Analysis of differential gene expression and alternative splicing is significantly influenced by choice of reference genome

ERIN SLABAUGH,[1,6] JIGAR S. DESAI,[1,6] RYAN C. SARTOR,[2] LOVELY MAE F. LAWAS,[3,4] S.V. KRISHNA JAGADISH,[3,5] and COLLEEN J. DOHERTY[1]

[1]Department of Molecular and Structural Biochemistry, North Carolina State University, Raleigh, North Carolina 27695, USA
[2]Crop and Soil Science Department, North Carolina State University, Raleigh, North Carolina 27695, USA
[3]International Rice Research Institute (IRRI), DAPO Box 7777, Metro Manila, Philippines
[4]Max Planck Institute of Molecular Plant Physiology, D-14476, Potsdam, Germany
[5]Department of Agronomy, Kansas State University, Manhattan, Kansas 66506, USA

ABSTRACT

RNA-seq analysis has enabled the evaluation of transcriptional changes in many species including nonmodel organisms. However, in most species only a single reference genome is available and RNA-seq reads from highly divergent varieties are typically aligned to this reference. Here, we quantify the impacts of the choice of mapping genome in rice where three high-quality reference genomes are available. We aligned RNA-seq data from a popular productive rice variety to three different reference genomes and found that the identification of differentially expressed genes differed depending on which reference genome was used for mapping. Furthermore, the ability to detect differentially used transcript isoforms was profoundly affected by the choice of reference genome: Only 30% of the differentially used splicing features were detected when reads were mapped to the more commonly used, but more distantly related reference genome. This demonstrated that gene expression and splicing analysis varies considerably depending on the mapping reference genome, and that analysis of individuals that are distantly related to an available reference genome may be improved by acquisition of new genomic reference material. We observed that these differences in transcriptome analysis are, in part, due to the presence of single nucleotide polymorphisms between the sequenced individual and each respective reference genome, as well as annotation differences between the reference genomes that exist even between syntenic orthologs. We conclude that even between two closely related genomes of similar quality, using the reference genome that is most closely related to the species being sampled significantly improves transcriptome analysis.

Keywords: Oryza sativa; differential gene expression; splicing; reference genome; RNA-seq; transcriptome analysis

INTRODUCTION

RNA-sequencing technologies have made it possible to evaluate genome-wide changes in the transcriptional state in any species from which quality RNA can be obtained. RNA-seq data from more than a thousand different species have been deposited into the NCBI Gene Expression Omnibus repository (Edgar et al. 2002; Barrett et al. 2013). Most commonly, RNA-seq data are used to identify transcripts whose abundance changes in response to environmental or developmental conditions through analysis of differentially expressed genes (DEGs). Additionally, RNA-seq analysis has facilitated the comparison of the relative abundance of isoforms for individual transcripts generated by alternative splicing (AS) of pre-mRNA molecules. As sequencing costs decrease, these techniques are being ever more widely applied to model and nonmodel species. For species for which a reference genome is available, the standard pipeline is to map the RNA-seq reads to that reference genome regardless of the accession being studied. However, the effect of genome relatedness between the species from which the RNA is derived and the species to which the RNA is aligned on downstream DEG and AS analyses has not been fully evaluated.

[6]These authors contributed equally to this work.
Corresponding author: cjdohert@ncsu.edu
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.070227.118.

For many agricultural species, there is a large amount of genetic variability between accessions in the same species. In practice, a diverse range of locally adapted high-performing varieties are grown and used for transcriptome and proteome experiments. Yet a single representative variety is often selected for genome mapping and serves as the reference genome for experiments performed in all varieties in that species. The effects of the evolutionary distance between the variety or individual being sequenced and the reference genome are rarely considered in downstream RNA-seq analysis. Here, we quantify the effects of choice of mapping genome in Asian rice (Oryza sativa) where three quality reference genomes are available.

There are two main groups of Asian cultivated rice, Oryza sativa ssp. japonica and Oryza sativa ssp. indica. Both groups are distinctive in the geographical locations in which they are grown, their genetic structure, and characteristics in grain quality and yield (Xu et al. 2015; Zhang et al. 2016). Rice derived from the indica subspecies accounts for more than 70% of worldwide production (Zhang et al. 2016); however, until recently, the only high-quality, publicly available reference genome for rice was for a temperate japonica variety called Nipponbare. As a result, most transcriptome data in rice has been aligned to the Nipponbare reference genome (Os-Nipponbare-Reference-IRGSP-1.0) regardless of the rice variety used in the study. However, thousands of rice germplasms are available to researchers, many of which are indica varieties (RiceVarMap; Zhao et al. 2015). Recently, two high-quality Oryza sativa ssp. indica genomes were published (Zhang et al. 2016), making it possible to analyze RNA-seq data from indica subspecies by alignment to a high-quality indica genome. We hypothesized that we would improve the accuracy of transcriptome studies by aligning RNA-seq reads to a more closely related reference genome. To this end, we aligned RNA-seq reads from the popular IR64 variety (Oryza sativa ssp. indica) to three high-quality genomes: Oryza sativa ssp. japonica cv Nipponbare (using the MSU annotation; Kawahara et al. 2013), and two Oryza sativa spp. indica lines, cv Minghui 63 and Zhenshan 97 (hereafter referred to as "MH63" and "ZS97," respectively; Zhang et al. 2016). IR64 is most closely related to MH63 as both are considered group II indica varieties (RiceVarMap; Zhao et al. 2015).

We determined that the reference genome used in transcriptome analysis had significant effects on read alignment, differential expression calling, and the identification of alternatively spliced transcripts. These effects were due to a combination of the presence of single nucleotide polymorphisms (SNPs) between IR64 and each reference genome and to annotation differences between reference genomes, which directly impacted the number of reads mapped to individual gene loci. We determine that the overall percentage of reads mapped is not a reliable indicator for optimizing the choice of reference genome. Care must be taken when interpreting DEGs and AS analysis when evaluating genotypes that diverge from the reference genome used for mapping. These results suggest that continued efforts to improve annotation and to provide additional individual genome sequences will have effects on discovery and evaluation of transcriptomes in both nonmodel and model organism.

RESULTS

Alignment of IR64 RNA-seq reads to three Oryza sativa genomes

To determine the effects of the choice of mapping genome on downstream transcriptome analysis, RNA-seq reads were aligned to three different Oryza genomes, and differences in differential gene expression and splicing analysis was compared for each data set (Table 1; Fig. 1). RNA was isolated from panicle tissue of field-grown IR64 rice, when 50% of the spikelets in the panicle had flowered. Four replicate biological samples were collected at two time points (dawn and dusk). Significant differences in expression levels and alternative splicing between dawn and dusk have previously been reported in rice and other plant species (Michael et al. 2008; Filichkin et al. 2010; Jonéźyk et al. 2011; Filichkin and Mockler 2012; Fu et al. 2012; James et al. 2012; Wang et al. 2012a; Reddy et al. 2013). Therefore, we chose to compare the effects of the choice of reference genome selection on the ability to identify DEGs and AS between dawn and dusk time points. The IR64 RNA-seq reads were mapped to either the MH63, ZS97, or Nipponbare reference genome using three different alignment programs: STAR (Dobin et al. 2013), HISAT2 (Kim et al. 2015), and Segemehl (Hoffmann et al. 2009, 2014). To analyze similarities and differences in counts for individual genes across genomes, we compared reciprocal best BLAST genes that also had syntenic orthologs in all three genomes. This additional synteny requirement was added to increase the stringency of orthologs identified by reciprocal best BLAST since using reciprocal best BLAST methods alone tend to generate many false

TABLE 1. Summary of transcriptome analysis of IR64 RNA-seq reads mapped to three different Oryza sativa genomes

	MH63	ZS97	Nipponbare
Total expressed genes	29,804	28,946	30,964
Syntenic DEGs	1845	1860	1825
Syntenic genes with DU splicing features	338	ND	119

The total number of expressed syntenic orthologs that were present in all three data sets was 17,039 loci. This pool of syntenic orthologs was used to compare DEGs and genes with DU splicing features identified by mapping to all three reference genomes.

