Comparative analysis of the transcriptomes of two rice subspecies reveals differentially expressed genes associated with phenotypic differences and reproductive isolation

hongbo pang
Shenyang Normal University https://orcid.org/0000-0002-1372-7685
Junrui Wang
Chinese Academy of Agricultural Sciences institute of crop sciences
Qiang Chen
Shenyang Normal University
Jiaqi Li
Shenyang Normal University
Yueying Li
Shenyang normal University
Longkun Wu
Shenyang Normal University
Qingwen Yang
Chinese Academy of Agricultural Science institute of crop sciences
Xiaoming Zheng (zhengxiaoming@caas.cn)
https://orcid.org/0000-0001-8343-7846

Original article

Keywords: Rice subspecies, Transcriptome, Differentially expressed genes, Functional annotation

Posted Date: December 31st, 2019

DOI: https://doi.org/10.21203/rs.2.19655/v1

Abstract

Background: Rice has been used as a model plant to study adaptation, genome evolution and reproductive isolation among species and the genetics and evolution of complex traits. Two subspecies of cultivated rice, *Oryza sativa* ssp. *indica* and *O. sativa* ssp. *japonica*, with reproductive isolation and differences in morphology and phenotypic differences, were established during the process of rice domestication.

Results: To understand how domestication has changed the transcriptomes of the two rice subspecies and given rise to the phenotypic differences, we obtained approximately 700 Gb RNA-Seq data from 26 *indica* and 25 *japonica* plants, and identified 97,005 transcribed fragments and 7702 novel transcriptionally active regions. We also identified 1857 (4.58% in all genes) differentially expressed genes (DEGs) between *indica* and *japonica* rice. According to previous population genetic analyses, these DEGs may associate with the phenotypic differences between the two subspecies. Functional annotation of these DEGs indicates that they are involved in cell wall biosynthesis and reproductive processes. Furthermore, compared with the non-DEGs, the DEGs from both subspecies had more 5' flanking regions with low polymorphism to divergence ratios, indicating a stronger positive selection pressure on the regulation of the DEGs.

Conclusion: This study improves our understanding of the rice genome by comparatively analyzing the transcriptomes of *indica* and *japonica* rice and identifies DEGs that may be responsible for the reproductive isolation and phenotypic differences between the two rice subspecies.

Background

In addition to being a staple food that feeds over 50% of the world's population, rice has also been used as a model plant for molecular, genetics and comparative genomic studies (Khush 1997; Sasaki and Burr 2000; Huang et al. 2012). Two major rice subspecies, Oryza sativa ssp. indica and O. sativa ssp. Japonica, were established during the long process of rice domestication. These two rice subspecies are different in terms of yield, grain quality and stress resistance (Vaughan et al. 2003; Garris et al. 2005; Zhu et al. 2007; Huang et al. 2012; Wang et al. 2014). Indica rice is mainly cultivated in tropical and subtropical regions, whereas japonica rice is planted mainly in temperate regions or at higher altitudes in tropical and subtropical regions. During the process of domestication and breeding, these two subspecies have evolved characteristic morphological and agronomic traits that may contribute to intraspecific phenotypic adaptations (Vaughan et al. 2003; Wu et al. 2003; Zheng et al. 2016). The morphological features, including leaf color, seed size and apiculus hair length, cannot be used alone to definitively distinguish between the two subspecies due to the presence of morphological variants (Kato 1928; Matsuo 1997; Oka 1988). According to a previous report, the hybrid progenies of these two subspecies are sterile, and it is difficult to utilize the heterosis of their hybrids (Yang et al. 2012). However, the molecular mechanisms underlying the reproductive isolation and phenotypic differences remain largely unknown.

Thus far, there have been many studies on the molecular basis of the phenotypic differences between the two subspecies using different methods (Beukert et al. 2017; Yuan et al. 2017). Among the methods used to search for candidate genes from hybrids of indica and japonica rice, the most important one is the quantitative trait locus (QTL) (Huang et al. 2010; Zheng et al. 2017). Using this method, several studies have managed to isolate and characterize candidate genes that are expressed in key tissues or at certain developmental stages from the two subspecies, indicating that gene expression variations contribute greatly to the phenotypic differences between the two subspecies (Lu et al. 2010; Liu et al. 2010; Jung et al. 2013; Sato et al. 2013; Yang et al. 2014; Horiuchi et al. 2015). Although there have been a number of transcriptomic studies and microarray analyses in several rice varieties, there are currently no transcriptomic studies carried out with multiple rice plants to establish a statistically significant genotype-phenotype correlation (Guo et al. 2016). It has been suggested that nucleotide diversity in gene regulatory regions has a major impact on the expression of genes associated with the phenotypic differences between the two subspecies. In addition, it has been speculated that these gene regulatory regions may have evolved under positive selection pressure (Jones et al. 2012; Nosil and Feder 2012; Hanikenne et al. 2013; Guo et al. 2016). Although the gene-coding regions have been compared between indica and japonica rice, there is no method to effectively investigate the gene regulatory regions.

Studies on gene expression profiles and gene regulatory regions are limited by the high cost and inherent cloning bias of conventional high-throughput sequencing approaches and incomplete cDNA or expressed sequence tag (EST) libraries (Sato et al. 2013). With the high sequencing efficiency and quality of paired-end-tag next generation sequencing, we are able to obtain large amounts of RNA-Seq data from a large number of samples and use these data to investigate gene expression profiles and study gene regulatory regions (Furutani et al. 2006; Li et al. 2006; Li et al. 2007; Satoh et al. 2007; Zhang et al. 2008; Wang et al. 2010). In this study, we analyzed the transcriptomes in the young panicle of 26 indica and 25 japonica plants using RNA- Seq and identified 97,005 transcribed fragments and 7702 novel transcriptionally active regions. On this basis, we compared the transcriptomes of the two subspecies and identified 1857 differentially expressed genes (DEGs) between indica and japonica rice. Some of the identified DEGs were then confirmed by reverse transcription polymerase chain reactions (RT-PCR). This study has the following conclusions: (1) we report that the two rice subspecies have different gene expression profiles; (2) we propose the relationship between the DEGs and the phenotypic differences; and (3) we speculate that nucleotide diversity in gene regulatory regions may associate with the phenotypic differences and play an important role in crop domestication. Collectively, the results of this study improve our understanding of the rice genome.

Materials And Methods

Sample Collection and Total RNAIsolation

In a previous study, we collected seeds from 825 *Oryza sativa* ssp. *japonica* and *indica* plants in a rice cultivation area and assessed the morphological and genetic differences between the two subspecies

(Wang et al. 2014). Based on this previous study, we selected 26 *indica* and 25 *japonica* plants and cultivated them in an artificial climate chamber under conditions of 12 h light and 12 h dark (Table S1).

For total RNA isolation, three 40-mm-long young panicles were collected from each rice plant. These panicles were subjected to RNA isolation using TRIzol reagent (Invitrogen, USA), according to the manufacturer's instructions. The purity and integrity of the RNA were determined by the NanoPhotometer® spectrophotometer (IMPLEN, USA) and the RNA Nano 6000 Assay Kit of the Bioanalyzer 2100 system (Agilent Technologies, USA), respectively. Thereafter, rRNA was removed from the total RNA samples and the rRNA-depleted total RNA was subjected to library construction using the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (NEB, USA), according to the manufacturer's instructions.

RNA-Seq Data Filtering and Assembly

Quality control of the raw data was performed with an in-house-developed PERL script. Low quality reads, reads with adaptor sequences and reads containing poly-N stretches were removed. The number of remaining clean reads for each sample is shown in Table S2. At the same time, Q20, Q30 and GC content of the clean reads were calculated (Table S2). To further evaluate the quality of the RNA-Seq data, we mapped reads from *japonica* and *indica* plants to a reference *japonica* genome (RGAP 7.0) and a pseudo-transcriptome of *indica*, respectively. The reference index was built using Bowtie v2.2.6 (Langmead and Salzberg 2012). In addition, paired-end clean reads were aligned to the reference genome with TopHat v2.1.0 (Kim et al. 2013). Duplicate reads were removed using SAMtools v1.2.5 (Li et al. 2009) and Picard v1.39. All the results are summarized in Table S2.

