Asymmetric distribution of gene expression in the centromeric region of rice chromosome 5

Hiroshi Mizuno1, Yoshihiro Kawahara2, Jianzhong Wu1, Yuichi Katayose3, Hiroyuki Kanamori4, Hiroshi Ikawa4, Takeshi Itoh2, Takiji Sasaki3† and Takashi Matsumoto1*

1 Plant Genome Research Unit, Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan
2 Bioinformatics Research Unit, Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan
3 Soybean Genome Research Team, Division of Genome and Biodiversity Research, National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan
4 Institute of the Society for Techno-innovation of Agriculture, Forestry and Fisheries, Tsukuba, Ibaraki, Japan
5 National Institute of Agrobiological Sciences, Tsukuba, Ibaraki, Japan

There is controversy as to whether gene expression is silenced in the functional centromere. The complete genomic sequences of the centromeric regions in higher eukaryotes have not been fully elucidated, because the presence of highly repetitive sequences complicates many aspects of genomic sequencing. We performed resequencing, assembly, and sequence finishing of two P1-derived artificial chromosome clones in the centromeric region of rice (Oryza sativa L.) chromosome 5 (Cen5). The pericentromeric region, where meiotic recombination is silenced, is located at the center of chromosome 5 and is 2.14 Mb long; a total of six restriction-fragment-length polymorphism markers (R448, C1388, S20487S, E3103S, C53260S, and R2059) genetically mapped at 54.6 cM were located in this region. In the pericentromeric region, 28 genes were annotated on the short arm and 45 genes on the long arm. To quantify all transcripts in this region, we performed massive parallel sequencing of mRNA. Transcriptional density (total length of transcribed region/length of the genomic region) and expression level (number of uniquely mapped reads/length of transcribed region) were calculated on the basis of the mapped reads on the rice genome. Transcriptional density and expression level were significantly lower in Cen5 than in the average of the other chromosomal regions. Moreover, transcriptional density in Cen5 was significantly lower on the short arm than on the long arm; the distribution of transcriptional density was asymmetric. The genomic sequence of Cen5 has been integrated into the most updated reference rice genome sequence constructed by the International Rice Genome Sequencing Project.

Keywords: genome sequencing, mRNA-Seq, International Rice Genome Sequencing Project, P1-derived artificial chromosome, centromere

INTRODUCTION

The centromere is essential for the correct segregation of chromosomes in dividing cells. The functional centromere complex is composed of proteins binding to highly repetitive centromere-specific DNA sequences (Houben and Schubert, 2003; Dawe and Hiatt, 2004; Hall et al., 2004; Sharma and Raina, 2005; Lamb et al., 2007; Ma et al., 2007; Gill et al., 2008). Centromere-specific histone-H3-like protein (CENH3) defines the boundaries of the functional centromeric region of DNA; CENH3 replaces the canonical histone H3 to form a specific type of nucleosome that is essential for kinetochore formation (Henikoff et al., 2001; Blower et al., 2002). The kinetochore links the chromosome to microtubule polymers, which are attached to the mitotic spindle during mitosis and meiosis.

However, the genomic sequences of the centromeric regions are diverse and have not yet been fully elucidated in higher eukaryotes, even in the case of the so-called “completely sequenced” genomes (Hosouchi et al., 2002; Mizuno et al., 2008b; Torras-Llort et al., 2009; Buscaino et al., 2010). Because the presence of highly repetitive sequences complicates many aspects of genomic sequencing (including cloning, mapping, chromosome walking, and computer-assisted assembly of the fragments of DNA sequences), sequencing of the centromeric regions of higher eukaryotes is extremely difficult.

Nevertheless, substantial progress in sequencing of the centromere region has been made in rice (Oryza sativa L.). As some rice centromeres have exceptionally small numbers of tandem repeats (IRGSP, 2005), rice is suitable for the comprehensive analysis of centromeric sequence composition and organization in eukaryotes. From 1998 to 2004, the International Rice Genome Sequencing Project (IRGSP) succeeded in constructing a P1-derived artificial chromosome (PAC) and bacterial artificial chromosome (BAC) clone contig including the centromere regions of three chromosomes. Initial Sanger dyeoxy sequencing of these clones revealed, for the first time, the overall structure of the centromeric regions of higher eukaryotes (IRGSP, 2005). To date, of the 12 rice chromosome centromeric regions, Cen3 (containing gaps; Yan et al., 2006), Cen4 (Zhang et al., 2004), and Cen8 (Wu et al., 2004) have been almost completely sequenced. In the case of Cen5, a PAC/BAC contig has been constructed by chromosome walking (Cheng et al., 2005); however, the contig is only partially sequenced (IRGSP, 2005).
In the core region of each rice centromere is a tandem array of a key sequence, the 155-bp RCS2/CentO sequence (Dong et al., 1998). Around the RCS2/CentO array is distributed the pericentromeric region in which meiotic recombination is suppressed. Genes have been computationally predicted in pericentromeric regions (Nagaki et al., 2004; Wu et al., 2004). Twenty-seven of the predicted genes in Cen8 are conserved in the japonica rice Nipponbare and the indica rice Kasalath (Wu et al., 2009). Although the centromere has been considered to be a highly heterochromatic and transcriptionally silent chromosomal domain, active genes have been found in the 750-kb core domain of Cen8 (Nagaki et al., 2004). There is therefore controversy as to whether gene expression is silenced in the functional centromere. To assess the functional importance of the expression of these centromeric genes, it is important to characterize them and quantify their transcripts.

Here, we performed sequence improvement and comprehensive expression analysis of rice Nipponbare chromosome 5 at single-nucleotide resolution. First, we used a Sanger sequencing-based finishing procedure to bridge the short and long arm chromosome 5 sequences in the public reference rice genome sequence constructed by the IRGSP. Second, we applied Illumina massive parallel sequencing technology to mRNA sequencing, revealing the distribution of gene expression in Cen5. We discovered that the distribution was asymmetric. We discuss the importance of gene expression in centromeric regions and the evolutionary history of the asymmetric distribution of expressed genes in Cen5.