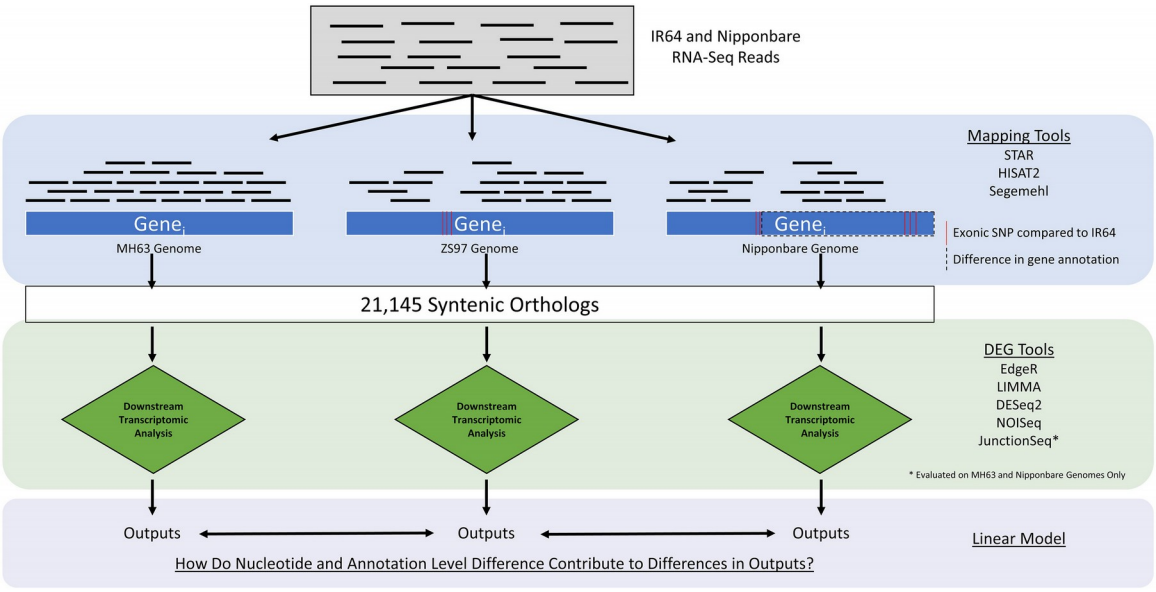


FIGURE 1. Schematic of genome alignment and transcriptome analysis. RNA-seq reads obtained from IR64 (Oryza sativa ssp. indica) panicle tissue or Nipponbare (Oryza sativa ssp. Japonica) seedlings were aligned to three high-quality Oryza sativa genomes. Each transcriptome alignment was performed using the aligning tools STAR, HISAT2, and Segemehl. The STAR alignment was analyzed by EdgeR, LIMMA, DESeq2, and NOISeq to identify DEGs. The IR64 alignments to MH63 and Nipponbare were analyzed by JunctionSeq to identify DU splicing features. The output from each transcriptome analysis was compared between syntenic orthologs. Differences between the number of reads mapped were observed for some loci, which can be attributed, in part, to differences in gene annotation and the presence of exonic SNPs. For the example depicted here (compare Gene$_i$ across all three genomes), syntenic genes that are annotated as being longer in one genome result in more reads mapped to that gene compared to the other genomes. Likewise, differences in the number of exonic SNPs may affect the number of reads mapped to a syntenic gene in one genome compared to another.

positives and are less accurate than phylogenetic approaches (Fulton et al. 2006; Dalquen and Dessimoz 2013; Lechner et al. 2014). Therefore, we combined reciprocal best BLAST with a syntenic approach to identify conservative orthologous relationships to ensure that comparisons across genomes were being made between the same gene. A total of 21,145 syntenic, orthologous genes were identified among all three genomes using MCScanX (Wang et al. 2012b; Supplemental Table S1, see Materials and Methods). This set of syntenic orthologs represents a highly conserved subset (~38%) of the rice genome. Scatterplots of the counts per ortholog show a substantial variation in the mapping depending on the reference genome (Fig. 2A–C). We observe a similar variation in counts per ortholog when mapping Nipponbare RNA to these three reference genomes (Supplemental Fig. S1). In contrast, when using a single reference genome, MH63, and comparing the alignment algorithms we observe a strong correlation between the results of the three aligners (Fig. 2D). While there is some variation between the aligners, particularly of reads with lower counts, there is more consistency between the alignment results when using three different aligners than when using the same aligner on the three different reference genomes.

IR64, a high-yielding, premium indica rice variety, is most closely related to MH63, while Nipponbare, a japonica subspecies, is the most distantly related (Zhao et al. 2015; Zhang et al. 2016). The percent of uniquely aligned IR64 RNA-seq reads to each genome varied between 83% and 88% (Fig. 3). The highest percent alignment was observed for the MSU annotation of the Nipponbare genome while the lowest was observed for the ZS97 genome. Differences were also observed for the mismatch rate per base for each genome, with the lowest mismatch rate occurring when mapped to the MH63 genome (Fig. 3). The total number of genes with mapped reads was similar between genomes with ~52% of annotated genes being detected (Supplemental Table S2). These results suggest that the choice of reference genome has a greater impact on the alignment of RNA-seq reads than the choice of alignment program. Furthermore, differences in the percent alignment of a transcriptome to a reference genome and the total number of genes with mapped reads do not necessarily reflect evolutionary relatedness indicating that the metric of percent alignment alone may not be a good indicator of mapping success.

Differential gene expression analysis is influenced by the reference genome

The majority of DEG analysis in rice has been performed using the Nipponbare genome. However, two new indica

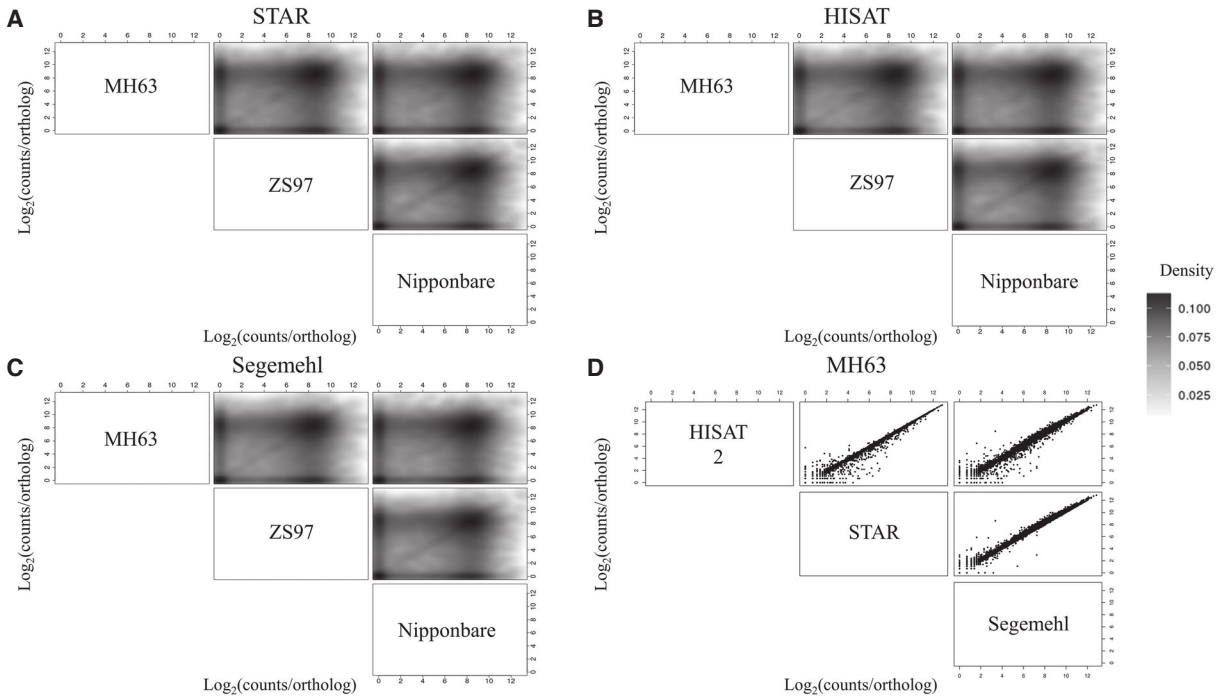


FIGURE 2. Genome alignment of RNA-seq reads derived from IR64 panicle. RNA-seq reads from field-grown IR64 rice were aligned using different alignment software to each of the three different annotated rice reference genomes: Nipponbare (Oryza sativa ssp. japonica using the MSU annotation, "MSU"), Minghui 63 (Oryza sativa ssp. indica, "MH63"), and Zhenshan 97 (Oryza sativa ssp. indica, "ZS97"). The effects of the reference genome are compared using three different aligners: (A) STAR, R^2 values: MH63-ZS97 (0.130), MH63-Nipponbare (0.104), ZS97-Nipponbare (0.152); (B) HISAT2, R^2 values: MH63-ZS97 (0.131), MH63-Nipponbare (0.106), ZS97-Nipponbare (0.153); or (C) Segemehl, R^2 values: – MH63-ZS97 (0.130), MH63-Nipponbare (0.105), ZS97-Nipponbare (0.151). (D) The effect of the alignment software is compared using the single MH63 reference genome. R^2 values: HISAT2-STAR (0.997), HISAT2-Segemehl (0.983), STAR-Segemehl (0.986). Scatterplots indicate the $\log_2$ of the counts of the reads aligned to syntenic orthologs in these comparisons.

genomes were recently published using a bacterial artificial chromosome (BAC)-by-BAC approach supplemented with Illumina and PacBio reads (Zhang et al. 2016). To determine if there would be substantial differences in identifying DEGs based on the choice of mapping genome, we identified DEGs between dawn and dusk samples for the IR64 panicle transcriptome mapped to all three reference genomes with the STAR aligner using four approaches to identifying DEGs: DESeq2 (Love et al. 2014), EdgeR (Robinson et al. 2010), LIMMA (Ritchie et al. 2015), and NOISeq (Tarazona et al. 2011, 2015). DEGs were identified using cutoff values of FDR-adjusted $P < 0.05$ for DESeq2, EdgeR, and LIMMA. For NOISeq, genes were considered DEGs when the probability of differential expression (q) >0.95. The total number of syntenic orthologs identified as DEGs was similar for all three genomes for each analysis method (Table 1; Supplemental Table S3). However, the four analysis methods identified different total DEGs, with LIMMA identifying the fewest DEGs across mappings to all three reference genomes, EdgeR and DESeq2 identifying slightly more than LIMMA, and NOISeq identifying the most DEGs. Previous reports

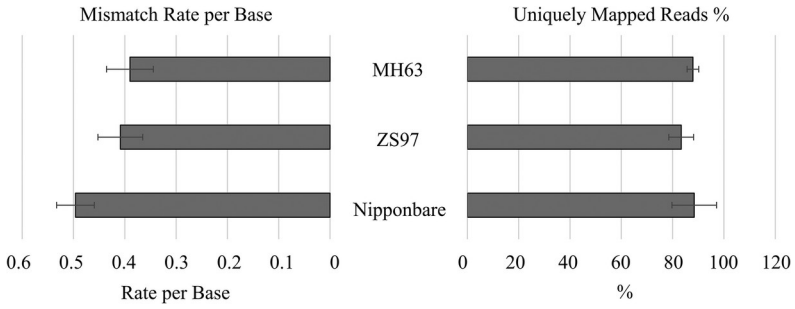


FIGURE 3. Alignment comparisons of RNA-seq reads derived from IR64 panicle mapped to different reference genomes. RNA-seq reads from field-grown IR64 rice were aligned to three different annotated rice genomes: Nipponbare (Oryza sativa ssp. japonica using the MSU annotation, "MSU"), Minghui 63 (Oryza sativa ssp. indica, "MH63"), and Zhenshan 97 (Oryza sativa ssp. indica, "ZS97") genomes. The Nipponbare genome shows the highest percent alignment, while the MH63 genome showed the lowest percent of multiple mapped reads. Percentages are representative of the average for all eight RNA-seq samples mapped to each genome and error bars represent standard deviation.