The mapped reads of each sample were assembled using Scripture beta2 (Guttman et al. 2010) and Cufflinks v2.1.1 (Trapnell et al. 2010). Cufflinks was also be used to estimate the expression of a transcript based on the mapping results (Cabili et al. 2011).

Identification of DEGs

The FPKM of each gene was calculated using Cufflinks v2.1.1 (Trapnell et al. 2010). We performed principal variance component analysis to evaluate the differences in gene expression levels between the two subspecies, and the Pearson correlation coefficients between RNA-Seq data and RT-PCR results were also calculated. To identify a gene as differentially expressed, we used the Cuffdiff program within the Cufflinks. With the application of a model based on the negative binomial distribution (Trapnell et al. 2010), Cuffdiff presents statistical routines for the determining of differential expression in the transcripts of gene expression data. To control the false discovery rate, we filtered out genes that were expressed at a level of less than 20 mapped reads in any of the 51 plants. We also calculated the Bayesian posterior *P*-value of each gene using a previously described method (Storey 2004).

Determination of the Effect of Selection Pressure Based on Resequencing Data

To identify polymorphisms in all the DEGs and non-DEGs, we explored the resequencing data obtained from 12 *japonica*, 12 indica and 5 *O. rufipogon* plants grown in China (Xu et al. 2011). The genomic DNA from the 29 individuals was sequenced at an 18–20 bp resolution and approximately 5.23 million single-nucleotide polymorphism (SNP) loci were obtained. We calculated the nucleotide diversity for *japonica* (π_j), *indica* (π_i) and *O. rufipogon* (π_w) populations using Vcftools v0.1.13. A rank-based method was adopted to determine the nucleotide diversity in the 5' flanking regions (2000 bp from start coden), gene-coding regions and 3' flanking (2000 bp from stop coden) regions of the DEGs and non-DEGs between *O. rufipogon* and *japonica* or between *O. rufipogon* and *indica*. The results were confirmed by a resampling test, which was repeated 1000 times.

GO Analysis

To annotate the functions of the DEGs, we performed Gene Ontology (GO) analysis using a GO Seq R package, in which the gene length bias was corrected. We also calculated the FDR-corrected *P*-values for the GO terms using hypergeometrical distribution. GO terms with corrected *P*-values less than 0.05 were considered to be significantly enriched.

Quantitative RT-PCR Analysis

Quantitative RT-PCR was performed using the SYBR® Premix Ex Taq™ Kit (TaKaRa, Japan) on an ABI PRISM 7900HT platform (Applied Biosystems, USA), according to the manufacturer's instructions. For each sample, two biological replicates were analyzed by three technical repeats. The rice ubiquitin-1 gene (LOC_Os03g13170) was used as the internal reference gene. The primers for qRT-PCR analysis are listed in Table S5.

Results

High-throughput Transcriptome Sequencing and Read Mapping in Two Rice Subspecies

To characterize the gene expression profiles at the reproductive stage of the two rice subspecies, we extracted total RNA from 40-mm-long young panicles of 26 *indica* and 25 *japonica* plants (Table S1). For each plant, equal amounts of total RNA isolated from three panicles was mixed to establish an RNA pool, which was then further processed using a previously described method (Marioni et al. 2008) with minor modifications, and the samples were subjected to shotgun sequencing by an Illumina GAII instrument. We obtained 2.45 billion reads for the 26 *indica* plants and 2.41 billion reads for the 25 *japonica* plants at a 125 base-pair resolution (Table S2). After filtering out low-quality reads and reads containing adaptor sequences (7% of all reads), we obtained approximately 4.52 billion clean reads (345.48 and 331.99 Gb for *indica* and *japonica* plants, respectively) (Table S2). These clean reads were then mapped to the reference *japonica* rice genome (Nipponbare; IRGSP v7.0), with at most two mismatches allowed for each read, while ignoring the reads that were mapped to more than two locations in the reference genome (multi-mapped reads). According to these criteria, we filtered out (~15%) of the clean reads and mapped 70.36–89.88% of the clean reads to the reference genome (Fig. 1a; Table S3). Among the mapped clean

reads, 62.47–78.92% were mapped uniquely to one genomic location (Table S3). In addition, the rates of non-splice genes were 41.08-56.60% for *indica* plants and 44.68–60.83% for *japonica* plants (Fig. 1b; Table S3). We then calculated the mapping rates for each sample and the average mapping rates for the two subspecies. The results showed a significant difference in the average mapping rate between the two subspecies. The average mapping rate for *indica* plants was 77.66%, whereas that for *japonica* was 83.88% (t=19.275 and p<0.001). Due to the fact that we used the *japonica* genome as the reference, we suspect that this difference may have been caused by mapping biases. Therefore, we decided to employ a previously constructed pseudo-transcriptome (Koenig et al 2013) of *indica* rice as the reference genome for the reads obtained from the *indica* rice samples. As a result, there was no significant difference in the average mapping rate between the two subspecies. The newly calculated average mapping rate for *indica* plants was 84.01%, whereas that for *japonica* remained 83.88% (t=0.12 and p=0.91).

RNA-Seq Read Assembly and Identification of Novel Transcriptionally Active Regions

We subsequently assembled the RNA-Seq clean reads into 97,005 transcribed fragments with a mean length of 884 bp (ranging from 50 to 20674 bp) (Fig 2a). These fragments were aligned to the sequences of known rice genes (MSU 7.0) The sequence alignment results are shown in Fig 2b. Among the assembled fragments, 26.31% overlapped with exons, 36.31% fell into annotated exons, 16.57% fell into introns, and the remaining 20.77% were non-gene sequences (Fig 2b). We also found that 42% of rice genes had alternative transcripts. This percentage was much lower than that previously estimated based on microarray results (Jung et al. 2013).

In total, we identified 4579 novel transcriptionally active regions based on the sequences of rice genes and non-coding RNAs (ncRNAs) deposited in the Ensembl ncRNA and National Center for Biotechnology Information (NCBI) EST “others” databases (E-value<1e-6, Table S4). Among these regions, 61.7% were unknown protein-coding regions, 67.2% could be transcribed into single-exon transcripts, 42.71% shared more than 90% sequence identity with at least one known EST, and 12.14% were likely to encode non-coding RNAs. In addition, our RNA-Seq data expanded the untranslated regions (UTR) of 11.23% of rice genes.

Identification of DEGs between the Two Rice Subspecies

According to the MSU rice genome annotation database, 74.71–88.49% of the mapped reads overlapped with annotated genes for each sample. The number of reads mapped to a certain annotated gene ranged from 2 to more than 300,000, with a median number of 23 and 32 for *indica* and *japonica* plants, respectively. There were 40,580 genes mapped by two reads in at least two plants. In addition, we were able to establish linear relationships between the number of reads mapped to a certain gene and the gene expression level (FPKM) in all the samples ($0.77 < R^2 < 0.94$; Fig. 3a). Given that we found significant differences in the number of reads mapped to a certain gene between the *indica* and *japonica* plants (t-test; P=0.012), we concluded that the two rice subspecies had significantly different gene expression

profiles. Furthermore, principal variance component analysis also revealed significant differences in gene expression levels between *japonica* and *indica* individuals (Fig 3b).

