Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



# Function annotation of the rice transcriptome at single-nucleotide resolution b RNA-seq

Tingting Lu,[1,4] Guojun Lu,[1,4] Danlin Fan,[1] Chuanrang Zhu,[1] Wei Li,[1] Qiang Zhao,[1,2] Qi Feng,[1] Yan Zhao,[1] Yunli Guo,[1] Wenjun Li,[1] Xuehui Huang,[1] and Bin Han[1,3,5]

[1]*National Center for Gene Research & Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China;* [2]*College of Life Science & Biotechnology, Shanghai Jiaotong University, Shanghai 200240, China;* [3]*Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China*

The functional comple it of the rice transcriptome is not et full elucidated, despite man studies ha ing reported the use of DNA microarra s. Ne t-generation DNA sequencing technologies pro ide a po erful approach for mapping and quantif ing the transcriptome, termed RNA sequencing (RNA-seq). In this stud , e applied RNA-seq to globall sample transcripts of the culti ated rice *Oryza sativa indica* and *japonica* subspecies for resol ing the hole-genome transcription profiles. We identified 15,708 no el transcriptional acti e regions (nTARs), of hich 51.7% ha e no homolog to public protein data and >63% are putati e single-e on transcripts, hich are highl different from protein-coding genes (<20%). We found that 48% of rice genes sho alternati e splicing patterns, a percentage considerabl higher than pre ious estimations. On the basis of the a ailable rice gene models, 83.1% (46,472 genes) of the current rice gene models ere alidated b RNA-seq, and 6228 genes ere identified to be e tended at the 5′ and/or 3′ ends b at least 50 bp. Comparati e transcriptome anal sis demonstrated that 3464 genes e hibited differential e pression patterns. The ratio of SNPs ith nons non mous/s non mous mutations as nearl 1:1.06. In total, e interrogated and compared transcriptomes of the t o rice subspecies to re eal the o erall transcriptional landscape at ma imal resolution.

[Supplemental material is a ailable online at http:// .genome.org. The RNA-seq data from this stud ha e been deposited in the EMBL Sequence Read Archi e (SRA) under accession no. ERA000212 (http:// .ebi.ac.uk/ena/data/ ie /ERA000212) and are a ailable in a genome bro ser at http:// .ncgr.ac.cn/rrs. The sequence data set of continuous transcribed fragments, the detailed list of identified splicing junctions, all identified SNP lists, the SPSS binar code, and Perl scripts are freel a ailable at http:// .ncgr.ac.cn/english/edatabase.htm.]

R ce e f e a c a da ce e - c ed de a T e ba a c e f d- fe e a e a bee e ed de b e e- de a a f - e cDNAs (FL-cDNAs) a d e ed e ce a (ESTs) (T e R ce F -Le cDNA C 2003; Z a e a . 2005; L e a . 2007). H e , ea ac f a e-ca e c a d e e c cDNA EST ba e a e c e c ba , a d e- e a a a e. W e a ab f ce cDNAs a d e e e ce (G ff e a . 2002; Y e a . 2002; I e a a R ce Ge e Se e c P ec 2005), cDNA c a a d -de e ce de a a a e bee de e ed de- d ce d a f ce a c e f a e e- de ca e (Ja e a . 2005; Ma e a . 2005; F a e a . 2006; L e a . 2006, 2007; Sa e a . 2007; Z a e a . 2008; Wa e a . 2010b). T e e ac e ed ce a c e a a a e ab ed e ea c e de a d e e e ce e e de e e a e a d e e a f c a d e a e a a c bed e a e a e e a be ee ba a c a d c ca c e . De e e- d de , e ab f c a a de f e c c e ed e e e a e ac e e- e e

[4]**These authors contributed equally to this work.**
[5]**Corresponding author.**
**E-mail bhan@ncgr.ac.cn; fax 86-21-64825775.**
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.106120.110. Freely available online through the *Genome Research* Open Access option.

a c e f . I e e a f e ba e a e a e a c (AS) a e e a c a ac e e ( TAR ), d ff e a e e ed e e (DEG ), a d d e c a be ee b ec e e a a f e ce e e.

Rece , e de e e f e e - e a - DNA e e c ec e dedica e d f b a a d a f a c e (RNA-e ); e e e ec e a a e e a c e e ab e- e ed a . RNA-e ec a bee a ed a , ea e a d *Arabidopsis*, e e e a c a a d ca e f e e ac a d AS a - a d a a e a e (C a e a . 2008; L e a . 2008; M a a e a . 2008; Na a a e a . 2008; Pa e a . 2008; S a e a . 2008; Wa e a . 2008; W e e a . 2008; F c e a . 2010). T e da ed e ea c f c ed e a ed c a- a e a a e f e e e a d e a a c a f ce ec ca b d (He e a . 2010). RNA-e da a e e d c be, fe e a c d c e a ce a ec ca e ca e (Ma e a . 2008). T e a e a ed-e d a e e c a e f RNA-e f e e DNA e e c e ff ce c a d e a d a ead e a e be e de a d a c e (F d e a . 2009).

He e, e a ed e RNA-e a ac a a e e ba a c e f ce a e be e . C a ed e a a e f ce a c e e ea c , a d f a f c ed e de f ca f e - c c b e e dee e e c f e a c e a a e-ba e e - b RNA-e . We de e ed SPSS ( ea c


20:1238–1249 © 2010 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/10; www.genome.org

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



c b RNA-e da a), c e ed de ec e b ea f a ed-e d ead . We e a e a ab a f e ce e e a e AS a e . Sec d, e de f ed 15,708 TAR a d a da ed e a d b e e e a c (RT)-PCR. T e e - e e a a e ed a ~10% f e a a ed e (UTR) b da e f ce e e a e bee e - e ded b RNA-e da a. T e be f de f ed SNP dec ea ed a e c d e e ce a d UTR . O e a e ea ed a ~83% f e e e de c d be de ec ed f a 2- ee d e ce e e a ed e I a Ge e A a e II (GAII).

## Results

### E perimental design

T a a e e f c a c e f e ce a c e, e ed e ec f a c e ea e a d d b ed ead a ba e- a e b RNA-e . We e a ed e cDNA acc d e c f Ma e a. (2008) e d f - ca . B e f , a RNA a e ac ed f - ee - d ee d f A a c a ed ce *Oryza sativa* L. *indica* a d *japonica* a e e, e *japonica* (N ) a d *indica* (G a4 a d 93-11); e (A) RNA a f ed f e RNA a d ea ed be f e a ed e a e cDNA. T e cDNA a ce ed a d e e ed I a GAII. We e a ed b ca e ca e f eac a e a d e e ced eac a e ee e ( a e f 2 × 40 b , e a e f 2 × 76 b ), a ac ee f e ac e.

### High-throughput sequencing and mapping of the rice transcriptome

We e e a ed 30.9 , 27.9 , a d 29.7 40-b - a ed-e d ead a d 23.6 , 26.1 , a d 24.8 76-b - a ed-e d ead f e 93-11, G a4, a d N a e e, e- e c e (S e e a Tab e 1). Tec ca e ca e , b ca e ca e , a d 40-b - . 76-b - a ed-e d ead f e ee a - e e e e a a ee e e e e e e e e 0.70 < $R^2$ < 0.88 (S e e a F . 1). We ed a d a ed e e d d a a e ' da a a a a e e e e ce e e e e ce f *japonica* c . N ba e (IRGSP 4.0) (Tab e 1). T e a ce e e e a a a ac e f 40-b ead a d a ac e f 76 b eac a e ead e a e e e ed. M e , e a e ac ed 20 36 b (f 40 b × 2) a d 38 71 b (f 76 b × 2) c a d e a ed ead a ca d da e e c ead . U e e c e a, 38.8% 57.3% f e ead a ed e a e- c ca a d 20.2% 24.7% f e ead e e f e ed a e- a ed - a ead (F . 1A). O f e a ed ec (22.4% 30.4%), 6.4% 7.7% f e ead c d be e fec a c ed e e e e e ce f *indica* c . 93-11, d ca e e e e ce a e a . Ba ed e ce MSU e e ( e 6.0), 69.3% 73.8% f e a ed ead ca ed a a ed e ce ( 1 b ea ); f e, 3% 4.9% a ed a a ed e . N ab , e e a 26.2% 30.7% ead a ed a a ed e c ca , a a a e TAR e a de f ed.