MATERIALS AND METHODS

SEQUENCE IMPROVEMENT OF PAC/BAC CLONES BY USING A FINISHING PROCEDURE

P1-derived artificial chromosome (P) and BAC (B) libraries were constructed from genomic DNA derived from the rice cultivar Nipponbare (JP 229579 in the National Institute of Agrobiological Sciences Genebank; O. sativa L. ss. japonica) and generated by the Rice Genome Research Program of Japan. The BAC library (OJINBa) was constructed by the Arizona Genomics Institute (Ammiraju et al., 2006). Details of the method used for Southern hybridization and PCR screening of the PAC/BAC libraries have been given previously (Wu et al., 2003). Two PAC clones (P0587F01, P0697B04) were resequenced in accordance with the IRGSP sequencing guidelines (IRGSP, 2005). Briefly, about 2000 subclone plasmid libraries from each PAC clone were end-sequenced, and these sequences were assembled with Phred–Phrap software. For the gap regions within PAC/BAC clones, bridging subclones were fully sequenced by primer walking. To resolve misassembly in the repeat regions, several subclones (~7 kb) were fully sequenced, and these continuous sequences were used as a guide for the reassembly process. Finally, the clone sequences were combined, taking into account overlaps.

PREPARATION OF cDNA, ILLUMINA SEQUENCING, AND MAPPING OF SHORT READS

Nipponbare rice was grown in a growth chamber at 28°C. After the seedlings had been grown for 7 days, total RNA was extracted from the shoots and roots by using an RNeasy Plant Kit (Qiagen, Hilden, Germany). RNA quality was calculated by using a Bioanalyzer 2100 algorithm (Agilent Technologies, USA); high-quality RNA (RNA integrity number >8) was used. Oligo(dT) magnetic beads were used to isolate poly(A) RNA from the total RNA samples. Poly(A) RNA was converted to cDNA for massive parallel sequencing in an Illumina Genome Analyzer IIX (Illumina, San Diego, CA, USA), in accordance with the protocol for the mRNA-Seq sample preparation kit (Illumina). All primary mRNA sequence read data had been previously submitted to the DNA Data Bank of Japan (DDBJ; DRA000159; Mizuno et al., 2010). Normal shoot and normal root reads that passed the filter were mapped onto the Nipponbare reference genome (Build 5.0) by using Bowtie (version 0.12.7; Langmead et al., 2009) and TopHat (version 1.2.0; Trapnell et al., 2009) software, with the default parameters. Uniquely mapped reads were used for further analysis. Differences in transcriptional density [total length of transcribed region (bp)/length of the genomic region (bp)] and expression level [number of uniquely mapped reads/length of transcribed region (bp)] were assessed statistically by Fisher’s exact test. The length of the genomic region was calculated on the basis of the Nipponbare reference genomic sequence (Build 5.0). A “transcribed region” was defined as a region in which at least one read derived from mRNA was mapped.

RESULTS

GENOMIC SEQUENCING OF Cen5

P1-derived artificial chromosome/BAC clone-based sequencing was adopted for genomic sequencing of Cen5. A PAC/BAC contig was constructed by chromosome walking to cover the genetically defined centromeric region of chromosome 5 (Cheng et al., 2005). The PAC/BAC contig was mapped by using restriction-fragment-length polymorphism (RFLP) markers S20487S and E3103S, located on the short and long arms, respectively, of chromosome 5 at 54.6 cM; the contig bridged the sequence between the short and long arms of chromosome 5 (Figure 1). Because a version of the sequences of two PAC clones (P0587F01, P0697B04) had already been published in draft status, these clones were divided into a number of pieces (12 in the case of AC146339 and 7 for AC137984; Table 1). To obtain more accurate information on Cen5, these PAC clones were resequenced by Sanger-based sequencing technology, reassembled, and finished (see Materials and Methods). Clone P0587F01 was reassembled into one contig and the sequence was submitted to the PLN (plant, fungal, and algal sequences) division of DDBJ.
(52,858 bp, AP011109; Table 1). In the case of P0697B04, all the gaps were filled, but because the center of this clone was occupied by the RCS2/CentO repeats the exact number and orientation of RCS2/CentO repeats were not determined; the sequence was submitted as an incomplete status high-throughput genomic sequence (HTGS)_PHASE2 (147,577 bp, AP011110; Table 1). Cent5 had two different-sized clusters of 155-bp RCS2/CentO satellite repeats (Figure 1). After removing redundant sequences from the regions overlapping between the neighboring PAC/BAC clones, we generated a continuous, high-quality DNA sequence covering the entire region of Cent5. The genomic sequence of Cent5 was integrated into the latest reference genomic sequence of rice constructed by the IRGSP (IRGSP Build 5.0 pseudomolecules).

**IDENTIFICATION OF EXPRESSED REGION BY USING mRNA-SEQ**

We defined pericentromeric regions as recombinational cold spots proximal to RCS2/CentO, as in a previous rice analysis (Wu et al., 2003). A total of six RFLP markers (R448, C1388, S20487S, E3103S, C53260S, and R2059) genetically mapped at 54.6 cM were located in the 2.14-Mb defined as the pericentromeric region of chromosome 5 (Figure 2). A total of five RFLP markers (R288, S2106, C53648S, C1794, and C954) were mapped at 19.6 cM in the 2.09-Mb pericentromeric region of Cent4 (Figure A1A in Appendix); and a total of six RFLP markers (C1374, R2381, E20691S, S21882S, C1115, and R2466) were mapped at 54.3 cM in the 2.43-Mb pericentromeric region of Cent8 (Figure A1B in Appendix).

