# A pilot study of transcription unit analysis in rice using oligonucleotide tiling-path microarray

Viktor Stolc[1,2,†], Lei Li[1,†], Xiangfeng Wang[3,4,5,†], Xueyong Li[1,6], Ning Su[1], Waraporn Tongprasit[8], Bin Han[8], Yongbiao Xue[9], Jiayang Li[9], Michael Snyder[1], Mark Gerstein[10], Jun Wang[4] and Xing Wang Deng[1,*]

[1]Department of Molecular Cellular and Developmental Biology Yale University, New Haven, CT, 06520, USA (*author for correspondence; e-mail xingwang.deng@yale.edu); [2]Genome Research Facility NASA Ames Research Center, MS 239–11, Moffett Field, CA, 94035, USA; [3]National Institute of Biological Sciences Zhongguancun Life Science Park, 102206, Beijing, China; [4]Beijing Institute of Genomics Chinese Academy of Sciences, 101300, Beijing, China; [5]Peking-Yale Joint Research Center of Plant Molecular Genetics and Agrobiotechnology, College of Life Sciences Peking University, 100871, Beijing, China; [6]National Center of Crop Design China Bioway Biotech Group Co., LTD, 100085 , Beijing, China; [7]Eloret Corporation at NASA Ames Research Center, MS 229–3, Moffett Field, CA, 94035 , USA; [8]National Center for Gene Research, Shanghai Institutes for Biological Sciences Chinese Academy of Sciences, 500 Caobao Road, 200233, Shanghai, China; [9]Institute of Genetics and Developmental Biology Chinese Academy of Sciences, Datun Road, 100101, Beijing, China; [10]Department of Molecular Biophysics and Biochemistry Yale University, 06511, New Haven, CT, USA; [†]These authors contributed equally to this work

## Abstract

As the international efforts to sequence the rice genome are completed, an immediate challenge and opportunity is to comprehensively and accurately define all transcription units in the rice genome. Here we describe a strategy of using high-density oligonucleotide tiling-path microarrays to map transcription of the *japonica* rice genome. In a pilot experiment to test this approach, one array representing the reverse strand of the last 11.2 Mb sequence of chromosome 10 was analyzed in detail based on a mathematical model developed in this study. Analysis of the array data detected 77% of the reference gene models in a mixture of four RNA populations. Moreover, significant transcriptional activities were found in many of the previously annotated intergenic regions. These preliminary results demonstrate the utility of genome tiling microarrays in evaluating annotated rice gene models and in identifying novel transcription units that will facilitate rice genome annotation.

## Introduction

Providing staple food for over half of the human population, rice (*Oryza sativa*) is the world's most important crop species (Hoshikawa, 1993). Its compact genome size, amenity to genetic analysis, and close relationship with other cereal species, all make rice an ideal model system to study grass physiology, development, agronomics, and genomics of the grass family (Gale and Devos, 1998b; Goff, 1999; Shimamoto and Kyozuka, 2002; Paterson *et al.*, 2003). Draft genome

sequences of the two major subspecies of rice, *indica* and *japonica*, were published in 2002 (Goff *et al.*, 2002; Yu *et al.*, 2002). The continuous and anchored (finished phase) sequences for *japonica* chromosomes 1, 4 and 10 were later been obtained (Feng *et al.*, 2002; Sasaki *et al.*, 2002; The Rice Chromosome 10 Sequencing Consortium, 2003). Finished whole genome sequences for *japonica* and *indica* rice have recently been determined by the International Rice Genome Sequencing Project (IRGSP, http://rgp.dna.affrc.go.jp/IRGSP/Build3/build3.html) and the Beijing Genome Institute (BGI, Yu *et al.*, 2005), respectively.

The completion of rice genome sequencing is foreseen to immediately enhance, or make possible, many aspects of rice research. For instance, one such effort is the identification and characterization of the abundant, novel rice- or cereal-specific genes revealed by genome sequencing. Genomic data indicate that between one third and one half of the predicted proteins in rice have no recognizable homologues in the genome of the other major model plant, *Arabidopsis thaliana*. In contrast, extensive homology has been detected between the rice genome and sequences or expressed sequence tags (ESTs) from other cereals (Feng *et al.*, 2002; Goff *et al.*, 2002; Kikuchi *et al.*, 2003; Bedell *et al.*, 2005; Yu *et al.*, 2005), indicating the presence of a large number of cereal-specific genes. Thus, rice functional genomics is expected to yield a wealth of information on cereal-specific proteins and the biochemical pathways in which they participate.

Another area of research is to address how rice genes are coordinated at the transcriptional level to carry out complex biological processes. Conventional approaches to study plant transcription factors and networks (reviewed in Schwechheimer *et al.*, 1998) are often labor intensive and can only be used to focus on a few targets at a time. The availability of the complete genomic sequence and the development of rice whole genome microarrays indicate that the computational approaches, such as *in silico* identification of transcriptional factor binding sites, successfully used in other model systems will soon be available in rice (Ren *et al.*, 2000; Bussemaker *et al.*, 2001; Iyer *et al.*, 2001; Li and Wang, 2003; Synder and Gerstein, 2003). The existence of a vast reservoir of possible cereal-specific genes and the unique life cycle of cereals suggest that this area of research will likely generate new insight into the mechanisms underlining plant gene function, expression, and regulation.

Comparative genomics, especially efforts to utilize rice as a reference to facilitate definition of gene order and regulatory patterns in other cereals, will greatly benefit from a complete rice genome sequence and annotation. Though extensive synteny among the cereal genomes has long been observed (Gale and Devos, 1998a; Moore, 2000), detailed comparison at the sequence level is still lacking (Bennetzen and Ma, 2003). Comparisons of the rice Chromosome 1 and 10 sequence with available genetic maps and sequences from sorghum and maize largely confirmed these syntenic relationships at the chromosomal level (Bowers *et al.*, 2003; Klein *et al.*, 2003). Hence, the complete rice sequence provides an opportunity to combine data generated from genomic and genetic studies in different cereals to interpolate gene order and to trace genome evolution in diverse cereals (Moore, 2000; Yuan *et al.*, 2001; Paterson *et al.*, 2003).

