4 contains roughly equal-sized heterochromatin and euchromatin that border one another at around 16 Mb (**Fig. 4c**). Fluorescence *in situ* hybridization (FISH) using two PCR-generated probes flanking the transcriptionally defined border (P1 and P2, **Fig. 4c**) showed that these were located precisely at the heterochromatin-euchromatin junction. These results indicate that tiling microarray analysis provides a high-fidelity map of the repression of transcriptional activities associated with heterochromatin in the rice genome.

The *indica* rice genome contains 18 distinct pairs of duplicated chromosome segments that together (~240 Mb) cover approximately two-thirds of the assembled genome sequence[3]. Seventeen pairs of these segments are thought to be the remnants of a whole-genome duplication predating the divergence of grasses, whereas the remaining pair (duplication 17; **Supplementary Table 1**) is more recent in origin[3]. We found that the general transcriptional activity, measured by the average probe fluorescence intensity in 20-kb windows across each segment, is often similar between duplicated segments. Analysis of one such duplication involving portions of chromosome 1 and 5 (duplication 1; **Supplementary Table 1**) is illustrated (**Fig. 5**).

To quantify this phenomenon, we examined transcriptional relationship between 1,217 marker pairs—that is, among a set of full-length cDNAs and their counterparts used to identify the duplicated segments[3]. The hybridization rate in genomic regions corresponding to the marker pairs was calculated and their Pearson correlation coefficient determined. When we assessed the correlation coefficient between the hybridization rate in the cDNA marker regions and that in randomly extracted genomic sequences of the same size, we found a significant positive correlation for 14 of the 18 segmental pairs. Not surprisingly, duplication 17, which has the most marker pairs per unit length of sequence, showed the highest correlation (**Supplementary Table 1**). We further calculated the relative probe fluorescence intensity in the genomic regions corresponding to the marker pairs. Most of these regions show the same average probe intensity (**Fig. 5** and data not shown), which seemed unlikely to be due to probe cross-hybridization because none of the 32,495 probes in the cDNA regions has ≥70% identity against the 28,694 probes in the homologous regions and vice versa. Further analysis relating transcription to the divergence of gene function should open up a new avenue for understanding the transcriptional components of genome regulation and evolution following genome duplication.

METHODS

MAS microarray design, production and hybridization. We designed a minimal tiling strategy using 36-mer oligonucleotides so as to represent the nonrepetitive sequences of the *indica* rice genome, based on the improved WGS sequencing result[3]. We synthesized the oligonucleotide probes in the maskless array synthesizer at a density of 389,000 oligos per array[10–13] and hybridized the microarrays to Cy3-labeled cDNA mixtures derived from seedling root, seedling shoot, panicle and suspension-cultured cells of *indica* rice (*Oryza sativa* L. ssp. *indica* cv. 9311). Target preparation, array hybridization and hybridization intensity value acquisition were done as previously described[11–13] (see **Supplementary Note** for details).

Gene model compilation. A total of 45,797 non–transposable element gene models have been predicted from the 374 Mb of assembled *indica* WGS sequences[3]. We aligned all these gene models to a collection of rice full-length cDNA sequences by BLAT[30] using cutoff criteria of 100 bp overlap and 90% identity over the entire length of each match. In cases in which multiple gene models intersected with the same full-length cDNA, only the gene model with the best match with the cDNA was chosen for further analysis. After this filtering step, a set of 43,914 nonredundant gene models was identified. The gene models matching the full-length cDNAs were called CG models. The remaining models were further matched with rice ESTs using the same criteria and the matched models were called EG models. The predicted genes without matches to cDNA and EST sequences were classified as UG models. To identify common and unique gene models between *indica* and *japonica*, *japonica* gene models (TIGR Rice Pseudomolecules Release 3, January 2005, http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml) were positioned to the *indica* genome sequence using BLAT. The *indica* models that mapped to the same loci as their *japonica* counterparts with over 100 bp of overlap in their predicted open reading frames were identified as common gene models, and the remaining models were identified as unique.

Identification and analysis of TARs. To identify TARs in the intergenic space of all non–transposable element gene models, we selected genomic regions represented by at least four consecutive signal probes (covering a region roughly equal in size to the median size of the annotated exons). We discarded those containing any probe that occurred more than once in the genome. Using these criteria, we identified 7,120 TARs. We aligned the sequences of these TARs to the annotated *indica* transposable element models and the unique *japonica* gene models mapped on the *indica* genome to identify the 5,464 unique TARs used in all latter analysis. The TAR sequences were compared to rice ESTs using BLAST. TARs and the major plant TCs derived from structured ESTs were compared at the protein level using TBLASTX in all six possible open reading frames. SSR identification and SSR motif determination were done using a program written in the Perl scripting language as previously reported[23]. Cutoffs for the number of repeats were nine, six and five for the dinucleotide, trinucleotide and tetranucleotide motifs, respectively.