We compared the averages of gene density, transcriptional density, and expression level in the centromeric region with those in other chromosomal regions. The average gene density in the centromeric region was the lowest in the whole chromosomal region (Figure 3). The average transcriptional density in the centromeric region was lower than that in other chromosomal regions, but the average expression level in the centromeric region was not (Figure 3). Gene expression in the centromeric region was compared by statistical analysis, which was independent of gene annotation. First, transcriptional density was compared. The transcriptional density of Cent5 was 0.070 (shoot) and 0.065 (root), whereas that of the other regions of the same chromosome was 0.168 (shoot) and 0.170 (root); transcriptional density was significantly lower ($P < 0.0001$) in Cent5 than in the average of the other regions by Fisher’s exact test (Table 2). The transcriptional densities in Cent4 and Cent8 were also significantly lower than in the averages of the other regions (Table 2). Second, expression level was compared. The expression level in Cent5 was 234.4 (shoot) and 177.5 (root), whereas that in the other regions was 264.8 (shoot) and 239.5 (root); the expression level in Cent5 was significantly lower than that in the other regions ($P < 0.0001$). However, in Cent4, expression of the gene Os04g0234600 (similar DNA sequence to that encoding sedoheptulose-bisphosphatase) was extremely high in the shoot (Figure A1A in Appendix), resulting in a high average expression level in Cent4 (data not shown). With the exception of the expression of Os04g0234600 in Cent4, expression levels were also significantly lower in Cent4 and Cent8 than in the other regions.

**FIGURE 2 | Distribution of expressed regions in Cent5.** The positions of restriction-fragment-length polymorphism (RFLP) markers mapped at 54.3–55.4 cM are indicated. The region in which RFLP markers are mapped at 54.6 cM is shown (gray box). The distribution of 36-bp mapped reads on the rice genome was graphed in GBrowse (Stein et al., 2002). The graph indicates the average depths of reads from mRNA-Seq for samples obtained from shoots (green) or roots (red). Only depths <50 are shown (Depths ≥50 are shown as 50i). The level of expression is normalized to that of the shoot (standard). Gene models based on Rice Annotation Project (RAP) representative loci (RAP_rep) and RAP predicted genes (RAP_pred) are shown. Expression of an RAP predicted gene is shown (white triangle). The position of Os05g00303000, a homolog of the wheat PSR161 gene mapped on wheat Cen1B (see text), is also indicated (black triangle). Red boxes represent RCS2/CentO clusters.
regions (Table 2). Thus, gene expression (transcriptional density and expression level) was significantly lower in the centromeric region than in the other regions.

We also compared transcription in the short and long arms in the pericentromeric regions. In Cen5, transcriptional density was 0.039 (shoot) and 0.035 (root) on the short arm and 0.110 (shoot), 0.103 (root) on the long arm. Transcriptional density was significantly ($P < 0.0001$) lower on the short arm than on the long arm by Fisher’s exact test (Table 3); the distribution of transcriptional density was asymmetric in Cen5. The expression level of Cen5 in shoots was significantly ($P < 0.0001$) lower on the short arm than on the long arm, whereas the expression level of Cen5 in roots was significantly ($P < 0.0001$) lower on the long arm than on the short arm (Table 3). Thus, the distribution of expression level of Cen5 was asymmetric, but the tendency was in the opposite directions in the shoots and roots.

**CHARACTERIZATION OF GENES EXPRESSED IN Cen5**

The annotated genes in Cen5 were characterized by using the Rice Annotation Project Database (RAP-DB; Rice Annotation Project, 2008); 28 genes were annotated in the pericentromeric region on the short arm of Cen5 (~1.06 Mb), whereas 45 genes were annotated on the long arm (~0.97 Mb; Table A1 in Appendix; Table 3). On the short arm close to RCS2/CentO (C1388 to S20487S), most of the genes encoding hypothetical proteins were hardly expressed (Table A1 in Appendix). On the long arm, genes encoding proteins similar to transcription factor IIA large subunit (Os05g0292200), acetyl-coenzyme A carboxylase (Os05g0295500), glyoxalase I (Os05g0295800), and zinc-finger-like protein (Os05g0299700) were expressed at relatively high levels (RPKM > 20; Table A1 in Appendix) in both shoots and roots. Analysis of the mapped reads also gave evidence of the expression of genes computationally predicted by the RAP (Figure 2). A non-protein-coding transcript (Os05g0296600) was also expressed (Table A1 in Appendix). Most of the genes highly expressed on the long arm were similar to genes encoding functional – not hypothetical – proteins.

The distribution of transcription of each gene was identified by using Illumina mRNA-Seq technology. We adopted the RPKM (reads per kilobase of exon models per million mapped reads) method (Mortazavi et al., 2008) for transcript quantification on the basis of the number of sequence reads mapped on each gene. The RPKM and signal intensity from microarray analysis of the same RNA materials as used in this study had been compared previously; these two independent measures of transcript abundance were correlated ($r = 0.75–0.77$; Mizuno et al., 2010). Dot plot analysis of the RPKM and the chromosomal position of each gene suggested that gene expression was low in the centromeric regions (Figure 4).

A putative gene conserved in the rice centromere and wheat centromere was found: Os05g0303000 was mapped only 90 kb distal to the marker R2059 on Cen5 and was highly expressed in shoots and roots (Figure 2). Os05g0303000 had 82.6% DNA sequence identity to PSR161 (data not shown). PSR161 is the only actively transcribed gene express of the genes highly expressed on the long arm were similar to genes encoding functional – not hypothetical – proteins.

The distribution of transcription of each gene was identified by using Illumina mRNA-Seq technology. We adopted the RPKM (reads per kilobase of exon models per million mapped reads) method (Mortazavi et al., 2008) for transcript quantification on the basis of the number of sequence reads mapped on each gene. The RPKM and signal intensity from microarray analysis of the same RNA materials as used in this study had been compared previously; these two independent measures of transcript abundance were correlated ($r = 0.75–0.77$; Mizuno et al., 2010). Dot plot analysis of the RPKM and the chromosomal position of each gene suggested that gene expression was low in the centromeric regions (Figure 4).

A putative gene conserved in the rice centromere and wheat centromere was found: Os05g0303000 was mapped only 90 kb distal to the marker R2059 on Cen5 and was highly expressed in shoots and roots (Figure 2). Os05g0303000 had 82.6% DNA sequence identity to PSR161 (data not shown). PSR161 is the only actively transcribed gene that has been mapped on the functional centromere of wheat chromosome 1B (Francki et al., 2002), suggesting that the location of this homolog is conserved in rice Cen5 and wheat Cen1B.