All of the research efforts aforementioned urgently require and eloquently justify the need for accurate account of the total gene number, precise definition of individual gene structures and their expression profiles in rice. Similar to other sequenced genomes, identification of genes in the rice genome has so far heavily relied on non-experimental methods such as *ab initio* gene prediction and sequence homology and motif analysis. These efforts are limited by the insufficient ability of current gene-finding programs to effectively identify and annotate genes from complex genomes (Guigó *et al.*, 2000; Mathé *et al.*, 2002; Zhang, 2002). This problem is particularly acute in rice due to unusual genome features such as high GC content near the 5′ end of the open reading frame (ORF) of many rice genes (Wong *et al.*, 2002; Yu *et al.*, 2002). Current approaches for mapping mRNA-coding regions on a genome scale in rice have focused on EST and full-length cDNA analysis (Kikuchi *et al.*, 2003). However, these techniques do not comprehensively reveal all genomic coding information as they are biased towards the highly expressed genes. Not surprisingly, exhaustive efforts have only represented less than half of the predicted genes (Feng *et al.*, 2002; Sasaki *et al.*, 2002; Yu *et al.*, 2003; Zhao *et al.*, 2004).

Tiling-path microarrays based on PCR-generated genomic fragments have been used to monitor transcriptional activities of human chromosome 22 and rice chromosome 4 (Rinn *et al.*, 2003; Jiao *et al.*, 2005). Meanwhile, new developments in microarray technology have made it possible to prepare high-density oligonucleotide microarrays such that complete genomic sequences can be represented in a small number of arrays. Probed with either labeled cDNA or cRNA, the chromosome or genome tiling-path arrays enable researchers to detect transcriptional activity without any presumption of gene structures. This approach has been used successfully in examining the transcriptional activity of human chromosomes (Frazer *et al.*, 2001; Shoemaker *et al.*, 2001; Kapranov *et al.*, 2002) and the Arabidopsis genome (Yamada *et al.*, 2003). Results from these studies indicate that a significant portion of the transcriptional activity resides outside the predicted coding regions and, thus, is not represented by current genome annotations. In addition, these studies show that the genome tiling microarray approach is able to confirm a large fraction of the computer predicted genes and to verify or correct their structures.

Here we report a tiling strategy to interrogate the non-repetitive sequence of the *japonica* genome on a set of 32 Maskless Array Synthesizer (MAS) generated arrays. In a pilot experiment, one of the arrays containing approximately 390,000 individual probes representing a portion of chromosome 10 was analyzed using a mathematical model developed in this study. Analysis of the array data confirmed expression of 77% of the reference gene models in a mixture of four RNA populations, namely, seedling root, seedling shoot, panicle and suspension-cultured cells. Moreover, transcriptional activity was detected in many of the intergenic regions, pointing to the existence of many putative novel transcripts or uncharacterized portions of the already annotated gene models.

## Materials and methods

### Plant materials and treatments

*Oryza sativa* ssp. *japonica* cv. Nipponbare was used for all experiments unless otherwise indicated. Seeds were surface-sterilized and imbibed at 37 °C for 2 days to break seed dormancy and then sowed on MS medium solidified with 0.8% (w/v) agar. Seedlings were kept under continuous light at 28 °C for 7 days before being harvested for RNA isolation. Alternatively, seven-day-old seedlings were transferred into soil and maintained in long-day conditions (13 h light/11 h dark) at 26–28 °C until mature. Heading and filling stage panicles were then collected from these plants. Suspension-cultured cells were prepared using the AA medium as previously described (Baba *et al.*, 1986). For RNA isolation, plant materials (seedling root, seedling shoot, panicle, and suspension-cultured cells) were frozen in liquid nitrogen and homogenized. Total RNA and mRNA were isolated using the RNeasy Plant Mini kit (Qiagen, Valencia, CA) and the Oligotex mRNA kit (Qiagen) according to the manufacturer's specifications, respectively.

### Design, production of MAS microarrays, and data extraction

Thirty-six-mer oligonucleotide tiling microarrays used in this study were produced through combined procedures of *in silico* array design and *in situ* oligonucleotide synthesis at the NASA Ames Genome Research Facility (www.phenomorph. arc.nasa.gov) using Maskless Array Synthesizer (Nuwaysir *et al.*, 2002) and custom NASA software. Microarray design was carried out using the ArrayScribe software from NimbleGen (www.nimblegen.com). The *japonica* genomic sequences (released in June 2003) were imported into ArrayScribe, and 36-mer probes were designed to cover the genome every 46 bp. The 36-mer oligonucleotide probe sequences were obtained using an algorithm based on sequence-dependent factors such as length, extent of complementarity and overall base composition. Probes that could form a stem-loop structure with stem length greater than seven bases were excluded. To mask those with high probability of cross hybridization, all probe were scored using the following method. For each probe, there were 20 consecutive 17-mer oligonucleotides, each of which was searched against the whole genome. A score was assigned to each 17-mer based on its copy number in the genome. For example, if three copies of a 17-mer were found in the genome, a score of three was determined. Finally, the average score of all the 20 different 17-mers of a probe was calculated

(Oligo Index) and a cutoff value of five was used for inclusion.

The density of the arrays was at 389,287 probes per array, which means a "chessboard" layout of the arrays where positive features (containing an interrogating oligonucleotide) and negative features (no oligonucleotide synthesized) were separated by each other. A control probe, CPK6 (5′-AAGGAGAGCGACATTACTGGA-GAT) was included in the probes and its placement specified. The CPK6, which is visualized by including in the array hybridization targets a complementary oligonucleotide labeled with Cy3, forms a specific "grid" that aids in identifying the edges of the array and helps in aligning a reference grid when extracting intensity values from the scanned image of the array. The final step for tiling array design was the creation of "digital masks," where all the above information was stored as specific electronic files for each array that will be used to control the patterning of the micromirrors in the array synthesizer during array production.

Unlike traditional photolithographic oligo arrays, which require physical masks, MAS arrays are produced on glass slides (rather than controlled pore glass supports) using the Digital Micromirror Device (DMD). The DMD is an array of 786,000 tiny aluminum mirrors, arranged on a computer chip, where each mirror is individually addressable and controllable. Using these tiny aluminum mirrors to shine UV light in specific patterns, coupled with photolithography, oligonucleotide arrays were produced. All of our arrays included the CPK6 reference probe. All arrays were scanned at a 5-micron resolution, and the images were saved as a standard TIFF file. Scanned images were imported into NimbleScan, where the image files were matched with corresponding array description files, the feature grid within each array automatically located, and the hybridization intensity value of individual features quantified.

