the plant journal

SEB Society for Experimental Biology

The Plant Journal (2015) **84,** 491–503 doi: 10.1111/tpj.13014

Genomic limitations to RNA sequencing expression profiling

Cory D. Hirsch[1], Nathan M. Springer[1] and Candice N. Hirsch[2,*]

[1]Department of Plant Biology, University of Minnesota, St Paul, MN 55108, USA, and

[2]Department of Agronomy and Plant Genetics, University of Minnesota, St Paul, MN 55108, USA

Received 28 April 2015; accepted 25 August 2015; published online 02 September 2015.

*For correspondence (e-mail cnhirsch@umn.edu).

SUMMARY

The field of genomics has grown rapidly with the advent of massively parallel sequencing technologies, allowing for novel biological insights with regards to genomic, transcriptomic, and epigenomic variation. One widely utilized application of high-throughput sequencing is transcriptional profiling using RNA sequencing (RNAseq). Understanding the limitations of a technology is critical for accurate biological interpretations, and clear interpretation of RNAseq data can be difficult in species with complex genomes. To understand the limitations of accurate profiling of expression levels we simulated RNAseq reads from annotated gene models in several plant species including Arabidopsis, brachypodium, maize, potato, rice, soybean, and tomato. The simulated reads were aligned using various parameters such as unique versus multiple read alignments. This allowed the identification of genes recalcitrant to RNAseq analyses by having over- and/or under-estimated expression levels. In maize, over 25% of genes deviated by more than 20% from the expected count values, suggesting the need for cautious interpretation of RNAseq data for certain genes. The reasons identified for deviation from expected expression varied between species due to differences in genome structure including, but not limited to, genes encoding short transcripts, overlapping gene models, and gene family size. Utilizing existing empirical datasets we demonstrate the potential for biological misinterpretation resulting from inclusion of 'flagged genes' in analyses. While RNAseq is a powerful tool for understanding biology, there are limitations to this technology that need to be understood in order to improve our biological interpretations.

Keywords: RNAseq, expression profile, structural annotation, Arabidopsis, maize.

INTRODUCTION

Profiling of transcript abundance is fundamental for functional genomics. Ideally, a researcher could determine the steady-state abundance of transcripts for each gene in a given sample. There have been rapid changes in the technologies employed to study transcript levels (Wang *et al*., 2009). Initial experiments relied upon Northern blots, assessing the abundance of a single gene at a time. In addition to the low-throughput nature of Northern blots, the ability to monitor a single gene is limited by the ability to identify a probe and conducive hybridization conditions for the desired gene, but not to other related genes. Quantitative real-time PCR (qRT-PCR) provided a faster method for analyzing transcript levels for a small number of genes. Again, the ability to resolve a single gene is limited by the ability to design gene-specific primers for amplification. Genomic-scale profiling of gene expression was first performed using microarray technology. Microarrays were produced containing bound DNA fragments of various lengths. Fluorescently labeled RNA samples for tissues of interest were hybridized to the arrays and the fluorescence of each feature was measured. By comparing the relative fluorescence of different dyes or different arrays, researchers were able to compare the expression level for a particular gene in different samples. Resolution to a single gene was again limited by the ability to design probes that would hybridize to a specific gene and no others. In all of these technologies, there was substantial variation in the hybridization strength or primer efficiency, such that it was quite difficult to compare the relative expression level for different genes. These techniques also required prior knowledge of the genes or genomes being examined.

The rapid development of low-cost high-throughput sequencing has enabled a shift towards a dependence upon RNA sequencing (RNAseq) for genome-wide expression profiles. RNA is collected from a sample and used to generate sequencing libraries. Typically, millions of molecules from these libraries are sequenced and aligned to a reference genome, then counting algorithms are used to determine the number of molecules derived from each feature to allow for quantitative whole-transcriptome profiling. A non-comprehensive evaluation of papers employing RNAseq for gene expression profiling in plant species revealed that a variety of different approaches are used for read alignment, read counting, and differential expression analysis (Severin *et al.*, 2010; Massa *et al.*, 2011; Davidson *et al.*, 2012; Paschold *et al.*, 2012; Gao *et al.*, 2013; Huan *et al.*, 2013; Koenig *et al.*, 2013; Li *et al.*, 2013, 2014; Song *et al.*, 2013; Yazawa *et al.*, 2013; Zhai *et al.*, 2013; Chen *et al.*, 2014; Gelli *et al.*, 2014; Gordon *et al.*, 2014; Hirsch *et al.*, 2014; Leisner *et al.*, 2014; Martin *et al.*, 2014; Nguyen *et al.*, 2014; Wakasa *et al.*, 2014). The vast majority of approaches used a splice-site-aware aligner, most often TopHat (Trapnell *et al.*, 2009; Kim *et al.*, 2013), but no standard software package dominated read counting and differential expression analyses. Several papers have compared the efficiency and accuracy of various different approaches for mapping, counting, and identifying differentially expressed (DE) genes (Lindner and Friedel, 2012; Nookaew *et al.*, 2012; Chandramohan *et al.*, 2013; Engstrom *et al.*, 2013; Guo *et al.*, 2013; Rapaport *et al.*, 2013; Zhang *et al.*, 2014; Seyednasrollah *et al.*, 2015). However, the issue of identifying genes recalcitrant to accurate counting using RNAseq has not yet been addressed in plant species.

Several issues can cause difficulties in accurately estimating gene expression using RNAseq. First, small transcripts can be more difficult to count due to the standard size selection implemented during construction of RNAseq libraries. Second, in some cases two different genes have overlapping transcripts. In this case it is difficult to determine to which gene the read should be assigned. There are related issues in precisely estimating the abundance of different transcripts from the same gene (Trapnell *et al.*, 2010). In this paper we have chosen to ignore alternative splicing in order to simply focus on estimating abundance per gene, rather than per isoform. Third, the presence of multiple related sequences within the genome can cause significant issues in accurately assigning expression to the correct gene. Multiple related sequences can arise from whole-genome duplication (WGD) events, local duplication events, or transposition events. The frequency of these events is quite variable among different plant species (Arabidopsis Genome, 2000; International Rice Genome Sequencing, 2005; Flagel and Wendel, 2009; Freeling, 2009; Schnable *et al.*, 2009; International Brachypodium, 2010; Schmutz *et al.*, 2010; Potato Genome Sequencing *et al.*, 2011; Tomato Genome, 2012). Due to the presences of these duplicate sequences, researchers are faced with the difficult choice of whether to allow a sequence read to be counted if it can be aligned to multiple places in the genome or whether to count a read only if it has a unique best alignment in the genome, which can have a dramatic impact on results.

In order to understand how these issues affect the implementation of RNAseq and the potential errors in estimating gene expression we simulated RNAseq reads from the transcriptomes of seven plant species with high-quality reference genomes including Arabidopsis (*Arabidopsis thaliana*), brachypodium (*Brachypodium disctachyon*), maize (*Zea mays*), potato (*Solanum tuberosum*), rice (*Oryza sativa*), soybean (*Glycine max*), and tomato (*Solanum lycopersicum*). These species were selected to represent diverse families (Brassicaceae, Fabaceae, Poaceae, and Solanaceae) as well as diversity in genome size, levels of WGD retention, and genome complexity. We surveyed multiple alignment and counting methodologies using the simulated reads for each of the seven species, identified transcripts that were under- or over-estimated with different methods, determined common characteristics of these genes, and demonstrated how deviations from expected values can potentially result in biological misinterpretations using empirical sequence reads.