**DISCUSSION**

**GENE EXPRESSION IN PERICENTROMERIC REGIONS**

To assess the functional importance of gene expression in the centromeric region, we performed genomic sequencing of Cen5 (Figure 1; Table 1) and expression analysis (Figure 2). Gene expression (transcriptional density and expression level) was significantly lower in the pericentromeric regions of Cen4, Cen5, and Cen8 than in the other regions (Table 2; Figures 3 and 4). Low transcriptional density was asymmetric in Cen5; the distribution of transcriptional density was asymmetric in Cen5. The expression level of Cen5 in shoots was significantly ($P < 0.0001$) lower on the long arm than on the short arm, whereas the expression level of Cen5 in roots was significantly ($P < 0.0001$) lower on the long arm than on the short arm (Table 3). Thus, the distribution of expression level of Cen5 was asymmetric, but the tendency was in the opposite directions in the shoots and roots.

**FIGURE 3 | Relationships of gene density, transcriptional density, and expression levels.** Average scores of gene density, transcriptional density, and expression level in each 1 Mb of sliding windows on chromosomes 4, 5, and 8 are shown. Horizontal axis indicates position on the reference rice genome (Build 5.0 pseudomolecules) constructed by the International Rice Genome Sequencing Project. Lines indicate the boundaries of each pericentromeric region.
Table 2 | Comparison of transcription in centromeric regions and in the whole genomic region.

| Genomic region (bp) | Tissue | Transcribed region (bp) | No. of uniquely mapped reads | Transcriptional density | Expression level |
|---------------------|--------|-------------------------|----------------------------|------------------------|-----------------|
|                     | Centromere | Other | Centromere | Other | Centromere | Other | Centromere | Other | Centromere | Other | Centromere | Other | P      | Centromere | Other | P      |
| Cen4                | 2,088,655  | 33,973,212 | 115,495  | 5,005,925 | 48,875  | 1,368,096 | 0.055  | 0.147 | <0.0001 | 214.7 | 273.3 | <0.0001 |
|                     |         |         | 101,453  | 5,096,940 | 15,409  | 1,084,734 | 0.049  | 0.150 | <0.0001 | 151.9 | 212.8 | <0.0001 |
| Cen5                | 2,139,098 | 27,934,342 | 149,160  | 4,687,562 | 34,964  | 1,241,332 | 0.070  | 0.168 | <0.0001 | 234.4 | 264.8 | <0.0001 |
|                     |         |         | 138,618  | 4,758,232 | 24,607  | 1,139,487 | 0.065  | 0.170 | <0.0001 | 177.5 | 239.5 | <0.0001 |
| Cen8                | 2,431,594 | 26,098,435 | 123,866  | 3,862,076 | 36,813  | 1,197,700 | 0.095  | 0.148 | <0.0001 | 158.8 | 310.1 | <0.0001 |
|                     |         |         | 239,917  | 3,843,450 | 35,396  | 787,908  | 0.099  | 0.147 | <0.0001 | 147.5 | 205.0 | <0.0001 |

Statistical significance (P) was based on Fisher’s exact test. Expression levels in the centromeric region of chromosome 4 were calculated without the gene Os04g0234600 (see text). The centromeric region was defined as from the start position of the short arm of the pericentromeric region to the end position of the long arm of the pericentromeric regions. Transcribed region, transcriptional density, and expression level are defined in Section “Materials and Methods.”

Table 3 | Comparison of transcription in RCS2/CentO core region and pericentromeric regions.

| Genomic region (bp) | Tissue | Transcribed region (bp) | No. of uniquely mapped reads | Transcriptional density | Expression level |
|---------------------|--------|-------------------------|----------------------------|------------------------|-----------------|
|                     | Pericent. short arm | RCS2/ CentO | Percent. long arm | Pericent. short arm | RCS2/ CentO | Percent. long arm | Pericent. short arm | RCS2/ CentO | Percent. long arm | Pericent. short arm | RCS2/ CentO | Percent. long arm | P | Percent. short arm | RCS2/ CentO | Percent. long arm | P | P      |
| Cen4                | 1,779,938  | 124,271  | 184,446  | Shoot | 88,498  | 140  | 26,857  | 40,984  | 5  | 7896  | 0.050  | 0.001  | 0.146 | <0.0001 | 463.1 | 35.7 | 293.6 | <0.0001 |
|                     |         |         |         | Root | 75,230  | 288  | 25,935  | 10,869  | 10  | 4,500 | 0.042  | 0.002  | 0.141 | <0.0001 | 144.5 | 34.7 | 174.7 | <0.0001 |
| Cen5                | 1,063,874 | 97,181  | 978,043  | Shoot | 41,629  | 0  | 107,531 | 4,332  | 0  | 30,632 | 0.039  | 0.000  | 0.110 | <0.0001 | 104.1 | 0.0  | 284.9 | <0.0001 |
|                     |         |         |         | Root | 37,507  | 0  | 101,111 | 9,725  | 0  | 14,882 | 0.035  | 0.000  | 0.103 | <0.0001 | 259.3 | 0.0  | 1472  | <0.0001 |
| Cen8                | 935,763  | 76,165  | 1,419,666 | Shoot | 94,769  | 178  | 136,919 | 14,495  | 5  | 22,313 | 0.101  | 0.002  | 0.096 | <0.0001 | 153.0 | 28.1 | 163.0 | <0.0001 |
|                     |         |         |         | Root | 95,763  | 176  | 143,978 | 14,069  | 5  | 21,322 | 0.102  | 0.002  | 0.101 | 0.496  | 146.9 | 28.4 | 148.1 | 0.0224 |

Statistical significance of the difference in gene expression between the short arm and long arm (P) was based on Fisher’s exact test. Transcribed region, transcriptional density, and expression level are defined in Section “Materials and Methods.”
density could be partly explained by the low gene density (Figure 3), as centromeric regions contain repetitive sequences such as the centromere-specific retrotransposon RIRE7/CRR and the tandem repetitive sequence RCS2/CentO. The high expression observed only under specific conditions (e.g., of Os04g0234600 in shoots, Figure A1A in Appendix) could be explained by the occurrence of permissive transcriptional activity through pockets of DNA hypomethylation (Wong et al., 2006) and/or mosaics of histone modification in the centromeric region (Stimpson and Sullivan, 2010): the presence of methylated histone H3 at Lys9 leads to heterochromatin assembly, whereas methylated histone H3 at Lys4 leads to euchromatin assembly. Thus, gene expression was generally low in the centromeric region, but the suppression could be selectively released in specific tissues and under specific cell conditions.