*Target preparation, slide hybridization, washing and visualization*

Hybridization targets were prepared using the post-transcriptional modification method. Briefly, an oligo(dT) primer was used to selectively synthesize and label cDNA from poly(A)$^+$ mRNA by direct incorporation of amino-allyl-modified dUTP (aa-dUTP) during reverse transcription essentially as previously described (Ma et al., 2001; Bertone et al., 2004; Stolc et al., 2004). The aa-dUTP decorated cDNAs were purified using the CyScribe GFX purification kit (Amersham Biosciences, Piscataway, NJ) and eluted with 0.1 M NaHCO$_3$, pH 9.0. Fluorescent labeling was achieved by conjugating the monofunctional Cy3 dye (Amersham Biosciences) to the amino-allyl functional groups in the cDNA at room temperature for 90 min and terminated by adding 0.5 M hydroxylamine. The fluorescent dye-labeled targets were purified using the CyScribe GFX purification kit. Labeled cDNA derived from 2 μg mRNA from various rice tissues were mixed with 50 mM MES (pH 6.5–6,7), 500 mM NaCl, 10 mM EDTA, 0.005% Tween-20, 0.5 mg/ml BSA, 0.1 mg/ml herring sperm DNA and 1 μM Cy3-modified oligonucleotide complementary to CPK6 in a final volume of 300 μl to make the hybridization solution.

Slide hybridization was carried out as previously described (Nuwaysir et al., 2002; Bertone et al., 2004) with a few modifications. The microarray slides were covered with the SA200 Secure-Seal Hybridization Chamber (Grace Bio-Labs, Bend, OR), which creates a small hybridization space over the surface of the slide. The slides were pre-hybridized in a 400 μl solution containing 50 mM MES (pH 6.5–6.7), 500 mM NaCl, 10 mM EDTA, 0.005% Tween-20, 0.5 mg/ml BSA, and 0.1 mg/ml herring sperm DNA at 50 °C for 15 min. During this period, the 300 μl hybridization solution was denatured at 95 °C for 10 min, chilled on ice, spun at 14,000 g for 5 min. The slide was then hybridized with the hybridization solution at 50 °C for 16 h with gentle agitation. After hybridization, slides were washed twice in low stringent washing buffer (6 × SSPE, 1 mM DTT, and 0.01% Tween-20) for 5 min at room temperature, six times in high stringent washing buffer (100 mM MES, pH 6.5–6.7, 0.1 M NaCl, 1 mM DTT, and 0.01% Tween-20) for 5 min each at 42 °C, and twice in low stringent washing buffer at room temperature again. After washing, slides were rinsed with 0.1 × SSC, dried by centrifugation and scanned by a GenePix 4000B scanner (Axon, Foster City, CA) using the 532 nm channel and visualized with the GenePix Pro 3 image analysis software (Axon).

## Results

### A workflow for genome wide transcription analysis using tiling-path microarrays

Genome annotation mandates generation, validation, and integration of information from multiple sources. One powerful and proven annotation approach is whole genome tiling-path microarray analysis coupled with array-guided cDNA cloning (Shoemaker et al., 2001; Kapranov et al., 2002; Yamada et al., 2003). Figure 1 illustrates the strategy of our effort to improve the rice genome annotation using an integrative approach centered on tiling microarrays. Essential to our approach is the design and development of custom tiling microarrays that represent the whole rice genome. Given the size of the rice genome and the repetitive nature of our experiments, suitable tiling microarray should offer high feature density, versatility
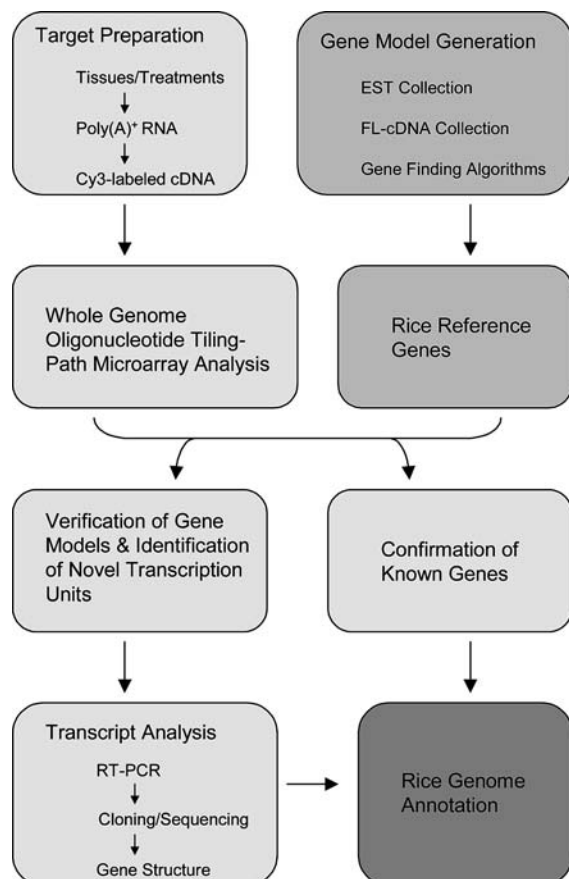


*Figure 1.* Diagram of the workflow for whole genome tiling-path microarray analysis in rice.

of modification and compatibility with our existing conventional microarray facility. We therefore chose the MAS platform developed by NimbleGen (Nuwaysir et al., 2002; Bertone et al., 2004; Stolc et al., 2004).

To effectively decode the tiling microarray hybridization data, reference gene models are needed in which a comprehensive set of rice gene models is precisely anchored in the genomic sequences. A unique advantage in rice is the availability of multiple annotations such as the TIGR *japonica* annotation (http://www.tigr.org/ tdb/e2k1/osa1/), the BGI *indica* and *japonica* annotations (Yu et al., 2005). While each annotation may have its own forte and weakness, comparison and analysis of these annotations should yield a more comprehensive inventory of gene models to be evaluated by tiling arrays than any single annotation can offer (Yu et al., 2005). Our reference gene compilation also takes advantage of the rice EST and full-length cDNA collections (Kikuchi et al., 2003), which provides value gene expression information. The full-length cDNA confirmed genes could naturally serve as positive controls to test array design, hybridization, and transcript detection.

The hybridization data of the tiling microarrays can be used to further rice genome annotation in several ways (Figure 1). For example, tiling array detection provides support or verification for gene models without prior experimental support. The array-generated transcription data can be used to identify the said gene models by means of array-guided cDNA cloning and analysis. Likewise, for array-revealed novel transcription units (i.e., those that were not annotated but detected by tiling arrays), the array data again could serve as a guidepost for cloning the corresponding cDNA (Figure 1). Comparing sequencing library-based full-length cDNA clones, which is considered the gold standard of gene annotation (Hass et al., 2002; Kikuchi et al., 2003), array-assisted cDNA cloning and analysis will be targeted, and, thus, more cost-effective and inclusive to cover the remaining portion of the rice genome that lacks expressed sequences.