RESULTS

Read simulation

Actual RNAseq data will exhibit variation in the distribution of reads within and among genes. There will also be variation in sequencing quality in true RNAseq reads. In order to evaluate different methods and identify genes that are not properly represented, we generated simulated RNAseq reads under a best-case scenario. A single representative transcript was identified for each gene in Arabidopsis, brachypodium, maize, potato, rice, soybean, and tomato (Figure 1a,b, Table S1 in Supporting Information). All 50 possible nucleotide reads were generated for the representative transcripts, simulating single-end reads with no mismatches, resulting in 50 × read depth across the body of the transcript with lower coverage at both ends of the transcript (Figure 1c,d). This bias in coverage over the transcript is due to the fact that there is only one unique read that included the first and last base of the transcript, but up to 50 unique reads that included bases in the middle of the transcript. The sampling bias mirrors what is observed when RNA fragmentation is used during library preparation (Wang *et al.*, 2009). The total number of simulated reads varied for each species depending on the size of the annotated transcriptome (Table 1 and S2). Importantly, for the simulated RNAseq reads the exact number of reads generated for each transcript was known and could be used to compare the number of observed aligned and

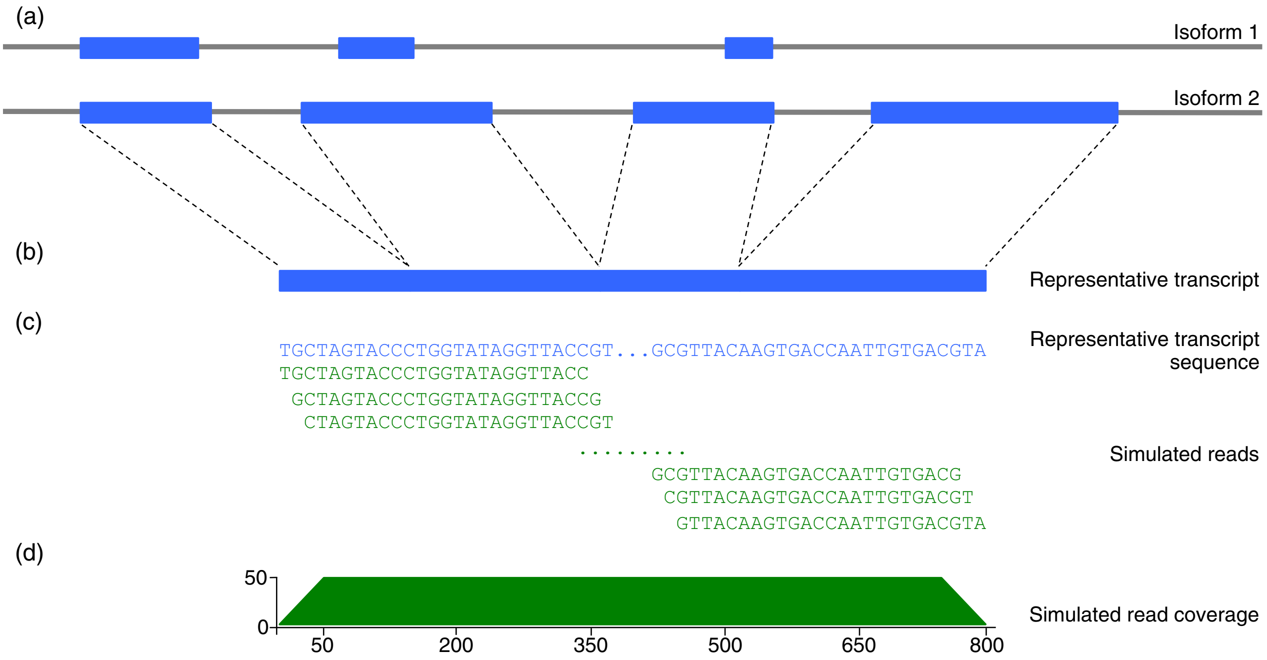

(a)
Isoform 1
Isoform 2
(b)
Representative transcript
(c)
Representative transcript sequence
TGCTAGTACCCTGGTATAGGTTACCGT...GCGTTACAAGTGACCAATTGTGACGTA
TGCTAGTACCCTGGTATAGGTTACC
GCTAGTACCCTGGTATAGGTTACCG
CTAGTACCCTGGTATAGGTTACCGT
....
GCGTTACAAGTGACCAATTGTGACG
CGTTACAAGTGACCAATTGTGACGT
GTTACAAGTGACCAATTGTGACGTA
Simulated reads
(d)
50
0
50
200
350
500
650
800
Simulated read coverage

Figure 1. Pictorial diagram of RNAseq read simulation for an example gene model.
(a) Two possible isoforms for the gene model.
(b) Identification of the representative transcript.
(c) Generation of 50-nucleotide simulated reads shifting one base for each read.
(d) Read depth distribution across the representative transcript.

Table 1 Summary of genome and transcriptome features of the representative sequences used for the simulated species

Species	Genome size (MB)	Gene space (MB)	Transcriptome size (MB)	Number of genes	Gene density (genome size in kb/no. of genes)	Number of gene families[a]	Number of singleton genes[b]	Average number of genes per family	Number of genes with an overlap[c]
Arabidopsis	119.7	60.0	40.2	27 416	4.36	3531	3433	6.8	308
Brachypodium	272.0	103.5	37.5	31 694	8.58	3842	4113	7.2	329
Maize	2066.4	156.8	61.5	39 656	52.11	5084	3592	7.1	2
Potato	647.2	106.0	46.4	35 119	18.43	3508	3354	9.1	556
Rice	374.5	110.2	55.5	39 049	9.59	4578	7194	7.0	170
Soybean	978.5	212.9	64.6	56 044	17.46	6728	2229	8.0	571
Tomato	823.9	109.9	42.0	34 725	23.73	3686	5152	8.0	678

[a] Number of OrthoMCL groups containing two or more representative transcripts in within-species analysis.
[b] Number of representative transcripts not contained within an OrthoMCL group in within-species analysis.
[c] Number of genes with at least one nucleotide of exon sequence in the representative transcript that overlaps at least one nucleotide of exon sequence in the representative transcript of another gene from the same strand.

counted reads with the true number of reads generated from each transcript.

Comparison of alignment and counting methods for a large and small plant genome

Our goal in this analysis was not to provide a robust comparison of different methods for alignment and counting of reads, but instead to focus on determining the set of genes that are difficult to properly analyze using RNAseq for expression analysis. A large number of RNAseq analyses in plant species have used the tuxedo suite of programs to perform splice-site-aware alignments and estimate transcript abundances (Trapnell *et al.*, 2012). When performing alignments using Bowtie2 and TopHat2 it is possible to align to the full genome, to the full genome with genome annotation guidance, or to the transcript sequences only. All three of these approaches were evaluated for both Arabidopsis and maize, which represent plant species with small and large genomes, respectively (Table 1). In both species, mapping to the full genome without an annotation file resulted in the greatest proportion of un-mapped reads. However, in all three cases more than 99% of the simulated reads were aligned at least once (Table S2). Reads that were not aligned to the genome primarily spanned introns with minimal overhang on one side of the intron. While minor gains were achieved by mapping with annotation guidance or to only the transcriptome, these approaches make assumptions that the structural annotation of the gene is high quality and can inflate the number of uniquely mapped reads. Some of the reads that aligned uniquely when using annotation files as guidance or transcript sequences are not actually unique in the genome assembly and thus could result in false inferences. Additionally, extensive differences in annotation between individuals within a species have been documented (Gan *et al.*, 2011), and this variation will not be captured in a single reference annotation. Finally, aligning reads to the full genome allows the study of un-annotated transcripts if desired. Bowtie2 and TopHat2 are widely used by plant scientists, and there are multiple benefits of mapping to the genome without guidance. Thus, for the remainder of the analyses, RNAseq reads were aligned to the genome without guidance using Bowtie2 and TopHat2.

Following the alignment of sequences to the genome it is necessary to count the number of reads derived from each annotated feature. The Tuxedo suite of tools (Trapnell *et al.*, 2012) implements Cufflinks for this step. Cufflinks provides a statistical approach to estimate the fragments per kilobase of exon model per million fragments mapped (FPKM) abundance for each transcript, rather than providing specific counts of aligned reads per feature. Other tools, such as HTSeq (Anders *et al.*, 2015), process alignments to count the number of reads derived from annotated genomic features, which can be converted to FPKM values. A comparison of the counts or FPKM values derived from HTSeq with those derived from Cufflinks revealed a complex relationship. This is mainly seen where HTSeq count values are low and FPKM values from Cufflinks are generally extremely high, where a linear relationship is expected (Figure S1, Tables S3 and S4). Shorter genes were highly over-estimated by Cufflinks (Figure S1a, b). When HTSeq counts were converted to FPKM values, a relatively linear relationship was observed for longer genes (>600 bp, $R = 0.853$; Figure S1c). One likely contributing factor is that Cufflinks assumes that shorter genes are more difficult to recover due to the size selection step during the preparation of the RNAseq library, and inflates the expression for these genes based on their length. The complexities that corrections of this nature introduce were not conducive to comparisons of observed and expected counts to identify genes recalcitrant to RNAseq-based estimates of transcript abundance. As such, we utilized HTSeq for the remaining analyses.