The distribution of gene expression was asymmetric in Cen5; genes were rarely expressed on the short arm and highly expressed on the long arm (Figure 2; Table 3). The size of the rarely expressed region C1388 to S20487S (∼700 kb; Figure 2) was almost the same as that of the kinetochore region on Cen8 (750 kb; Nagaki et al., 2004; Wu et al., 2004), suggesting that these rarely expressed gene regions are related to the formation of kinetochores in Cen5. In the 700-kb region, most of the genes were annotated as hypothetical and were hardly expressed (Table A1 in Appendix), suggesting that these genes do not have specific functions. On the long arm of Cen5, genes with similarity to those encoding known functional proteins were highly expressed (RPKM > 20; Table A1 in Appendix); the statistical median of the RPKM for all RAP2 annotated genes was 3.399 in the shoots and 4.241 in the roots (Mizuno et al., 2010). Moreover, rice Os05g0303000 had a DNA sequence similar to that of wheat PSR161. Os05g0303000 and PSR161 have been mapped in the centromeric regions of rice Cen5 (Figure 2) and wheat Cent1B (Francki et al., 2002), respectively; their chromosomal positions are consistent with the chromosomal synteny between these two crops (Devos, 2005). The results of application of a molecular–cytogenetic method have also suggested synteny between the centromeric regions of wheat and rice (Qi et al., 2009). PSR161 encodes HSP70, which is thought to function as a molecular chaperone. As HSP70 is also conserved in Pisum sativum, Cucumis sativus, Spinacia oleracea, and Chlamydomonas reinhardtii (Francki et al., 2002), HSP70 gene silencing is likely to have serious effects. Therefore, because of the existence of highly expressed regions proximal to RCS2/CentO on the long arm, including the conserved HSP70 homolog, we consider that kinetochore formation on Cen5 on an evolutionary time scale was restricted to the short arm.

The RCS2/CentO sequence is tandemly arrayed in the core region of Cen5. The length of a unit of rice RCS2/CentO is 155 bp (Dong et al., 1998); this length is considered to be related to the formation of the nucleosomal unit required for kinetochore formation (Houben and Schubert, 2003; Dawe and Hiatt, 2004; Ma et al., 2007). Cen5 had two clusters of RCS2/CentO repeats (Figure A2 in Appendix). In comparison, Cen8 has three large clusters (Wu et al., 2004) and Cen4 has 18 clusters (Zhang et al., 2004); thus the amount and organization of RCS2/CentO clusters differ markedly among Cen4, Cen5, and Cen8 (Figure A2 in Appendix). No genes were annotated (Figure A2 in Appendix), and expression was hardly detected, in the sequence separating the RCS2/CentO arrays (Table 2), suggesting that gene expression did not occur in the core region of the centromeric region. The sequences separating the RCS2/CentO array are derived from repetitive sequences, such as the centromere-specific gypsy-like retrotransposon RIRE7 (Kumekawa et al., 2001), that are fragmented and have nucleotide substitutions (Wu et al., 2004; Zhang et al., 2004). Even though Cen8 has other small RCS2/CentO sequences that have the Os08g0319450 gene within the RCS2/CentO array, Os08g0319450
was not expressed in the shoots or roots (Figure A1B in Appendix). Therefore, the region separating the RCS2/CentO array had little expression activity.

REMAINING GAP IN THE REFERENCE RICE GENOME SEQUENCE
The published rice genomic sequence covers 95.3% of the estimated 390-Mb total genome sequence, and it contains 36 gaps (IRGSP, 2005). The 36 gaps have been gradually sequenced since the completion of the IRGSP. This sequencing has included telomeres, subtelomeres, and the ribosomal DNA cluster (Mizuno et al., 2008a). However, the latest rice genomic sequence contains only a portion of the centromeric regions. Here, we performed resequencing, assembly, and finishing of PAC clones in rice Cen5 (Figure 1; Table 1). In the remaining centromeric regions of rice chromosomes, interference by repetitive sequences has prevented further chromosome walking and subsequent genomic sequencing (Wu et al., 2003; IRGSP, 2005). In an in situ hybridization analysis, unsequenced centromeres had relatively large clusters of repetitive sequences (Cheng et al., 2002). Moreover, RCS2/CentO repetitive DNA inserted into PAC/BAC clones is easily deleted: 47.2% of centromeric PAC clones have inserts <60 kb in length, compared with 13.6% in the total library (Mizuno et al., 2006), suggesting that these clones are unstable in Escherichia coli. Thus, complete genomic sequencing of the remaining centromeric regions will be a challenging problem.

Our work has primarily helped to bridge the short arm and long arm of chromosome 5 of the reference rice genome sequence constructed by the IRGSP. By using the reference genomic sequence, massive parallel sequencing of mRNA was used to generate transcript maps. Recently, the massive parallel sequencing technique has also been applied to the analysis of DNA methylation, histone modification, and protein binding. Thus, high-quality reference genomic sequencing will play pivotal roles in further sequence-based functional analysis of centromeric regions in the next-generation sequencing era.

ACKNOWLEDGMENTS
The authors thank F. Aota, K. Ohtsu, and K. Yamada for their technical assistance in sample preparation. This work was supported by the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics for Agricultural Innovation, RTR-0001).