have shown similar performance of these algorithms (Seyednasrollah et al. 2013; Khang and Lau 2015). Using EdgeR, 2084 of the 21,145 syntenic genes were identified as DEGs when mapped to any of the three reference genomes. Only 1595 of those syntenic orthologs (~76%) were commonly identified as DEGs when mapped to all three genomes (Fig. 4). An additional 353 syntenic DEGs (17% of all identified DEGs) were identified when mapped to the indica genomes (MH63 and ZS97) that were not identified when the RNA-seq data were mapped to the Nipponbare genome. Of these 353 syntenic DEGs detected only when mapping to the indica genomes, 161 of these (~46%) were commonly identified when mapped to both MH63 and ZS97 (Fig. 4). Furthermore, 235 syntenic orthologs (~12% of DEGs identified using any reference) were only identified as differentially expressed when mapped to one of the reference genomes. Differences between syntenic DEGs were observed even between the two indica genomes (Fig. 4), suggesting that differences between even closely related reference genomes that were assembled using an identical pipeline influence DEG analysis. The DE identification methods varied in the total DEGs called, but did not vary in their sensitivity to the reference genome (Supplemental Fig. S2). Of the 2084 DEGs identified across all genomes identified using EdgeR, 490 DEGs were uniquely identified when mapping to only one or two genomes (23.5%). This pattern was similar for all DE methods; LIMMA (24.3%), NOISeq (23%), and DESeq2 (31.8%) (Supplemental Fig. S2B), indicating that this observation is not method dependent, but rather an effect from mapping differences using the three different reference genomes.

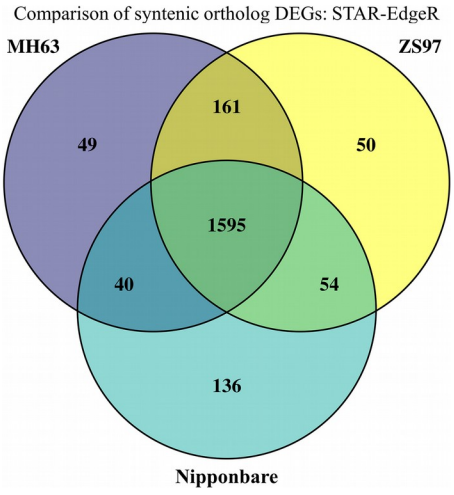


FIGURE 4. Comparison of DEGs that have syntenic orthologs. MCScanX was used to identify 21,145 syntenic orthologs, of which 17,039 were expressed in all three data sets. STAR was used to align the RNA-seq reads to each reference genome, and EdgeR was used to identify DEGs. Of the DEGs that had syntenic orthologs in all three genomes (2085 total syntenic orthologs), approximately 76% were commonly identified using all three reference genomes. Approximately 8% (161 genes) were identified as differentially expressed using both indica genomes.

A previous study demonstrated that identifying genes that have greater than a fourfold change in expression can help control for variability in the DEGs identified for RNA-seq data sets with less than six biological replicates (Schurch et al. 2016). Therefore, we compared DEGs that showed greater than a fourfold change. With a greater than fourfold cutoff using EdgeR, 481 syntenic orthologs were identified as DEGs. The proportion of genes commonly identified by all three genomes increased slightly (~78%; Supplemental Fig. S2) compared with DEGs that were selected based on statistical significance alone (76%; Fig. 4). However, the percent of genes identified as DEG in only one reference genome was similar using either significance alone or significance and the fourfold cutoff (Fig. 4; Supplemental Fig. S2). Therefore, although increasing the stringency of LFC cutoff values substantially reduces the number of identified DEGs, it does not account for the inter-genome variability observed when analyzing DEGs.

Differences in genome annotation and SNP density influences the number of reads mapped and the DEGs identified between syntenic genes

We further investigated the differences between loci identified as differentially expressed to determine if the effects of the reference genome on identified DEGs were due to mapping differences in individual transcripts, through single-nucleotide polymorphisms (SNPs) or the differences in gene annotation in the reference genome. The EdgeR significance score (adjusted P-values) for the differential expression of the IR64 transcripts between dusk and dawn when mapped to different reference genomes was compared (Fig. 5; Supplemental Figs. S3, S4). We observe that for transcripts uniquely identified in one reference genome, many are well below the significance score in the other genome (Fig. 5; Supplemental Fig. S3). We compared the adjusted P-values of syntenic orthologs identified as DEG only when mapped to the MH63 genome (49 loci) compared to their syntenic orthologs when mapped to the ZS97 or the most commonly used Nipponbare genome. By definition, the DEGs identified when mapped to the MH63 genome have adjusted P-values that are less than 0.05, while the adjusted P-values of the same loci when mapped to the ZS97 or Nipponbare genomes have a population of adjusted P-values that approach the 0.05 cutoff, and a population that have adjusted P-values of 1.0 (Supplemental Fig. S3A). This indicates that only some of the genes identified as DEGs when mapped to the MH63 genome show similar trends using the other reference genomes, but are just below the significance threshold. Likewise, syntenic orthologs that were

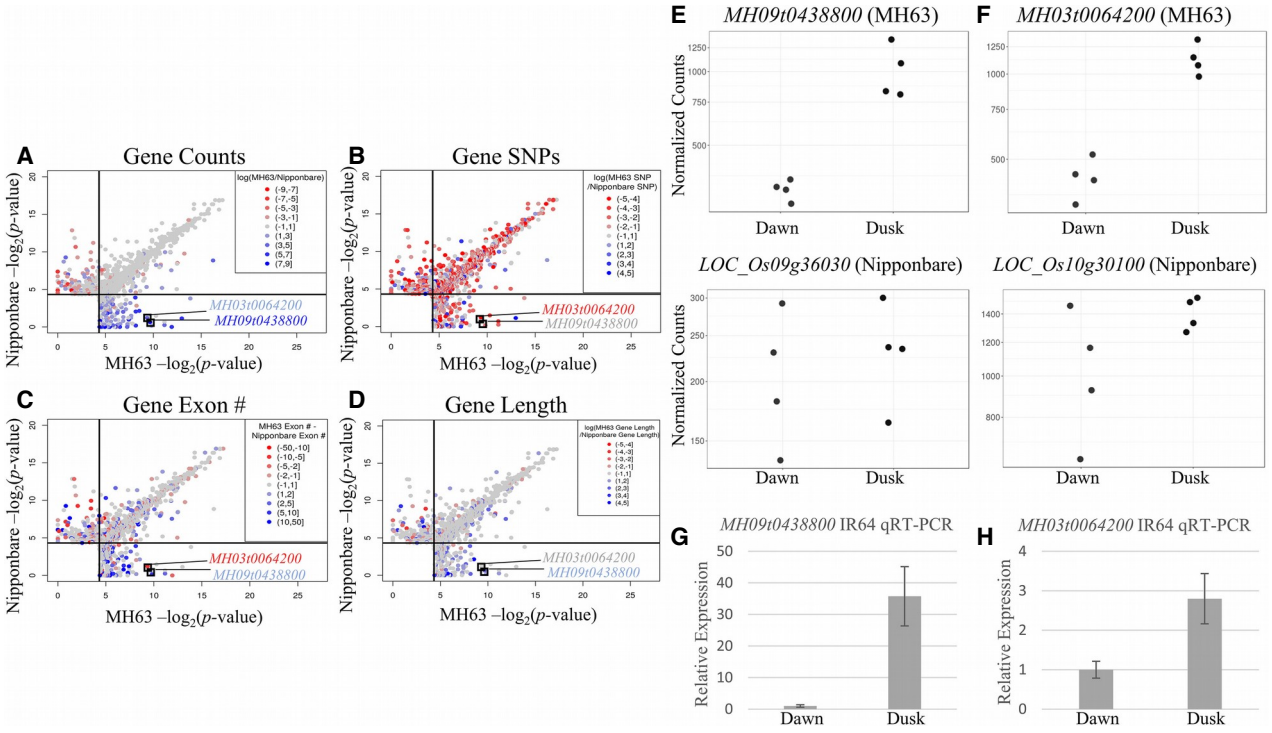


FIGURE 5. Comparison of features that contribute to differences in identifying IR64 DEGs between syntenic orthologs when mapped to different genomes. The significance of differential expression between dawn and dusk for each of the 21,145 syntenic genes as identified by EdgeR (−log of the adjusted P-value) is plotted for IR64 transcripts mapped using STAR to either Nipponbare (Oryza sativa ssp. japonica using the MSU annotation, "MSU") or Minghui 63 (Oryza sativa ssp. indica, "MH63"). The solid lines indicate the significance cutoff of adjusted P-value <0.05. The transcripts points are colored to highlight different features that could contribute to the observed differences in DEG identification. (A) Point colors are based on the ratio of total counts for each transcript. Red indicates higher counts in the MSU alignment and blue indicates higher counts in the MH63 aligned transcripts. (B) Point colors are based on the number of SNPs in each gene between the reference genome and the IR64 transcript from the RNA-seq read sequence. Red indicates more SNPs between the MSU genome and IR64, blue indicates more SNPs between MH63 and IR64. (C) The points are colored based on the number of exons in each genome annotation. Red indicates more exons per transcript in MSU genome and blue indicates a higher exon number in the MH63 genome. (D) The transcript points are colored based on the annotated gene length in each reference genome. Red indicates that the annotation for a transcript is longer in the MSU reference genome and blue indicates that the annotated length is longer in the MH63 genome. (E) Normalized counts of MH09t0438800 when mapped to MH63 and its syntenic ortholog LOC_Os09g36030 mapped to Nipponbare. (F) Normalized counts of MH03t006400 when mapped to MH63 and its syntenic ortholog LOC_Os10g30100 when mapped Nipponbare. (G) Relative expression of MH09t0438800 measured by qRT-PCR. (H) Relative expression of MH03t006400 measured by qRT-PCR.

only called differentially expressed when mapped to the ZS97 genome (50 loci), when mapped to the MH63 or Nipponbare genome had adjusted P-values that either approached the cutoff value or had an adjusted P-value of 1.0 (Supplemental Fig. S3B; Supplemental Table S5). These data indicate that the differences observed between DEG calls are not solely a product of statistical cutoff selection. The syntenic orthologs that were uniquely identified when mapping to the Nipponbare genome were loci that mostly approached the cutoff adjusted P-values when mapped to either of the indica reference genomes, MH63 or ZS97, suggesting that these may be a consequence of the statistical analysis and cutoff (Supplemental Fig. S3C; Supplemental Table S6).