Identification of and causes for recalcitrance of transcripts to RNAseq analyses

There are many different strategies to estimate and present transcript abundance, such as read counts normalized by transcript length and number of reads mapped (FPKM). Here, we focus specifically on the actual number of reads aligning to each transcript, rather than a derived measure of expression such as an FPKM value, as we have *a priori* knowledge of the exact number of reads generated for each transcript (expected counts). Genes recalcitrant to RNAseq analyses were identified in each species by comparing observed and expected counts. In each of the seven species the observed counts were determined using conditions that only counted uniquely aligned reads. Many plant genomes have undergone tandem, segmental, or whole-genome duplication events resulting in gene families with members that are identical or nearly identical in the genome (Flagel and Wendel, 2009; Freeling, 2009). These duplication events can potentially result in genes that cannot be assayed at all, or only partially assayed, using only uniquely mapping reads. In maize, at least 1% of genes have nearly identical paralogs (defined as having ≥98% identity between the paralogs) (Emrich *et al.*, 2007). To determine the impact on estimates of transcript abundance of allowing multiple mapping, particularly for genes in gene families, we developed a method to work around the base parameters of HTSeq and allow reads that align to multiple genomic locations to be counted as well. The proportion of reads aligned to multiple genomic locations varied from 3% to 15% depending upon the species (Table S2). Per gene deviation from expected counts is provided for each alignment condition for all species in Tables S3–S9. In all species, a large number of genes were estimated very well (Figure 2a). Examples of such genes are shown in Figure S2(a,b). However, clear outliers with observed counts that deviated from the expected count by more than 20% were seen in all species, and were classified as recalcitrant to RNAseq expression profiling. Furthermore, the proportion of recalcitrant genes varied quite substantially between the different species (Figures 2a and S3). In total, between 4.8% (Arabidopsis) and 25.7% (maize) of genes were recalcitrant to current short-read RNAseq-based approaches using a unique mapping pipeline. Transcript abundance is a potential factor influencing the estimation of expression, and therefore the detection of genes recalcitrant to current short-read RNAseq-based approaches. To test the impact of abundance variation we simulated reads with 50 ×, 100 ×, and 200 × coverage of each of the maize transcripts. Of the 39 656 maize genes, 38 549 had exactly the same observed/expected value to three decimal places. Additionally, the

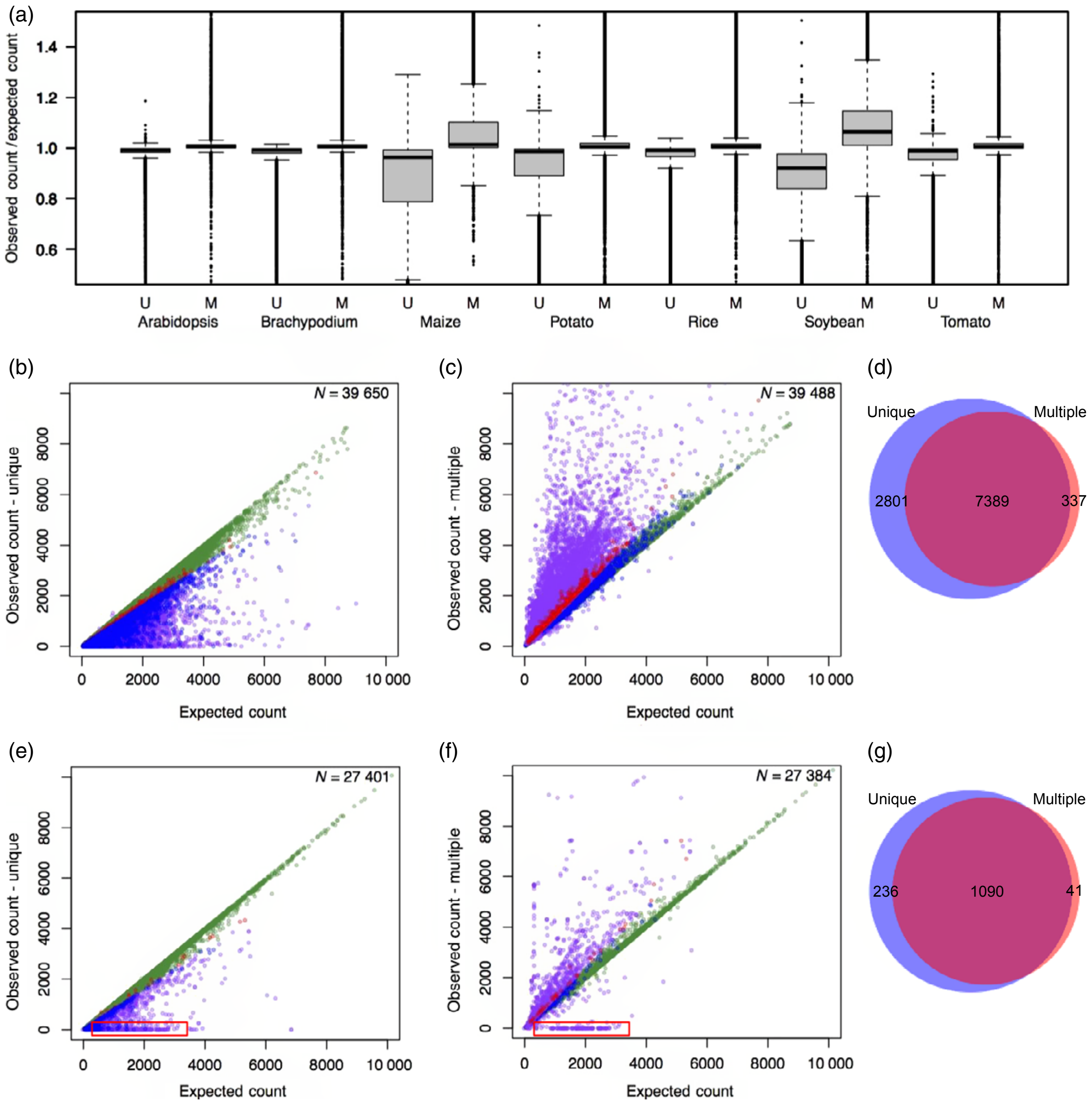


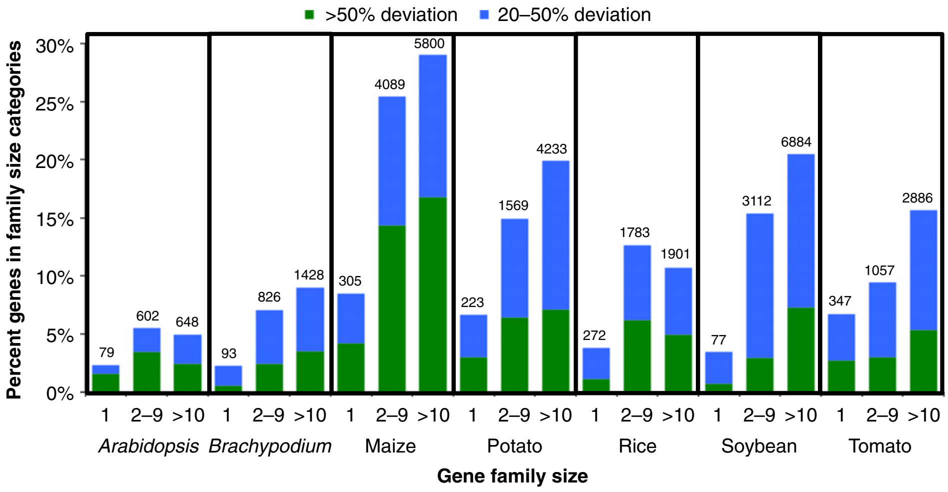
Figure 2. Distribution of observed versus expected counts based on alignment methodology across seven species.
Reads were aligned requiring either a unique best match (Unique/U) or allowing multiple valid alignments (Multiple/M). (a) Distribution of observed/expected values with unique and multiple alignments across seven species. The *y*-axis was truncated and excluded extreme outliers. (b–d) Maize and (e–g) Arabidopsis. (b, e) Observed versus expected count plots for only unique aligned reads. (c, f) Observed versus expected count for multiple aligned reads. (b, c) and (e, f) Color indicates genes that deviated by >20% from expected in only unique alignments (blue), only multiple alignments (red), in both alignment methods (purple), or neither alignment method (green). The plots are scaled to show the majority of the distribution and exclude extreme outliers (N indicates the number of data points per graph). The red boxes indicate genes with zero counts whether counting unique or multiple alignments. (d, g) Venn diagram of genes that deviated by >20% from the expected count for unique and multiple aligned reads.

classification of only one gene changed from recalcitrant to non-recalcitrant at the 100 and 200× depth relative to the original 50× depth.