REFERENCES
Amiriraj, J. S., Luo, M., Goicoechea, J. L., Wang, W., Kudrna, D., Mueller, C., Talagi, J., Kim, H., Sisneros, N. B., Blackmon, B., Fang, E., Tomkins, J. B., Brar, D., Mackill, D., McCouch, S., Kurata, N., Lambert, G., Galbraith, D. W., Arumuganathan, K., Rao, K., Walling, J. G., Gill, N., Yu, Y., SanMiguel, P., Soderlund, C., Jackson, S., and Wing, R. A. (2006). The Orzya bacterial artificial chromosome library resource: construction and analysis of 12 deep-coverage large-insert BAC libraries that represent the 10 genome types of the genus Oryza. Genome Res. 16, 140–147.
Blower, M. D., Sullivan, B. A., and Karpen, G. H. (2002). Conserved organization of centromeric chiasmata in flies and humans. Dev. Cell 2, 319–330.
Buscaino, A., Allshire, R., and Pidoux, A. (2010). Building centromeres: home sweet home or a nomadic existence? Curr. Opin. Genet. Dev. 20, 118–126.
Cheng, C. H., Chung, M. C., Liu, S. M., Chen, S. K., Kao, E. Y., Lin, S. J., Hisiao, S. H., Tseng, I. C., Hsing, Y. L., Wu, H.-P., Chen, C. S., Shaw, J. F., Wu, J., Matsumoto, T., Sasaki, T., Chen, H. H., and Choy, T. Y. (2005). A fine physical map of the rice chromosome 5. Mol. Genet. Genomics 274, 337–345.
Cheng, Z., Dong, F., Langdon, T., Ouyang, S., Buell, C. R., Gu, M., Blattner, F. R., and Jiang, J. (2002). Functional rice centromeres are marked by a satellite-repeat and a centrome-specific retrotransposon. Plant Cell 14, 1691–1704.
Dawe, R. K., and Hiatt, E. N. (2004). Plant neocentromeres: fast, focused, and driven. Chromosome Res. 12, 655–669.
Devis, K. M. (2005). Updating the “crop circle”. Curr. Opin. Plant Biol. 8, 155–162.
Dong, F., Miller, I. T., Jackson, S. A., Wang, G. L., Ronald, P. C., and Jiang, J. (1998). Rice (Oryza sativa) centromeric regions consist of complex DNA. Proc. Natl. Acad. Sci. U.S.A. 95, 8135–8140.
Francki, M. G., Berzonsky, W. A., Ohm, H. W., and Anderson, J. M. (2002). Physical location of a HSP70 gene homologe on the centromere of chromosome 1B of wheat (Triticum aestivum L.). Theor. Appl. Genet. 104, 184–191.
Gill, N., Hans, C. S., and Jackson, S. (2008). An overview of plant chromosome structure. Cytogenet. Genome Res. 130, 192–201.
Hall, A. E., Keith, K. C., Hall, S. E., Copenhagen, G. P., and Preuss, D. (2004). The rapidly evolving field of plant centromeres. Curr. Opin. Plant Biol. 7, 108–114.
Henikoff, S., Ahmad, K., and Malik, H. S. (2001). The centromere paradox: stable inheritance with rapidly evolving DNA. Science 293, 1098–1102.
Hosouchi, T., Kumezawa, N., Tsurooka, H., and Kotani, H. (2002). Physical map-size of the centromeric regions of Arabidopsis thaliana chromos-omes 1, 2, and 3. DNA Res. 9, 117–121.
Houben, A., and Schubert, I. (2003). DNA and proteins of rice centromeres. Curr. Op. Plant Biol. 6, 554–560.
IRGSP (2005). The map-based sequence of the rice genome. Nature 436, 793–800.
Kumezawa, N., Ohmido, N., Fukui, K., Ohtsubo, E., and Ohtsubo, H. (2001). A new gypsy-type retrotransposon, RIRE7: preferential insertion into the tandem repeat sequence TrsD in pericentromeric heterochromatin regions of rice chromosomes. Mol. Genet. Genomics 265, 480–488.
Lamb, J. C., Yu, W., Han, F., and Bircher, J. A. (2007). Plant chromosomes from end to end: telomeres, heterochromatin and centromeres. Curr Opin. Plant Biol. 10, 116–122.
Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 10, R25.
Ma, J., Wing, R. A., Benetzen, J. L., and Jackson, S. A. (2007). Plant centromere organization: a dynamic structure with conserved functions. Trends Genet. 23, 134–139.
Mizuno, H., Ito, K., Wu, J., Tanaka, T., Kanamori, H., Katayose, Y., Sasaki, T., and Matsumoto, T. (2006). Identification and mapping of expressed genes, simple sequence repeats and transposable elements in centromeric regions of rice chromosomes. DNA Res. 13, 267–274.
Mizuno, H., Kawahara, Y., Sakai, H., Kanamori, H., Wakimoto, H., Yamagata, H., Oono, Y., Wu, J., Ikawa, H., Itoh, T., and Matsumoto, T. (2010). Massive parallel sequencing of mRNA in identification of unannotated salinity stress-inducible transcripts in rice (Oryza sativa L.). BMC Genomics 11, 683. doi: 10.1186/1471-2164-11-683.
Mizuno, H., Sasaki, T., and Matsumoto, T. (2008a). Characterization of internal structure of the nucleolar organizing region in rice (Oryza sativa L.). Cytogenet. Genome Res. 121, 282–285.
Mizuno, H., Wu, J., Katayose, Y., Kanamori, H., Sasaki, T., and Matsumoto, T. (2008b). Characterization of chromo-
some ends on the basis of the structure of TrsA subtelomeric repeats in rice (Oryza sativa L.). Mol. Genet. Genomics 280, 19–24.
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008). Mapping and quantifying mamma-
lian transcriptomes by RNA-Seq. Nat. Methods 5, 621–628.
Nagaki, K., Cheng, Z., Ouyang, S., Talbert, P. B., Kim, M., Jones, K. M., Henikoff, S., Buell, C. R., and Jiang, J. (2004). Sequencing of a rice centromere uncovers active genes. Nat. Genet. 36, 138–145.
Qi, L., Friebe, B., Zhang, P., and Gill, B. S. (2009). A molecular-cytogenetic method for locating genes to pericen-
tric regions facilitates a genomewide comparison of synteny between the centromeric regions of wheat and rice. Genetica 183, 1233–1247.
Rice_Annotation_Project. (2008). The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 36, D1028–D1033.
Sharma, S., and Raina, S. N. (2005). Organization and evolution of highly repeated satellite DNA sequences in plant chromosomes. Cytogenet. Genome Res. 109, 15–26.
Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., and Lewis, S. (2002). The genomic gateway browser: a building block for a model organism system database. Genome Res. 12, 1599–1610.
Stimpson, K. M., and Sullivan, B. A. (2010). Epigenomics of centromere
asymmetric distribution of gene expression in the centromeric region of rice chromosome 5. Front. Plant Sci. 2:16. doi: 10.3389/fpls.2011.00016

This article was submitted to Frontiers in Plant Genetics and Genomics, a specialty of Frontiers in Plant Science.

