Bioinformatics, 33(3), 2017, 327–333
doi: 10.1093/bioinformatics/btw625
Advance Access Publication Date: 30 September 2016
Original Paper

Genome analysis

Comprehensive evaluation of *de novo* transcriptome assembly programs and their effects on differential gene expression analysis

Sufang Wang[1] and Michael Gribskov[1,2,*]

[1]Department of Biological Sciences and [2]Department of Computer Science, Purdue University, West Lafayette, IN, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on June 2, 2016; revised on September 8, 2016; accepted on September 23, 2016

Abstract

Motivation: With the decreased cost of RNA-Seq, an increasing number of non-model organisms have been sequenced. Due to the lack of reference genomes, *de novo* transcriptome assembly is required. However, there is limited systematic research evaluating the quality of *de novo* transcriptome assemblies and how the assembly quality influences downstream analysis.

Results: We used two authentic RNA-Seq datasets from *Arabidopsis thaliana*, and produced transcriptome assemblies using eight programs with a series of *k-mer* sizes (from 25 to 71), including BinPacker, Bridger, IDBA-tran, Oases-Velvet, SOAPdenovo-Trans, SSP, Trans-ABySS and Trinity. We measured the assembly quality in terms of reference genome base and gene coverage, transcriptome assembly base coverage, number of chimeras and number of recovered full-length transcripts. SOAPdenovo-Trans performed best in base coverage, while Trans-ABySS performed best in gene coverage and number of recovered full-length transcripts. In terms of chimeric sequences, BinPacker and Oases-Velvet were the worst, while IDBA-tran, SOAPdenovo-Trans, Trans-ABySS and Trinity produced fewer chimeras across all single *k-mer* assemblies. In differential gene expression analysis, about 70% of the significantly differentially expressed genes (DEG) were the same using reference genome and *de novo* assemblies. We further identify four reasons for the differences in significant DEG between reference genome and *de novo* transcriptome assemblies: incomplete annotation, exon level differences, transcript fragmentation and incorrect gene annotation, which we suggest that *de novo* assembly is beneficial even when a reference genome is available.

Availability and Implementation: Software used in this study are publicly available at the authors' websites.

Contact: gribskov@purdue.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Next-generation sequencing of RNA, is a sequencing technology that directly determines the sequence cDNA, and is often used to quantify the expression level of an entire transcriptome in a single experiment (Wang *et al.*, 2009). As the cost of sequencing has plummeted in recent years, an increasing number of non-model organisms, for which there are no available reference genomes, have been the subject of RNA-Seq experiments, generally with the goal of identifying differentially expressed genes (DEG). Identifying DEG requires a reference transcriptome, a catalog of all of the transcripts that may be relevant to the particular experiment. This reference transcriptome may be derived from a whole genome assembly, an approach used by programs such as Cufflinks (Trapnell *et al.*, 2010). We call this type of analysis a reference genome-based analysis. When a reference genome is not available, a *de novo* transcriptome assembly can be used. There are many programs available for *de novo* transcriptome assembly including BinPacker (Liu *et al.*, 2016), Bridger (Chang *et al.*, 2015), IDBA-tran (Peng *et al.*, 2013), Mira (Chevreux *et al.*, 1999), Oases-Velvet (Schulz *et al.*, 2012; Zerbino and Birney, 2008), SOAPdenovo-Trans (Luo *et al.*, 2012; Xie *et al.*, 2014), SSP (Safikhani *et al.*, 2013), Trans-ABySS (Robertson *et al.*, 2010; Simpson *et al.*, 2009) and Trinity (Grabherr *et al.*, 2011). These programs generally use similar approaches based on de Bruijn graphs for transcriptome assembly, but differ widely in the number of transcripts and genes they predict. The quality and character of the assembled transcriptome, naturally, has a great effect on the downstream analysis of differential gene expression.