Many of the same genes are classified as recalcitrant under both unique and multiple alignment conditions; with transcripts that were under-counted with unique alignments becoming over-counted when multiple alignments were allowed. Indeed, this was observed in maize (Figure 2b,c) as well as the other species. Analysis of Arabidopsis, however, revealed an interesting set of genes with zero observed read counts in both unique and multiple alignment conditions (red box in Figure 2e,f). Closer examination revealed that each of these genes had complete overlap with another transcript at the exon level (example of overlapping genes are shown in Figure S2c). When genes overlap, HTSeq will not count reads aligned to the overlapping region because it cannot confidently determine which gene the read should be ascribed to and denotes the read as ambiguously aligned (Figure S4a). The number of genes with overlapping annotations varied strongly among the different species, ranging from just two genes (maize) to more than 600 genes (tomato) (Figure S4b). However, it is not clear if this variation is driven by biological differences or by differences in the annotation approaches that were used for each species. A comparison of the observed/expected read counts and percentage overlap found that the amount of overlap is a good predictor for the deviation in observed read counts for many genes (Figure S4c). However, a subset of 161 genes out of the 2614 genes across the seven species that had more than 70% overlap with another gene at the exon level, had an observed count/expected count of <0.2 with unique alignments (red box in Figure S4c). These genes are likely to be repeated elsewhere in the genome and have fewer observed reads because only unique alignment counts are analyzed. Thus, the percentage overlap of a gene is of concern for under-estimated gene counts but it cannot solely explain all deviations.

The presence of multiple related sequences in a genome can cause significant issues with accurate estimation of transcript abundances. In our simulated dataset, a large set of genes were under-represented when focusing on uniquely aligned reads and over-estimated when allowing reads to align to multiple locations (Figure 2d,g). Many of these genes have nearly identical transcript sequences elsewhere in the genome (Figure S2d). There are several situations that could result in multiple genes having nearly identical sequences, such as WGD events. These events can culminate in many genes having a second nearly identical copy. The two copies are expected to diverge over time, such that more recent WGD events are more likely to cause problems in mapping relative to older WGD events. Alternatively, gene family expansion through tandem duplication or transposition of genes to unlinked positions could result in multiple highly similar sequences for specific genes. In general, there are considerably more genes that deviate substantially from the expected value in species with more recent WGD events such as maize and soybean (Lee *et al.*, 2013) (Figure 2a). We assessed the relationship between gene-family size and the percentage of genes with highly deviated observed counts (Figure 3). In general, the larger the size of the gene family, the more likely it was for genes to be under-represented by requiring uniquely aligned reads. However, it is worth noting that even in large gene families only 10–30% of the genes were severely underrepresented.

In order to assess whether similar genes are flagged as recalcitrant in multiple species we focused on the three grass species included in this study, maize, rice and brachypodium. Genes that are present in maize sub-genome 1 (Schnable *et al.*, 2011) and have syntenic orthologs (Schnable *et al.*, 2012) in each of the other species were identified. The subset of genes within this set that were flagged in at least one species were identified and compared (Figure S5a). In general, most genes were only flagged in a single species. Only a small number of genes were flagged in two or three of these species. A Gene Ontology (GO) analysis of the flagged genes was also performed to assess whether similar functions were enriched for recalcitrant genes (Figure S5b). The majority (84%) of



	Total number of genes		
	Single gene family	2–9 gene family	>10 gene family
Arabidopsis	3433	10 899	13 084
Brachypodium	4113	11 709	15 872
Maize	3592	16 092	19 972
Potato	3354	10 510	21 255
Rice	7194	14 806	17 769
Soybean	2229	20 221	33 594
Tomato	5152	11 171	18 402

Figure 3. Percentage deviation from expected values by gene family size for seven species.
Reads were aligned requiring a unique alignment. Percentage deviation is the deviation from the expected count of one. The inset table denotes the total number of genes in each family size for each species.

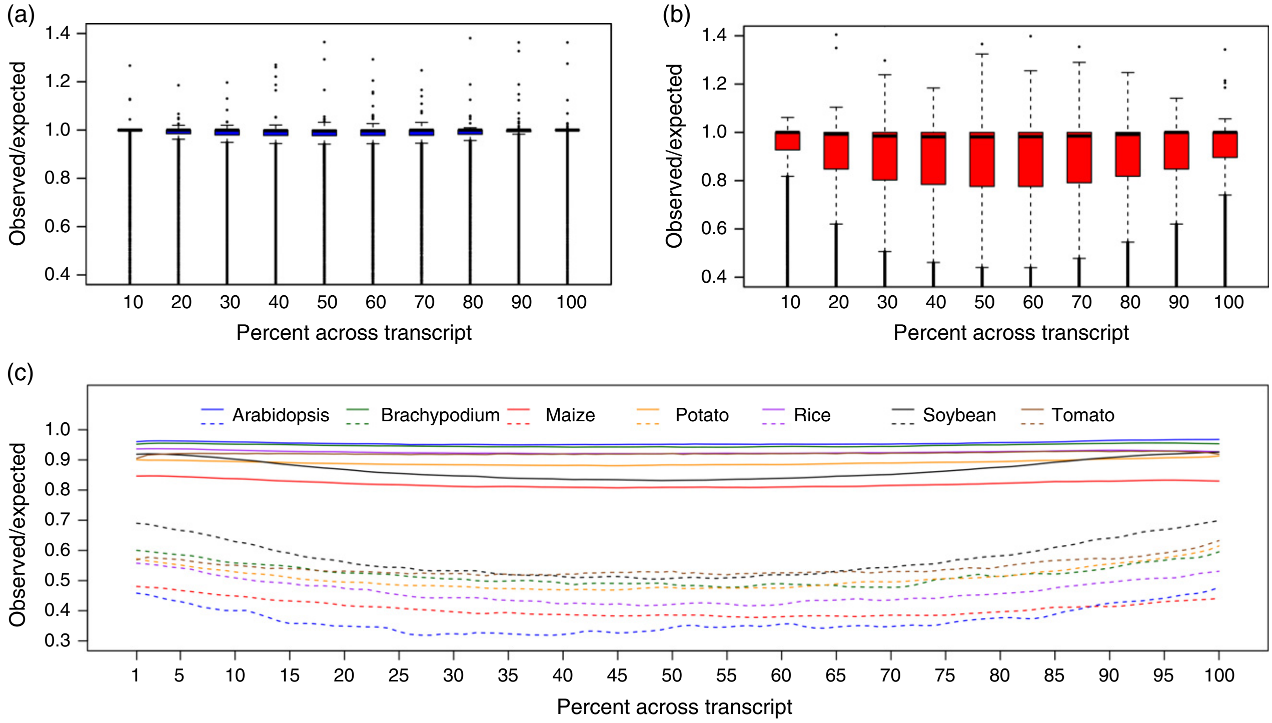


Figure 4. Deviation in observed/expected read counts by position for unique alignments.
(a) The distribution of all transcripts observed/expected values for every 10% across each transcript for Arabidopsis.
(b) The distribution for all transcripts of observed/expected values for every 10% across each transcript for maize.
(c) The average observed/expected value for all transcripts (solid lines) and transcripts deviating by >20% across the whole transcript (dashed lines) for every 1% across the transcript for each species.

biological process GO terms that were significantly enriched were only identified in a single species. The few GO terms that were found in more than one species were found in multiple grass species. The analysis of both syntenic genes and GO terms suggest limited conservation of genes or functions that are recalcitrant to RNAseq expression profiling across species.

Genes recalcitrant to RNAseq expression profiling were identified on a whole transcript basis. However, it is possible that some regions of the transcript have a greater effect on the observed deviations than others. To determine if there were regions of transcripts that showed more extreme deviations we evaluated observed versus expected counts by position across transcripts. In the largest genome of the simulated species (maize) a larger positional effect of observed/expected values was seen in the middle of transcripts than in the species with the smallest genome, Arabidopsis (Figure 4a,b). However, in all species, the 5′ and 3′ ends of transcripts were on average closer to expected values than the middle of the transcripts. This trend was even more apparent in flagged genes (Figure 4c). However, it is noteworthy that while the 5′ and 3′ ends of transcripts were closer to expected count values, on average the ends of flagged transcripts were still well beyond the acceptable threshold of <20% deviation.

Potential biological misinterpretation in empirical RNAseq datasets

We utilized two existing maize empirical RNAseq datasets to evaluate the potential for biological misinterpretation resulting from the inclusion of 'flagged genes' in analyses and subsequent biological conclusions. The first dataset explored was the maize B73 expression atlas (Stelpflug *et al.*, 2015). This dataset consisted of RNAseq transcriptome profiling of 80 tissues throughout the development of maize with either two or three biological replicates per tissue. Using the previously calculated Cufflinks-derived FPKM values averaged across replicates (Stelpflug *et al.*, 2015), a significant difference in average expression values was observed between small, moderate, and larger transcripts (Figure S6). Overall, transcripts of <600 bp displayed lower average expression across the 80 tissues than transcripts of more than 600 bp. While it is possible this is biologically relevant (i.e. fragmented pseudogenes), or the result of inaccurate gene prediction, it is most likely a product of the size selection implemented during construction of the RNAseq library. This further highlights the need for caution in biological interpretation of transcriptional variation for genes encoding short transcripts.