We hypothesized that the differences in differential gene expression may be due to differences in transcriptome alignment and annotation differences between the three reference genomes (MH63, ZS97, and Nipponbare). We also hypothesized that genome relatedness, measured by the presence of SNPs, may partly influence the number of reads mapped per gene (Degner et al. 2009; Stevenson et al. 2013; Raghupathy et al. 2018). Because RNA-seq data provides the actual DNA sequence for each read, we can determine the number of exonic SNPs between IR64 and the three reference genomes. We developed an algorithm to identify the number of exonic SNPs between the mapped IR64 RNA-seq reads and each reference genome (see Materials and Methods). Consistent with the known evolutionary relationship between IR64 and the three reference genomes, MH63 had the fewest number of exonic SNPs in total compared to the IR64 reads (41,213), while the Nipponbare genome had the most (72,329; Supplemental Table S7). These results support that IR64 is most closely related to MH63, as previously determined

(Zhao et al. 2015; Zhang et al. 2016). Differences in annotation of the reference genomes could also influence mapping. Of the 21,145 syntenic orthologous genes only 11,654 (55%) had the same number of exons and only 598 (3%) were the same length. We sought to evaluate how these factors associate with loci differentially identified as DEG when using different reference genomes. We generated scatterplots of all syntenic orthologous genes comparing the EdgeR determined P-value for differential expression of IR64 reads mapped to one reference genome versus another. We colored each gene by ratios of counts, number of SNPs, number of exons, or gene length (Fig. 5; Supplemental Fig. S4). In the comparison between MH63 and Nipponbare, both annotation and sequence differences contribute to differences in DEG identification (Fig. 5). Many of the genes identified as uniquely DE in MH63 had higher counts in MH63 compared to Nipponbare (Fig. 5A, blue colored genes); a higher exon number in the MH63 genome (Fig. 5C, blue colored genes); and were longer genes in MH63 (Fig. 5D, blue colored genes). The DEGs uniquely identified in Nipponbare compared to MH63 also had more SNPs between the IR64 read sequences and the Nipponbare genome than the MH63 genome (Fig. 5B, red colored genes).

When we compare the uniquely identified DEGs between mapping to either of the two indica genomes, MH63 or ZS97, we observed that count differences and SNPs contributed to the uniquely identified DEGs (Supplemental Fig. S4A,B). However, most uniquely identified DEGs did not show annotation differences in the number of gene exons or the gene length (Supplemental Fig. S4C,D). This may be due to the similarity between the method of determining the annotation for these two genomes or similarity between these two indica genomes.

Because none of the reference genomes are a perfect representation of the IR64 samples we are analyzing, we evaluated syntenic orthologous genes that were uniquely identified as DEG depending on the reference genomes by qRT-PCR with specific primers designed for IR64 sequences. For example, MH09t0438800, had higher counts when mapped to MH63 than the syntenic ortholog when mapped to Nipponbare, and was identified as a DEG using the MH63 genome, but not the Nipponbare genome (Fig. 5A,E). Analysis of the IR64 transcript by qRT-PCR indicates that it is correctly identified as a DEG (Fig. 5G). A second gene, MH03t0064200 identified as a DEG when mapped to the MH63 genome, showed similar counts when mapped to MH63 as the syntenic ortholog mapped to the Nipponbare genome (Fig. 5A,F). Although the counts were in a similar range for MH03t0064200, we observed that there were more SNPs between the IR64 RNA-seq reads that mapped to this region and the Nipponbare genome than the MH63 genome (Fig. 5B), and there were more exons in the syntenic ortholog of this gene in the Nipponbare genome than in the MH63 genome (Fig. 5C). These genes are identified as DEG only in the MH63 genome, no matter which mapping software we used, STAR, HISAT, or Segemehl (Supplemental Fig. S5). qRT-PCR analysis of the IR64 RNA indicates that this locus is differentially expressed between dawn and dusk, thus mapping to the commonly used Nipponbare reference would have failed to identify these two DEGs.

To further understand if our observation of the effects of the reference genome on DEG identification holds true with a rice species other than IR64, we also evaluated the ability to identify DEGs between dawn and dusk from RNA-seq data obtained from Nipponbare seedings (Supplemental Fig. S6). In this case, since the transcripts are from the Nipponbare genome, DEGs identified when mapped to Nipponbare should be more reliable. As expected, when mapping to the Nipponbare reference genome, the Nipponbare seedlings reads had the largest number of unique DEGs and the fewest number of SNPs, compared to mapping to either MH63 or ZS97 genomes (Supplemental Figs. S6, S7). Furthermore, as we observed for IR64, both SNPs and annotation differences contributed to the identification of DEGs (Supplemental Fig. S7).

Genome relatedness contributes significantly to the confidence of DEG identification

To further investigate the effects of exonic SNPs and annotation features on transcriptome analysis, we sought to identify the significant features that contribute to the difference in identifying DEGs (based on P-values) when mapping to the three reference genomes using linear regression. For each syntenic ortholog, we determined the significance of DEGs (adjusted P-value) for each gene when the IR64 RNA-seq data were mapped to the MH63, ZS97, or Nipponbare genomes. We then sought to explain the differences in the P-values of the DEGs identified when mapped to each reference genome (response variable) using the following features (explanatory variables): counts, gene length, SNPs between IR64 transcripts and the reference genome, number of exons, sequence identity, alignment length, genomic SNPs, and sequence gaps. These explanatory variables were calculated for each reference genome. We then performed pairwise comparisons between individual reference genomes of the response and explanatory variables, which was then used as input for the linear model. Depending on the reference genomes compared, different explanatory variables contributed to the observed differences seen in DEGs identified between genomes (Fig. 6; Supplemental Table S10). Combined, these features explained 55%–67% of the variation in DEGs between reference genomes. Counts (the log2 difference in total counts per syntenic ortholog) contributed the most in explaining the variance in DEG identification in all genome comparisons. Two related features, sequence identity (the percent sequence similarity between syntenic

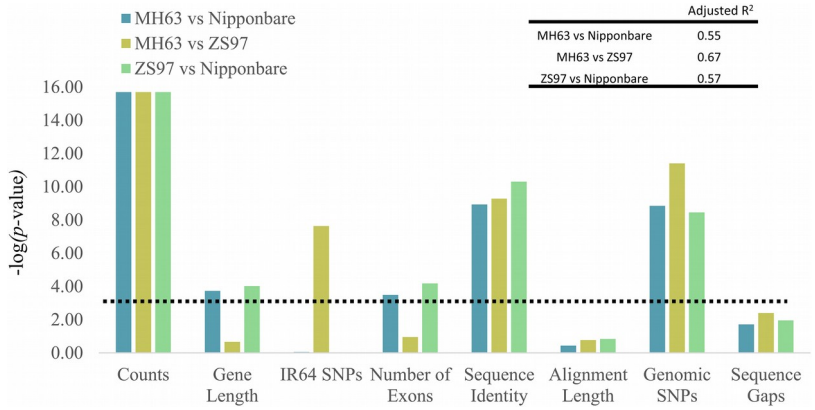


FIGURE 6. Explanatory features that contribute to differences in DEG identification from the linear model output. Linear regression was used to identify significant relationships between DEG significance value and selected explanatory variables. Counts between IR64 reads mapped to the MH63, ZS97, and Nipponbare (MSU annotation) genome were obtained from the HTSeq-count output. Gene length and exon length were obtained from reference GTF files. Exon length and the total number of exons were calculated using the longest transcript model. Mismatches (RNA) are the number of SNPs identified by GATK when mapping reads to each genome. Sequence identity, Alignment length, Mismatches (DNA), and Sequence gaps were obtained from BLAST results between transcript sequences for each genome. –Log(P-value) indicates the significance of the explanatory variable to DEG.