### Development of rice tiling-path microarrays

The tiling microarray experiments described here focus on analyzing the *japonica* genome. This is in

part because several complete *japonica* chromosomal sequences were already available at the commencement of this project (Feng *et al.*, 2002; Sasaki *et al.*, 2002; The Rice Chromosome 10 Sequencing Consortium, 2003), permitting pilot experiments to streamline and quality control the workflow for whole genome work (Figure 1). Additionally, the *japonica* genome is the target for several publically funded rice genomics projects. Consequently, results from our analysis can be readily integrated with these efforts to maximize the detection of rice genes expressed at different developmental stages or under diverse environmental conditions.

The current version of the MAS technology allows up to 786,000 features to be simultaneously synthesized onto a $1.74 \times 1.31$ cm$^2$ area of a single glass slide (Singh-Gasson *et al.*, 1999; Nuwaysir *et al.*, 2002). To represent effectively the entire *japonica* sequences, we designed perfect-match 36-mer oligonucleotide probes every 46 nucleotides in a "head-to-tail" fashion along the genome, with every two probes separated by an average of 10 nucleotides. This design was applied to both strands of the DNA, resulting in pairs of complementary probes that represent sequence information on both strands of the DNA molecule. These probes were then mathematically calculated to filter out probes that represent sequences with high copy numbers or high-degree of complementarity (see Materials and methods). In total, 12,254,374 oligonucleotides were retained in the final collection to cover approximately 282 Mb of "non-repeated" sequences of the *japonica* genome. Table 1 shows the number of tiling probes for each of the 12 chromosomes. These oligonucleotides were designed for synthesis at a density of 389, 287 probes per slide. Using this design, the interrogating probes that tile the *japonica* genome can be accommodated onto a set of 32 MAS arrays.

### Determination of hybridization signal and noise

As a pilot experiment to test the workflow, four custom MAS arrays were synthesized that cover portions of *japonica* chromosome 1 and 10. The 36-mer probes in the two chromosome 1 arrays were randomly selected from the probe set that represents the whole chromosome. These two slides were subsequently hybridized with labeled

*Table 1.* Number of 36-mer oligonucleotides tiling the *japonica* genome.

| Chromosome | Oligos |
|---|---|
| 1 | 1,516,954 |
| 2 | 1,277,332 |
| 3 | 1,237,874 |
| 4 | 1,127,386 |
| 5 | 913,674 |
| 6 | 1,053,222 |
| 7 | 974,822 |
| 8 | 938,738 |
| 9 | 734,364 |
| 10 | 750,282 |
| 11 | 848,758 |
| 12 | 880,428 |
| Total | 12,254,374 |

cDNA derived from seedling shoot to test hybridization conditions and data acquisition procedures (data not shown). The two chromosome 10 arrays contained probes that tile the second half of the chromosome ($\sim$11.2 Mb). These two arrays were hybridized with targets prepared from equal amounts of four selected RNA populations, namely, seedling root, seedling shoot, panicle, and suspension-cultured cells. The hybridized arrays were scanned and the hybridization signal values of all features (positive and negative) extracted.

One of the arrays, which represents the reverse strand of the 11.2 Mb chromosome 10 sequence, was chosen for this pilot analysis. The scanned image of this array is shown in Figure 2. The adjusted intensity value of a positive feature was determined by subtracting the mean intensity of the four surrounding negative features from the raw intensity of the positive feature. A bimodal distribution of the adjusted intensity was observed with one peak (noise) formed at 6.5 Log$_2$ and the other (signal) around 12.5 Log$_2$ (Figure 3; in red). To statistically determine whether a given interrogating probe represents noise or signal, we modeled the noise as a normal distribution, assuming that probes with very low intensity (1–6 Log$_2$) were exclusively noise (Figure 3; in blue). A cutoff of 8.79 Log$_2$ could then be set where probes with intensity lower than 8.79 Log$_2$ would be regarded as noise with a 90% confidence (Figure 3, green line). Using this cutoff, approximately 16% of the probes in the array were considered signal, or that they detected a transcript and are referred to as
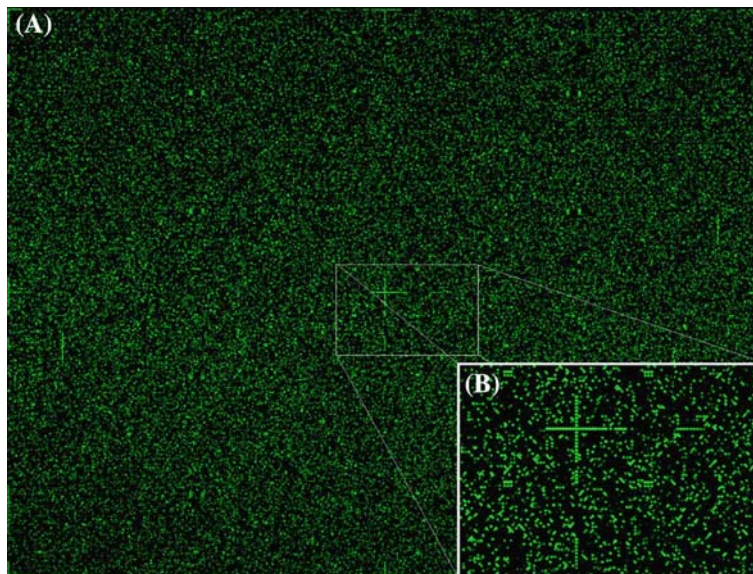
*Figure 2.* Scanned image of a high-density MAS oligonucleotide tiling microarray. (A) Overview of the feature area $(1.74 \times 1.31 \text{ cm}^2)$ of a MAS microarray slide hybridized with Cy3-labeled cDNA target derived from rice seedlings, panicles and suspension-cultured cells. (B) An enlarged image of a small section of the feature area showing individual hybridized features. Note the grid formed by the control *CPK6* oligonucleotide.

signal probes hereafter. When the hybridized slides were striped and re-hybridized with the same cDNA targets, the bimodal distribution drifted toward the center (data not shown), indicating that
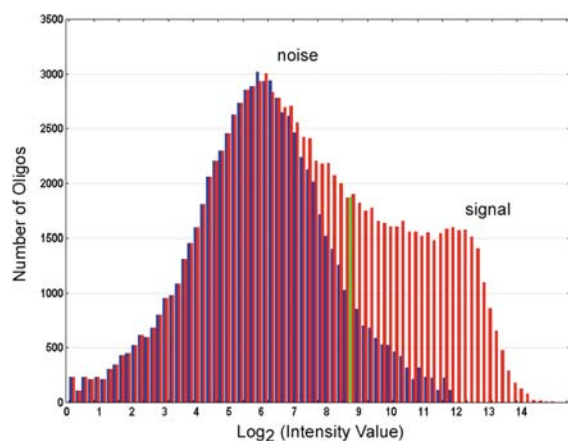


*Figure 3.* Bimodal distribution of probe hybridization intensity of a high-density MAS tiling microarray. The microarray slide contains probes representing the reverse strand of the last 11.2 Mb of *japonica* chromosome 10 and was hybridized with Cy3-labeled cDNA samples. The $\text{Log}_2$ hybridization intensity of all probes forms two easily discernable peaks that are considered to represent the noise and the signal, respectively (red color). The noise is modeled as a normal distribution based on the distribution of all probes at low intensity (blue color). The cutoff value of 8.79 $\text{Log}_2$ corresponding to a 90% confidence is shown in green.

re-use of the MAS array degrades data quality and may not, thus, be suitable for high-quality data generation.