The ability to accurately identify differentially expressed genes is predicated on the ability to accurately determine transcript abundance. To evaluate the impact of flagged genes in differential expression analysis, a second dataset consisting of maize B73 seedling tissue from plants grown under control and cold treatments (Makarevitch *et al.*, 2015) was investigated. Using this dataset, we evaluated the extent to which genes flagged by various attributes in the simulated data are called differentially expressed (DE) in empirical data. Transcript abundances and differential expression analysis was conducted using unique alignments and allowing multiple alignments (Table S10). Using only the unique alignments, 5113 DE genes were identified, whereas 5707 DE genes were identified using multiple alignments. A large overlap between the two pipelines was observed, with 4874 genes detected as DE in both pipelines. However, a substantial number of genes (833) were only detected in the multiple alignment pipeline. Interestingly, genome wide, in the simulated maize data 10 190 genes deviated >20% from the expected counts when analyzing unique alignments, while only 7726 were found when multiple alignments were allowed. In the empirical data, however, there were only 473 flagged DE genes with unique alignments compared with 781 with multiple alignments (Figure 5a). Thus, while fewer genes in total were flagged in the simulated multiple alignment pipeline for maize, in the empirical data, nearly twice as many flagged genes were identified as DE in the multiple alignment pipeline. In addition, the level of deviation from expected was larger in flagged DE genes when multiple alignments were counted compared with unique alignment counts (Figure 5b).

The greater number of DE genes detected only in the multiple alignment pipeline were probably primarily false positives, as differential expression of one member of a family will be dispersed across other members of the gene family. An example of this is shown for a two-member maize gene family in Figure 6. In this example, one member of the family (GRMZM2G055178) shares complete sequence homology with a portion of the other member (GRMZM2G031721) that contains some unique sequences as well. Both of these genes were flagged in the simulated RNAseq dataset in the unique and multiple alignment pipelines (Table S4). In the empirical dataset comparing control and cold-treated seedlings, allowing unique alignments resulted in one member of the family being called DE (Figure 6a). However, when multiple alignments were used, both family members were designated as DE. This highlights two important points. First, if differential expression were determined based on unique alignments only, GRMZM2055178 would never be interrogated, regardless of the expression level. In addition, the expression of GRMZM2G031721 is based on only about 33% of the transcript, which would lead to lower FPKM values and reduced transcript information available for conducting differential expression analysis (Figure 6b,c). Second, when multiple alignments were used the expressions from the two genes were overlaid for each transcript. In this case it is highly likely that only one gene was transcribed (GRMZM2G031721), as it has reads aligning to its unique regions and the shared region of the gene did not show extra expression under multiple mapping. Thus, the determined expression of GRMZM2G055178 and the DE between the treatments with multiple mapping is probably a false positive result. This example highlights the potential for multiple biological misinterpretations with RNAseq expression analyses.

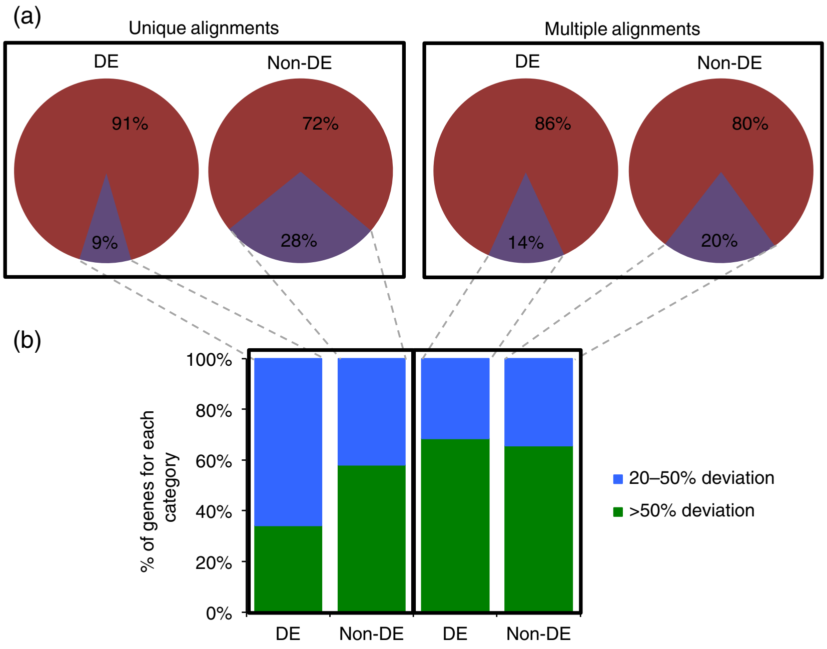


Figure 5. Differential expression analysis between three replicates of B73 control and cold-treated seedlings using unique and multiple alignments.
(a) Pie charts of the percentage of differentially expressed (DE) and non-DE genes flagged due to deviation of >20% in observed counts compared with expected counts in the maize simulated RNA-seq data.
(b) Breakdown of observed deviation in the simulated data for flagged DE genes. Reads were obtained from a previously published experiment (Makarevitch *et al.*, 2015).

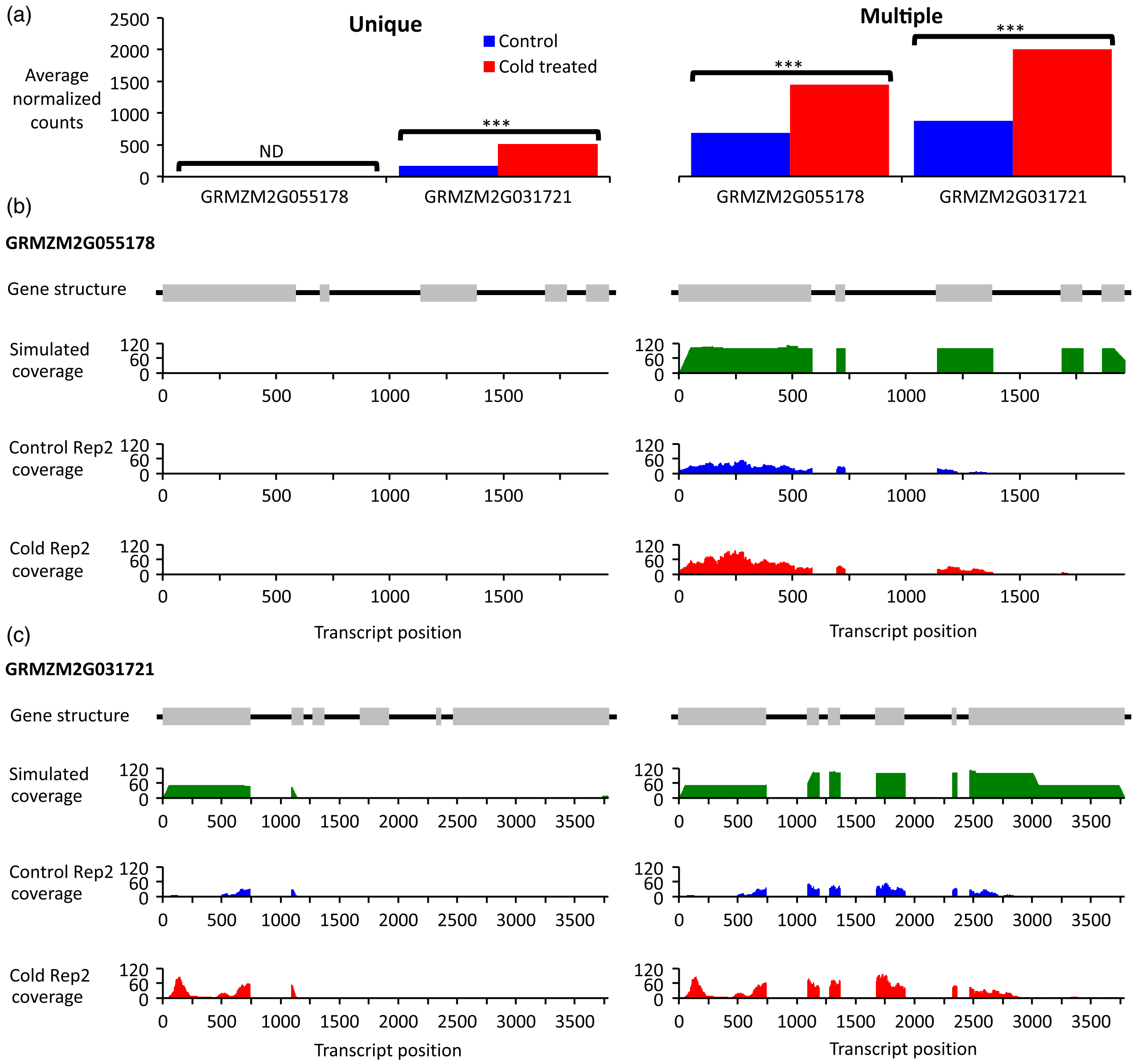


Figure 6. Transcript abundance counts and differential expression statistics for a two-member gene family (GRMZM2G055178 and GRMZM2G031721) for maize B73 control and cold-treated seedling data.