We then employed fragments per kilobase of transcript per million mapped reads (FPKM) values to estimate the expression level of various genes. To validate the estimation results, we examined the relative expression levels of three genes with clear functional annotations (Table S5) by quantitative real-time PCR (qRT-PCR) analysis. We found statistically significant correlations between the relative expression levels determined by qRT-PCR and the FPKMs of these genes (Spearman's correlation coefficient, r=0.8924, $P<0.05$; Fig 4), indicating that the gene expression levels revealed by our RNA-Seq data were reliable. Among the 41,080 expressed genes, 38911 (94.72%) were expressed in both rice subspecies, 623 genes were expressed only in *indica* plants and 1546 genes were expressed only in *japonica* plants. Thereafter, we analyzed our RNA-Seq data using a generalized linear model and identified 1857 (4.52% of all the expressed genes) DEGs between *indica* and *japonica* rice (false discovery rate [FDR] <0.05; Fig. 5a). Of these DEGs, 576 and 1281 were expressed at lower and higher levels, respectively, in *indica* rice than in *japonica* rice (Fig. 5a). We then mapped these DEGs to the rice genome and found they are distributed throughout the 12 chromosomes ($P>0.05$; Fig. 5b). Furthermore, using REEF software (sliding window size was set to 500 kb to 1 Mb and the step size was set to 10 kb), we found that the DEGs tended to group in close proximity along the chromosomes (FDR <0.05) (Copper et al. 2006). However, we could not conclude that the DEGs were clustered.

Functional Annotation of the DEGs

We annotated the DEGs using enrichment analysis tool. According to the annotations, 21 Gene Ontology (GO) terms were significantly enriched ($P<0.05$; Fig. 6). These terms could be classified into two groups. The first group consisted of reproduction-related terms, including "reproduction", "recognition of pollen", "multi-organism reproductive process" and "cell recognition". The second group consisted of cell wall biosynthesis-related terms, including "cell wall organization", "cellulose biosynthetic process", "cellulose metabolic process", "cellulose synthase activity" and "UDP-glucosyl transferase activity". The GO analysis showed that the genes encoding the binding proteins were enriched in these co-expressed genes (Fig. 6). Interestingly, genes expressed at lower levels in *indica* rice than in *japonica* rice were annotated with reproduction-related terms, whereas those expressed at higher levels in *indica* rice than in *japonica* rice were annotated with cell wall biosynthesis-related terms. Notably, the DEGs annotated with reproduction-related terms included the MADS-box genes, which comprise a gene family associated with flower development and reproduction (Becker and Theissen 2003). We identified six MADS-box genes, namely *OsMADS26*, *OsMADS15*, *OsMADS37*, *OsMADS63*, *OsMADS74*, *and OsMADS8* that were expressed at significantly different levels between the two subspecies (Table S1). Consistent with these findings, among the ten most significant morphological differences between the two subspecies, six were reproduction-related traits.

Effect of Selection Pressure on the Regulation of the DEGs

To explore whether artificial selection may have affected the expression of the DEGs, we investigated the gene regulatory regions (i.e., 5’ and 3’ flanking regions) and gene-coding regions of the DEGs (Fig 7). Specifically, we searched for genomic regions with polymorphism to divergence (π/Dxy) ratios ranked in the lowest 5% (P<0.001), because such regions may have a greater chance of experiencing positive selection. Compared with non-DEGs, DEGs from both subspecies had more 5' flanking regions with low π/Dxy ratios (Fisher’s exact test, P<0.05). These results were then confirmed by a resampling test (P<0.05). Taken together, these results suggest that the 5' regulatory regions of the DEGs experienced positive selection during rice domestication, and that gene regulation plays an important role in the evolution of the two subspecies.

Discussion

In this study, the transcriptomes of two rice subspecies, indica and japonica, were characterized by the high-throughput RNA-Seq approach. The 31.58 billion RNA-Seq data provided us with valuable resources to discover novel transcriptionally active regions, identify DEGs and investigate the effect of selection pressure on the expression of the DEGs. According to the read mapping results, 25% of the clean reads were mapped to intergenic regions in the reference rice genome, suggesting the identification of 4579 novel transcriptionally active regions. This speculation was then confirmed by RT-PCR, which validated the expression of these transcripts in all the samples tested. Therefore, in addition to characterizing the transcriptome of annotated rice genes, we also identified 4579 novel transcriptionally active regions. More than 53% of these novel transcriptionally active regions may have been transcribed into ncRNAs because: (1) the transcribed fragments could be processed into mature transcripts with a mean length of 200 bp, which is the general length of long noncoding RNAs (Zheng et al. 2019); (2) 98.76% of the interval sequences between these transcribed regions were less than 5 kb in length, which is consistent with the distribution characteristics of intergenic pre-miRNAs (Zheng et al. 2019); and (3) the sequences of these transcribed regions shared high similarity with 37.2% of the ncRNAs deposited in the Ensembl ncRNA database. These results also highlight the significance of RNA-Seq in identifying the boundaries between sense-antisense gene pairs and quantitatively characterizing transcriptomes at a high resolution. Taken together, these results improve our understanding of the rice genome.

The expression patterns of different rice genes may vary during different developmental stages and in different plant individuals. Therefore, it is important to obtain RNA samples from a large number of plant individuals to reveal statistically significant differences in gene expression levels between rice subspecies. In this study, we obtained RNA samples from three young panicles of each of the 26 indica and 25 japonica plants and searched for DEGs between the two rice subspecies based on the RNA-Seq data. In a previous study, the percentage of DEGs between closely related species in all detected genes varied greatly from 2–25% (Pavey et al. 2010). In extreme situations, this percentage can reach up to 78% (Guo et al. 2016). The big differences in the percentage of DEGs among these studies may have been due to the fact that they use different gene expression quantification (e.g., qRT-PCR, microarray, and RNA-Seq) and data analysis methods (Guo et al. 2016). Nonetheless, these studies have demonstrated that the percentages of DEGs between two species are generally higher than those between two subspecies. For

example, a previous transcriptomic study found that 3.3% of the 18,242 genes detected were expressed at significantly different levels between maize and its progenitor teosinte (Swanson-Wagner et al. 2010), whereas in another study, the percentage of DEGs between wild and weedy sunflowers (Helianthus annuus) was 5% (Lai et al. 2008). In this study, we found that 4.5% of all the genes detected were expressed at significantly different levels between indica and japonica rice, suggesting that the strong artificial selection altered the expression of a small proportion of rice genes. According to our GO analysis, genes that were expressed at lower levels in indica rice than in japonica rice were annotated with reproduction-related terms, whereas those that were expressed at higher levels in indica rice than in japonica rice were annotated with cell wall biosynthesis-related terms. Therefore, it is reasonable to speculate that differences in gene expression patterns are one of the reasons for the reproductive isolation between the two rice subspecies.

The contribution of genetic drift and selection pressures to the genetic and phenotypic differences between species has become a research hotspot in evolutionary biology in the past few decades. During rice domestication, artificial selection has been reported to have a major impact on rice gene expression (Zheng et al. 2019). According to previous studies on domesticated plants, the 5' regulatory regions of many genes evolved under the pressure of artificial selection and that the 5' regulatory regions of DEGs evolved faster those of non-DEGs (Guo et al. 2016). Consistent with these results, multiple studies in model fish species found that nucleotide diversity in gene regulatory regions, rather than gene-coding regions, are important for speciation (Wang et al. 2015; Guo et al. 2016). The relationships between nucleotide diversity and gene expression during species evolution have also been investigated by transcriptomic studies in Drosophila, fire ants, Arabidopsis and maize, and by comparative transcriptomic studies in humans and chimpanzees, and humans and mice. Taken together, these results support the hypothesis that gene regulatory regions and gene-coding regions evolve independently during speciation. Meanwhile, divergence in 5' regulatory regions of DEGs is a strong indication for the involvement of positive selection during speciation. In this study we looked at the role of RNAs transcript in shaping the changes in cell wall biosynthesis and reproductive processes that occurred during rice domestication. We've found that they play an important role in this process. Despite almost 20 years of genomics and genome-enabled studies of crop domestication, we still know remarkably little about the genetic basis of most domestication traits in most crop species. Early studied tended to go for "low-hanging fruit" — simple traits that were controlled by just one or two genes with easily identifiable mutations. Far more difficult is figuring out the more subtle developmental changes that were critical for a lot of the changes during crop domestication. This study offers a step in that direction, by examining one regulatory mechanism that has been critical for modulating domestication-associated changes in rice cell wall biosynthesis and reproductive processes.