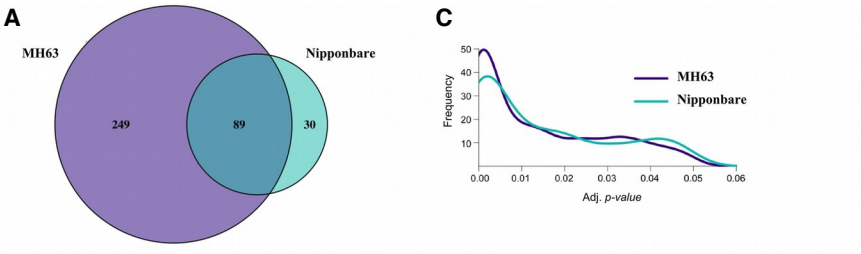
orthologs in the different reference genomes) and genomic SNPs (the number of nucleotide mismatches between syntenic orthologs), contributed significantly to the differences in DEG significance in all three comparisons. While sequence identity contributed similarly to all three pairwise reference genome comparisons, Genomic SNPs contributed more to the difference in DEGs identification when comparing the similarly generated references MH63 and ZS97, the two indica genomes. Another feature of sequence variation, the IR64 exonic SNP feature (the SNPs identified between the IR64 transcripts and the reference genomes) was only identified as a significant feature in the comparison between MH63 and ZS97. In contrast, two features related to annotation, number of exons (the difference in annotated exons between syntenic orthologs) and gene length (the difference in annotated transcript length) were significant features only in the comparisons between the indica and japonica genomes. This analysis suggests that count differences between syntenic orthologs is not the only factor that accounts for differences in DEG identification, as we observed in MH03t0064200 (Fig. 5F). These results show that both genome relatedness measured through nucleotide level differences and genome annotations influence DEG analysis and both factors account for the differences observed when analyzing differential gene expression across different reference genomes.

Splicing analysis is influenced by reference genome

To elucidate the effects that the choice of reference genome may have on AS analysis, we performed JunctionSeq (Hartley and Mullikin 2016) analysis on the IR64 RNA-seq data mapped to different rice genomes. We focused our splicing analysis on comparing differences between mapping to the MH63 and Nipponbare reference genomes. The reference genome for Nipponbare is the commonly used reference genome and between MH63 and ZS97, IR64 is most closely related to MH63. We chose to use the JunctionSeq package for splicing analysis as it can determine differential exon as well as splice junction usage and is built on the DEXSeq package (Hartley and Mullikin 2016). We identified features (exons or splice junctions) that were differentially used (DU) between dawn and dusk when RNA-seq reads were mapped to either the MH63 or Nipponbare genomes. By aligning reads to the MH63 genome, 931 DU features were identified, whereas only 276 DU features were identified by aligning the reads to the Nipponbare genome (Table 1; Supplemental Table S8). Of the total number of DU features identified when mapped to the MH63 genome, 66% corresponded to splice junctions, while only 40% of DU features accounted for splice junctions when mapped to the Nipponbare genome (Supplemental Table S8). This indicates that not only is the overall ability to detect DU features affected, but there is also a change in the distribution of the number of exons or splice junctions identified as DU.

To directly compare the effects of mapping to two different reference genomes on the ability to detect alternative isoforms using JunctionSeq, we compared syntenic orthologs of loci that contained at least one DU feature (a total of 368 genes) when mapped to either genome. Between the MH63 and Nipponbare reference genomes, 89 syntenic orthologs (~24%) were commonly identified as containing a DU feature (Fig. 7A). An additional 249 syntenic orthologs (~68%) were identified by mapping to the MH63 genome that were not identified when mapping to the Nipponbare genome, while 30 syntenic orthologs (~8%) were identified using the Nipponbare genome that were not identified when mapped to the MH63 genome. Furthermore, mapping to MH63 increased the total number of splicing features identified as DU by 3.4-fold, compared to the Nipponbare genome (Supplemental Table S8).

We compared the distribution of adjusted P-values of all features identified as DU when mapping the IR64 RNA-seq data to either genome to further investigate the differences on the JunctionSeq output based on the choice of the



B

SNPs/Boundary	0	1	2	3	4	5	Total Features	Average SNPs/Feature
Unique MH63	362	52	10	1	3	1	429	0.214
Common MH63 and Nipponbare	158	11	5	0	0	0	174	0.121

FIGURE 7. Comparison of syntenic loci that contain at least one DU splicing feature. (A) Syntenic orthologs that contain at least one DU exon or splice junction were compared between the MH63 and Nipponbare ("MSU") genomes. Of the total 368 syntenic orthologs that contain a DU feature, only ~24% (89 genes) were commonly identified when mapped to both reference genomes. The majority of syntenic orthologs (~92%; 338 genes) that contained at least one DU splicing feature between dawn and dusk were identified when mapped to the MH63 genome. (B) Table of the number of SNPs between Nipponbare and MH63 genomes identified at DU exon/intron boundary for each class of gene. (C) Adjusted P-values of significantly DU features were plotted from the JunctionSeq output when mapped to the MH63 and Nipponbare ("MSU") genomes. A higher frequency of lower adjusted P-values was observed when IR64 RNA-seq reads were mapped to the MH63 genome compared to the Nipponbare genome, while a higher frequency of adjusted P-values that approached the significance threshold (0.05) was observed when mapped to the Nipponbare genome.

reference genome. Adjusted P-values tended to be lower when mapped to the MH63 genome, with a higher frequency of features having adjusted P-values of less than 0.01, compared to the Nipponbare genome (Fig. 7B). In contrast, mapping to the Nipponbare genome resulted in a higher frequency of adjusted P-values between 0.04 and 0.05 (Fig. 7B). These data indicate that mapping IR64 RNA-seq data to the MH63 genome increased the confidence of exons and splice junctions identified as DU.

Differences in identifying DU features based on the reference genome could be due to differences in gene annotation or sequence variation. Gene annotation differences contribute to variation in identified DU features because missing an exon or splice junction would prevent identification. Genome relatedness may have an enhanced impact on identifying DU features because higher sequence variation at exon boundaries could impact the mapping of spliced reads used to determine DU features. To evaluate if identification of DU features is influenced by the number of SNPs between IR64 and the reference genomes, we calculated the SNP density at exon boundaries. We compared the frequency of SNPs between DU features uniquely identified when mapped to MH63 and those identified when mapped to both reference genomes. From the 249 unique MH63 genes, 429 DU features were identified and from the 89 genes identified using either reference genome, 174 DU features were identified. MH63 unique genes averaged 0.214 SNPs/boundary and commonly identified DU genes averaged 0.12 SNPs/boundary, a significant increase in the SNPs/boundary for uniquely identified DU genes (Student's t-test P-value 0.027). There is a higher distribution of exon junctions with >1 SNP in the set of genes with DU splicing features only identified when mapping to MH63 compared to the Nipponbare reference (Fig. 7B).

For validation of the differential splicing analysis, we arbitrarily chose a gene (MH12t0411000) that was identified as having a DU splice site (J011) only when mapped to the MH63 genome compared to its syntenic ortholog (LOC_Os12g39630; Fig. 8C,D) and was identified as a DEG in both genomes (Fig. 9A,B), indicating that enough counts were detectible to determine differences at the whole gene level. Validation for both the differential gene expression and splicing analysis was carried out using semiquantitative RT-PCR. We designed primers to detect differential expression at the gene level (Fig. 9A, upper panel), or differential usage of the splice site (Fig. 9A, middle panel) between dawn and dusk compared to UBC-E2 as a control (Fig. 9A, lower panel; Auler et al. 2017). Relative quantification of band intensity showed that the gene, as well as the splice junction, were both more highly expressed at dusk than at dawn (Fig. 9A). Furthermore, the ratio of relative expression at dusk compared with dawn (dusk: dawn) was higher for the splice junction (Fig. 9B), indicating that this locus in IR64 is both differentially expressed at the gene level and differentially spliced between dawn and dusk, consistent with DESeq2 and JunctionSeq analysis when mapped to the MH63 genome. The J011 splice junction in MH12t0411000 is an example of a DU splicing feature that would have been missed by mapping to the more commonly used, but more distantly related Nipponbare reference genome alone.

To further investigate why the J011 splice junction was identified as DU when mapped to the MH63 genome, but not the MSU genome, we analyzed differences in read counts of splicing features between the MH12t0411000 and LOC_Os12g39630 loci. We observed that total read counts varied only slightly between these syntenic orthologs, and read counts for the DU splice junction was identical between both genomes (Supplemental Table S9). Therefore, this change in detection may be due to differences in total read counts across the entire gene and information sharing across gene loci that influences dispersion estimates of individual features when analyzing differential

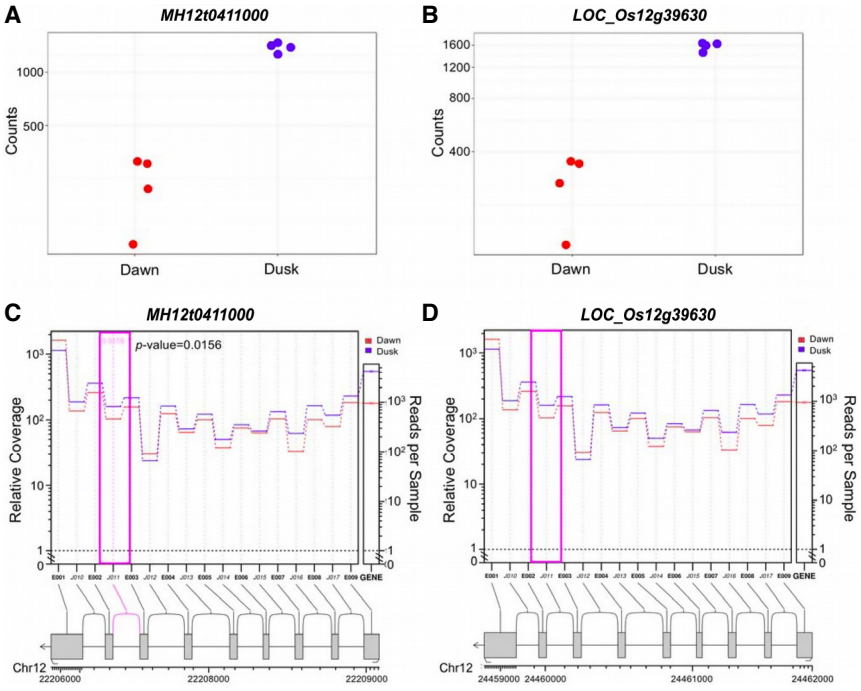


FIGURE 8. Differential gene expression and splice site usage of syntenic orthologs MH12t0411000 and LOC_Os12g39630. (A,B) Both MH12t0411000 and LOC_Os12g39630 are called differentially expressed between dawn and dusk at the gene level. Splice junction J011 is called DU when mapped to the MH63 genome (C), but not when mapped to the Nipponbare genome (D).

usage of exons and splice junctions (Hartley and Mullikin 2016).