(a) Average normalized counts across the three replicates of control and cold-treated seedlings for both the unique and multiple alignment pipelines. The brackets denote tests for differential expression between treatments (ND, no difference; ***$P < 0.001$). (b) Gene structural annotation, maize simulated read coverage, read coverage for control treatment, and read coverage for cold treatment across the GRMZM2G055178 transcript for uniquely aligned reads (left column) and multiple aligned reads (right column).

(c) Gene structural annotation, maize simulated read coverage, read coverage for control treatment, and read coverage for cold treatment across the GRMZM2G031721 transcript for uniquely aligned reads (left column) and multiple aligned reads (right column).

DISCUSSION

Transcriptome profiling is a fundamental tool used to understand genetic phenomena such as heterosis (Hansey *et al.*, 2012; Paschold *et al.*, 2012; Zhai *et al.*, 2013) and to further our understanding across biological disciplines such as developmental biology (Kang *et al.*, 2013), physiology (Nguyen *et al.*, 2014; Sekhon *et al.*, 2014), stress tolerance (Gao *et al.*, 2013), and evolution (Davidson *et al.*, 2012). The high-throughput and cost-effective nature of generating RNAseq data has resulted in a dramatic increase in its use for transcriptome profiling. As such, an understanding of the limitations of this technology as far as biological interpretations are concerned is essential.

The main goals of this study were to identify genes that are recalcitrant to the currently available technologies and tools across a set of diverse plant species, and to identify common attributes of recalcitrant genes to aid in biological interpretation of RNAseq-based transcriptome profiling. To identify these genes, we conducted simulated RNAseq experiments across seven diverse plant species. Simulation approaches mimicking current library preparation methods and biases have been developed (Grant *et al.*, 2011; Griebel *et al.*, 2012; Frazee *et al.*, 2015), and could have been used for this project. However, the intention of this study was to simulate an ideal situation with no bias during library preparation methods, as the recalcitrant genes identified with this approach will more likely remain problematic as library preparation and sequencing methods continue to advance and improve.

To make this analysis as broadly useful to plant scientists as possible, efforts were made to use standard methods such as aligning to the genome without guidance from a reference annotation. For species with highly complex genomes that have high-quality, manually curated annotations, aligning to the transcriptome rather than the genome may be advantageous. However, large amounts of structural variation (Cao *et al.*, 2011; Chia *et al.*, 2012; Anderson *et al.*, 2014; Hirsch *et al.*, 2014; Schatz *et al.*, 2014) and alternative annotations between individuals (Gan *et al.*, 2011) have been documented across plant species. Additionally, annotations are continually evolving as we expand our knowledge of splice variants, differences between individuals within a species, and the importance of non-coding RNAs. While there may be advantages in using the transcriptome for reference-to-reference aligning, the standard method at this time is to align to the genome (Severin *et al.*, 2010; Massa *et al.*, 2011; Davidson *et al.*, 2012; Paschold *et al.*, 2012; Gao *et al.*, 2013; Huan *et al.*, 2013; Koenig *et al.*, 2013; Li *et al.*, 2013, 2014; Song *et al.*, 2013; Yazawa *et al.*, 2013; Zhai *et al.*, 2013; Chen *et al.*, 2014; Nguyen *et al.*, 2014; Gelli *et al.*, 2014; Gordon *et al.*, 2014; Leisner *et al.*, 2014; Hirsch *et al.*, 2014; Wakasa *et al.*, 2014; Martin *et al.*, 2014). Similarly, for counting, annotations that have been developed and adopted by each research community were used. As a result, the pipelines used to generate the annotations varied between species (Arabidopsis Genome, 2000; International Rice Genome Sequencing, 2005; Schnable *et al.*, 2009; International Brachypodium, 2010; Schmutz *et al.*, 2010; Potato Genome Sequencing *et al.*, 2011; Tomato Genome, 2012). These differences may explain some observations between species such as the extremely low number of overlapping genes in maize and the relatively high number of overlapping genes in soybean.

By comparing different species it was possible to assess the variation in the frequency of genes that are recalcitrant to precise expression estimates by RNAseq among species. It was expected that species with smaller genomes and more ancient WGD events would have fewer genes that had problems with obtaining unique alignments, and we see evidence for this. However, knowing the proportion of problematic genes in each species is useful for understanding the extent of potential difficulties with precise application of RNAseq. The use of multiple species also provided the opportunity to determine whether the same genes or biological processes are affected in multiple species. There is very little evidence that flagged genes are conserved. Instead, the genes that were flagged in each species tended to be a fairly unique set. This has important implications when performing cross-species comparisons of gene expression. In some cases, the failure of orthologs to show similar transcriptional properties or responses in related species (Davidson *et al.*, 2012) may be due to technical issues in estimating the gene expression in these species rather than a lack of conserved responses.

Most current RNAseq experiments attempt to interrogate gene expression across the entire length of transcripts. New techniques profiling transcript expression are being developed, such as sequencing only the 3′ ends of transcripts (Beck *et al.*, 2010; Moll *et al.*, 2014). This approach of using the 3′ end of a transcript to estimate expression levels was also used in previous array technologies for estimating transcript abundance (http://www.affymetrix.com/estore/browse/level_three_category_and_children.jsp?category = 35871). Untranslated region (UTR) sequences have been shown to vary between gene family members (Lee *et al.*, 2012; Tao *et al.*, 2012), and as such may provide more accurate estimates of transcript abundance between family members. We observed a larger deviation and higher variation of observed/expected values in the middle of transcripts when requiring unique alignments. This could possibly be due to a greater proportion of unique sequences in UTR regions compared with gene bodies among members of gene families. Although the observed counts towards the ends of transcripts are closer to the expected ones, sequencing only one or both ends of transcripts will still not provide sufficiently accurate expression profiling for all transcripts, as genes that deviated by >20% from expected on the gene level still deviated >20% at their 3′ and 5′ ends on average.

Our simulated and empirical datasets highlighted real problems associated with small genes, stemming from both library preparation methods and corrections performed by the commonly used tuxedo package (Trapnell *et al.*, 2012). Variation in library preparation (i.e. 200 bp versus 350 bp insert size during size selection) could produce a significant batch effect. As a result, there will be an artificial bias against a certain transcript size in one library and not in the other, which can create a large number of false positive DE genes that encode short transcripts. Again, being cognizant of this issue is of great importance for biological interpretation of RNAseq data sets.

The intention of this study was to identify genes recalcitrant to RNAseq transcriptome profiling, and to characterize common attributes of these genes across diverse plant species. This information can be incorporated into analysis pipelines in a variety of ways. A conservative approach might be to retain only genes that are DE under both unique and multiple aligning pipelines, and that have not been flagged in the simulated RNAseq data set. Using these criteria, the number of DE genes between cold and control treatments in the empirical dataset used in this study would be reduced to about 4500 high-confidence DE genes. Another less conservative approach would be to simply note flagged DE and non-DE genes and proceed with caution in biological interpretations, as flagged genes have a higher likelihood of producing false positives or false negatives. Furthermore, information regarding the deviation from expected counts could potentially be used for correcting FPKM values to allow for more accurate comparison of differences in transcript abundance between genes. In any case, recognition of recalcitrant genes will improve the quality of biological interpretations that are drawn from RNAseq-based transcriptome profiling experiments.

EXPERIMENTAL PROCEDURES

Read simulation

A single representative transcript was identified for each annotated gene model (Tables S1, S3–S9). The representative transcript was defined as that reported by the sequencing consortium as the representative transcript, if available, or as the longest transcript if not provided by the consortium (Figure 1a,b). Simulated reads were generated from the representative transcript sequence. The first read was defined as the first 50 nucleotides of the transcript, and subsequent reads were generated by shifting over one position per read up to the transcript length –49 position of the transcript (Figure 1c). This resulted in a 50× simulated read depth in the body of the transcript and 1–49× coverage at the 5′ and 3′ ends of the transcript (Figure 1d). Reads were simulated for seven species including Arabidopsis, brachypodium, maize, potato, rice, soybean, and tomato.

Maize B73 control and cold-treated read preparation

Analysis of empirical data was done using previously published RNAseq reads from a study containing maize B73 control and stress-treated seedlings (Makarevitch *et al*., 2015). The RNAseq reads from B73 control and cold treatments were downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA; B73 control accessions SRR1238718, SRR1819617, SRR1819621; B73 cold accessions SRR1238717, SRR1819204, SRR1819205). Sequence adapters were removed using CUTADAPT version 1.7.1 (Martin, 2011) requiring a minimum read length of 50 and low-quality ends were trimmed requiring a minimum quality score of 10. For consistency with the simulated data, only read one was used with reads being treated as single-end reads for downstream analyses.