Conclusions

In summary, we identified 97,005 transcribed fragments, 7702 novel transcriptionally active regions and 1857 differentially expressed genes (DEGs) between indica and japonica rice. These results enhance our understanding of the rice genome by identifying DEGs that may be responsible for the reproductive

isolation and phenotypic differences between the two rice subspecies. The evolution of Asian cultivated rice, a major crop for more than half of the world's population, has been a key topic of study. Our expression variations and sequence diversity confirm of gene expression and regulation regions allowed us to deduce the relationships between genome evolution and phenotype diversity, providing much new insight into the origin and domestication process of rice.

Declarations

Acknowledgements

Not applicable.

Authors' contributions

Hongbo Pang and Xiaoming Zheng wrote the manuscript, Hongbo Pang Qingwen Yang, and Xiaoming Zheng designed the experiment, Hongbo Pang, Junrui Wang, Qiang Chen, Jiaqi Li, Yueying Li, and Longkun Wu do experiment, and Qingwen Yang revised the paper. All the authors have read and approved the final manuscript.

Funding

This work was supported by the National Natural Science Foundation of China (no. 31670211 and 31970237 to X.M.Z.) and Support Plan for Young and Middle-aged Scientific and Technological Innovation Talents of ShenYang (no. RC190223).

Availability of data and materials

All the data generated or analyzed during this study are included in the published version of this paper and its supplementary information files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Abbreviations

DEG: differentially expressed gene; QTL: quantitative trait locus; qRT-PCR: quantitative real-time polymerase chain reaction; EST: expressed sequence tag; NCBI: National Center for Biotechnology Information; UTR: untranslated region; GO: Gene Ontology; FPKM: reads per kilobase of exon model per million mapped reads; FDR: false discovery rate.

References

1. Becker A, Theissen G (2003) The major clades of MADS-box genes and their role in the development and evolution of flowering plants. Mol Phylogenet Evol 29 (3):464-489
2. Beukert U, Li Z, Liu G, Zhao Y, Ramachandra N, Mirdita V, Pita F, Pillen K, Reif JC (2017) Genome-Based Identification of Heterotic Patterns in Rice. Rice 10 (1):22
3. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Dev 25 (18):1915-1927
4. Coppe A, Danieli GA, Bortoluzzi S (2006) REEF: searching REgionally Enriched Features in genomes. BMC bioinformatics 7 (1):453
5. Derome N, Bougas B, Rogers SM, Whiteley AR, Labbe A, Laroche J, Bernatchez L (2008) Pervasive sex-linked effects on transcription regulation as revealed by expression quantitative trait loci mapping in lake whitefish species pairs (Coregonus sp., Salmonidae). Genetics 179 (4):1903-1917
6. Furutani I, Sukegawa S, Kyozuka J (2006) Genome-wide analysis of spatial and temporal gene expression in rice panicle development. The Plant J 46 (3):503-511
7. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S (2005) Genetic structure and diversity in Oryza sativa L. Genetics 169 (3):1631-1638
8. Guo J, Liu R, Huang L, Zheng XM, Liu PL, Du YS, Cai Z, Zhou L, Wei XH, Zhang FM, Ge S (2016) Widespread and Adaptive Alterations in Genome-Wide Gene Expression Associated with Ecological Divergence of Two Oryza Species. Mol Biol Evol 33 (1):62-78
9. Guo W, Sarkar S (2016) Adaptive Controls of FWER and FDR Under Block Dependence. arXiv:1611.03155v1
10. Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28 (5):503-510
11. Hanikenne M, Kroymann J, Trampczynska A, Bernal M, Motte P, Clemens S, Krämer U (2013) Hard Selective Sweep and Ectopic Gene Conversion in a Gene Cluster Affording Environmental Adaptation. PLoS Genet 9 (8):e1003707
12. He GH, Zheng JK, Yin GD, Yang ZL(1994) Gamete fertility of the between *indica* and *japonica.* Chin J Rice Sci.8: 177-180

13. Horiuchi Y, Harushima Y, Fujisawa H, Mochizuki T, Fujita M, Ohyanagi H, Kurata N (2015) Global expression differences and tissue specific expression differences in rice evolution result in two contrasting types of differentially expressed genes. BMC genomics 16:1099
14. Huang HZ, Lou SL, Wang HC (1982) Genetic basis and hybrid sterility in indica/japonica hybrids. J Xiamen Univ (Natural Science)21: 189-199
15. Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, Zhao Y, Liu K, Lu H, Li W, Guo Y, Lu Y, Zhou C, Fan D, Weng Q, Zhu C, Huang T, Zhang L, Wang Y, Feng L, Furuumi H, Kubo T, Miyabayashi T, Yuan X, Xu Q, Dong G, Zhan Q, Li C, Fujiyama A, Toyoda A, Lu T, Feng Q, Qian Q, Li J, Han B (2012) A map of rice genome variation reveals the origin of cultivated rice. Nature 490 (7421):497-501
16. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M, Fan D, Guo Y, Wang A, Wang L, Deng L, Li W, Lu Y, Weng Q, Liu K, Huang T, Zhou T, Jing Y, Li W, Lin Z, Buckler ES, Qian Q, Zhang QF, Li J, Han B (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42 (11):961-967
17. Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, Swofford R, Pirun M, Zody MC, White S, Birney E, Searle S, Schmutz J, Grimwood J, Dickson MC, Myers RM, Miller CT, Summers BR, Knecht AK, Brady SD, Zhang H, Pollen AA, Howes T, Amemiya C, Broad Institute Genome Sequencing P, Whole Genome Assembly T, Baldwin J, Bloom T, Jaffe DB, Nicol R, Wilkinson J, Lander ES, Di Palma F, Lindblad-Toh K, Kingsley DM (2012) The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484 (7392):55-61
18. Jung KH, Gho HJ, Giong HK, Chandran AK, Nguyen QN, Choi H, Zhang T, Wang W, Kim JH, Choi HK, An G (2013) Genome-wide identification and analysis of Japonica and Indica cultivar-preferred transcripts in rice using 983 Affymetrix array data. Rice 6 (1):19
19. Kato S (1928) On the affinity of rice varieties as shown by fertility of hybrid plants. Bull Sci Fac Agric Kyushu Univ 3:132-147
20. Khush GS (1997) Origin, dispersal, cultivation and variation of rice. Plant Mol Biol 35 (1-2):25-34
21. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14 (4):R36
22. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9 (4):357-359
23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25 (16):2078-2079
24. Li L, Wang X, Sasidharan R, Stolc V, Deng W, He H, Korbel J, Chen X, Tongprasit W, Ronald P, Chen R, Gerstein M, Deng XW (2007) Global identification and characterization of transcriptionally active regions in the rice genome. PloS One 2 (3):e294
25. Li L, Wang X, Stolc V, Li X, Zhang D, Su N, Tongprasit W, Li S, Cheng Z, Wang J, Deng XW (2006) Genome-wide transcription analyses in rice using tiling microarrays. Nat Genet 38 (1):124-129