While several collaborations have examined quality of genome assembly programs, for instance Assemblathon 1 and 2 (Bradnam *et al.*, 2013; Earl *et al.*, 2011) and the Genome Assembly Gold-standard Evaluation (GAGE) (Salzberg *et al.*, 2012), there have been relatively few efforts to evaluate transcriptome assembly. A comparison of 14 transcript reconstruction methods using RNA-Seq showed wide variations among programs (Steijger *et al.*, 2013), but most of the programs evaluated were reference genome-based; only Oases-Velvet was a *de novo* transcriptome assembly program. Another study compared *de novo* assemblies produced using Trinity and SOAPdenovo-Trans, using simulated data from Zebra finch (Vijay *et al.*, 2013), but did not examine the relative performance of the programs on real RNA-Seq data. Similarly, Neil and Emrich (2013) used simulated reads from *Drosophila melanogaster* to examine the performance and quality of assemblies using Newbler (Roche), a program originally designed for DNA sequence assembly. Only one study comparing the four most widely used packages, Oases, Trinity, Trans-ABySS and SOAPdenovo-Trans has appeared, but this study only focused on the quality of the assembly and not on downstream analysis (Yang and Smith, 2013).

More recently, two studies have developed software to evaluate transcriptome assemblies using probabilistic models (Li *et al.*, 2014; Smith-unna *et al.*, 2015). However, the scores from statistical models do not explain biological concepts. Taken together, there is only a limited amount of systematic research evaluating the quality of *de novo* transcriptome assemblies, and very little on the effect of the transcriptome assembly quality on downstream differential expression analysis. Furthermore, many of the available comparisons are based only on simulated data that is generally less complex and noisy than real biological data.

In this study, in order to better reflect the real world, we used authentic RNA-Seq data from two *Arabidopsis thaliana* studies (Lai *et al.*, 2014; Zhu *et al.*, 2014) to examine the effect of transcriptome assembly programs on both assembly quality and differential gene expression analysis. We chose Arabidopsis because its reference genome is well annotated, and more importantly, we obtained datasets for which biological analyses of differential gene expression have already been published (Lai *et al.*, 2014; Zhu *et al.*, 2014), providing us with an expertly curated gold standard. The first goal in this research is to compare the *de novo* transcriptome assemblies created by eight currently available programs, namely BinPacker (Liu *et al.*, 2016), Bridger (Chang *et al.*, 2015), IDBA-tran (Peng *et al.*, 2013), Oases-Velvet (Schulz *et al.*, 2012; Zerbino and Birney, 2008), SOAPdenovo-Trans (Luo *et al.*, 2012; Xie *et al.*, 2014), SSP (Safikhani *et al.*, 2013), Trans-ABySS (Robertson *et al.*, 2010; Simpson *et al.*, 2009) and Trinity (Grabherr *et al.*, 2011). We evaluated assembly quality through the number of predicted transcripts, base coverage, gene coverage, the presence of chimeric transcripts, number of recovered full-length genes, Detonate score and alignment rate. In addition, we also examined the effects of *k-mer* size in IDBA-tran, Oases-Velvet, SOAPdenovo-Trans, Trans-ABySS and Trinity in order to explore the best choice of *k-mer* size in transcriptome assembly. The second goal of this paper is to investigate the effect of transcriptome assembly on the downstream differential gene expression analysis. The main advantages of our approach are its coverage of a wide range of available software, use of authentic biological data, and evaluation of both assembly quality and the effects of the transcriptome assembly on differential gene expression analysis.

2 Methods

2.1 RNA-seq data

RNA-seq data from three *Arabidopsis thaliana* genotypes were used in this study: a wildtype Col-0 and two loss-of-function mutants, cdk8 and med18. RNA was extracted from two time points (0h and at 36h after pathogen treatment). There were two biological replicates at each time point for each sample. As a result, there were 12 libraries sequenced using the Illumina TruSeq approach. The sequencing data had been submitted to NCBI Sequence read archive under accession code SRP033777 (Lai *et al.*, 2014).

2.2 Quality control andtranscriptome assembly

The adapter sequences and low quality portions of reads were removed using Trimmomatic (version 0.32) program (Bolger *et al.*, 2014). Adapters and low quality read regions with average quality below 13 over a 4 base window were removed. Only reads with a trimmed length over 30 bases were used in further analysis. The number of paired-end reads and unpaired reads in each sample is shown in Supplementary Table S1. Cleaned reads were assembled by BinPacker (version 1.0), Bridger (r2014-12-01), IDBA-tran (version 1.1.1), Oases-Velvet (version velvet 1.2.10 and oases 0.02.08), SOAPdenovo-Trans (version 1.03), SSP (no version information), Trans-ABySS (version 1.5.1) and Trinity (version 2.0.6) using parameters recommended by the authors (except as noted in the text).