Read alignment and counting

Reads were aligned using BOWTIE2 version 2.1.0 (Langmead and Salzberg, 2012) and TOPHAT2 version 2.0.10 (Kim *et al*., 2013). Bowtie2 indices were either generated from the genome assembly sequence or the representative transcript sequences for aligning to the genome and transcriptome, respectively. For all species, the minimum intron size was set to 5 and the maximum intron size was set to 20 000 for Arabidopsis, brachypodium, potato, and soybean, and 60 000 for maize, rice, and tomato. For GFF guided mapping, a GFF file was provided containing only information for representative transcripts. All other mapping parameters were set to the default values.

Estimates of transcript abundance were generated with HTSEQ version 0.6.1p1 (Anders *et al*., 2015) and CUFFLINKS2 version 2.1.1 (Trapnell *et al*., 2010). For HTSeq, with unique alignments the union mode and a minimum mapping quality of 20 were used to determine read counts per transcript. For multiple alignments the NH:i:X tag was converted to NH:i:1 for all alignments in the bam file to allow counting of all alignments, and HTSeq was subsequently run using the same options as for unique alignments, except a minimum mapping quality of zero was used. For the simulated data counting was strand specific, and for the empirical data non-strand-specific counting was used. Cufflinks was run requiring a minimum and maximum intron size mirroring the TopHat2 command and providing a GFF file containing information for representative transcripts. All other parameters were set to the default parameters. For FPKM estimates using Cufflinks for the unique alignments pipeline, alignments with a mapping quality score of <50 were discarded and Cufflinks was run with the same options as for multiple mapping.

Additional bioinformatic analyses

The overlap between transcript sequences was determined at the exon level for transcripts on the same strand using the intersect program within BEDTOOLS version 2.19.0 (Quinlan and Hall, 2010) with the –wo parameter. The output was filtered and used to calculate percentage overlap.

Within each species an all-versus-all Blast with the representative transcript sequence was performed using WU TBLASTX (Altschul *et al*., 1990) requiring a minimum E-value of 1×10^{-5} and allowing up to 5000 hits per sequence. Sequence clustering was performed using ORTHOMCL version 1.4 (Li *et al*., 2003; Chen *et al*., 2007) in mode 5 with default parameters. The gene family size for each gene was determined based on the number of representative transcript sequences contained within each OrthoMCL group.

Differentially expressed genes were determined based on counts from unique and multiple alignments using DESeq (Anders and Huber, 2010) within R version 3.1.2 (R Development Core Team, 2011) using BIOCONDUCTOR version 3.0 (Gentleman *et al*., 2004). The default analysis method for replicated data as outlined in the program manual was used. Genes with an adjusted *P*-value <0.01 and a fold change greater than two in either direction were considered as DE.

The enrichment of GO terms was analyzed using AgriGO (Du *et al*., 2010). For analyzed species a singular enrichment analysis was conducted. Fisher's exact test was used to calculate an enrichment *P*-value with a value <0.05 being considered significant, using the Yekutieli method used for multiple test correction.

ACKNOWLEDGEMENTS

We would like to thank Patrick West for his contributions to this manuscript. CDH was supported by a National Science Foundation National Plant Genome Initiative Postdoctoral Fellowship in Biology Fellowship (grant no. 1202724).

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Effect of the counting method on estimates of transcript abundance for maize simulated RNAseq reads and the relationship with transcript size.

Figure S2. Genome browser view of example Arabidopsis non-flagged and flagged genes from simulated RNAseq reads.

Figure S3. Percentage deviation from expected values for simulated RNAseq reads in seven species.

Figure S4. Effect of overlapping gene models on estimates of transcript abundance.

Figure S5. Conservation of flagged genes across species.

Figure S6. Distribution of average expression values for a maize developmental expression atlas categorized by transcript length.

Table S1. Summary of the species and sequences used for simulating RNAseq reads.

Table S2. Mapping metrics for the simulated RNAseq reads across the seven simulated datasets and the empirical maize B73 control and cold-treated RNAseq reads.

Table S3. Gene metrics, expected and observed read counts, fragments per kilobase of exon model per million fragments mapped (FPKM) values, and flagged status for Arabidopsis simulated RNA-seq reads.

Table S4. Gene metrics, expected and observed read counts, fragments per kilobase of exon model per million fragments mapped (FPKM) values, and flagged status for maize simulated RNAseq reads.

Table S5. Gene metrics, expected and observed read counts, and flagged status for brachypodium simulated RNAseq reads.

Table S6. Gene metrics, expected and observed read counts, and flagged status for potato simulated RNAseq reads.

Table S7. Gene metrics, expected and observed read counts, and flagged status for rice simulated RNAseq reads.

Table S8. Gene metrics, expected and observed read counts, and flagged status for soybean simulated RNAseq reads.

Table S9. Gene metrics, expected and observed read counts, and flagged status for tomato simulated RNAseq reads.

Table S10. Gene metrics and read counts for B73 control and cold-treated seedling RNAseq reads.

REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Anders, S. and Huber, W. (2010) Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106.

Anders, S., Pyl, P.T. and Huber, W. (2015) HTSeq-a Python framework to work with high-throughput sequencing data. *Bioinformatics*, **31**, 166–169.

Anderson, J.E., Kantar, M.B., Kono, T.Y. *et al.* (2014) A roadmap for functional structural variants in the soybean genome. *G3: Genes – Genomes – Genetics*, **4**, 1307–1318.

Arabidopsis Genome I (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.

Beck, A.H., Weng, Z., Witten, D.M. *et al.* (2010) 3′-end sequencing for expression quantification (3SEQ) from archival tumor samples. *PLoS ONE*, **5**, e8768.

Cao, J., Schneeberger, K., Ossowski, S. *et al.* (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. *Nat. Genet.* **43**, 956–963.

Chandramohan, R., Wu, P.Y., Phan, J.H. and Wang, M.D. (2013) Benchmarking RNA-Seq quantification tools. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* **2013**, 647–650.

Chen, F., Mackey, A.J., Vermunt, J.K. and Roos, D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE*, **2**, e383.

Chen, J., Zeng, B., Zhang, M., Xie, S., Wang, G., Hauck, A. and Lai, J. (2014) Dynamic transcriptome landscape of maize embryo and endosperm development. *Plant Physiol.* **166**, 252–264.

Chia, J.M., Song, C., Bradbury, P.J. *et al.* (2012) Maize HapMap2 identifies extant variation from a genome in flux. *Nat. Genet.* **44**, 803–807.

Davidson, R.M., Gowda, M., Moghe, G., Lin, H., Vaillancourt, B., Shiu, S.H., Jiang, N. and Robin Buell, C. (2012) Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. *Plant J.* **71**, 492–502.

Du, Z., Zhou, X., Ling, Y., Zhang, Z. and Su, Z. (2010) agriGO: a GO analysis toolkit for the agricultural community. *Nucleic Acids Res.* **38**, W64–W70.

Emrich, S.J., Li, L., Wen, T.J., Yandeau-Nelson, M.D., Fu, Y., Guo, L., Chou, H.H., Aluru, S., Ashlock, D.A. and Schnable, P.S. (2007) Nearly identical paralogs: implications for maize (*Zea mays* L.) genome evolution. *Genetics*, **175**, 429–439.

Engstrom, P.G., Steijger, T., Sipos, B. *et al.* (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods*, **10**, 1185–1191.

Flagel, L.E. and Wendel, J.F. (2009) Gene duplication and evolutionary novelty in plants. *New Phytol.* **183**, 557–564.

Frazee, A.C., Jaffe, A.E., Langmead, B. and Leek, J. (2015) Polyester: simulating RNA-seq datasets with differential transcript expression. *Bioinformatics*, **31**, 2778–2784.

Freeling, M. (2009) Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. *Annu. Rev. Plant Biol.* **60**, 433–453.

Gan, X., Stegle, O., Behr, J. *et al.* (2011) Multiple reference genomes and transcriptomes *for Arabidopsis thaliana*. *Nature*, **477**, 419–423.

Gao, L., Tu, Z.J., Millett, B.P. and Bradeen, J.M. (2013) Insights into organ-specific pathogen defense responses in plants: RNA-seq analysis of potato tuber-Phytophthora infestans interactions. *BMC Genom.* **14**, 340.

Gelli, M., Duo, Y., Konda, A.R., Zhang, C., Holding, D. and Dweikat, I. (2014) Identification of differentially expressed genes between sorghum genotypes with contrasting nitrogen stress tolerance by genome-wide transcriptional profiling. *BMC Genom.* **15**, 179.