26. Liu F, Xu W, Wei Q, Zhang Z, Xing Z, Tan L, Di C, Yao D, Wang C, Tan Y, Yan H, Ling Y, Sun C, Xue Y, Su Z (2010) Gene expression profiles deciphering rice phenotypic variation between Nipponbare (Japonica) and 93-11 (Indica) during oxidative stress. PloS One 5 (1):e8632.
27. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W, Huang X, Han B (2010) Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res 20 (9):1238-1249
28. Matsuo T (1997) Origin and differentiation of cultivated rice. Genetics
29. Nosil P, Feder JL (2012) Genomic divergence during speciation: causes and consequences. Philosophical Transactions of the Royal Society B: Biological Sciences 367 (1587):332
30. Oka HI (1988) Origin of Cultivated Rice. Japan Sci Soc
31. Pavey SA, Collin H, Nosil P, Rogers SM (2010) The role of gene expression in ecological speciation. Ann N Y Acad Sci 1206:110-129
32. Sasaki T, Burr B (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol 3 (2):138-141
33. Sato Y, Takehisa H, Kamatsuki K, Minami H, Namiki N, Ikawa H, Ohyanagi H, Sugimoto K, Antonio BA, Nagamura Y (2013) RiceXPro version 3.0: expanding the informatics resource for rice transcriptome. Nucleic Acids Res 41 (Database issue):D1206-1213
34. Satoh K, Doi K, Nagata T, Kishimoto N, Suzuki K, Otomo Y, Kawai J, Nakamura M, Hirozane-Kishikawa T, Kanagawa S, Arakawa T, Takahashi-Iida J, Murata M, Ninomiya N, Sasaki D, Fukuda S, Tagami M, Yamagata H, Kurita K, Kamiya K, Yamamoto M, Kikuta A, Bito T, Fujitsuka N, Ito K, Kanamori H, Choi IR, Nagamura Y, Matsumoto T, Murakami K, Matsubara K, Carninci P, Hayashizaki Y, Kikuchi S (2007) Gene organization in rice revealed by full-length cDNA mapping and gene expression analysis through microarray. PloS One 2 (11):e1235
35. Storey JD (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates : a unified approach. J R Stat Soc Ser B Stat Methodol 66:187-205
36. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511
37. Vaughan DA, Morishima H, Kadowaki K (2003) Diversity in the Oryza genus. Curr Opin Plant Biol 6 (2):139-146
38. Wang CH, Zheng XM, Xu Q, Yuan XP, Huang L, Zhou HF, Wei XH, Ge S (2014) Genetic diversity and classification of Oryza sativa with emphasis on Chinese rice germplasm. Heredity 112 (5):489-496
39. Wang J, Zhang J, Li R, Zheng H, Li J, Zhang Y, Li H, Ni P, Li S, Li S, Wang J, Liu D, McDermott J, Samudrala R, Liu S, Wang J, Yang H, Yu J, Wong GK (2010) Evolutionary transients in the rice transcriptome. Genomics, proteomics & bioinformatics 8 (4):211-228
40. Whiteley AR, Derome N, Rogers SM, St-Cyr J, Laroche J, Labbe A, Nolte A, Renaut S, Jeukens J, Bernatchez L (2008) The phenomics and expression quantitative trait locus mapping of brain

transcriptomes regulating adaptive divergence in lake whitefish species pairs (Coregonus sp.). Genetics 180 (1):147-164

41. Wu C, Li X, Yuan W, Chen G, Kilian A, Li J, Xu C, Li X, Zhou DX, Wang S, Zhang Q (2003) Development of enhancer trap lines for functional analysis of the rice genome. Plant J 35 (3):418-427
42. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, Dong Y, Gutenkunst RN, Fang L, Huang L, Li J, He W, Zhang G, Zheng X, Zhang F, Li Y, Yu C, Kristiansen K, Zhang X, Wang J, Wright M, McCouch S, Nielsen R, Wang J, Wang W (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 30:105
43. Yang CC, Kawahara Y, Mizuno H, Wu J, Matsumoto T, Itoh T (2012) Independent domestication of Asian rice followed by gene flow from japonica to indica. Mol Biol Evol 29 (5):1471-1479
44. Yang Y, Zhu K, Xia H, Chen L, Chen K (2014) Comparative proteomic analysis of indica and japonica rice varieties. Genet Mol Biol 37 (4):652-661
45. Yin C, Li H, Zhao Z, Wang Z, Liu S, Chen L, Liu X, Tian Y, Ma J, Xu L, Zhang D, Zhu S, Li D, Wan J, Wang J (2017) Genetic dissection of top three leaf traits in rice using progenies from a japonica x indica cross. J Integr Plant Biol 59 (12):866-880
46. Yuan H, Fan S, Huang J, Zhan S, Wang S, Gao P, Chen W, Tu B, Ma B, Wang Y, Qin P, Li S (2017) 08SG2/OsBAK1 regulates grain size and number, and functions differently in Indica and Japonica backgrounds in rice. Rice 10 (1):25
47. Zhang G, Lu Y, Liu G, Yang J, Zhang H (1993) Genetic Studies of the Hybrid Sterility in Cultivated Rice (Oryza sativa)Ⅱ.Allele Differentiation of F_(1) Pollen Sterility in Different Types of Varieties. Yi chuan xue bao = Acta genetica Sinica 20 (6):541-551
48. Zhang Q, Li J, Xue Y, Han B, Deng XW (2008) Rice 2020: a call for an international coordinated effort in rice functional genomics. Mol Plant 1 (5):715-719Zheng XM, Gong T, Ou HL, Xue D, Qiao W, Wang J, Liu S, Yang Q, Olsen KM (2017) Genome-wide association study of rice grain width variation. Genome:1-8
49. Zheng Y, Crawford GW, Jiang L, Chen X (2016) Rice Domestication Revealed by Reduced Shattering of Archaeological rice from the Lower Yangtze valley. Sci Rep 6:28136
50. Zheng XM, Chen J, Pang HB, Liu S, Gao Q, Wang JR, Qiao WH, Wang H, Liu J, Olsen KM, Yang QW (2019) Genome-wide analyses reveal the role of non-coding variation in complex traits during rice domestication. Sci Adv 5: aax3619.
51. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S (2007) Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol 24 (3):875-888