*Mapping signal probes into the chromosomal sequence*

When signal probes were sorted by their actual sequences and mapped according to their chromosomal coordinates, they appeared in discrete clusters dispersed along the chromosome (Figure 4). To examine the tiling microarray hybridization data at the individual gene level, we compiled a set of reference gene models based on the BGI *japonica* annotation using FgeneSH (Yu *et al.*, 2005). For the reverse strand of 11.2 Mb chromosome 10 sequence, a set of 681 non-overlapping and non-redundant gene models was determined and mapped. Mapping with rice EST and full-length cDNA categorized these gene models into three groups: 247 (36.3%) cDNA confirmed gene (CG) models, 187 (27.5%) EST supported gene (EG) models, and 247 (36.3%) models lacking any previous experimental support (UG). In addition, the gene models predicted by another program called BGF (Yu *et al.*, 2002; Zhao *et al.*, 2004) that do not align with the FgeneSH models were also included, but only for
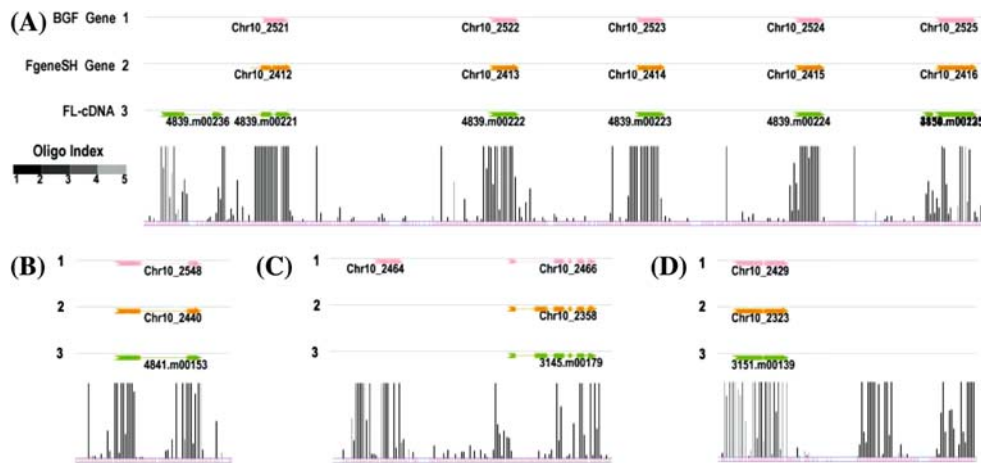
*Figure 4.* Alignment of tiling microarray hybridization data with genomic sequences. The gene models shown on the top of each panel are BGF genes, predicted from an in-house algorithm in BGI (line 1), and FgeneSH genes, predicted by the FGENESH program (line 2). FL-cDNA, full-length cDNA sequences downloaded from TIGR (line 3). The hybridization intensity value of each signal probe is depicted as a vertical line, the degree of gray shade of which represents the oligonucleotide repetition index (Oligo Index). Oligo Index was calculated based on frequency and extent of complementarity of the 36-mer oligonucleotide pairs in the genome (see Materials and methods). The short vertical lines underneath the hybridization intensity depict positions of the 36-mer oligonucleotides sorted by their sequences. (A and B) Signal probe clusters in predicted genic regions that have corresponding full-length cDNA. (C) Pattern of signal probe clusters in a computationally predicted genic region without a matching cDNA clone. (D) Clusters of signal probe representing putative new transcribed loci in an intergenic region.

comparison with the distribution pattern of signal probes (Figure 4).

When the hybridization data were compared with the reference gene models, a clear trend was seen where a majority of the annotated exons matched with signal probe clusters. A few specific examples are shown in Figure 4. For most CG models, the signal probe clusters matched well with the predicted gene structures (Figure 4A and B). Signal probe clusters were also detected in regions where UG models locate, clearly suggesting expression of these models (Figure 4C). This result provides experimental support to these predicted genes as they are likely transcribed, although further experiments such as RT-PCR and cDNA sequence analyses are needed to unambiguously validate the annotation of these gene models. Figure 4 also illustrates another important aspect of our tiling microarray analysis; that is, discovery of potentially novel transcription units in the intergenic regions (see below).

*Examination of transcriptional activity by hybridization rate*

An algorithm based on hybridization rate (HR), which is defined as the number of signal probes per kb genomic sequence, was developed to examine the transcriptional activity. By definition, HR value should range from 0 to 21.7, however, masking repeated sequences is expected to render HR in a narrower range. Figure 5 shows the distribution of HR in exons ($HR_E$) and introns ($HR_I$) of the chromosome 10 region examined. For $HR_I$, two peaks were observed with the major peak having a mode very close to 0, consistent with the fact that most introns are not expressed (Figure 5, blue line). The minor peak of $HR_I$, centered on 1.4, likely reflects transcriptional activities representing misidentified exons as introns or due to unknown alternative splicing events (Figure 5, blue line). Conversely, $HR_E$ formed a predominant peak centered at 5.9 with a significant portion of exons having $HR_E$ above 10. A minor peak for $HR_E$ distribution was observed around 0, suggesting that it was likely due to misinterpretation of introns as exons, or genes not expressed at all in the RNA populations used (Figure 5, red line).