2.3 Base coverage, gene coverage, chimeric sequences and number of recovered full length genes

We compared each *de novo* assembly to the *Arabidopsis thaliana* reference genome (http://plants.ensembl.org/Arabidopsis_thaliana, TAIR version 10.20 database) using BlastN (Altschul *et al.*, 1990) with default settings (Blast version 2.2.29+). We filtered hits by two criteria: identity score ≥ 97%; and aligned length ≥100 bases. We used a locally developed Perl script to calculate coverage and the number of chimeras, as described in Results.

2.4 Alignment and quantification

RSEM (Li *et al.*, 2010; Li and Dewey, 2011) program (version 1.2.11) was used to align and quantify gene level expression with bowtie2, and the very sensitive setting.

2.5 Differential expression gene analysis

The edgeR package (Robinson and Smyth, 2007) was used to determine differential expression. Only genes with observed counts bigger than 4 CPM across all samples were analyzed.

3 Results

3.1 Basic statistics of *de novo* transcriptome assembly

We assembled transcriptomes by BinPacker, Bridger, IDBA-tran, Oases-Velvet, SOAPdenovo-Trans, SSP and Trans-ABySS and Trinity. Genome and transcriptome assembly programs are strongly affected

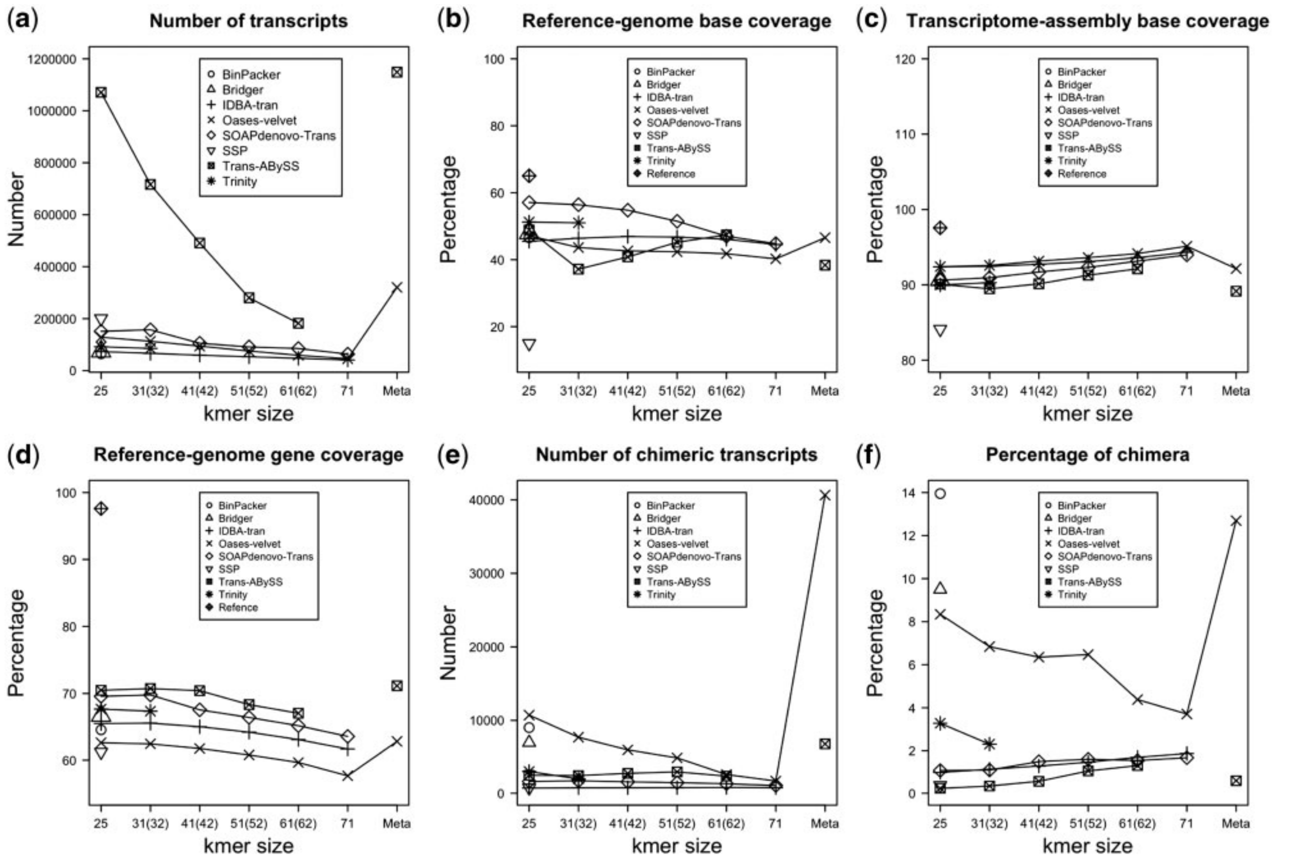

Fig. 1. Transcriptome assembly quality part 1. (**a**) Number of assembled transcripts. (**b**) Reference-genome base coverage. (**c**) Transcriptome-assembly base coverage. (**d**) Reference-genome gene coverage. (**e**) Number of chimeric sequences. (**f**) Percentage of chimeric sequences