Copyright © 2011 Mizuno, Kawahara, Wu, Katayose, Kanamori, Ikawa, Itoh, Sasaki and Matsumoto. This is an open-access article subject to a non-exclusive license between the authors and Frontiers Media SA, which permits use, distribution and reproduction in other forums, provided the original authors and source are credited and other Frontiers conditions are complied with.
FIGURE A1 | Distribution of expressed regions proximal to RCS2/CentO in Cen4 and Cen8. The distributions of reads mapped on Cen4 (A) and Cen8 (B) were graphed in GBrowse, as in Figure 2. Os04g0234600 in Cen4 is extremely highly expressed (black triangle). Os08g0319450 in Cen8 is located in the small RCS2/CentO sequence.
FIGURE A2 | Distributions of RCS2/CentO clusters. RCS2/CentO clusters in Cen4, Cen5, and Cen8 on rice genome sequence Build 5.0 (blue boxes) are shown. The small RCS2/CentO sequence in Cen8 (Figure A1B in Appendix) is not shown. Gene models based on Rice Annotation Project (RAP) representative genes are shown.
### Table A1 | Annotated genes in Cen5.

| Gene_ID | S/L | Start   | End     | Length | Strand | Description                                                                 |
|---------|-----|---------|---------|--------|--------|----------------------------------------------------------------------------|
| Os05g0276500 | S  | 11422127| 11423907| 1101   | −      | Expansin Os-EXPA3                                                          |
| Os05g0277000 | S  | 1144481 | 11448493| 354    | −      | Similar to Expansin Os-EXPA3                                              |
| Os05g0277200 | S  | 11463176| 11464421| 1246   | +      | Conserved hypothetical protein                                             |
| Os05g0277300 | S  | 11465333| 11469546| 3295   | −      | Similar to cDNA clone: 001-013-F11                                         |
| Os05g0277350 | S  | 11474153| 11475272| 691    | +      | Similar to leucine rich repeat family protein                              |
| Os05g0277500 | S  | 11496173| 11497090| 840    | +      | Similar to germin-like protein subfamily 2 member 4 precursor             |
| Os05g0277800 | S  | 11638503| 11644034| 1551   | −      | Transferase family protein                                                 |
| Os05g0277850 | S  | 11643456| 11644083| 628    | +      | Hypothetical gene                                                          |
| Os05g0278950 | S  | 11670044| 11672821| 715    | −      | Similar to ATP-dependent Clp protease proteolytic subunit                  |
| Os05g0279300 | S  | 11678665| 11685603| 1209   | −      | Similar to RNA pseudouridine synthase A                                    |
| Os05g0279400 | S  | 11689568| 11695386| 3310   | +      | Zinc-finger, RING-type domain containing protein                           |
| Os05g0279600 | S  | 11700491| 11709989| 1528   | +      | Endonuclease/exonuclease/phosphatase domain containing protein             |
| Os05g0279750 | S  | 11721971| 11726153| 4183   | +      | Hypothetical gene                                                          |
| Os05g0279900 | S  | 11728764| 11731789| 1475   | +      | Similar to Polygalacturonase A                                             |
| Os05g0280200 | S  | 11752678| 11754728| 672    | −      | Similar to Ras-related protein RGP2                                         |
| Os05g0280350 | S  | 11752728| 11754718| 990    | +      | Hypothetical gene                                                          |
| Os05g0280500 | S  | 11782293| 11785389| 1156   | −      | Phospholipid/glycerol acyltransferase domain containing protein            |
| Os05g0280700 | S  | 11817709| 11820877| 3169   | −      | Similar to resistance protein candidate                                     |
| Os05g0281400 | S  | 11920597| 11921702| 1013   | +      | Protein of unknown function DUF810 domain containing protein              |
| Os05g0282500 | S  | 12041129| 12043113| 600    | −      | Hypothetical conserved gene                                                |
| Os05g0282900 | S  | 12079928| 12081735| 1808   | +      | Conserved hypothetical protein                                              |
| Os05g0283000 | S  | 12088257| 12091692| 1435   | +      | Conserved hypothetical protein                                              |
| Os05g0283200 | S  | 12098520| 12099575| 1056   | +      | Pectinesterase inhibitor domain containing protein                        |
| Os05g0283600 | S  | 12122939| 12131356| 3569   | +      | Zinc-finger, CCHC-type domain containing protein                           |
| Os05g0285900 | S  | 12322935| 12327534| 1162   | +      | Conserved hypothetical protein                                              |
| Os05g0286100 | S  | 12337263| 12338299| 1037   | +      | Similar to zinc-finger protein KNUCKLES                                     |
| Os05g0286200 | S  | 12353858| 12356702| 744    | +      | Conserved hypothetical protein                                              |
| Os05g0287800 | S  | 12482678| 12486801| 1445   | +      | Conserved hypothetical protein                                              |
| Os05g0289100 | L  | 12601354| 12602492| 1058   | +      | Hypothetical conserved gene                                                |
| Os05g0289400 | L  | 12630181| 12635126| 2945   | −      | Similar to CRN (Crooked neck) protein                                       |
| Os05g0289700 | L  | 12650476| 12651976| 1395   | +      | Arbuscular mycorrhizal specific marker                                     |
| Os05g0290300 | L  | 12704171| 12705720| 1219   | −      | Hypothetical conserved gene                                                |
| Os05g0290400 | L  | 12704190| 12715035| 2635   | +      | Hypothetical gene                                                          |
| Os05g0291600 | L  | 12860254| 12860794| 541    | +      | Hypothetical conserved gene                                                |
| Os05g0291700 | L  | 12862505| 12868432| 1316   | −      | Similar to PTAC16                                                          |
| Os05g0291800 | L  | 12872863| 12873468| 526    | +      | Similar to predicted protein                                                |
| Os05g0292200 | L  | 12995006| 12901403| 1630   | +      | Similar to Transcription factor IIA large subunit (TFIIF-A-L1)             |