DISCUSSION

Choice of reference genome has a significant impact on downstream transcriptional analysis

Identification of DEGs and DU splicing features are both affected by choice of mapping genome. Although differences were observed between genes that were called differentially expressed between dawn and dusk, the majority of DEGs from syntenic orthologs (~75%) were commonly identified when mapped to all three genomes. To consistently compare the same orthologs between the sequences mapped to the three genomes, we focused our comparisons on syntenic orthologs. This subset of genes is most likely to be conserved and well annotated between the three genomes. In each genome, the syntenic orthologs are only ~70% of the expressed genes. We ignored between 7800–9800 expressed genes that were detected when mapped to each reference genome, because they lacked syntenic orthologs and direct comparisons would be challenging. Therefore, our estimates are likely an underrepresentation of all the differences that result from the choice of mapping genome when all genes are considered.

In comparison to differential gene expression, splicing analysis was affected to a greater extent by the choice of reference genome. When evaluating the effects of the choice of reference genome on identifying DU transcript isoforms, ~92% of syntenic loci identified as having at least one DU splicing feature was identified by mapping to the MH63 genome, while only 24% of syntenic loci were commonly identified by both genomes (Fig. 8). This larger effect on splicing analysis could be due to several factors. First, for splicing analysis, the features (e.g., exons and junctions) are smaller than DEG features (the entire gene length). As a result, smaller variations in reads per feature are of higher consequence in splicing analysis. Secondly, since there are differences in the total number of exons between each genome, the total number of statistical tests varies between each genome. For DEG analysis we could control this by limiting our focus to a set of syntenic

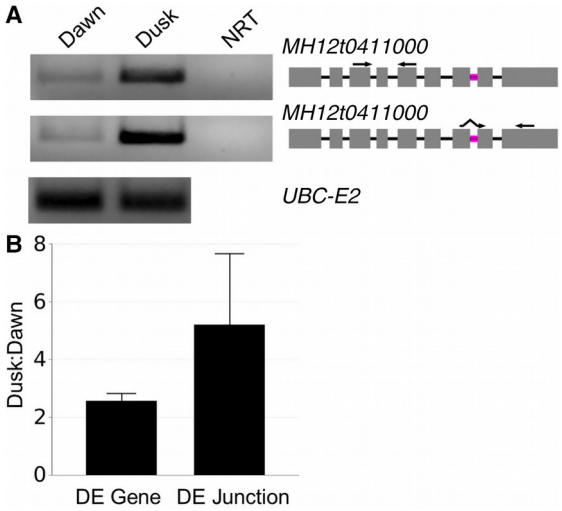


FIGURE 9. Validation of differential gene expression and splice site usage of the MH12t0411000 locus. Semiquantitative RT-PCR was used to confirm DESeq2 and JunctionSeq results of the MH12t0411000 locus. (A) Primers used to either amplify gene expression (upper panel) or expression of the splice junction J011 (middle panel) both showed increased band intensity at dusk, compared to UBC-E2 as the reference (lower panel; Auler et al. 2017). (B) Band intensity of UBC-E2 at dawn was used as a reference for relative quantification. The dusk to dawn ratio of pixel intensity of the MH12t0411000 locus was 2.6 at the gene level and 5.2 for splice junction J011, on average. Data are representative of three biological replicates. Error bars represent standard deviation.

orthologs. However, for splicing analysis, the variation in the number of features between each genome will impact the overall statistical analysis. Differences in the number of splice junctions identified as DU may also partially account for this increase, as 66% of DU features identified were splice junctions when mapped to the MH63 genome compared to 40% when mapped to the Nipponbare genome (Supplemental Table S8). JunctionSeq identifies splice junctions as places where contiguous reads span a non-contiguous region in the genome (e.g., when one read aligns to two exons that are separated by an intron; Hartley and Mullikin 2016). Because SNPs can be found at splice site donors and acceptors (Zhang et al. 2016), and the length of defined splicing junctions are relatively short, an increase in the frequency of SNPs between the sampled species and the reference genome may have a greater impact on the identification of splicing junctions. In support of this we observe that the DU features identified only in MH63 showed higher average SNPs per boundary than those commonly identified when either MH63 or Nipponbare reference genomes were used. Therefore, relatedness between the species being analyzed and the reference genome used may have a greater influence on splicing analysis compared with differential gene expression analysis. Again, this is likely an underestimation of the overall effect on splicing analysis since we draw conclusions from comparing syntenic loci, which are more likely to be conserved.

Differences of genome annotation between syntenic loci affects transcriptome mapping

We observed that differences in counts had the greatest influence on the DEG identification between theIR64 transcripts when mapped to the MH63, ZS97, and Nipponbare genomes. Count differences could arise from differences in annotation and polymorphisms due to genetic divergence between the transcript genome and the reference. These differences in genome annotation may arise from differences in genome assembly, gene prediction, and gene annotation methods used in the construction of these reference genomes. These differences in gene and exon annotation ultimately influence the number of reads mapped to each gene and the total number of genes identified. If a gene is improperly annotated as being shorter than it actually is, read counts that correspond to the missing annotated region will be lost for that gene (Fig. 1; Supplemental Table S11). In other words, if two syntenic loci vary in their annotated length, but biologically are similar in length, more reads will be mapped to the longer annotated gene compared to the shorter annotated gene, which will ultimately affect downstream analysis. Not only can gene annotation differences lead to variations in mapped reads per gene, but the effects of transcript misannotation can also lead to misinterpretation of AS products (Brown et al. 2015). Therefore, methods used for genome construction and annotation can influence transcriptome analysis independent of evolutionary relatedness. However, the observation that variation in gene annotation and SNP features we considered only account for 55%–67% of the observed differences in DEG identification indicates that there are factors we have not identified that also affect downstream analysis. Other potential contributors could be the distribution of SNPs within an exon, types of SNPs, sequencing errors, the distribution of reads within a gene, or the similarity of a gene to other genes in the genome that may influence mapping using default parameters, as well as other yet unknown factors.

Differences in counts for a single gene do not explain all the effects of the reference genome

We observed that not all effects of the reference genome on the identified DEGs or DU splicing features was due to differences in counts at that locus (e.g. Figs. 5F,H, 9; Supplemental Tables S4–S6). Count differences for these genes did not fully explain the differences in significance of DEGs (Figs. 5, 6). This suggests that the overall variation of the genome could influence the identification of downstream analysis even when the specific gene of interest is similar between the reference genomes. The algorithms for identifying DEGs or DU splicing features use information sharing across genes for variance estimation; therefore, the impacts of multiple mapping differences across many genes could contribute to the overall statistical evaluation and impact even syntenic orthologs that are similar in annotation. This difference may explain some of the enhanced effects of the reference genome we observe in splicing analysis since the total number of features will vary between genomes (Supplemental Table S7). For example, the splice junction J011 is called DU in MH12t0411000 when mapped to MH63, but is not identified as DU when mapped to Nipponbare (Fig. 8), the annotation and counts of this gene are similar and the gene is identified as DEG when using either reference genome. However, the impact of the total number of features is likely to have a greater impact on DEG identification than we observe in our analysis. Here we limited our analysis to a core set of syntenic orthologs, so the same total number of genes were analyzed for DEGs. Thus there were no differences in the number of annotated genes in each genome or the percent of those genes that map (Supplemental Table S2). However, in a more standard analysis, the gene set would not be restricted and the effects of the variation across the entire genome would likely be exacerbated for even genes with similar annotations due to the importance of the total expression in the normalization and variance calculations in the DEG identification algorithms (Dillies et al. 2013).

Using a closely related reference genome may still introduce bias into downstream data analyses

Although MH63 is the most closely related high-quality reference genome to IR64, the existence of >40k exonic SNPs between MH63 and IR64 may fail to identify significant differences in DEGs or exon and splice site usage. These differences may result in genes or exons and splice sites not being identified as significantly differentially expressed when in fact they would be identified as differentially expressed if using an IR64 reference genome. This may lead to an incomplete snapshot of RNA-seq data analyses, especially when analyzing AS. However, using an available mapping genome that is the closest relative to the sampled species greatly improves these downstream analyses. For example, mapping IR64 RNA-seq reads to the Nipponbare genome obscured the identification of 249 syntenic genes that contained at least one DU splicing feature (Fig. 7; Supplemental Table S8). Therefore, mapping RNA-seq reads of one rice species to the rice genome of the closest evolutionary relative can greatly improve AS analysis. Ideally, the reference genome and the sampled genome would be from the same subspecies, however, when a high-quality annotated genome is not available for the subspecies being studied, as may often be the case with the wide genetic architecture within nonmodel species that are amenable to experimental research, using the most closely related reference genome will increase the accuracy of downstream data analysis. These results underscore the need for generating more high-quality annotated genomes in subspecies where there is significant intra-species variation.

Parameters used for genome alignment may influence downstream RNA-seq data analysis

We used STAR to align IR64 RNA-seq reads to individual genomes. The default setting of STAR controls mismatch rate based on mismatches to either mapped read length or total read length. If the ratio of mismatches to the mapped read length are less than 0.3 and if the ratio of mismatches to the total read length is less than 1, this passes the criteria. We decided to use this default setting for a couple of reasons. First, many studies that analyze RNA-seq data also use this default setting, therefore, we wanted to accurately simulate data handling by other studies to demonstrate how using different reference genomes may realistically impact data analysis. Second, allowing for up to two mismatches increases the total number of reads mapped to the genome, as it helps to account for differences in SNPs as well as sequencing errors. We evaluated changing the tolerance for mismatches and the number of uniquely mapped reads plateaus in MH63 at one mismatch allowed and in Nipponbare between three to four mismatches (Supplemental Fig. S8). Therefore, altering the parameter of allowed mismatches for the genome alignment tool used may also influence downstream RNA-seq data analysis and should be taken into consideration during experimental design.