The results indicate that HR can be used as an indicator of the "expression status" of a given genomic fragment. Indeed, the mean $HR_E$ (3.51) was approximately twice that of $HR_I$ (1.62) across the chromosome 10 region investigated. This
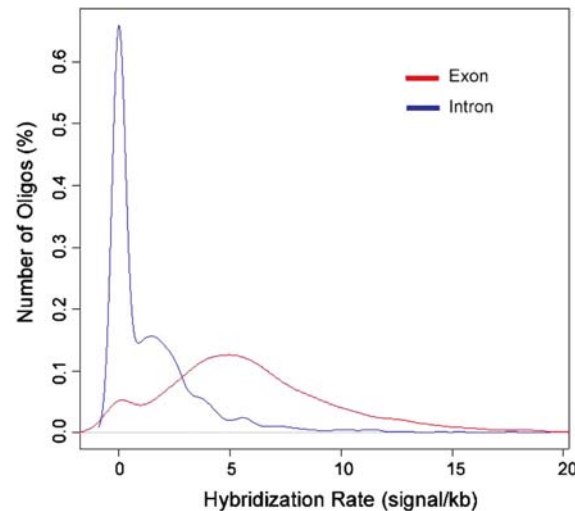
*Figure 5.* Distribution of hybridization rate in exonic and intronic regions. Hybridization rate (HR) is calculated as the number of signal probe per kb sequence. Distribution of HR values in the exonic regions according to the reference gene models is shown in red and intronic regions in blue. See Figure 2 and text for the description of signal probes.

prompted us to develop a simple algorithm to ascertain the expression of gene models using $2HR_I$ as a threshold, where $HR_I$ is the mean HR of all introns of a multiple exon gene model, or the mean HR of all chromosomal introns for single-exon gene models. When $HR_E \geq 2HR_I$, the gene model is considered transcribed; conversely, if $HR_E$ (2$HR_I$, the gene model is regarded as not transcribed. This algorithm also implies that when the structure of a gene model was incorrectly annotated, it would likely be regarded as unexpressed and, thus, flagged for re-examination.

### Detection of reference gene models

To verify reference gene models in the tested chromosome 10 region, we applied the above algorithm to the pilot tiling microarray data. Overall, transcription of 76.9% of the reference gene models was confirmed. Specifically, expression of 80.2% CG (198 out of 247), 80.7% of EG (151 out of 187), and 70.9% of UG (175 out of 247) was detected. The detection rate is very similar for reference genes with previous experimental support (CG and EG) to that found in a study in Arabidopsis (Yamada *et al.*, 2003). This is probably not surprising given that similar RNA populations were included in the microarray hybridization though different algorithms were used to determine the expression rate. However,

the detection rate of UG models was much higher in our analysis (71% compared with 37% in Arabidopsis, Yamada *et al.*, 2003). One possible explanation is that rice full-length cDNA and EST collections are less exhaustive than those in Arabidopsis.

As expected, lowering the threshold increases the number of gene models being detected. For instance, if $1HR_I$ was used instead of $2HR_I$, we could detect transcription for 95.5% (236 out of 247) of CG, 90.4% (169 out of 187) EG, 85.4% (211 out of 247) UG models, and an overall detection rate of 90.5% for all the reference genes. However, it is expected that lowering the cutoff will result in more false positive detection of transcription units in a sensitivity/specificity trade-off. Thus, further optimization of the gene detection algorithm based on HR could lead to more sensitive and reliable interpretation of the hybridization data that allows better evaluation of different rice genome annotations.

### Identification of novel transcription units

Visual inspection of the chromosome 10 tiling microarray data clearly indicated that many of the signal probe clusters locate in the intergenic regions (Figure 4D). These probe clusters could conceivably represent genomic features such as unrecognized exons or untranslated regions of the

annotated genes. There is also a possibility that they represent novel transcripts in those annotated intergenic regions that were simply missed from the current rice genome annotation. In fact, the average HR in intergenic regions (2.27) was significantly higher than $HR_I$ (1.62) in the chromosome 10 region examined, indicating stronger transcriptional activity in the intergenic regions. It should be possible to identify systematically putative new transcription units by an expression cutoff based on HR. For example, more than 500 transcribed loci could be identified from the pilot tiling microarray analysis using HR = 3.24 ($2HR_I$) as a cutoff. About 40% of these loci have a putative translation start codon and significant coding capacities ((50 amino acids; data not shown). When we selected a much more stringent expression cutoff (HR = 8), there still were more than 150 ''expressed'' loci recognized in the intergenic regions. Given that the median $HR_E$ was 5.9, these results suggest that there are potentially hundreds of novel transcribed loci in the intergenic region examined. Using sequences of these array-detected loci as a guidepost to clone, assemble and analyze the corresponding cDNA will then lead to the identification of novel genes and, thus, enhancement of rice genome annotations.

## Discussion

One of the immediate tasks after the finished rice sequences become available is to compile a precise catalogue and expression profile of all transcription units/genes. Estimation of the total gene number based upon the draft sequences ranges widely from 30 to 60 thousand (Goff *et al.*, 2002; Yu *et al.*, 2002). Complete sequences of chromosome 1, 4 and 10 allow a much fine-tuned estimation that puts the total gene number of rice between 57,000 and 62,500 (Feng *et al.*, 2002; Sasaki *et al.*, 2002; Yu *et al.*, 2003). However, these annotations are based on gene-finding computer programs and therefore need to be verified experimentally. Moreover, recent studies in other genomes using integrative approaches centered on microarray analysis revealed hundreds if not thousands of new genes in each of the genomes investigated (Shoemaker *et al.*, 2001; Kapranov *et al.*, 2002; Hild *et al.*, 2003; Rinn *et al.*, 2003; Yamada *et al.*, 2003). Therefore, experimental

efforts complementary to computer-based genome annotation are needed to verify predicted genes and to discover new genes in rice.

As shown in Figure 1, a key feature of our workflow of tiling microarray analysis is the cataloguing of the rice ''reference gene'' set by combining all available genomic information to represent the non-redundant gene models of the *japonica* genome. This is not trivial, as the outcome of the tiling microarray analysis impinges on the completeness and quality of the reference gene set (Figure 1). Rice sequences have been subjected to extensive analysis, resulting in multiple annotations of both the *indica* and the *japonica* genome. Close comparison among these annotations and between the two genomes should help to identify common and unique gene models and integrate multiple annotations to provide a more comprehensive representation of the genome content. For our first attempt to analyze the tiling microarray data, BGI *japonica* gene models were used for simplicity (Yu *et al.*, 2005). It should be noted that the 681 reference genes used in this study are slightly less than what have been predicted for this region in the TIGR annotation (The Rice Chromosome 10 Sequencing Consortium, 2003). This is due to the fact that we have masked gene models that (1) have over 100 bp perfect match with other gene models, (2) are extremely large (spanning ( 20 kb), and (3) have extensive homology in the genome. Therefore, this reference gene set excluded high copy number genes such as transposases and genes with overlap in their coding regions. For the whole genome tiling microarray analysis, individual annotations will be studied and compared.