We a a ed eac da a e a a e *indica* c . 93-11 e e e ce e a e f e a e (S e- e a F . 2A). M e ead (45.0% 62.2%) ed e fe e ce; 3.1% 5.4% ead e e e - a a d e e a 30.1% 46.7% ead e f e ed a d ead . We f e de ec ed e a a ed b f e a ed a d a ed ead (S e e a F . 2B). T e ce a e f - a a e (*phred* a e 2) f e a ed ead c ea ed ~32%, .e., f d e a a f a ed ead . T e - a ba e e e e d a d c ead , a d e e e e f f ed b e a e a a e . I a bab e e ea f e a ed ead .

### Identification of nTARs

F e e a ed e e- e e c a c bed f a e b a e a ed-e d ead bac e e f- e e ce e . T e ead ac a e f ≥100 b a d a a e a e e c de f ≥5 e /b a a e f 50 100 b a de f ≥10 e /b e e ac ed a e ab a c bed f a e . I a , e b a ed ~100,000 a c bed f a e a ea e f 220 300

**Table 1. Summary of mapping reads (Nipponbare genome sequence as reference), genes, splice junctions, alternative splicing genes, and novel transcribed regions identified by RNA-seq**

| | *Oryza sativa* varieties | | |
|---|---|---|---|
| | 93-11 | Guangluai-4 | Nipponbare |
| Total read pairs (PEs) | 54,446,080 | 54,016,813 | 54,542,233 |
| Unmapped PEs | 16,556,614 | 12,124,235 | 14,207,128 |
| PEs with unique and good matches to *indica* 93-11 shotgun genome sequence | 1,064,016 | 918,740 | 1,097,372 |
| PEs with multiple matches | 10,738,500 | 4,491,443 | 7,125,996 |
| PEs with unique but bad matches | 6,028,999 | 6,445,790 | 6,347,810 |
| PEs with unique and good matches | 21,121,967 | 30,955,345 | 26,861,299 |
| PEs in intergenic regions | 6,488,300 | 8,104,676 | 7,313,010 |
| PEs in pure intronic regions | 1,038,804 | 922,762 | 1,267,675 |
| PEs matched to MSU genes | 8,718,188 | 14,439,913 | 10,849,704 |
| MSU genes with mapped reads | 49,574 | 43,038 | 55,491 |
| Reads aligned to splice junctions | 764,568 | 1,352,152 | 1,066,921 |
| Identified unique junctions (depth ≥2) | 46,248 | 60,214 | 68,441 |
| Identical with known junctions | 17,985 | 23,313 | 26,245 |
| Previous unknown junctions in genes | 26,499 | 34,791 | 39,496 |
| Previous unknown junctions in novel transcribed regions | 971 | 1252 | 1515 |
| Previous unknown junctions in pure intergenic | 793 | 858 | 1185 |
| Genes with previous known junctions | 5553 | 6123 | 6421 |
| Genes with previous unknown junctions | 6358 | 6961 | 7393 |
| Continuous transcribed fragments | 102,985 | 99,965 | 118,064 |
| Mean length (bp) | 227 | 295 | 305 |
| N50 (bp) | 320 | 467 | 473 |
| Fragments fully in exons | 33,336 | 25,048 | 21,882 |
| Fragments overlapped with exons | 33,747 | 36,619 | 39,449 |
| Fragments in pure intronic regions | 18,249 | 24,464 | 31,741 |
| Fragments in intergenic regions | 17,653 | 13,834 | 24,992 |
| Novel identified AS genes | 12,081 | 11,713 | 14,428 |
| Known AS genes with novel AS patterns | 5399 | 5589 | 5941 |
| Genes with UTR extended (≥50 bp) | 2714 | 4971 | 6228 |
| Genes with 5′ UTR extended | 1093 | 1325 | 2118 |
| Genes with 3′ UTR extended | 1824 | 4007 | 4646 |

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



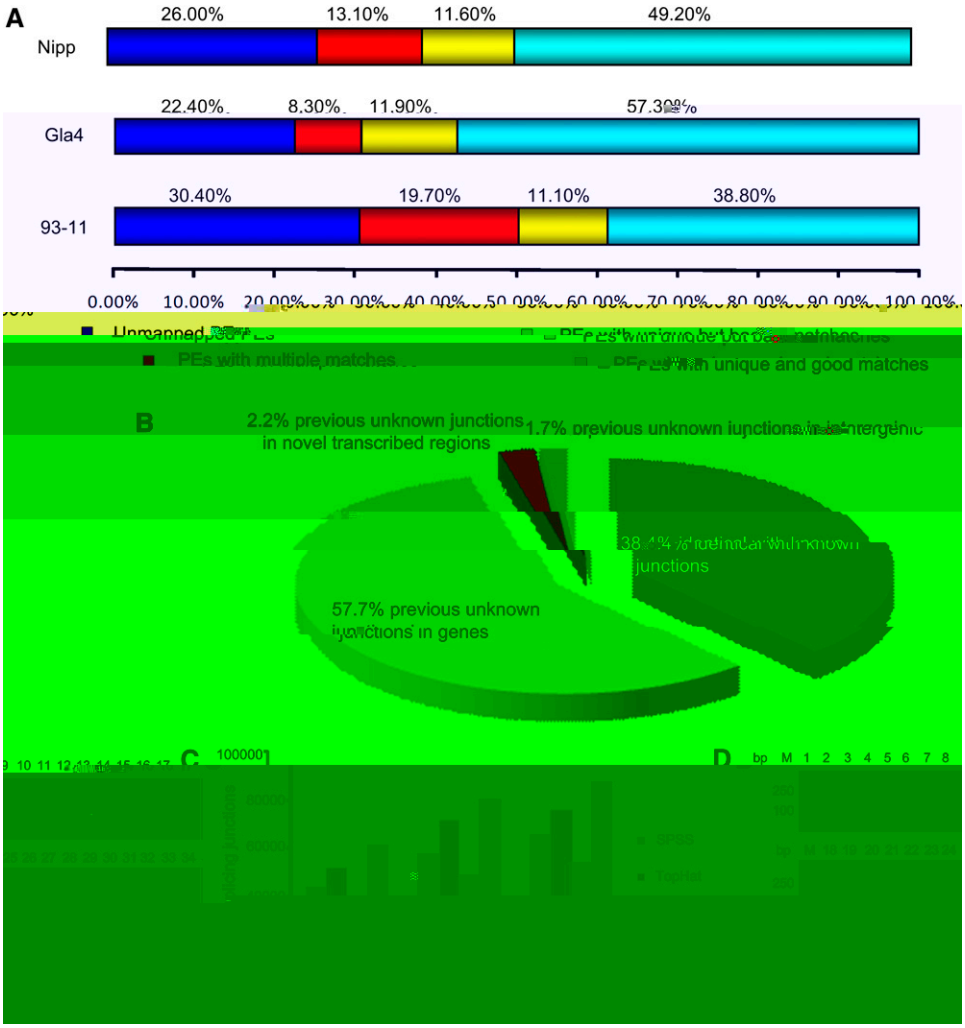


**Figure 1.** Summary of RNA-seq mapping data. (*A*) Overall mapping results of paired-end reads (PEs) referring to the Nipponbare genome sequence. (*B*) Classification of the identified exon–exon junctions based on the known splicing junctions of the MSU annotated gene models. (*C*) Splicing junctions identified by SPSS and TopHat. (*D*) RT-PCR validation of 40 randomly selected novel identified exon–exon junctions were carried out by using the total RNAs of Nipponbare. Amplification of the actin fragment in RT-PCR was used as control.