Gentleman, R.C., Carey, V.J., Bates, D.M. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80.

Gordon, S.P., Priest, H., Des Marais, D.L. *et al.* (2014) Genome diversity in *Brachypodium distachyon*: deep sequencing of highly diverse inbred lines. *Plant J.* **79**, 361–374.

Grant, G.R., Farkas, M.H., Pizarro, A.D., Lahens, N.F., Schug, J., Brunk, B.P., Stoeckert, C.J., Hogenesch, J.B. and Pierce, E.A. (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, **27**, 2518–2528.

Griebel, T., Zacher, B., Ribeca, P., Raineri, E., Lacroix, V., Guigo, R. and Sammeth, M. (2012) Modelling and simulating generic RNA-Seq experiments with the flux simulator. *Nucleic Acids Res.* **40**, 10073–10083.

Guo, Y., Li, C.I., Ye, F. and Shyr, Y. (2013) Evaluation of read count based RNAseq analysis methods. *BMC Genom.* **14**(Suppl 8), S2.

Hansey, C.N., Vaillancourt, B., Sekhon, R.S., de Leon, N., Kaeppler, S.M. and Buell, C.R. (2012) Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. *PLoS ONE*, **7**, e33071.

Hirsch, C.N., Foerster, J.M., Johnson, J.M. *et al.* (2014) Insights into the maize pan-genome and pan-transcriptome. *Plant Cell*, **26**, 121–135.

Huan, Q., Mao, Z., Zhang, J., Xu, Y. and Chong, K. (2013) Transcriptome-wide analysis of vernalization reveals conserved and species-specific mechanisms in Brachypodium. *J. Integr. Plant Biol.* **55**, 696–709.

International Brachypodium I (2010) Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, **463**, 763–768.

International Rice Genome Sequencing P (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Kang, C., Darwish, O., Geretz, A., Shahan, R., Alkharouf, N. and Liu, Z. (2013) Genome-scale transcriptomic insights into early-stage fruit development in woodland strawberry *Fragaria vesca*. *Plant Cell*, **25**, 1960–1978.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R. and Salzberg, S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36.

Koenig, D., Jimenez-Gomez, J.M., Kimura, S. *et al.* (2013) Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc. Natl Acad. Sci. USA*, **110**, E2655–E2662.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Lee, J., Mun, S., Meyer, T.J. and Han, K. (2012) High levels of sequence diversity in the 5′ UTRs of human-specific L1 elements. *Comp. Funct. Genomics*, **2012**, 1–8.

Lee, T.H., Tang, H., Wang, X. and Paterson, A.H. (2013) PGDD: a database of gene and genome duplication in plants. *Nucleic Acids Res.* **41**, D1152–D1158.

Leisner, C.P., Ming, R. and Ainsworth, E.A. (2014) Distinct transcriptional profiles of ozone stress in soybean (*Glycine max*) flowers and pods. *BMC Plant Biol.* **14**, 335.

Li, L., Stoeckert, C.J. Jr and Roos, D.S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189.

Li, L., Petsch, K., Shimizu, R. *et al.* (2013) Mendelian and non-Mendelian regulation of gene expression in maize. *PLoS Genet.* **9**, e1003202.

Li, G., Wang, D., Yang, R. *et al.* (2014) Temporal patterns of gene expression in developing maize endosperm identified through transcriptome sequencing. *Proc. Natl Acad. Sci. USA*, **111**, 7582–7587.

Lindner, R. and Friedel, C.C. (2012) A comprehensive evaluation of alignment algorithms in the context of RNA-seq. *PLoS ONE*, **7**, e52403.

Makarevitch, I., Waters, A.J., West, P.T., Stitzer, M., Hirsch, C.N., Ross-Ibarra, J. and Springer, N.M. (2015) Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet.* **11**, e1004915.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12.

Martin, J.A., Johnson, N.V., Gross, S.M. *et al.* (2014) A near complete snapshot of the Zea mays seedling transcriptome revealed from ultra-deep sequencing. *Sci. Rep.* **4**, 4519.

Massa, A.N., Childs, K.L., Lin, H., Bryan, G.J., Giuliano, G. and Buell, C.R. (2011) The transcriptome of the reference potato genome *Solanum tuberosum* Group Phureja clone DM1-3 516R44. *PLoS ONE*, **6**, e26801.

Moll, P., Ante, M., Seitz, A. and Reda, T. (2014) QuantSeq 3′ mRNA sequencing for RNA quantification. *Nat. Methods*, **11**, i–iii.

Nguyen, C.V., Vrebalov, J.T., Gapper, N.E., Zheng, Y., Zhong, S., Fei, Z. and Giovannoni, J.J. (2014) Tomato GOLDEN2-LIKE transcription factors reveal molecular gradients that function during fruit development and ripening. *Plant Cell*, **26**, 585–601.

Nookaew, I., Papini, M., Pornputtapong, N., Scalcinati, G., Fagerberg, L., Uhlen, M. and Nielsen, J. (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic Acids Res.* **40**, 10084–10097.

Paschold, A., Jia, Y., Marcon, C. *et al.* (2012) Complementation contributes to transcriptome complexity in maize (*Zea mays* L.) hybrids relative to their inbred parents. *Genome Res.* **22**, 2445–2454.

Potato Genome Sequencing C, Xu, X., Pan, S. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C.E., Socci, N.D. and Betel, D. (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. *Genome Biol.* **14**, R95.

Schatz, M.C., Maron, L.G., Stein, J.C. *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of aus and indica. *Genome Biol.* **15**, 506.

Schmutz, J., Cannon, S.B., Schlueter, J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.

Schnable, P.S., Ware, D., Fulton, R.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Schnable, J.C., Springer, N.M. and Freeling, M. (2011) Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss. *Proc. Natl Acad. Sci. USA*, **108**, 4069–4074.

Schnable, J.C., Freeling, M. and Lyons, E. (2012) Genome-wide analysis of syntenic gene deletion in the grasses. *Genome Biol. Evol.* **4**, 265–277.

Sekhon, R.S., Hirsch, C.N., Childs, K.L., Breitzman, M.W., Kell, P., Duvick, S., Spalding, E.P., Buell, C.R., de Leon, N. and Kaeppler, S.M. (2014) Phenotypic and transcriptional analysis of divergently selected maize populations reveals the role of developmental timing in seed size determination. *Plant Physiol.* **165**, 658–669.

Severin, A.J., Woody, J.L., Bolon, Y.T. *et al.* (2010) RNA-Seq Atlas of Glycine max: a guide to the soybean transcriptome. *BMC Plant Biol.* **10**, 160.

Seyednasrollah, F., Laiho, A. and Elo, L.L. (2015) Comparison of software packages for detecting differential expression in RNA-seq studies. *Brief Bioinform.* **16**, 59–70.

Song, G., Guo, Z., Liu, Z. *et al.* (2013) Global RNA sequencing reveals that genotype-dependent allele-specific expression contributes to differential expression in rice F1 hybrids. *BMC Plant Biol.* **13**, 221.

Stelpflug, S., Sekhon, R.S., Vaillancourt, B., Hirsch, C.N., Buell, C.R., de Leon, N. and Kaeppler, S. (2015) An expanded maize gene expression atlas based on RNA-sequenceing and its use to explore root development. *The Plant Genome*, doi:10.3835/plantgenome2015.3804.0025.

Tao, P., Peng, L., Huang, X. and Wang, J. (2012) Comparative analysis of the variable 3′ UTR and gene expression of the KIN and KIN-homologous LEA genes in Capsella bursa-pastoris. *Plant Cell Rep.* **31**, 1769–1777.

Tomato Genome C (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J., Salzberg, S.L., Wold, B.J. and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578.

Wakasa, Y., Oono, Y., Yazawa, T., Hayashi, S., Ozawa, K., Handa, H., Matsumoto, T. and Takaiwa, F. (2014) RNA sequencing-mediated transcriptome analysis of rice plants in endoplasmic reticulum stress conditions. *BMC Plant Biol.* **14**, 101.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63.

Yazawa, T., Kawahigashi, H., Matsumoto, T. and Mizuno, H. (2013) Simultaneous transcriptome analysis of Sorghum and *Bipolaris sorghicola* by using RNA-seq in combination with de novo transcriptome assembly. *PLoS ONE*, **8**, e62460.

Zhai, R., Feng, Y., Wang, H. *et al.* (2013) Transcriptome analysis of rice root heterosis by RNA-Seq. *BMC Genom.* **14**, 19.

Zhang, Z.H., Jhaveri, D.J., Marshall, V.M. *et al.* (2014) A comparative study of techniques for differential expression analysis on RNA-Seq data. *PLoS ONE*, **9**, e103207.