Figures

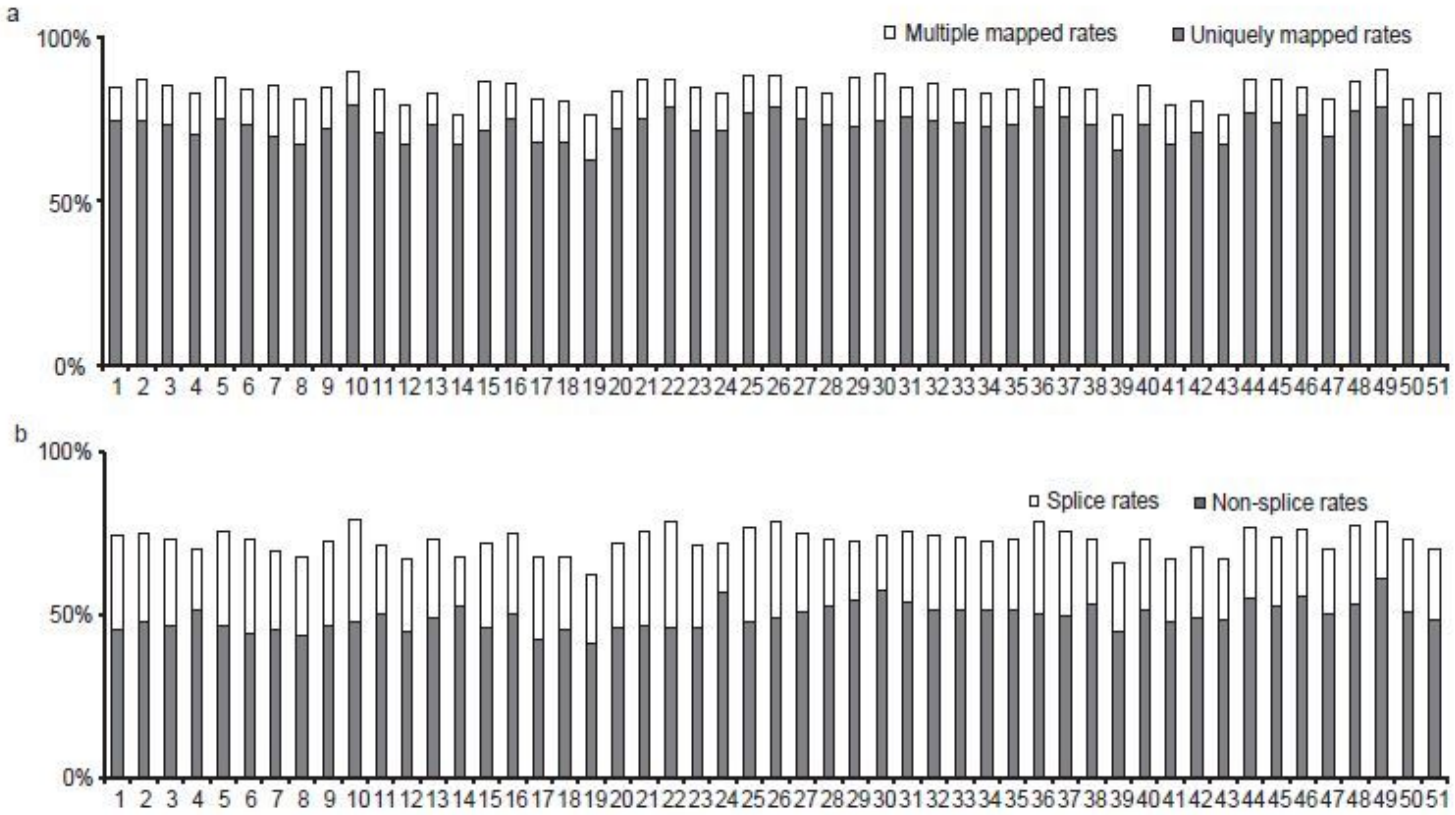

Figure 1

The mapping rates of the clean data in 26 indica and 25 japonica plants. (a) The mapping rates of the clean data. The white block represents the rate of multi-mapped reads, and the grey block indicates the rate of uniquely mapped reads. (b) The splice rate in 51 rice plants. The white block represents the splice rate, whereas the grey block indicates the non-splice rate. The numbers associated with the x-axis correspond to the sample name (see Table S1 for sample information).

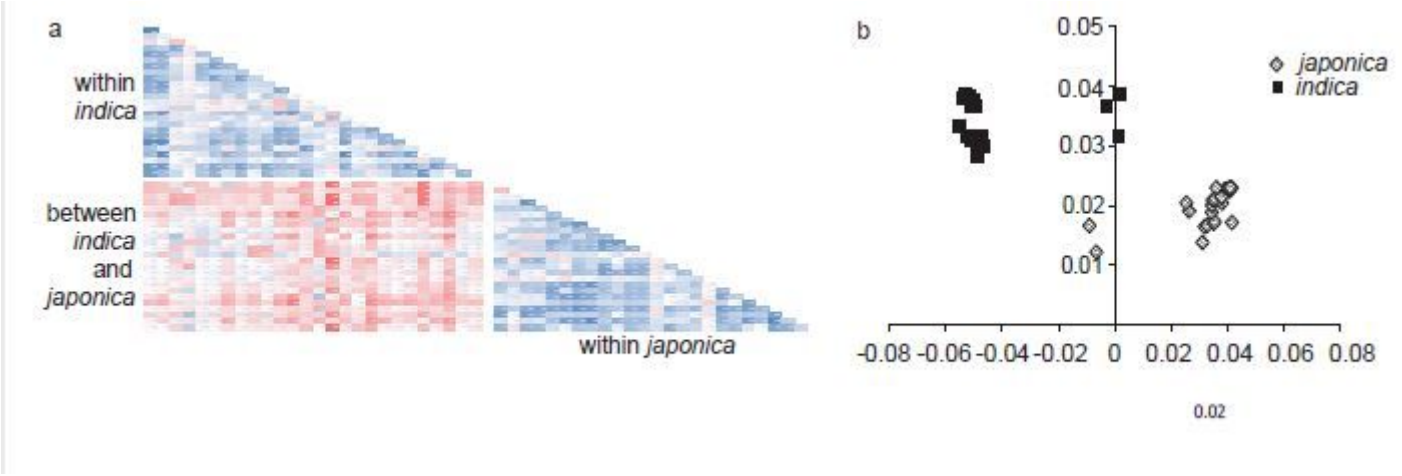

Figure 2

Significant differences in gene expression levels between indica and japonica. (a) Linear relationships in gene expression level between each sample. The R2 ranges from 0.75 (red) to 0.95 (blue). (b) The plot for principal variance component analysis of expression variations between indica and japonica plants.

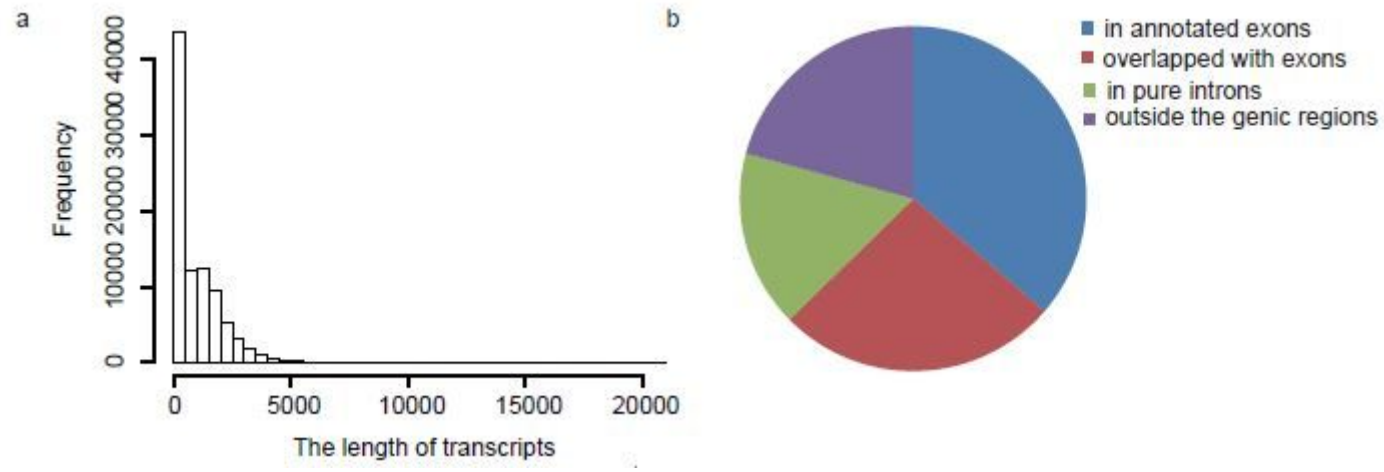

Figure 3

Summary of the assembled transcribed fragments. (a) Length distribution of the assembled transcribed fragments. (b) Classification of the assembled transcribed fragments. The transcribed fragments overlapping with annotated exons, exons, introns and non-gene regions are shown in blue, green, red and purple, respectively.