b a d a e f 50 8610 b eac a e (Tab e 1; :// . c .ac.c /e / eda aba e. ). F e 2A ed a 18.5% 32.4% f e f a e e a ed e ; 32.8% 33.4% e a ed e ; 17.7% 26.9% i ca ed e , a d e e a 13.8% 21.2% e e c e e de e e c e (*indica* 93-11: 17,653; *indica* G a4: 13,834; *japonica* N : 24,992), c e e c de ed a e e a c bed e e ce (S e e a Tab e 2). W e e e a c bed e e ce e e ed a a e RAP2 cRNA a d e da ed NCBI e - e da aba e (*E*- a e $\leq 1 \times 10^{-10}$), ~2.0% f e f a e ed i a - c d RNA , a d 36.5% 65.2% ad a ea e a c i EST a >90% e e ce de e e e e e (Tab e 2). Ne b ca e e c ca i f eac e a c bed f a e e e d e ad ace f a e (±5 b) e a c i a ac e e . I a , e de f ed 10,885, 8265, a d 14,300 TAR 93-11, G a4, a d N , e ec e (Tab e 2). T e e ed a >63% f TAR e e c ed f e a c bed a e . T d e d f e MSU da a, e e <20% f e e e a e - e a c . We c b ed a e e a c bed f a e a d de f ed 15,708 e ce TAR . O f e 15,708 e TAR , 8126 (51.7%) f d b c e da aba e e ea c a a NCBI DB (*E*- a e $\leq 1 \times 10^{-6}$). O i 713 TAR f e 8126 e e ed c ed e ead f a e >100 a ac d (aa). O f e 7582 i a c ed TAR , 3718 TAR ad e a e DB e e >100 aa a c e a d >60% a i . M f e (3355) e e a c ed - e ca e , i e , a d e e ed e ; e e e e e a d (57 TAR ), a - (10 TAR ), e (10 TAR ), a d c f e - e (e TAR ). We f e ca f ed e 15,708 TAR e e ca e e acc d e e e e f a c bed f a - e e e e ee a e e (F .2B). Ab 45.2% (7104) f e TAR e e ed b e e da a a ea a e ; f e 7104 TAR , 2994 e e de f ed a a e e . A e 2150, 1271, a d 5183 TAR e e de f ed 93-11, G a4, a d N , e ec e . O e a e f TAR e e ed

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



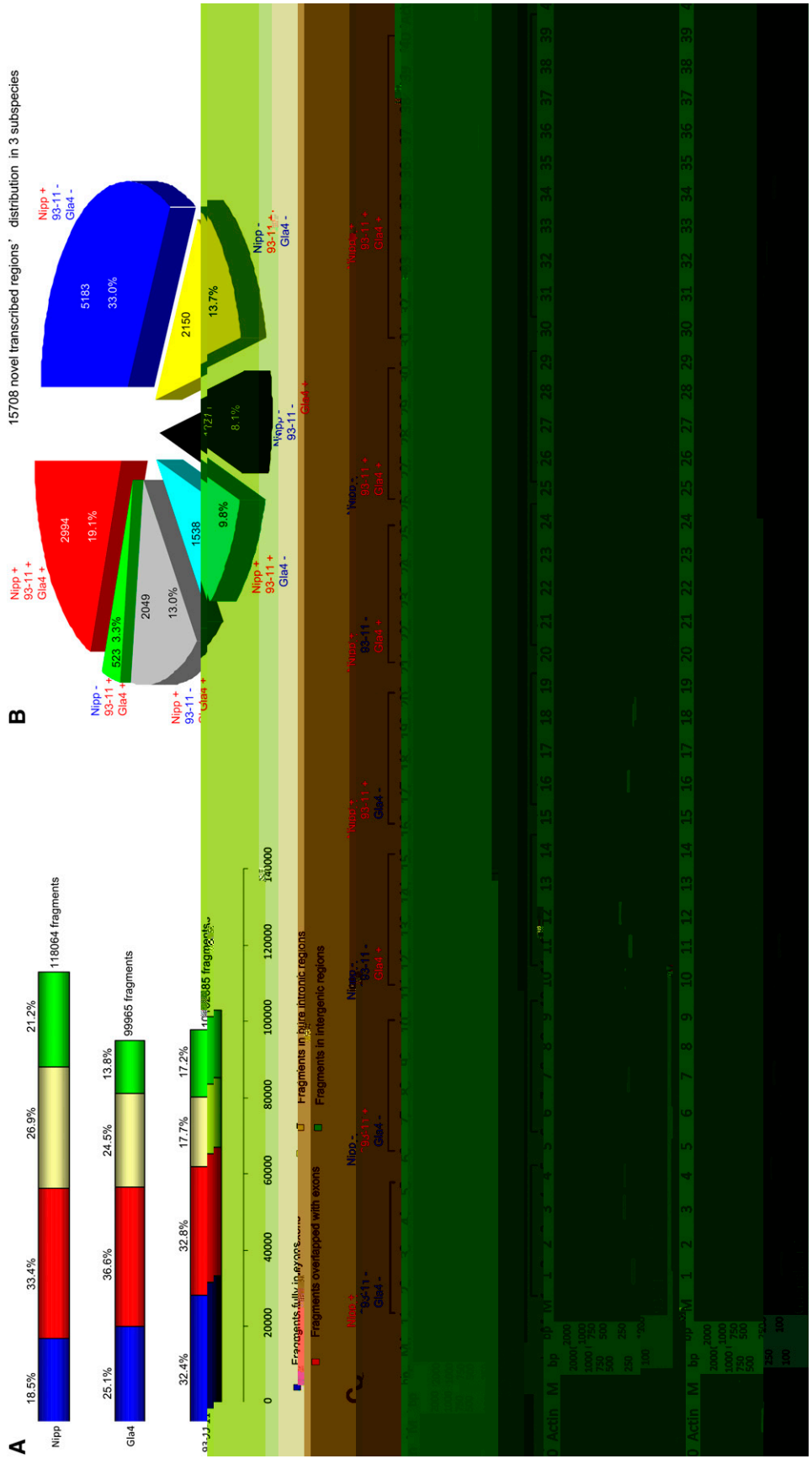


**Figure 2.** Characterization and classification of nTARs. (*A*) Distribution of continuous transcribed fragments according to the MSU annotated gene models. (*B*) The 15,708 nTARs were classified in seven categories based on the positive (+) or negative (–) detection of the transcribed fragments in three rice varieties, 93-11, Guangluai 4, and Nipponbare. (*C*) RT-PCR validation of 40 randomly selected transcripts from the seven categories (indicated at *top*) of nTARs was carried out using the total RNAs of the three rice varieties, Nipponbare, 93-11, and Guangluai 4, as indicated. Amplification of the actin fragment in RT-PCR was used as control.

F e 3C. RT-PCR a ca ed f e a da e e e e e d b b e ec 40 a d TAR f e e ca e e (F ! 2C).