Overall percent alignment may not be reflective of the most closely related genome

The overall percent alignment of IR64 RNA-seq reads differed when mapped to the MH63, ZS97, or Nipponbare reference genomes (Fig. 2). We observed that the highest percent alignment occurred with the Nipponbare genome while the lowest was with the ZS97 genome. In contrast, the lowest percentage of multiple mapped reads was observed when mapping to the MH63 genome. The parameter of fewest multiple mapped reads may be indicative of relative species relatedness that should be investigated further. Considering that MH63 is the most closely related reference genome to IR64, we conclude that the overall percent alignment of an RNA-seq data set to a reference genome is not an ideal benchmark to use when choosing between reference genomes.

In summary, the accuracy of identifying differentially expressed and alternatively spliced genes is significantly impacted by the choice of reference genome. The contribution of the reference genome is influenced by several factors. First, the genetic distance between the reference and the sample, which translates into SNPs that have a direct effect on mapping and therefore counts. Second, the quality and completeness of the genome, which also results in altered counts as missing or incorrect features will not be accurately mapped. Finally, as a consequence of these first two effects on counts, the modeling of the gene expression through the DEG-analysis algorithms can be altered, thus impacting even genes with similar counts when mapped to different reference genomes (e.g., Fig. 5F). This effect on genes which are similar across the reference genomes reveals that the DEG-analysis algorithms we used all learn their parameters from the entire data set. Therefore, the impacts of sequence relatedness and genome annotation quality persist even when restricting the analysis to conserved gene annotation models. We propose that when working with nonmodel species, if researchers have a choice of reference genomes of similar annotation quality, the reference genome of the closest evolutionary relative should be used to maximize gene discovery efforts. In many species, on-going efforts exist to sequence multiple individuals to quantify the genetic diversity within genus and species (The 1000 Genomes Project Consortium 2015; The 1001 Genomes Consortium 2016). Recent sequencing efforts show that major differences exist between

sequenced human genomes (Sherman et al. 2018). The large effects we observe of reference genome choice in transcriptional analysis, particularly for investigating isoform variant expression, underscores an additional benefit of these sequencing efforts, the improvement of downstream analysis. These results suggest that continued efforts to improve annotation and provide additional individual genome sequences will have effects on discovery and evaluation of transcriptomes in both nonmodel and model organisms.

MATERIALS AND METHODS

Plant material

IR64 rice plants were grown under field conditions at the International Rice Research Institute, Philippines (14° 13′N, 121° 15′E, 23 MASL) in 2014 during the dry season. Panicles from primary tillers were collected at 6:15 a.m. (dawn) and 6:00 p.m. (dusk) from plants when 50% of the middle portions of the panicles were flowering (i.e., the upper 50% of the panicle had finished flowering). Four biological replicates were collected for each time point.

RNA extraction and RNA-seq

For RNA extraction, the panicle samples were first ground in liquid nitrogen with a metal pestle. The tissue was then lyophilized at −60°C overnight before RNA extractions. Total RNA was extracted using RNeasy Plant Mini Kit (Qiagen) with the RLT lysis buffer. The provided RNA extraction protocol was followed with the inclusion of DNase treatment. After the RWI wash step, 3 µL of DNase I (Roche), 8 µL buffer (200 mM Tris, pH 8.0, 20 mM $MgCl_2$, 500 mM KCl), and 69 µL nuclease-free water was added to each column and incubated for 10 min. Following DNase treatment, the column was washed again with the RWI buffer from the Qiagen kit. RNA concentration was then measured with NANOdrop 2000 (Thermo Scientific). mRNA was isolated from 2 µg of total RNA using the NEBNext Poly(A) Magnetic mRNA Isolation Kit (NEB). Before library preparation, the mRNA was heated to 95°C for 15 min to achieve 150–200 bp fragment sizes. NEBNext Ultra RNA Library Prep Kit for Illumina was then used to generate directional libraries for sequencing. First strand cDNA was primed with random hexamers using Protoscript II reverse transcriptase and followed by second strand synthesis. The cDNA was purified using AMPure beads. End repair, adaptor ligation, and size selection with AMPure beads were performed as described to recover 150–200 bp fragments and removed adaptors. Fifteen cycles of PCR using USER for strand specificity were performed. Concentration and size verification of the libraries was performed on an Agilent Bioanalyzer high sensitivity DNA chip after a 1:4 (or 1:10) dilution. Concentrations were verified using the NEBNext Library Quant Kit for Illumina. Raw single-end sequencing reads were generated from libraries diluted to 10 nmol/µL concentrations using the Illumina HiSeq2000 platform at North Carolina State University's Genomics Science Laboratory.

Quality control and transcriptome alignment

For quality control, seqtk (https://github.com/lh3/seqtk) and FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) were used to generate high-quality trimmed reads. Trimmed fastq files were uploaded to NCBI GEO (Series GSE92302, Samples GSM2425416-GSM245419 and GSM242 5432-GSM2425435). STAR (version 2.5.3a), TopHat2 (v2.0.4), and Segmehl (0.2.0-418) were used to align trimmed reads to either the Minghui 63, Zhenshan 97, or Nipponbare (MSU annotation) genomes. For all aligners, the option for a reverse stranded library was used; all other parameters used were default. GFF and genome sequence files were obtained from RIGW (http://rice.hzau.edu.cn/rice/) or the Rice Genome Annotation Project websites (http://rice.plantbiology.msu.edu/). Genome sequence and annotation files were downloaded from their respective websites in September 2016. Original GFF files were parsed and reformatted to GTF files that matched input requirements for DESeq2, EdgeR, NOISeq, LIMMA, and JunctionSeq packages (scripts available at www.github/DohertyLab).

Differential gene expression and splicing analysis

The DESeq2 package (Love et al. 2014), EdgeR (Robinson et al. 2010), NOISeq (Tarazona et al. 2011, 2015), and LIMMA (Ritchie et al. 2015) were used to identify DEGs for each of the three genomes independently. Significance cutoff values were set to adjusted P-value <0.05 for DESeq2, EdgeR, and LIMMA. For NOISeq, genes were considered DEGs with a probability of differential expression (q) >0.95. The JunctionSeq package (Hartley and Mullikin 2016) was used for differential splicing analysis for the Minghui 63 and Nipponbare genomes (MSU annotation), using the companion package QoRTs to generate raw counts (Hartley and Mullikin 2015). Analysis was performed for both known and novel splice junctions. For JunctionSeq analysis, the FDR was set to 0.05; all other parameters used were default.

Identification of syntenic orthologs

The MCScanX package (Wang et al. 2012b) was used to identify conservative syntenic orthologs between MH63, ZS97, and Nipponbare (MSU annotation) genomes. MCScanX allows for the comparison of multiple custom genomes and identifies orthologs using pairwise best reciprocal BLAST and syntenic relationships. GFF files were parsed and reformatted to match input requirements. A data frame of the MCScanX collinearity results was compiled using a customized script in R (available at www.github/DohertyLab). Only gene loci that were filtered by reciprocal best BLAST and had syntenic orthologs in all three genomes (21,145 loci; Supplemental Table S1) were used to make cross genome comparisons.

qRT-PCR (for DEG validation)

Reverse transcription was performed, following manufacturer instructions, from 1 µg of total RNA using the Bio-Rad iScript Reverse Transcription Supermix. cDNA was diluted to 1:100. Bio-Rad SYBR Green Master Mix was used for qPCR using a Bio-Rad CFX instrument. Gene-specific primers were selected

that produced only a single, sharp inflection using a dissociation curve. Each biological replicate was measured by the average of four technical replicates; outlier technical replicates were removed from the analysis. Data analysis was performed using Bio-Rad CFX Software, dawn samples were set as the reference, and UBC-E2 was used as a reference housekeeping gene.

Semiquantitative RT-PCR (for splicing validation)

cDNA was synthesized from total RNA samples using the iScript Advanced cDNA Synthesis Kit (Bio-Rad) with 600 ng of RNA as the input for each sample. One microliter of undiluted cDNA was used as the template for each PCR reaction. Primer sequences can be found in Supplemental Table S13. Relative quantification of gel bands was performed using a Bio-Rad Gel Doc EZ Imager using the UBC-E2 band from the dawn sample as the reference. Semiquantitative PCR was carried out on three biological replicates for both dawn and dusk.

SNP identification

A custom script was developed to identify SNPs between IR64 RNA-seq reads and the MH63, ZS97, and Nipponbare (MSU annotation) genomes (available at www.github/DohertyLab). The pipeline uses picard (version 2.10.2; https://github.com/broadinstitute/picard) to order and remove duplicated sequence alignments and GATK functions (version 3.7; https://software.broadinstitute.org/gatk/) for SNP identification. Exonic SNPs per transcript were assessed by counting the number of SNPs per exon based on the longest transcript model for each syntenic loci. Total read counts per gene were obtained from the DESeq2 output. Exonic SNP density was calculated by dividing the number of exonic SNPs per gene by total gene length.