The results from the pilot experiment in which the reverse strand of the last 11.2 Mb sequence of chromosome 10 was analyzed in detail are reported here. This analysis showed a 77% detection rate of all reference genes in the mixture of four RNA populations by a moderate cutoff ($HR_E \geq 2HR_I$), demonstrating the feasibility of our approach. The transcribed loci detected based on HR, besides supported gene models (e.g. CG and EG), belong to two general groups: those that have been predicted by current genome annotations (UG) and those that have not (Figure 4). The relatively lower detection rate for the UG models than the supported suggests that these models are expressed at lower levels or restricted to specific

cell types/developmental stages. Alternatively, some of these predicted gene models might be false models that do not exist *in planta*. Nevertheless, array detection provides experimental support to most (70.9%) of the UG models, as detection of hybridization signals is strongly indicative of RNA synthesis directed by the genome segment represented by the probes. Therefore, these results can be used to assist in targeted transcript cloning to further verify and correct the annotation of UG models.

By the same criteria, the second group of transcribed loci identified from tiling array analysis represents putative novel transcripts or portions of other genes. In the chromosome 10 region examined, expression of numerous intergenic regions was detected (Figure 4D and data not shown). The resolution of our pilot tiling microarrays was less than sufficient to define precise exon/intron junctions of transcription units and their 5′ and 3′ ends, especially for sequences without prior annotation. Thus, the array-revealed new transcribed loci will need to be verified by obtaining cDNA clones and their sequence information. One strategy to achieve this goal is to utilize array data-guided RT-PCR cloning (Figure 1). Comparison and assembly of the cDNA clones and their mapping to the genomic sequences will help in identifying the corresponding novel transcription units and in defining their structure. These results, when integrated with information from gene ontologies and homology searching, will, in our hope, greatly enhance on-going rice genome annotation efforts.

It is clear from our pilot experiments that new computational tools need to be developed and validated to facilitate tiling data analysis, storage, retrieval, and distribution. For example, our algorithm, based on HR to detect transcription, needs to be improved by incorporating more statistical parameters. To present the results from our tiling microarray analysis in an easily accessible form, we will design and develop an interactive database in which all the reference genes and transcription units identified by tiling arrays will be displayed in relationship with the complete genome sequence, and linked to cDNA/EST sequences, along with other information. This database will also include functions that make it easy to update when new transcription units or genome features become available as our whole genome analysis proceeds.

Whole genome tiling microarray coupled with array-guided cDNA analysis are emerging as a powerful approach to characterize eukaryotic genomes (Frazer *et al.*, 2001; Shoemaker *et al.*, 2001; Kapranov *et al.*, 2002; Yamada *et al.*, 2003). Applying these approaches in rice should aid in the current genomic efforts to provide a complete and accurate genome annotation, the success of which will be invaluable to research of the grass family in areas such as elucidating biochemical pathways and developmental processes, defining the molecular mechanisms of stress responses, and understanding plant evolutionary genomics. These advancements will, in turn, provide greater insights into basic biology of the agriculturally and economically important cereal crops.

## References

Baba, A., Hasezawa, S. and Syono, K. 1986. Cultivation of rice protoplasts and their transformation mediated by *Agrobacterium spheroplasts* Plant Cell Physiol. 27: 463–471.

Bedell, J.A., Budiman, M.A., Nunberg, A., Citek, R.W., Robbins, D., Jones, J., Flick, E., Rohlfing, T., Fries, J. and Bradford, K. 2005. Sorghum genome sequencing by methylation filtration PLOS. Biol. 3: e13.

Bennetzen, J.L. and Ma, J. 2003. The genetic colinearity of rice and other cereals on the basis of genomic sequence analysis Curr. Opin. Plant Biol. 6: 128–133.

Bertone, P., Stolc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Tongprasit, W., Samanta, M., Weissman, S. and Gerstein, M. 2004. Global identification of human transcribed sequences with genome tiling arrays Science 306: 2242–2246.

Bowers, J.E., Abbey, C., Anderson, S., Chang, C., Draye, X., Hoppe, A.H., Jessup, R., Lemke, C., Lennington, J. and Li, Z.K. 2003. A high-density genetic recombination map of sequence-tagged sites for Sorghum, as a framework for comparative structural and evolutionary genomics of tropical grains and grasses Genetics 165: 367–386.

Bussemaker, H.J., Li, H. and Siggia, E.D. 2001. Regulatory element detection using correlation with expression Nature Genetics 27: 167–174.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. 1998. Cluster analysis and display of genome-wide expression pattern Proc. Natl. Acad. Sci. USA 95: 14863–14868.

Feng, Q., Zhang, Y.J., Hao, P., Wang, S.Y., Fu, G., Huang, Y.C., Li, Y., Zhu, J.J., Liu, Y.L., Hu, X. and , 2002. Sequence and analysis of rice chromosome 4 Nature 420: 316–320.

Frazer, K.A., Sheehan, J.B., Stokowski, R.P., Chen, X., Hosseini, R., Cheng, J-F., Fodor, S.P.A., Cox, D.R. and Patil, N. 2001. Evolutionarily conserved sequences on human Chromosome 21 Genome Res. 11: 1651–1659.

Gale, M.D. and Devos, K.M. 1998a. Comparative genetics in the grasses Proc. Natl. Acad. Sci. USA 95: 1971–1974.

Gale, M.D. and Devos, K.M. 1998b. Plant comparative genetics after 10 years Science 282: 656–659.

Goff, S.A. 1999. Rice as a model for cereal genomics Curr. Opin. Plant Biol. 2: 86–89.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R.L., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P. and Varma, H. 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp japonica) Science 296: 92–100.

Guigó, R., Agarwal, P.A., Abril, J.F., Burset, M. and Fickett, J.W. 2000. An assessment of gene prediction accuracy in large DNA sequences Genome Res. 10: 1631–1642.

Hass B.J., Volfovsky N., Town C.D., Troukhan M., Alexandrov N., Feldmann K.A., Flavell R.B., White O. and Salzberg S.L. 2002. Full-length messenger RNA sequences greatly improve genome annotation. Genome Biol. 3: research0029.1–0029.12.

Hild, M., Beckman, B., Hass, S.A., Koch, B., Solovyev, V., Busold, C., Fellenberg, K., Boutros, M., Vingron, M. and Sauer, F. 2003. An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome Genome Biol. 5: R3.

Hoshikawa, K. 1993. Science of the Rice Plant, vol 1: Morphology. Nobunkyo, Tokyo, pp. 133–186.

Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M. and Brown, P.O. 2001. Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF Nature 409: 533–538.

Jiao Y., Jia P., Wang X., Su N., Yu S., Zhang D., Ma L., Feng Q., Jin Z., Li L., *et al.* 2005. A tiling microarray expression analysis of rice chromosome 4 suggests a chromosomal level regulation of transcription. Plant Cell: In Press.

Kapranov, P., Cawley, S.E., Drenkow, J., Bekiranov, S., Strausberg, R.L., Fodor, S.P.A. and Gingeras, T.R. 2002. Large-scale transcriptional activity in chromosomes 21 and 22 Science 296: 916–919.

Kikuchi, S., Satoh, K., Nagata, T., Kawagashira, N., Doi, K., Kishimoto, N., Yazaki, J., Ishikawa, M., Yamada, H., Ooka, H. and , 2003. Collection, mapping, and annotation of over 28,000 cDNA Clones from *japonica* Rice Science 300: 1566–1569.

Klein, P.E., Klein, R.R., Vrebalov, J. and Mullet, J.E. 2003. Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement Plant J. 34: 605–621.

Li, H. and Wang, W. 2003. Dissecting the transcription networks of a cell using computational genomics Curr. Opin. Gen. Dev. 13: 611–616.

Luscombe, N.M., Royce, T.E., Bertone, P., Echols, N., Horak, C.E., Chang, J.T., Snyder, M. and Gerstein, M. 2003. ExpressYourself: a modular platform for processing and visualizingmicroarray data Nucl. Acids Res. 31: 3477–3482.

Ma, L., Li, J., Qu, L., Hager, J., Chen, Z., Zhao, H. and Deng, X.W. 2001. Light control of Arabidopsis development entails coordinated regulation of genome expression and cellular pathways Plant Cell 13: 2589–2607.

Mathé, C., Sagot, M-F., Schiex, T. and Rouzé, P. 2002. Current methods of gene prediction, their strengths and weaknesses Nucl. Acids. Res. 30: 4103–4117.

Moore, G. 2000. Cereal chromosome structure, evolution, and pairing Annu. Rev. Plant Physiol. Mol. Biol. 51: 195–222.

Nuwaysir, E.F., Huang, W., Albert, T.J., Singh, J., Nuwaysir, K., Pitas, A., Richmond, T., Gorski, T., Berg, J.P. and Ballin, J. 2002. Gene expression analysis using oligonucleotide arrays produced by maskless photolithography Genome Res. 12: 1749–1755.

Paterson, A.H., Bowers, J.E., Peterson, D.G., Estill, J.C. and Chapman, B.A. 2003. Structure and evolution of cereal genomes Curr. Opin. Genet. Dev.13: 644–650.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E. and , 2000. Genome-wide location and function of DNA binding proteins Science 290: 2306–2309.

Rinn, J.L., Euskirchen, G., Bertone, P., Martone, R., Luscombe, N.M., Hartman, S., Harrison, P.M., Nelson, F.K., Miller, P. and Gerstein, M. 2003. The transcriptional activity of human Chromosome 22 Genes Dev. 17: 529–540.

Sasaki, T., Matsumoto, T., Yamamoto, K., Sakata, K., Baba, T., Katayose, Y., Wu, J.Z., Niimura, Y., Cheng, Z.K., Nagamura, Y. and , 2002. The genome sequence and structure of rice chromosome 1 Nature 420: 312–316.

Schwechheimer, C., Zourelidou, M. and Bevan, M.W. 1998. Plant transcription factor studies Annu. Rev. Plant Physiol. Plant Mol. Biol. 49: 127–150.

Shoemaker, D.D., Schadt, E.E., Armour, C.D., He, Y.D., Garrett-Engele, P., McDonagh, P.D., Loerch, P.M., Leonardson, A., Lum, P.Y. and Cavet, G. 2001. Experimental annotation of the human genome using microarray technology Nature 409: 922–927.

Shimamoto, K. and Kyozuka, J. 2002. Rice as a model for comparative genomics of plants Annu. Rev. Plant Biol. 53: 399–419.

Singh-Gasson, S., Green, R.D., Yue, Y.J., Nelson, C., Blattner, F., Sussman, M.R. and Cerrina, F. 1999. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array Nat. Biotech. 17: 974–978.

Snyder, M. and Gerstein, M. 2003. Defining genes in the genomics era Science 300: 258–260.

Stolc, V., Gauhar, Z., Mason, C., Halasz, G., Batenburg, M.F.van, Rifkin, S.A., Hua, S., Herreman, T., Tongprasit, W. and Barbano, P.E. 2004. A gene expression map for the euchromatic genome of Drosophila melanogaster Science 306: 655–660.

The Rice Chromosome 10 Sequencing Consortium 2003. In-depth view of structure, activity, and evolution of rice Chromosome 10 Science 300: 1566–1569.

Wang, W., Vinocur, B. and Altman, A. 2003. Plant response to drought, salinity and extreme temperatures: towards genetic engineering for stress tolerance Planta 218: 1–14.

Wong, G.K., Wang, J., Tao, L., Tan, J., Zhang, J., Douglas, A. and Yu, J. 2002. Compositional gradients in Gramineae genes Genome Res. 12: 851–856.

Yamada, K., Lim, J., Dale, J.M., Chen, H., Shinn, P., Palm, C.J., Southwick, A.M., Wu, H.C., Kim, C. and Nguyen, M. 2003. Empirical analysis of transcriptional activity in the *Arabidopsis* genome Science 302: 842–846.

Yu, J., Hu, S., Wang, J., Shu, W.G, Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y. and Zhang, X. 2002. A Draft Sequence of the rice genome (*Oryza sativa* L. ssp. *indica*) Science 296: 79–92.

Yu, J., Wang, J., Lin, W., Li, S., Li, H., Zhou, J., Ni, P., Dong, W., Hu, S., Zeng, C. and , 2005. The genomes of *Oryza sativa*: A history of duplications PLOS Biol. 3: e38.

Yuan, Q.P., Quackenbush, J., Sultana, R., Pertea, M., Salzberg., S.L. and Buell, C.R. 2001. Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species Plant Physiol. 125: 1166–1174.

Zhang, M.Q. 2002. Computational prediction of eukaryotic protein-coding genes Nature Rev. Gent. 3: 698–710.

Zhao, W.M., Wang, J., He, X., Huang, X., Jiao, Y., Dai, M., Wei, S., Fu, J., Chen, Y. and Ren, X. 2004. BGI-RIS: an integrated information resource and comparative analysis workbench for rice genomics Nucl. Acids Res. 32: D377–D382.