## E tension of gene boundaries

T e ec e de f e e 5′ a d 3′ e e b da e , e - e a ed e ea a d d ea e f a a c e c a c bed f a e . I a , 1093, 1325, a d 2118 e e e e e ded a e 5′ e d 93-11, G a4, a d N , b a ea 50 b , e ea 1824, 4007, a d 4646 e e e e e ded a e 3′ e d. O 203, 361, a d 536 e e e e e - e ded a b e d 93-11, G a4, a d N (S e e a Tab e 3). We f e c a ed e UTR- e d b f a a a f Ge e O (GO) ca e e (S e e a F . 3). T a c e c d b d e ad e 5′ UTR , e ea e c a ec a ac a d ca a c ac e e ad e 5′

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press

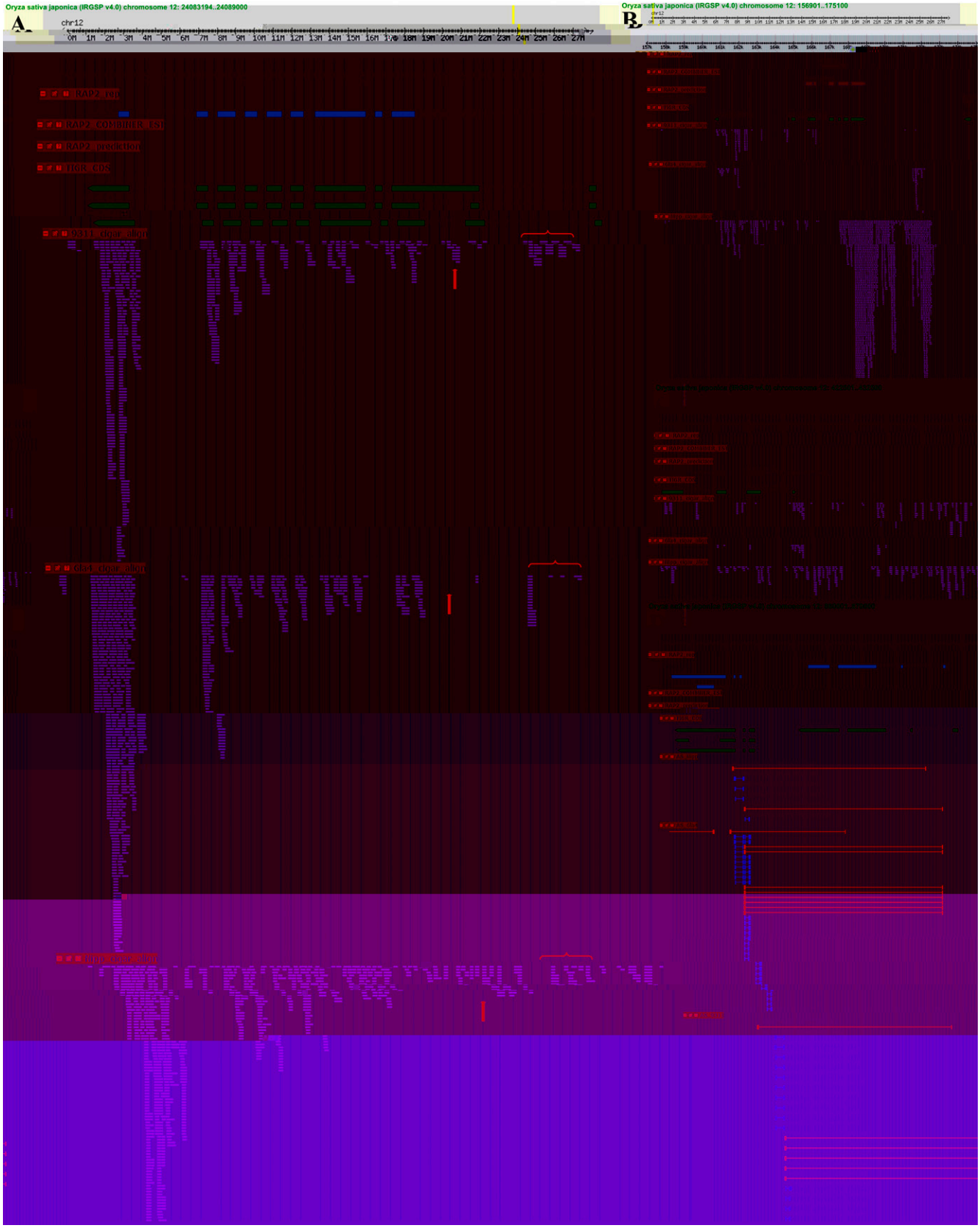


**Figure 3.** Examples of identified novel AS patterns and nTARs. (*A*) Novel exons (brackets) and alternative exons (vertical arrows) in the LOC_Os12g38850 transcript among 93-11, Guangluai 4, and Nipponbare are shown. Exons (filled boxes and horizontal arrows) and introns (lines) predicted by gene models (RAP2 and TIGR) were indicated. The RNA-seq short reads were indicated by purple lines. (*B*) Strong transcriptional activity was detected by the RNA-seq reads (purple lines) in the Nipponbare nearby gene LOC_Os12g01290. (*C*) nTARs identified by RNA-seq reads (purple lines) in Nipponbare and 93-11 as compared with the gene prediction models. (*D*) Novel exon–exon splicing junctions were identified in the three varieties.