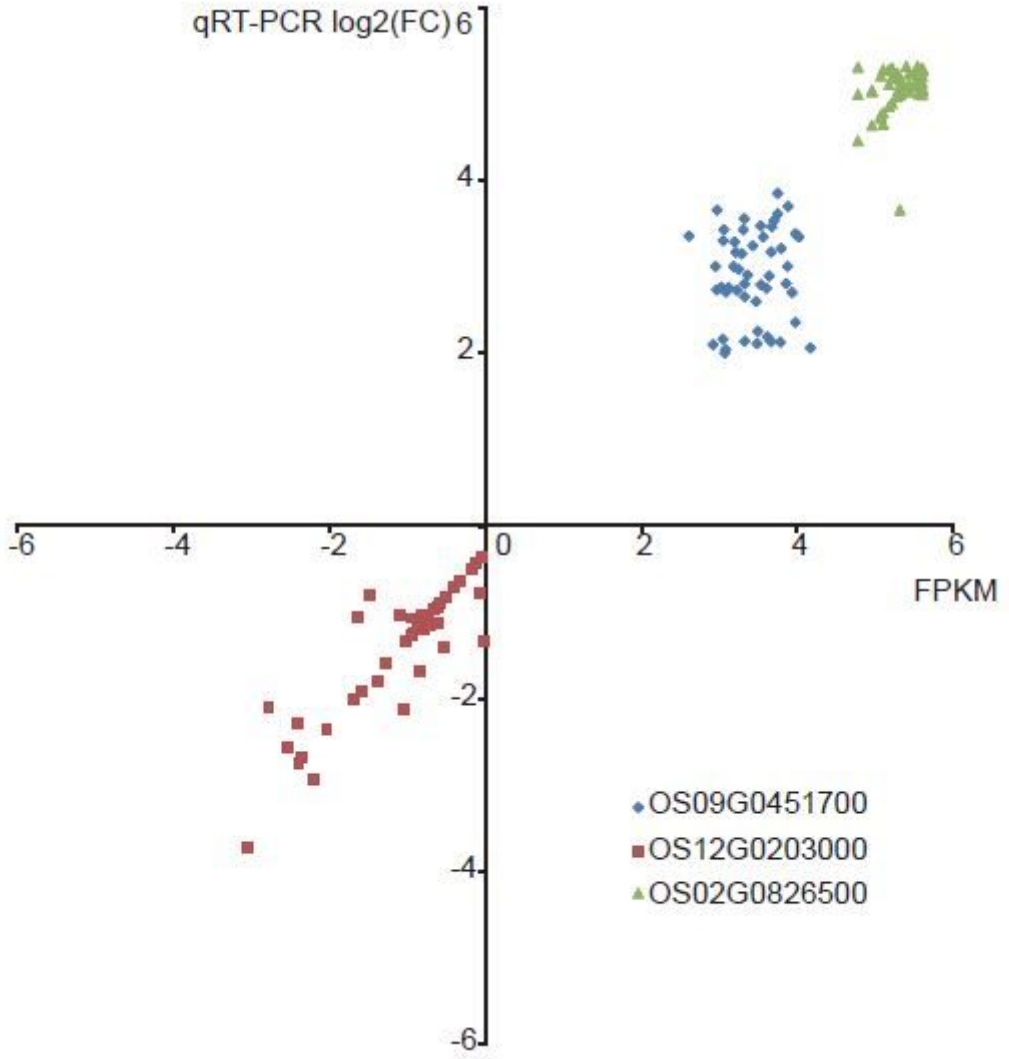

Figure 4

Correlation between qPCR results and FPKM values for three genes in the 51 rice individuals. Each point represents the relative expression level of a gene. Fold-change values were log2-transformed. The blue diamond represents Os09g0451700; the red square represents Os12g0203000; and the green triangle represents Os02g0826500.

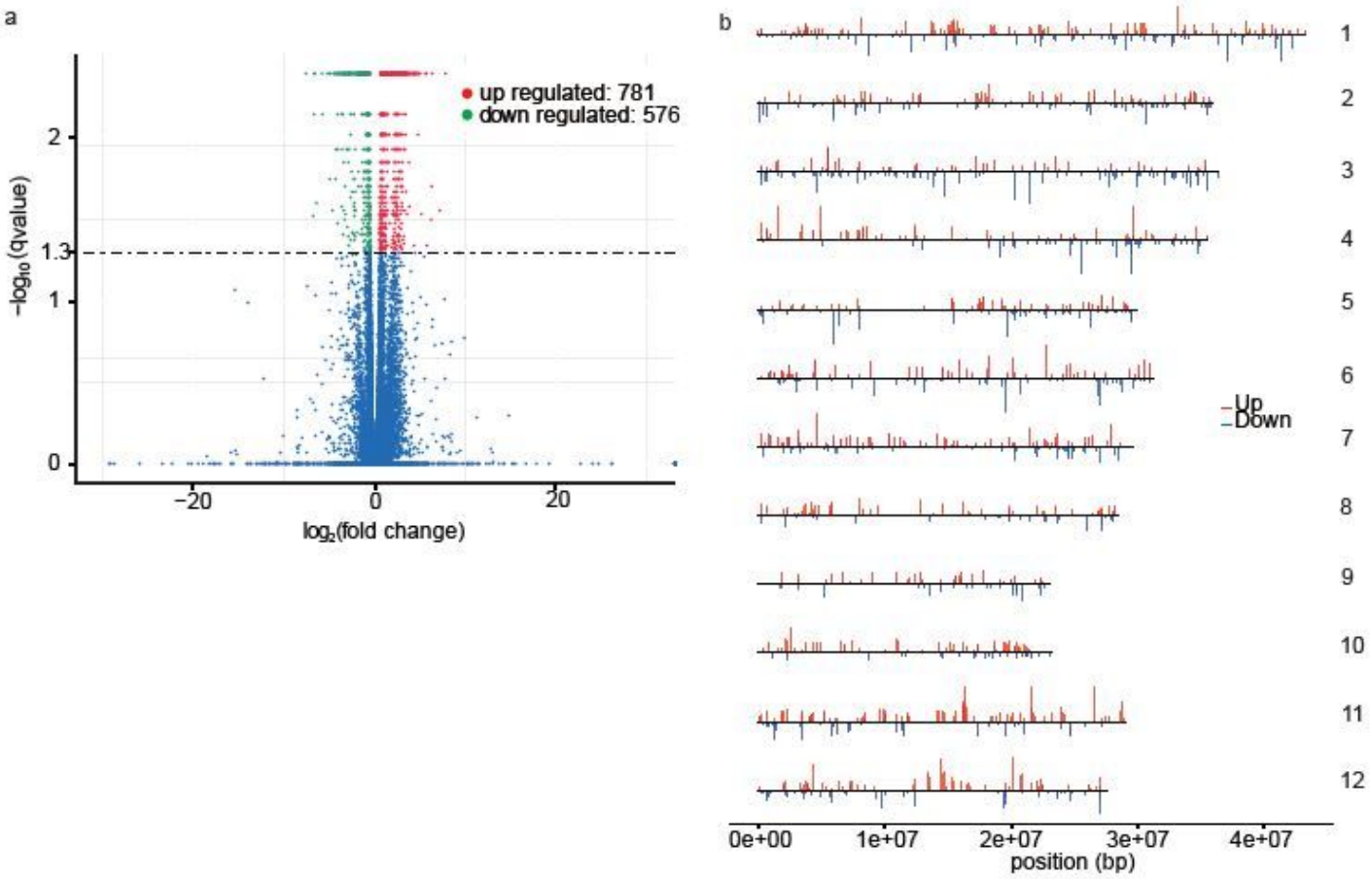

Figure 5

Summary of DEGs between indica and japonica rice. (a) Genes that were expressed at significantly different levels between indica and japonica rice. Red dots represent the up-regulated genes and green dots show the down-regulated genes. The blue dots represent non-DEGs. The volcano spots show 1857 unigenes, including 1281 up-regulated unigenes and 576 down-regulated unigenes, which were identified as DEGs. (b) Distribution of the DEGs along rice chromosomes.