by the length of overlapping sequence used to align the sequence reads. This parameter is commonly referred to as the *k-mer*. In BinPacker, Bridger, SSP, only one *k-mer* size (25) was used. In Trinity, two *k-mer* sizes (25, 32) are available. IDBA-tran, Oases-Velvet and SOAPdenovo-Trans support multiple *k-mer* choices (we used 25, 31, 41, 51, 61, 71). For Trans-ABySS, we used *k-mer* 25, 32, 42, 52 and 62. In addition, we also produced a meta-assembly, an assembly in which all the assemblies at different *k-mer* lengths are combined into a single assembly, for both Oases-Velvet and Trans-ABySS. In total, we performed 30 *de novo* transcriptome assemblies.

A summary of the characteristics of the assemblies produced by each program (and *k-mer*) is shown in Supplementary Table S2. The number of predicted transcripts generally decreases when *k-mer* size is increased.

For example, for Oases-Velvet, the number of predicted transcripts dropped from 128 483 with *k-mer* 25 to 45 397 with *k-mer* 71 (Fig. 1a, Supplementary Table S2). Trans-ABySS generally produced the highest number of predicted transcripts (Fig. 1a), but the distribution of predicted transcript length in SSP, SOAPdenovo-Trans and Trans-ABySS assemblies was left skewed due to a large number of short predicted transcripts (Supplementary Table S2). As a result, the minimum, first quartile, median, mean and third quartile length in SOAPdenovo-Trans, SSP and Trans-ABySS were shorter than that in Binpacker, Bridger, IDBA-tran, Oases-Velvet and Trinity, but the maximum predicted transcripts lengths were all similar except for the SSP assembly, which has a very short maximum predicted transcript length (Supplementary Table S2).

There is a clear relationship between genes and isoforms in BinPacker, Bridger, Oases-Velvet, SOAPdenovo-Trans, SSP and Trinity, which is an advantage in the downstream analysis. For example, in Trinity (version 2.0.6), a gene unit might be designated TR1|c0_g1, and the multiple isoforms designated as TR1|c0_g1_i1, TR1|c0_g1_i2 and so on. This relationship is not clear in IDBA-tran and Trans-ABySS. Comparing the number of predicted genes in each assembly, Oases-Velvet produced a number of predicted genes ranging from 15 353 (long *k-mer*) to 28 346 (short *k-mer*); the latter number is closer to the expected number of genes based on the *Arabidopsis thaliana* reference genome annotation (Supplementary Table S2). BinPacker, Bridger, SOAPdenovo-Trans, SSP and Trinity tend to predict more genes than seem plausible; there are 31 775 transcripts listed in the Arabidopsis reference gene annotation file, and one does not expect all transcripts to be present in every sample.

3.2 Base and gene coverage of *de novo* transcriptome assemblies

In order to estimate the quality of the *de novo* assemblies, we compared each assembly with the *Arabidopsis thaliana* reference genome annotation. We defined three measures of transcriptome quality: reference-genome base coverage (fraction of bases of the reference genome covered by the transcriptome assemblies), transcriptome-assembly base coverage (fraction of bases of the transcriptome assembly covered by the reference genome) and reference-genome gene coverage (fraction of annotated genes in the reference genome covered, completely or in part, by transcriptome assembly). We also used the *Arabidopsis thaliana* cDNA file (TAIR 10.20) as a control in calculating coverage.