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



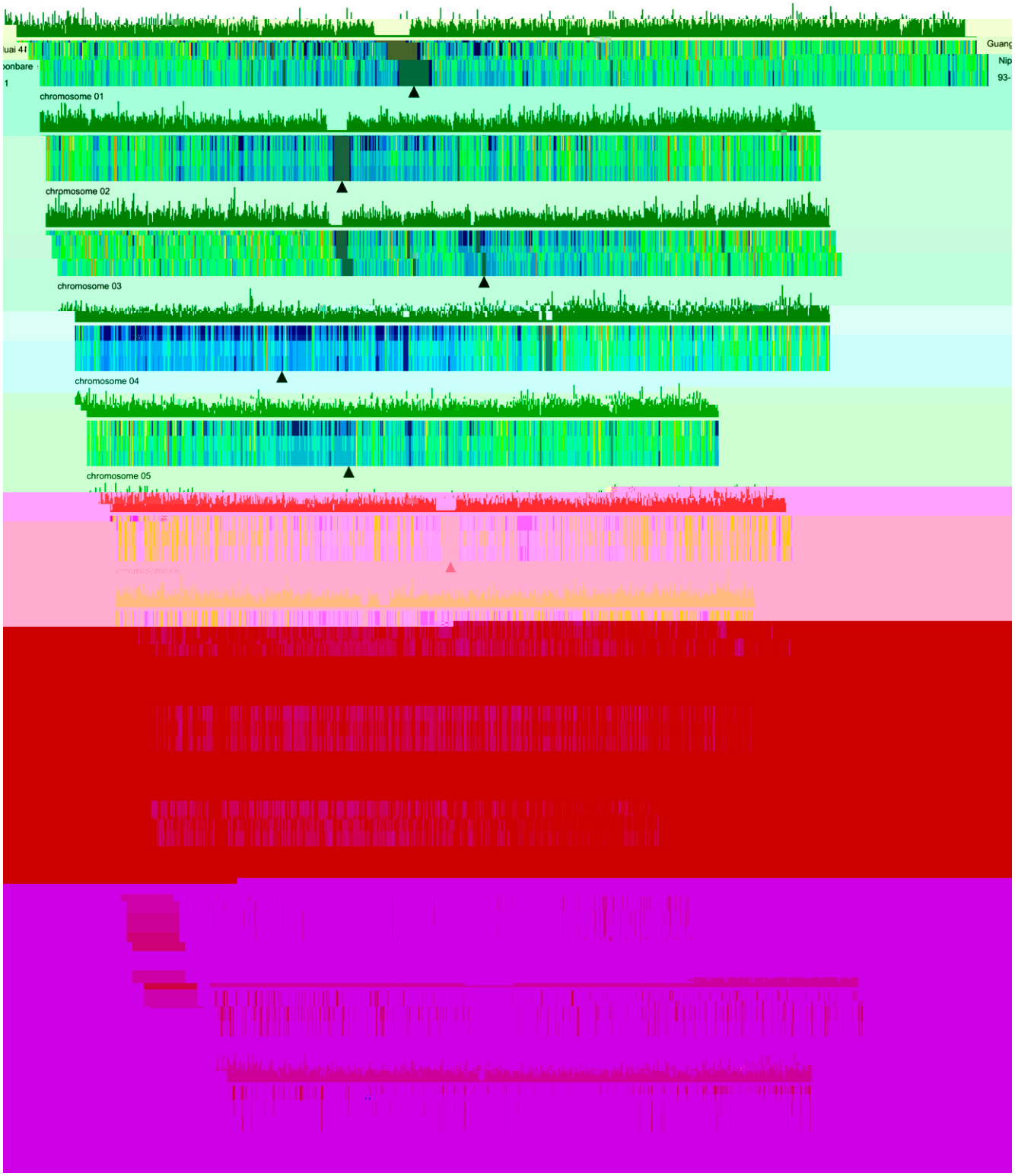


**Figure 4.** The genome distribution of transcribed regions in rice. Plots showing the number of mapped paired-end reads that was calculated in 50-kb windows along the 12 rice chromosomes are shown as color-coded vertical bars (see color index).

e a . 2008). Re ed a , e DEG f 93-11 . G a4, 93-11 . N , a d G a4 . N e de ed a 803, 838, a d 1028, e ec e (S e e a Tab e 7). We d a 69.2% 74.0% f e e RPKM-ba ed DEG e a ed e DEG a e de ed b e d ed d .

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



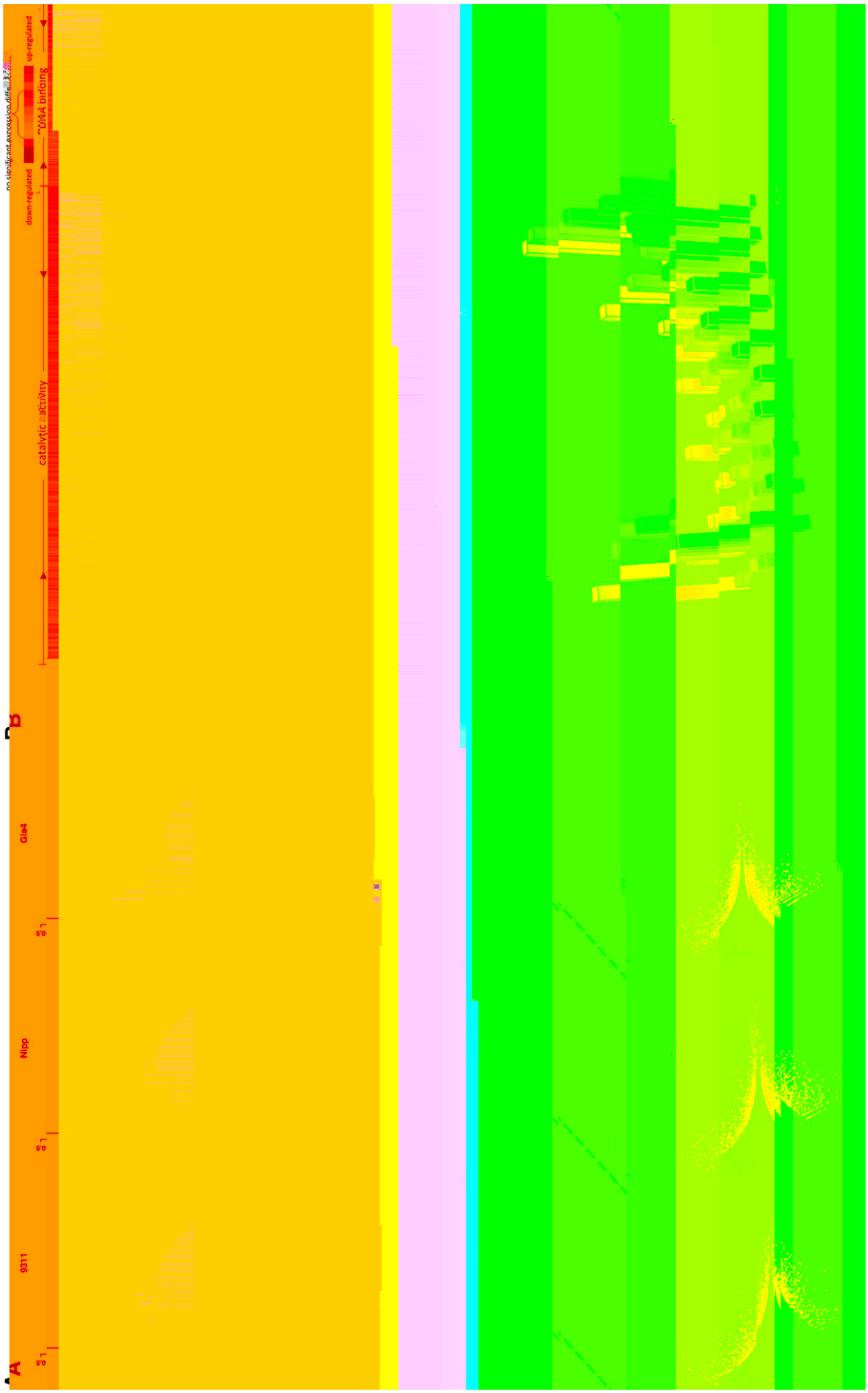


**Figure 5.** Analyses of DEGs between *indica* and *japonica*. (*A*) The *top row* of graphs show $\log_2$ ratios of the expressed genes in 93-11, Nipponbare, and Guangluai 4, respectively. The *second row* of box-and-whisker plots show $\log_2$ transformed values of the expressed genes in the three rice varieties. The *third row* of graphs show the scatter plot comparing the gene expression levels between 93-11 and Gla4, between 93-11 and Nipp, and between Gla4 and Nipp, respectively. The *bottom row* of graphs show the genes identified as differentially expressed by Fisher's exact test (red dots). (*B*) Cluster display of 34,738 coexpressed genes by GO analysis. Differentially expressed genes are shown in red (up-regulated) and blue (down-regulated). Genes with similar expression levels are shown in gray. (*C*) Genes are categorized on the basis of expression support. (MSU) MSU annotated ; (co-genes) coexpressed genes; (either-DEG) differentially expressed genes in either two varieties; (all-DEG) differentially expressed genes in all three varieties. The percentage of each molecular function for the gene categories is shown.