Chromosome preparation and FISH. We carried out chromosome preparation and FISH essentially as previously reported[29]. However, the probes used in this study were derived from PCR products 2 kb in length amplified from selected genomic regions of specific 100-kb windows at the euchromatin-heterochromatin boundary of chromosome 4 (**Fig. 3**). The PCR template sequences were identified by fragmenting the 100-kb windows into 2-kb fragments *in silico* and then identifying those fragments unique in the genome with BLAST. Digoxigenin-labeled probes were detected by fluorescein isothiocyanate–conjugated sheep anti-digoxigenin (Roche Diagnostics). Chromosomes were counterstained with 4′,6-diamidino-phenylindole (DAPI) in an antifade solution (Vector Laboratories). Chromosomes and fluorescence signal images were captured using the Olympus BX61 fluorescence microscope conjunct with a CCD camera. Gray-scale images were captured for each color channel and then merged using the software of Image-Pro Plus.

Transcription analysis of duplicated segments. We identified a total of 1,217 markers (one full-length cDNA) and their homologous counterparts to represent the 18 segmental duplications in *indica* rice. These homologs were defined by TBLASTN searching of the rice genome for putative homologs in any of the six possible reading frames[3]. Because the homologous regions do not necessarily represent annotated genes, we assayed transcription in the genomic regions corresponding to the markers. In the tiling arrays, the cDNA regions contain 32,495 probes whereas their homologous regions contain 28,694 probes. The hybridization rate was calculated in these regions and the Pearson correlation coefficient of the hybridization rate between the duplicated segments determined. Average probe fluorescence intensity in each cDNA region and its counterpart were calculated and their relative log2 ratio determined.

Accession codes. Tiling microarray design and experimental data are available in the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/projects/geo/index.cgi) under series GSE3452.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The collaborative *indica* rice tiling microarray project in China was supported by the 863 rice functional genomics program from the Ministry of Science and Technology of China and by the National Institute of Biological Sciences at Beijing. The rice tiling microarray project at Yale University was supported by a grant from the US National Science Foundation Plant Genome Program.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Yu, J. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* **296**, 79–92 (2002).

2. Goff, S.A. *et al.* A draft sequence of the rice genome (*Oryza sativa* L. *ssp japonica*). *Science* **296**, 92–100 (2002).
3. Yu, J. *et al.* The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* **3**, e38 (2005).
4. Sasaki, T. *et al.* The genome sequence and structure of rice chromosome 1. *Nature* **420**, 312–316 (2002).
5. Feng, Q. *et al.* Sequence and analysis of rice chromosome 4. *Nature* **420**, 316–320 (2002).
6. Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice Chromosome 10. *Science* **300**, 1566–1569 (2003).
7. Wong, G.K. *et al.* Compositional gradients in *Gramineae* genes. *Genome Res.* **12**, 851–856 (2002).
8. Bennetzen, J.L., Coleman, C., Liu, R., Ma, J. & Ramakrishna, W. Consistent over-estimation of gene number in complex plant genomes. *Curr. Opin. Plant Biol.* **7**, 732–736 (2004).
9. Jabbari, K., Cruveiller, S., Clay, O., Le Saux, J. & Bernardi, G. The new genes of rice: a closer look. *Trends Plant Sci.* **9**, 281–285 (2004).
10. Nuwaysir, E.F. *et al.* Gene expression analysis using oligonucleotide arrays produced by maskless photolithography. *Genome Res.* **12**, 1749–1755 (2002).
11. Stolc, V. *et al.* A gene expression map for the euchromatic genome of *Drosophila melanogaster. Science* **306**, 655–660 (2004).
12. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
13. Li, L. *et al.* Tiling microarray analysis of rice chromosome 10 to identify the transcriptome and relate its expression to chromosomal architecture. *Genome Biol.* **6**, R52 (2005).
14. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
15. Kikuchi, S. *et al.* Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **300**, 1566–1569 (2003).
16. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant. *Arabidopsis thaliana. Nature* **408**, 796–815 (2000).
17. Bedell, J.A. *et al.* Sorghum genome sequencing by methylation filtration. *PLoS Biol.* **3**, e1 (2005).
18. Kapranov, P. *et al.* Large-scale transcriptional activity in chromosomes 21 and 22. *Science* **296**, 916–919 (2002).
19. Kampa, D. *et al.* Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* **14**, 331–342 (2004).
20. Cheng, J. *et al.* Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science* **308**, 1149–1154 (2005).
21. Rinn, J.L. *et al.* The transcriptional activity of human Chromosome 22. *Genes Dev.* **17**, 529–540 (2003).
22. Messing, J. *et al.* Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci. USA* **101**, 14349–14354 (2004).
23. Temnykh, S. *et al.* Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res.* **11**, 1441–1452 (2001).
24. Li, C. *et al.* Sequence variations of simple sequence repeats on chromosome-4 in two subspecies of the Asian cultivated rice. *Theor. Appl. Genet.* **108**, 392–400 (2004).
25. Nagaki, K. *et al.* Sequencing of a rice centromere uncovers active genes. *Nat. Genet.* **36**, 138–145 (2004).
26. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).
27. She, X. *et al.* Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* **431**, 927–930 (2004).
28. Jiao, Y. *et al.* A tiling microarray expression analysis of rice chromosome 4 suggests a chromosomal level regulation of transcription. *Plant Cell* **17**, 1641–1657 (2005).
29. Cheng, Z. *et al.* Toward a cytological characterization of the rice genome. *Genome Res.* **11**, 2133–2141 (2001).
30. Kent, W.J. BLAT-the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).