Linear model regression

Counts represents the $\log_2$ difference in total counts between syntenic orthologs when mapped to the compared reference genomes. Gene Length represents the difference in annotated transcript length syntenic orthologs between genomes. IR64 SNPs represents the difference in SNP density identified by mapping IR64 RNA-seq reads on syntenic orthologs between genomes. Number of Exons represents the differences in annotated exons in the syntenic orthologs between genomes. Sequence Identity represents the percentage of sequence similarity calculated by nBLAST of syntenic orthologs between genomes. Alignment Length represents the length aligned by nBLAST of syntenic orthologs between genomes. Genomic SNPs represents the number of nucleotide mismatches found by nBLAST of syntenic orthologs between genomes. Sequence gaps represent the length of gaps found by nBLAST of syntenic orthologs between genomes. Counts between IR64 reads mapped to the MH63 and Nipponbare (MSU annotation) genome were obtained from the DESeq2 output. A custom script was developed to calculate gene length and exon length. Exon length and the total number of exons were calculated using the longest transcript model. For exonic SNP density, 1.0×10^{-10} was added to each value for all genes to avoid dividing by zero for genes that had zero SNPs. Values used for the response and explanatory variables in the linear model were generated by taking the $\log_2$ of the ratio between the values for the MH63 and Nipponbare genomes [$\log_2$(MH63/Nipponbare)]. A linear model was generated with the lm function in R using the difference in counts of the two genomes as the response variable.

Altering mismatch rate of STAR alignment

To vary the allowed mismatch parameter in STAR (Supplemental Fig. S8), we changed the following parameters in the STAR alignment command: –outFilterMismatchNmax 100 –outFilterMultimapNmax 2 –outFilterMismatchNoverLmax 999 –outFilterMismatchNoverReadLmax 999 –outFilterMatchNmin 0 –outFilterMatchNminOverLread 0 –outFilterScoreMinOverLread 0. For each mismatch setting, 20 million IR64 reads were mapped to either the MH63 or Nipponbare genome. We used –outFilterScoreMax as the mismatch allowed cutoff and this was varied from 0–10.

Custom scripts

Please refer to our github page (www.github/DohertyLab) for access to all custom scripts developed by the Doherty laboratory used in this study.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We would like to thank Scientific Studios (www.scientificstudios.com) for assistance with the design and layout of figures and preparation of Figure 9. We would also like to thank Olivia Wilkins for editing the manuscript and providing thoughtful comments. This project was supported by the Agriculture and Food Research Initiative competitive grant number 2015-67013-22814 of the USDA National Institute of Food and Agriculture and the USDA National Institute of Food and Agriculture project 1002035.

Received January 5, 2019; accepted March 6, 2019.

REFERENCES

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature 526: 68–74.

The 1001 Genomes Consortium. 2016. 1,135 genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166: 481–491.

Auler PA, Benitez LC, Nogueira M, Vighi IL, Rodrigues S, Carlos L, Jacira E, Braga B. 2017. Evaluation of stability and validation of reference genes for RT-qPCR expression studies in rice plants under water deficit. J Appl Genet 58: 163–177. doi:10.1007/s13353-016-0374-1

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM,

Holko M, et al. 2013. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41: D991–D995. doi:10.1093/nar/gks1193

Brown JWS, Simpson CG, Marquez Y, Gadd GM, Barta A, Kalyna M. 2015. Lost in translation: pitfalls in deciphering plant alternative splicing transcripts. Plant Cell 27: 2083–2087. doi:10.1105/tpc.15.00572

Dalquen DA, Dessimoz C. 2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. Genome Biol Evol 5: 1800–1806. doi:10.1093/gbe/evt132

Degner JF, Marioni JC, Pai AA, Pickrell JK, Nkadori E, Gilad Y, Pritchard JK. 2009. Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics 25: 3207–3212. doi:10.1093/bioinformatics/btp579

Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot NS, Castel D, Estelle J, et al. 2013. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14: 671–683. doi:10.1093/bib/bbs046

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21. doi:10.1093/bioinformatics/bts635

Edgar R, Domrachev M, Lash AE. 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 30: 207–210. doi:10.1093/nar/30.1.207

Filichkin SA, Mockler TC. 2012. Unproductive alternative splicing and nonsense mRNAs: a widespread phenomenon among plant circadian clock genes. Biol Direct 7: 20. doi:10.1186/1745-6150-7-20

Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W, Mockler TC. 2010. Genome-wide mapping of alternative splicing in Arabidopsis thaliana. Genome Res 20: 45–58. doi:10.1101/gr.093302.109

Fu XZ, Gong XQ, Zhang YX, Wang Y, Liu JH. 2012. Different transcriptional response to Xanthomonas citri subsp. citri between kumquat and sweet orange with contrasting canker tolerance. PLoS One 7: e41790. doi:10.1371/journal.pone.0041790

Fulton DL, Li YY, Laird MR, Horsman BGS, Roche FM, Brinkman FSL. 2006. Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics 16: 1–16.

Hartley SW, Mullikin JC. 2015. QoRTs: a comprehensive toolset for quality control and data processing of RNA-seq experiments. BMC Bioinformatics 16: 1–7.

Hartley SW, Mullikin JC. 2016. Detection and visualization of differential exon and splice junction usage in RNA-seq data with JunctionSeq. Nucleic Acids Res 44: e127.

Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, Stadler PF, Hackermüller J. 2009. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol 5: 1–10. doi:10.1371/journal.pcbi.1000502

Hoffmann S, Otto C, Doose G, Tanzer A, Langenberger D, Christ S, Kunz M, Holdt LM, Teupser D, Hackermüller J, et al. 2014. A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. Genome Biol 15: R34. doi:10.1186/gb-2014-15-2-r34

James AB, Syed NH, Bordage S, Marshall J, Nimmo GA, Jenkins GI, Herzyk P, Brown JWS, Nimmo HG. 2012. Alternative splicing mediates responses of the Arabidopsis circadian clock to temperature changes. Plant Cell 24: 961–981. doi:10.1105/tpc.111.093948

Jończyk M, Sobkowiak A, Siedlecki P, Biecek P, Trzcinska-Danielewicz J, Tiuryn J, Fronk J, Sowiński P. 2011. Rhythmic diel pattern of gene expression in juvenile maize leaf. PLoS One 6: e23628. doi:10.1371/journal.pone.0023628

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, Mccombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S, et al. 2013. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice 6: 4. doi:10.1186/1939-8433-6-4

Khang TF, Lau CY. 2015. Getting the most out of RNA-seq data analysis. PeerJ 3: e1360. doi:10.7717/peerj.1360

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12: 357–360. doi:10.1038/nmeth.3317

Lechner M, Hernandez-rosales M, Doerr D, Wieseke N, Stoye J, Hartmann RK, Prohaska SJ. 2014. Orthology detection combining clustering and synteny for very large datasets. PLoS One 9: e105015. doi:10.1371/journal.pone.0105015

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15: 550. doi:10.1186/s13059-014-0550-8

Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM, et al. 2008. Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules. PLoS Genet 4: e14. doi:10.1371/journal.pgen.0040014

Raghupathy N, Choi K, Vincent MJ, Beane GL, Munger SC, Korstanje R, Pardo-Manual de Villena F. 2018. Hierarchical analysis of multi-mapping RNA-seq reads improves the accuracy of allele-specific expression. Bioinformatics 34: 2177–2184. doi:10.1093/bioinformatics/bty078

Reddy ASN, Marquez Y, Kalyna M, Barta A. 2013. Complexity of the alternative splicing landscape in plants. Plant Cell 25: 3657–3683. doi:10.1105/tpc.113.117523

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43: e47. doi:10.1093/nar/gkv007

Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26: 139–140. doi:10.1093/bioinformatics/btp616

Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-hughes TOM, et al. 2016. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? RNA 22: 839–851. doi:10.1261/rna.053959.115

Seyednasrollah F, Laiho A, Elo LL. 2013. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinform 16: 59–70. doi:10.1093/bib/bbt086

Sherman RM, Forman J, Antonescu V, Puiu D, Daya M, Rafaels N, Boorgula MP, Chavan S, Vergara C, Ortega VE, et al. 2018. Assembly of a pan-genome from deep sequencing of 910 humans of African descent. Nat Genet 51: 30–35. doi:10.1038/s41588-018-0273-y

Stevenson KR, Coolon JD, Wittkopp PJ. 2013. Sources of bias in measures of allele-specific expression derived from RNA-seq data aligned to a single reference genome. BMC Genomics 14: 536. doi:10.1186/1471-2164-14-536

Tarazona S,García-Alcalde F, Dopazo J, Ferrer A, Conesa A. 2011. Differential expression in RNA-seq: a matter of depth. Genome Res 21: 2213–2223. doi:10.1101/gr.124321.111

Tarazona S,Furió-Tarí P, Turrà D, Di Pietro A, Nueda MJ, Ferrer A, Conesa A. 2015. Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43: e140. doi:10.1093/nar/gkv711

Wang X, Wu F, Xie Q, Wang H, Wang Y, Yue Y, Gahura O, Liu L, Cao Y, Jiao Y, et al. 2012a. SKIP is a component of the spliceosome linking alternative splicing and the circadian clock in

Arabidopsis. Plant Cell 24: 3278–3295. doi:10.1105/tpc.112.100081

Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H, et al. 2012b. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res 40: 1–14. doi:10.1093/nar/gkr648

Xu Q, Chen W, Xu Z. 2015. Relationship between grain yield and quality in rice germplasms grown across different growing areas. Breed Sci 232: 226–232. doi:10.1270/jsbbs.65.226

Zhang J, Chen L-L, Xing F, Kudrna DA, Yao W, Copetti D, Mu T, Li W, Song J-M, Xie W, et al. 2016. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci 113: E5163–E5171. doi:10.1073/pnas.1611012113

Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, Xing Y, Chen L, Xie W, Variation R, et al. 2015. RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res 43: 1018–1022. doi:10.1093/nar/gku894

Analysis of differential gene expression and alternative splicing is significantly influenced by choice of reference genome

Erin Slabaugh, Jigar S. Desai, Ryan C. Sartor, et al.

RNA 2019 25: 669-684 originally published online March 14, 2019
Access the most recent version at doi:10.1261/rna.070227.118