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



### Identification of SNPs and comparati e anal sis

RNA-se ec a d e se f c a d e ab de ec i e- c e de i (SNPs). We ade a- c i e c a be ee ce bs ec es, *indica* a d *japonica*, SNPs de ec ed f 93-11 a d G a4 RNA-se da a. W e e e *japonica* N ba e e se e ces, e de ed 67,011 SNPs be ee 93-11 a d N , a d 64,481 SNPs be ee G a4 a d N . We a de ed 26,480 SNPs be ee e N a c a d e N e e se e ce. I add , 8729, 2825, a d 14,263 e a e e ab e e e SNPs e e de ed 93-11, G a4, a d N , es ec e ( ://. c .ac.c /e /eda abase. ). O a f e se SNPs (32,920 SNPs f 93-11 . N a d 39,809 SNPs f G a4 . N ) e e ca ed 16,597 a a ed e e de . H SNP f e- ec es a ea ed a e f 200 b ea f e a a a, a e a 800 b f c d e ces, a d 400 b d ea f es c d , es ec e (S e e a F . 4A). Ab 60.8% f e SNPs e e ca ed c d e , 29.1% e e f d e 3′ UTR a d 10.1% e 5′ UTR. F 93-11 . N , e a be f SNPs a d SNPs e e 8622 a d 9144, es ec e . F G a4 . N , e a be f SNPs a d SNPs e e 9684 a d 10,303, es ec e . T e e f e, e a f . SNPs a ea 1:1.06. M f e SNPs be ed e f a d ec d c e de f ec d e e a f SNPs e e e d base. We d a e a c- e a ac , c b d , a d c a b d ad e a a be ee *indica* a d *japonica* as e e c d e es ad e SNPs; e c a ec a ac , a e ac , a a d ce ac , e ese ac , a d ea se ac e e c e ed as e e c d e es ad e SNPs (S e e a F . 4B).

### Identification of SNPs and comparati e anal sis

RNA-se ec a d e se f c a d e ab de ec i e- c e de i (SNPs). We ade a- c i e c a be ee ce bs ec es, *indica* a d *japonica*, SNPs de ec ed f 93-11 a d G a4 RNA-se da a. W e e e *japonica* N ba e e se e ces, e de ed 67,011 SNPs be ee 93-11 a d N , a d 64,481 SNPs be ee G a4 a d N . We a de ed 26,480 SNPs be ee e N a c a d e N e e se e ce. I add , 8729, 2825, a d 14,263 e a e e ab e e e SNPs e e de ed 93-11, G a4, a d N , es ec e ( ://. c .ac.c /e /eda abase. ). O a f e se SNPs (32,920 SNPs f 93-11 . N a d 39,809 SNPs f G a4 . N ) e e ca ed 16,597 a a ed e e de . H SNP f e- ec es a ea ed a e f 200 b ea f e a a a, a e a 800 b f c d e ces, a d 400 b d ea f es c d , es ec e (S e e a F . 4A). Ab 60.8% f e SNPs e e ca ed c d e , 29.1% e e f d e 3′ UTR a d 10.1% e 5′ UTR. F 93-11 . N , e a be f SNPs a d SNPs e e 8622 a d 9144, es ec e . F G a4 . N , e a be f SNPs a d SNPs e e 9684 a d 10,303, es ec e . T e e f e, e a f . SNPs a ea 1:1.06. M f e SNPs be ed e f a d ec d c e de f ec d e e a f SNPs e e e d base. We d a e a c- e a ac , c b d , a d c a b d ad e a a be ee *indica* a d *japonica* as e e c d e es ad e SNPs; e c a ec a ac , a e ac , a a d ce ac , e ese ac , a d ea se ac e e c e ed as e e c d e es ad e SNPs (S e e a F . 4B).

## Discussion

O es de a ed e ba a c e f f ce bs ec es, *indica* a d *japonica*, a- se e c ec . We e a ed e a cab f e ed d a a a es, c d e a c bed e d c e , c e de ec , e e e a ese e a d SNP de f ca .

T e da a e f 42 (40 b × 2) a d 35 (76 b × 2) e a ed ead ded c es ces f e e a a c e e e. O es ed a a e f ese ead a ed e e ce , ce- e a ed 15,708 TARs. Ra d RT-PCR a da ed e e es a RNA a es a d e ce de s a e abe es a f add a a c bed e c be d e a a ed e es. Ha f f e se TARs c d be e a c d RNAs f e f ea s: (1) O e 63% TARs be ed s e a c bed f a e s a ea e f ~220 b ±5- b e e se e ce a es (Re e MSU a a ed da a, 99.76% f s es e e <5 b.); (2) 47.2% f RAP2 c d RNAs c d be a ed ~2.0% f e e a- c bed f a e s; (3) 51.7% f TARs a e e NCBI DB da abase.

I add , e ass e d ced ead s ade ss be de f e a c s f eac d d a a e. O es d ca ed a 2150, 1271, a d 5183 TARs e e e de- f ed 93-11, G a4, a d N , es ec e . S e f e e TARs f eac a e c d be a da ed b RT-PCR s ce e- f f e e d c a ea s es a e a as ec ed (“N +, 93-11−, G a4−”; “N −, 93-11+, G a4−”, a d “N −, 93-11−, G a4+” F . 2C). T e RT-PCR es s es ed a e e es s f a f ese e es as de ec ed de e e e a c d s ed s es a .

O da a a s ded e de ce f s a d f e e ded e de ed 5′ a d 3′ UTRs. T e e ded 5′ UTRs e s f e s a d a e a e e s, c be ed e e ac a s e a f ac s. M e a e a e e s f a c s e c d ATP b d , e b d , a d c e c ac d b d e s ed beca se e e e ad e 5′ UTRs. T e e ded 5′ UTRs a d 3′ UTRs d be a ab e f de e e ac e e b da es, c s a s e f f f d e c c a c s b s a d s (e ed a a a se se a c s [NATs]). I add- d , ese e ded 3′ UTRs i be e e a e e ss be es c RNA- ed a ed c f a a a d s a- c a RNA e a . M e e, i b d a c e e a d ec se a f UTR-e e ded se e ces (es ec a f d s - se se-a se se e b da es), s a d-s ec f c RNA-se (ssRNA-se ) da a a e e ed.

AS s a e f c e a f e es c de f e a- c s. T e de f ca f AS a e s f c ce e es be e f be e de s a d e ec a s f e a c c . P e s s d es es a ed a 20% 30% f e ce e es EST/cDNA da a de AS (Ca be e a . 2006; Wa a d B e de 2006). He e, RNA-se da a, c b ed e MSU a a ed AS a e s, d ca e a a ea s 48% f e e es ce *japonica* N ba e a e a e a e s ced. T e da ed RNA-se da a f *Arabidopsis* d ca e a ~42% f e i- c a e es a e s b ec AS (F c e a . 2010). I e ea ed a AS a s c s de ab e des ead a e s ed c ed. As e ec ed, e c se s s GT AG i b da es c s e e es i (>99.0%). Dee e a- i i e e c , e a ed-e d ead s, e se e c- i e s f ce a c e da a, a d e RNA sa es f d ffe e s es a a s de e e a s a es a e ess e a f f e c i e e s e a a ce AS a e s d ce a e e a ed s c a e a . A s , ssRNA-se e c- be add a f a a e. I F e 3B f e a a e, ssRNA-se da a, d be d ff c de e e e e a c de f ed N s a ed a a f e e LOC_Os12 01290, a AS a e , a d d a e a c e s e e a . M e e, e e e a c f a i d be a f f i e a da ese c i a a e- d c s a d de s a d e b i ca f c s.