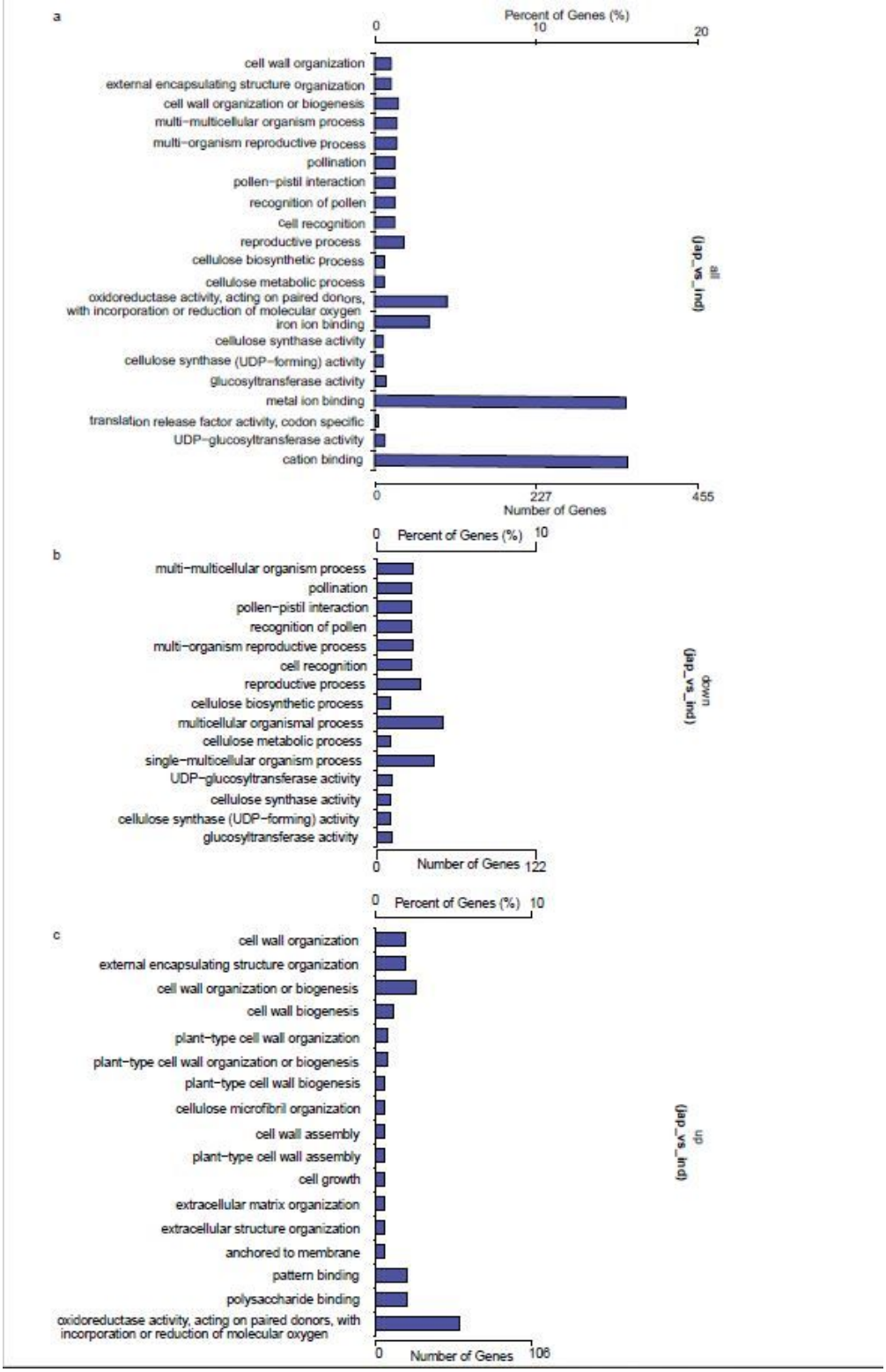

Figure 6

Histogram of Gene Ontology classifications. The letters associated with the x-axis indicate the GO categories, and the y-axis indicates the number of unigenes in each category. (a) GO analysis of all the DEGs. (b) GO analysis of the down-regulated genes. (c) GO analysis of the up-regulated genes.

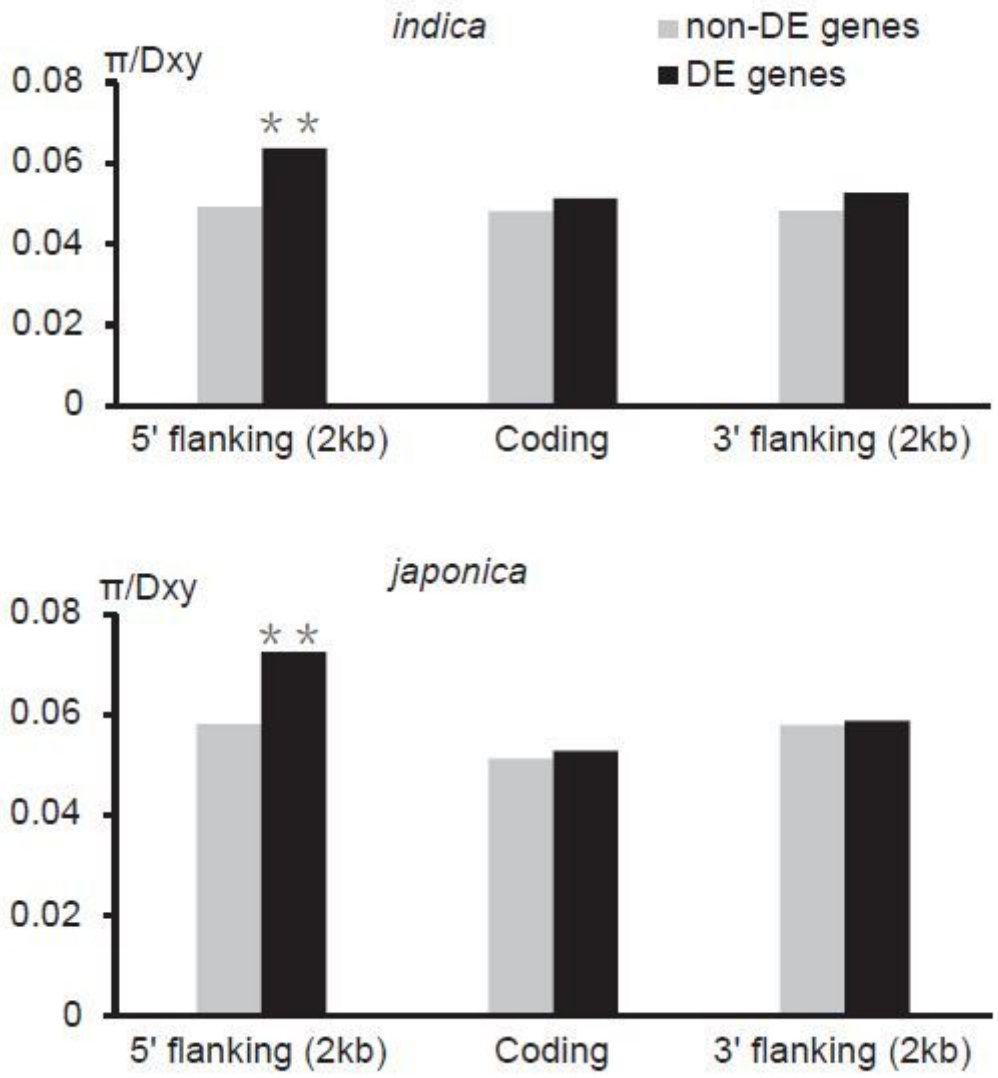

Figure 7

Polymorphism and divergence in gene-coding and regulatory regions of genes in indica and japonica rice. Y-axis indicates the ratio of polymorphism to divergence (π/Dxy), and bars in the x-axis represent the 5' regulatory regions, gene-coding regions and 3' regulatory regions of genes. The grey blocks indicate the non-DEGs, and the black blocks represent the DEGs. ** indicates P-value <0.05.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS2squencequality.xlsx
- TableS4RTpcrprimer.xlsx
- TableS3mappingrate.xlsx

- TableS1samplelist.xls