(Continued)
| Gene_ID     | S/L | Start       | End         | Length | Strand | Description                                                                 | RPKM_shoot | RPKM_root |
|------------|-----|-------------|-------------|--------|--------|------------------------------------------------------------------------------|------------|-----------|
| Os05g0293500 | L   | 12962105    | 12967380    | 1237   | –      | Similar to Pectate lyase B                                                  | 0          | 0         |
| Os05g0293600 | L   | 12978536    | 12984017    | 5482   | +      | Similar to RNA polymerase beta’ chain                                       | 0          | 0         |
| Os05g0294600 | L   | 13018766    | 13021491    | 2425   | –      | Pentatricopeptide repeat domain containing protein                          | 14.97      | 2.73      |
| Os05g0294800 | L   | 13035304    | 13039195    | 2262   | +      | Hypothetical gene                                                           | 10.52      | 10.5      |
| Os05g0295100 | L   | 13056572    | 13075697    | 2031   | +      | Hypothetical conserved gene                                                 | 0.99       | 2.73      |
| Os05g0295200 | L   | 13086136    | 13089196    | 2181   | –      | Conserved hypothetical protein                                              | 10.32      | 1.34      |
| Os05g0295300 | L   | 13093233    | 13094329    | 952    | –      | Similar to acetyl-coenzyme A carboxylase                                    | 40.12      | 45.31     |
| Os05g0295700 | L   | 1317580     | 13121926    | 2251   | –      | Similar to homoserine dehydrogenase-like protein                            | 10.22      | 11.75     |
| Os05g0295800 | L   | 13212310    | 13217786    | 1052   | –      | Similar to glyoxalase I                                                    | 39.12      | 36.36     |
| Os05g0295900 | L   | 13135652    | 13144818    | 3064   | –      | Conserved hypothetical protein                                              | 0.71       | 2.97      |
| Os05g0296000 | L   | 13169380    | 13171753    | 2374   | +      | Conserved hypothetical protein                                              | 0          | 0         |
| Os05g0296600 | L   | 13216923    | 13217232    | 310    | +      | Non-protein coding transcript                                               | 23.77      | 62.26     |
| Os05g0296700 | L   | 13221667    | 13222206    | 540    | –      | Similar to small heat shock protein                                         | 3.62       | 3.24      |
| Os05g0296750 | L   | 13221730    | 13223528    | 623    | +      | Hypothetical gene                                                           | 3.23       | 2.34      |
| Os05g0296800 | L   | 13226261    | 13226572    | 897    | –      | Hypothetical protein                                                        | 0.31       | 0.32      |
| Os05g0296900 | L   | 13259004    | 13259727    | 508    | –      | Conserved hypothetical protein                                              | 0          | 0         |
| Os05g0297001 | L   | 13261758    | 13263921    | 2164   | +      | Similar to predicted protein                                                | 0          | 0         |
| Os05g0297300 | L   | 13287199    | 13288934    | 1736   | +      | Protein of unknown function DUF1618 domain containing protein              | 0          | 0         |
| Os05g0297400 | L   | 13289996    | 13290998    | 992    | –      | Similar to CXIP4                                                            | 0          | 0         |
| Os05g0297800 | L   | 13304340    | 13307779    | 2408   | –      | Conserved hypothetical protein                                              | 0.77       | 0.21      |
| Os05g0297850 | L   | 13309305    | 13309728    | 424    | –      | Hypothetical conserved gene                                                 | 0          | 0         |
| Os05g0297900 | L   | 13311413    | 13315153    | 1034   | +      | Similar to signal peptidase 18 subunit                                      | 9.67       | 17.76     |
| Os05g0298200 | L   | 13337235    | 13341454    | 2401   | +      | Ankyrin repeat containing protein                                           | 14.93      | 9.83      |
| Os05g0298600 | L   | 13349202    | 13351414    | 2213   | –      | Hypothetical conserved gene                                                 | 3.43       | 5.1       |
| Os05g0298700 | L   | 13357011    | 13359346    | 1220   | –      | Similar to xylan endohydrolase isoenzyme X-I                               | 0          | 0         |
| Os05g0298800 | L   | 13395955    | 13396672    | 718    | +      | Conserved hypothetical protein                                              | 6.84       | 14.11     |
| Os05g0299000 | L   | 13400919    | 13401654    | 736    | +      | Hypothetical protein                                                        | 0.08       | 0.1       |
| Os05g0299101 | L   | 13402647    | 13403283    | 550    | –      | Hypothetical gene                                                           | 0.41       | 0         |
| Os05g0299200 | L   | 13407527    | 13412497    | 1491   | –      | Hypothetical conserved gene                                                 | 10.04      | 2.98      |
| Os05g0299300 | L   | 13414154    | 13420043    | 3226   | –      | WD40 repeat-like domain containing protein                                  | 4.48       | 5.13      |
| Os05g0299500 | L   | 13434338    | 13439817    | 1563   | +      | Protein of unknown function DUF914                                          | 6.64       | 15.66     |
| Os05g0299600 | L   | 13440049    | 13442337    | 2171   | –      | Protein of unknown function DUF1677                                         | 1.18       | 0.67      |
| Os05g0299700 | L   | 13450657    | 13453015    | 2359   | –      | Similar to expressed protein (zinc-finger-like protein)                     | 38.25      | 39.38     |
| Os05g0300700 | L   | 13504825    | 13512070    | 2425   | +      | Cell division cycle-associated protein domain containing protein           | 9.19       | 16.98     |
| Os05g0301500 | L   | 13558563    | 13563304    | 2162   | +      | Similar to ribophorin I                                                    | 18.88      | 33.4      |

R2059

**Genes located between restriction-fragment-length polymorphism (RFLP) markers R448 and R2059 on chromosome 5 are listed.** Gene ID (gene_ID); mapped on short arm or long arm (S/L); start position (start); end position (end); total nucleotide length of each transcript (length); coding strand (strand); description in Rice Annotation Project Database (description); RPKM in shoot (RPKM_shoot); and RPKM in root (RPKM_root) are listed. The position of RFLP markers and RCS2/CentO repeats are also shown in bold letter.