F d a e e e i s c c s s ac a e as as e RNA-se ead s a e s se e c e s a a s cc , a d a s ce c s a be s a ed b e fe ead s. He e, SPSS a s ed 82.5% e ff c e c 93.0% ac- c ac . T e s- a d e - de f ed s e a e bab d e e f ea s: (1) Pa f e ead s c d be a ed c- s e e fe e ce beca se f se e c i e s; (2) se- e ce s f e s c c s e e ed e e a e e s e fe e ce; (3) e e s e ce f s e se- e ce e ea s (SSRs), e a e e s s f ead s be e e e ca s. T e ff c e c f c d s c e i as i as as a f T Ha H e e, SPSS c d d s c e s eadd a s ce i c s. T e d s c e a ce s e e b se ed

bab beca se d ffe e a e s f a e ac a es e e sed a d d ffe e a a e e s e e a ed b SPSS a d T Ha H - t se e c s a e f a d a a e e d f e -de t a a s s f a sc e s a a t l es- t . C s e t e s e t s (L e a . 2006), e RNA- se da a ded a a a es t a f e d ffe e a sc t a ac es ass c a ed t e e c t a a d t e c a e ce e e (F . 4). T e d ffe e e e es e e s f e a sc s c d be de e s d ffe e ces e a e f a de e e ; e b t ca d ffe e ces bab a s t e s ed be ee e a t e es. O a a se s f DEGs 93-11, G a4, a d N t ded a a t e a e a s ec a e c c a a e ese a c . Us e $\chi^2$ d es- f t es t e RNA- se es s c d be e ab t eas ed de f t DEGs a b ased t a e b e a e a e ff ec t Less d s f a f ca f RNA t - t se e c d es t ab e a e ec se es t a f t e e e e- es e e .

We f d 75,740, 67,306, a d 40,743 SNPs ( c d e ab e e s SNPs) be ee e a sc s f 93-11 a d N , be t ee e a sc s f G a4 a d N , a d be ee e a - sc s f N t a d N , es ec e . T s e se s a t ea e be f SNPs a e ec ed. P s b e eas s f e se SNPs a e as f s: T e a e f ec t e ea SNPs be ee ce s b ec es; e a be aff ec ed b RNA ed ; e t e be ed se- t e ce e s s e f se e ce a f a ; e t a be de ed f e a e , e t e e s t s a e b a e f e e ce se e c e . I s d be ed a t ac- c d e se da a, e c d t c ea de e t e t a SNPs a e d b es d e RNA ed t. Us b t e c DNA a d RNA f t e sa e t a s t f e sa e a e a e f dee - se e c t e base a t be a be e t a de- ec a d a a e RNA ed . t t t t T da e, e e ac ed e a ed eads (acc d e f e ce IRGSP 4.0) f t s s d . A e 8.3% 19.7% f e t e a ed eads t RNA- se da a e a ed a a ed. M s t f e - eads c d be a b ed d ca ed e es t a d se e a d ca s. I add t , 22.4% 30.4% f e eads c d t be a ed e N t ba e e e se- e t ce. T e a eas t s f e se t a ed eads a e se e c e s (S e e a F . 2B) a d e e e a f e sa e e a a s; a t e eas s t e s ca a t f t e e f e- e ce 6.4% 7.7% eads f s s ec t c d be a c ed e 93-11 e e se e ce; f t e eas s c de e f e t e ce e t s a d e de f ed a c e s. W e e sed e *indica* 93- 11 e t e se e ce as a a t e f e e ce, a t ed sec- as ~45% 62%, d ca a e 93-11 e e se e ce s t c e e. t t

I s a , e RNA- se a ac as e as a e ff c e e d f a sc t e f a a ses. I ca be de a - ed f a t s ea c ses, a f t de f a sc bed ac e e s, s c s es, a d AS a e t s f ea- s t e RNA e ess f de ec t DEGs f d ffe e sa - es e c a s bs ec es. t t t

## Data anal sis

Se e ce a e .12.8 e e e f ed b SSAHA2 2.3 (Se e ce Sea c a d A e b Has A ) f 146.3 40.4(s f a e)]TJ0-1.2 sed e I a FASTQ 146.3240.8(Sa e )-228(s a da d)-238(F)26.3(AST IRGSP t e d ec es f ce *japonica* c . N ba e (B d 4.0 a d 5.0, :// .d a.aff c. . IRGSP/B d4/b d4. ), BGI ce *in.1.di2.4caJ101f2.0220dcv2.1.2.2.11.1124.co2.nt24.2i22.1s,1.an2no22.ta1.t1.* *results, reads that.21alined1to multiple enomic locations ere i nored. Of.24.2the24.uniuely22.mapped24.reads,24.hs ere set to allo*

## Methods

Plant.3h04.3(materials.3h04.8(and)-25.2(gro th)-307)l(conditions)]T / T101Tf7.702007.7025.7543106.1858Tm[(Seeds)-268.6(from)-265.8(the)-26.2(c tation.3h18(kit)3h15.2((Ambion.12.5())-322h)146.3h(first)3h15.4(cDNA)-30.2(strand)-313.( as)-315.(s nl46sied)-30.3(using)]TT[(random)-317.(he am constructed b follo ing manufacturers instructions (Illu- mina). Fragments of ~300 b e e e c s eda d e c ed b PCR f 18 c c es. T e d c e e aded f ce c a e s a a c ce a f 2 M a ed-e .12.5(d)-360.2(40)-349.7(b )]TJ/T1_31Tf21.77330Td(3)T /T1_01Tf0.3548Tc1.18790Td[(2 )354.8( )-3 a ed f t a e a a s s a d base

a ed ead: f e 40-b a ed-e d ead:, a d 38 71-b c - a ed ead: f 76-b a ed-e d ead: e e a: ac ed a: ca d da e e e c ead:.

Pa ed-e d ead: a e e a ed e e e e ce e e f e a ed e : e e ed f MSU 6.0. T e be f ead: a e e f ca ed e a: c ed. T e e e: e e f eac e e a: de e ed a: e : f a e .

## Identification of no el transcribed acti e regions

T e e e e c e (e c d a UTR:, e , a d ) e e b de ed e ce MSU cDNA da a a d e N - ba e e d ec e:. S e c e: f c e e: e e c e e e f e de f ed. T e a c - a e f ≥100 b a a e a e e c de f ≥5 e:/b a a e f 50 100 b a de f ≥10 e:/b e e e ed f a e e a c bed f a - e . W e a f 95.4% f e ce e e: e: be <8 b a d 99.76% e: be <5 b ( e ed ce MSU da a), e c b ed e a c bed f a e c ca ed e ad a- ce e c e (±5 b) e d d a a c bed N e a c bed e e a: ed a: TAR: a d e e ea c ed a a DB a d e e a d e ec ed f f e a da b RT-PCR.

## Gene boundar determination using continuous transcribed fragments

T e c a c bed e e e e ed e e f e a e be ee a c a f a e : a: <30 b (a: ~0.1% f ce e e: ad a e f <30b ). T e , e ce MSU e e b da e: e e c a ed e c b ed a- c bed e de e e e a e ded e e b d- a e: b c ee f ab ea e a c bed e e a e a ed UTR:. T e UTR a: c de ed e ded f e - ea /d ea e e ce e f e a a c bed f a e : a: >50 b .

## Detection of splicing sites

T e a SSAHA2 e: e a ed a e f e e ce ead: a a a c ed e e e ce e e e ce a d ed - e a ced ead:. We e ec ed ca d da e *trans*- ead: f f e a da e f c e a: F 40-b a ed-e d ead:, e e ac ed e 20 36 b c a d e a c ed e e e e ce, f 76-b a ed-e d ead:, e e ac ed e 38 71 b . We de e ed SPSS, e C++, de e e b ea e f c e:. T e a a e f a a b e f de: bed e e. F e ca f e a e *trans*- ead: e e e e ce e ceeded be de e ed; a: a a: b e ec e e e a e f ead. Sec d, e ea c ed e a ed a f e *trans*- ead a a a f 50 b 6 b f e a: e e ce a e f e ea d a ea e e ce, acc d ca a d e a e e e e ce e e ce (f ece:: a , e e e ce d be e ed). T d, e f e ed e a c e: b ca e GT/AG ( e e e, CT/AC) f eac a c ed *trans*- ead. F a , a de f eac de ed c ≥2 a: e ed. A e GC/AG e a: a e: ed de ec AS a e . We a a - ed T Ha a a ab e f a e ac a e, d c e ce c a e a a e e e c e f SPSS (T a e e a . 2009). T a da e e e c c a, e de: ed 40 e a d ed e a RNA: f N ba e f RT-PCR (S - e e a Tab e 6).

## Differentiall e pressed gene assessment

We de ec ed e e e e: e e f f e DEG de f ca a e ca c a : (Ma c ed PE ead: e e e/ a c ed PE ead: f e a e ) × ( a a c ed PE: a e a e e:/3). We e de f ed DEG: f d f e e a e e: (93-11, G 34, a d N ) acc d a R ac a e a ed “DEG e ” (Wa e a . 2010a). T e Pea $\chi^2$ e: a: a ed a e e a e e ec F eac e e, e *P*

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press



G ff SA, R ce D, La TH, P e G, Wa R, D M, G a eb J, Se A, Oe e P, Va a H, e a. 2002. A d a e ce f e ce e e (*Oryza sativa* L. . *japonica*). *Science* **296:** 92 100.

He G, Z X, E AA, C e L, Wa X, G L, L a M, He H, Z a H, C e F, e a. 2010. G ba e e e ca d a c a e d a ce b ec e a d e ec ca b d. *Plant Cell* **22:** 17 38.

I e a a R ce Ge e Se e c P ec 2005. T e a -based e ce f e ce e e. *Nature* **436:** 793 800.

J a Y, J a P, Wa K, S N, Y S, Z a D, Ma L, Fe Q, J Z, L L, e a. 2005. A c a a e e a a f ce c e 4 e a c e-e e e a f a c . *Plant Cell* **17:** 1641 1657.

L L, Wa X, S cV, L X, Z a D, S N, T a W, L S, C e Z, Wa J, e a. 2006. Ge e- de a c a a e ce c a a . *Nat Genet* **38:** 124 129.

L M, X W, Ya W, K Z, X e Y. 2007. Ge e- de e e e f e ea c e ed a d e ec a f c f e a ce. *Plant Physiol* **144:** 1797 1812.

L H, R a J, D b R. 2008. Ma DNA e e c ead a d ca a a a a c e. *Genome Res* **18:** 1851 1858.

L e R, O'Ma e RC, T -F J, G e BD, Be CC, M a AH, Ec e JR. 2008. H e a ed e-ba e e a f e e e e *Arabidopsis*. *Cell* **133:** 523 536.

L X, L T, Y S, L Y, H a Y, H a T, Z a L, Z J, Z a Q, M J, e a. 2007. A c ec f 10,096 *indica* ce f - e cDNA e ea e e ed e e ce be ee *Oryza sativa indica* a d *japonica* b ec e. *Plant Mol Biol* **65:** 403 415.

Ma L, C e C, L X, J a Y, S N, L L, Wa X, Ca M, S N, Z a X, e a. 2005. A c a a a a f e ce a c ea d c a *Arabidopsis*. *Genome Res* **15:** 1274 1283.

Ma J, Ma C, Ma e S, S e M, G ad Y. 2008. RNA- e : A e ec ca e d c b a d c a e e e a a. *Genome Res* **18:** 1509 1517.

M a a A, W a BA, McC e K, Sc ae e L, W d B. 2008. Ma a d a f a a a c e b RNA-Se . *Nat Methods* **5:** 621 628.

Na a a U, Wa Z, Wae K, S C, Ra a D, Ge e M, S de M. 2008. T e a c a a d ca e f e ea e e de ed b RNA e e c . *Science* **320:** 1344 1349.

Pa Q, S a O, Lee LJ, F e BJ, B e c e BJ. 2008. Dee e a e a e c c e e a a c e b - e e c . *Nat Genet* **40:** 1413 1415.

T e R ce F -Le cDNA C . 2003. C ec , a , a d a a f e 28,000 cDNA e f a ca ce. *Science* **301:** 376 379.

Sa K, D K, Na a T, K N, S K, O Y, Ka a J, Na a a M, H a e-K a T, Ka a a S, e a. 2007. Ge e a a ce e ea ed b f - e cDNA a a d e e e a a c a a . *PLoS ONE* **2:** e1235. d : 10.1371/ a . e.0001235.

S a M, Sc MH, R c a d H, Ma e A, K e A, Sc e f M, Se e M, B d a T, S da A, Pa c D, e a. 2008. A ba e f e e ac a d a e a c b dee e c f e a a c e. *Science* **321:** 956 960.

T a e C, Pac e L, Sa be S. 2009. T Ha D c e ce c RNA-Se ! *Bioinformatics* **25:** 1105 1111.

Wa BB, B e de V. 2006. Ge e dec a a ea a a f a e a e c a . *Proc Natl Acad Sci* **103:** 7175 7180.

Wa ET, Sa dbe R L S, K eb a I, Z a L, Ma C, K e SF, Sc GP, B e CB. 2008. A e a e f e a a e a e la c e. *Nature* **456:** 470 476.

Wa L, Fe Z, Wa X, Wa X, Z a X. 2010a. DEG e : A R ac a e f de f d e a e ed e f RNA- e da a. *Bioinformatics* **26:** 136 138.

Wa L, X e W, C e Y, Ta W, Ya J, Ye R, L L, L Y, X C, X a J, e a. 2010b. A d a c e e e a c e e e f ec ce f ce. *Plant J* **61:** 752 766.

W e BT, Ma e a S, Wa S, Sc be F, W d V, G d ead I, Pe e CJ, R e J, Ba e J. 2008. D a c e e e f a e a c a c e ed a e- c e de e . *Nature* **453:** 1239 1243.

Y J, H S, Wa J, W GK, L S, L B, De Y, Da L, Z Y, Z a X, e a. 2002. A d a e e ce f e ce e e (*Oryza sativa* L. . *indica*). *Science* **296:** 79 92.

Z a J, Fe Q, J C, Q D, Z a L, X e K, Y a D, Ha B, Z a Q, Wa S. 2005. Fea e f e e e ed e ce e ea ed b a a e- ca e a a f ESTs f a a ed cDNA b a f e e e *indica* ce c a M 63. *Plant J* **42:** 772 780.

Z a HY, He H, C e LB, L L, L a MZ, Wa XF, L XG, He GM, C e RS, Ma LG, e a. 2008. A Ge e- de a c a a e ea a c e ea f e INDEL b a d e e c e e e e ce b d. *Mol Plant* **1:** 720 731.

*Received February 4, 2010; accepted in revised form July 12, 2010.*

Downloaded from genome.cshlp.org on June 28, 2017 - Published by Cold Spring Harbor Laboratory Press

# Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq

Tingting Lu, Guojun Lu, Danlin Fan, et al.

*Genome Res.* 2010 20: 1238-1249 originally published online July 13, 2010
Access the most recent version at doi:10.1101/gr.106120.110

**Supplemental Material** http://genome.cshlp.org/content/suppl/2010/07/14/gr.106120.110.DC1

**Related Content** **Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome**
Guojie Zhang, Guangwu Guo, Xueda Hu, et al.
Genome Res. May , 2010 20: 646-654

**References** This article cites 33 articles, 18 of which can be accessed free at:
http://genome.cshlp.org/content/20/9/1238.full.html#ref-list-1

Articles cited in:
http://genome.cshlp.org/content/20/9/1238.full.html#related-urls

**Open Access** Freely available online through the *Genome Research* Open Access option.

**Creative Commons License** This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email Alerting Service** Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or **click here.**

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

Copyright © 2010 by Cold Spring Harbor Laboratory Press