Overall, SOAPdenovo-Trans had the highest reference-genome base coverage at all *k-mer* sizes (Fig. 1b, Supplementary Table S3). When *k-mer* was 25, Trinity ranked the second highest in reference

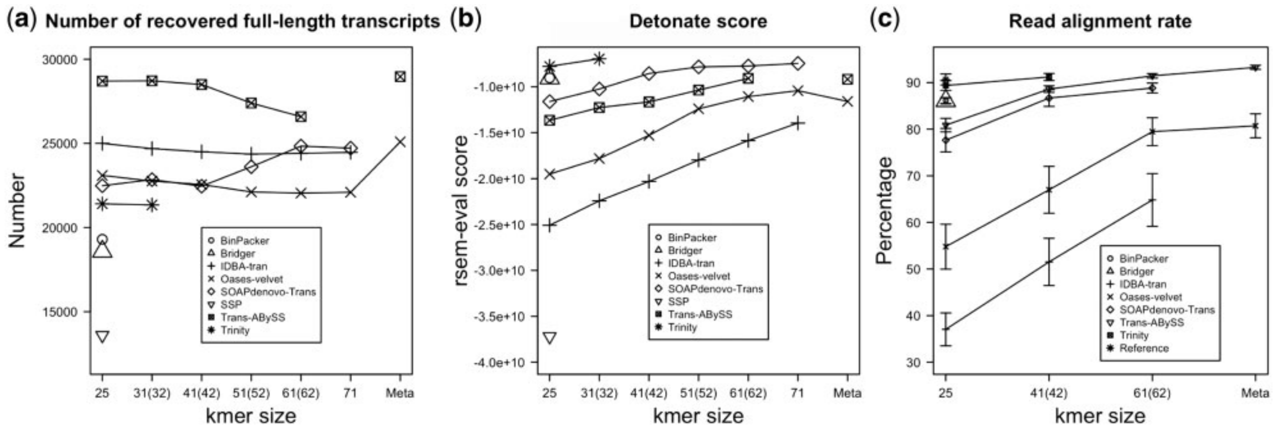

Fig. 2. Transcriptome assembly quality part 2. (a) The number of recovered full-length transcripts. (b) The score produced by Detonate program. (c) Read alignment rate. Each dot represents the average alignment rate over 12 libraries. The error bar represents the standard deviation of the mean

genome base coverage. Both SOAPdenovo-Trans and Oases-Velvet showed decreases in reference-genome base coverage as *k-mer* size increased. However, Trans-ABySS showed an opposite trend. Reference genome base coverage decreased from *k-mer* 25 to *k-mer* 32, and then increased from *k-mer* 32 to *k-mer* 62, but dropped again in the meta-assembly (Fig. 1b, Supplementary Table S3). Bridger and IDBA-tran had reference genome base coverage similar to Trans-ABySS and Oases-Velvet at *k-mer* 25, but SSP had a very low coverage (14.99%), which probably is due to small number of predicted transcripts in the SSP assembly.

The trend of transcriptome-assembly base coverage is uniform. Coverage increases with *k-mer* size, but drops in both meta-assemblies. At every *k-mer* length, Oases-Velvet assemblies had higher coverage than assemblies made by other methods. Actually, all of the *de novo* assemblies had similar transcriptome-assembly base coverage of around 90% with the exception of SSP, which only showed 84% (Fig 1c, Supplementary Table S3).

In the Arabidopsis gene annotation, some annotated genes have overlapping regions (for instance, convergent genes on different strands), which will lead to assembly programs producing a single predicted transcript in this region when the data is not strand-specific, as is the case here. Thus, in order to better estimate the coverage, we combined overlapping genes into gene blocks. A total of 30 039 blocks were produced from the original 31 775 genes. Reference-genome gene coverage ranged from 57.64% to 71.14% across the 30 assemblies (Supplementary Table S3). Since the RNA samples were obtained from the entire aerial portion of the plant, this degree of coverage seems reasonable. Trans-ABySS showed the highest gene coverage in all single *k-mer* assemblies and in the meta-assemblies (Fig. 1d). Again, IDBA-tran, Oases-Velvet, SOAPdenovo-Trans and Trinity showed decreases in gene coverage as *k-mer* size increased, which follows the same trend as the reference-genome base coverage. Both meta-assemblies (Trans-ABySS and Oases-Velvet) showed higher coverage than any single *k-mer* size.

3.3 Chimeric sequences in *de novo* transcriptome assemblies

Chimeric transcripts, where reads from two different genes are combined into a single transcript, are common in *de novo* assemblies. We do not consider transcripts that have overlapping regions, as discussed above, to be chimeric. We define two scenarios for chimeric predicted transcripts. First, a predicted transcript that has multiple matches to the reference genome on different chromosomes is clearly chimeric. Second, an assembled transcript that has matches on the same chromosome, but the matches are separated by 10 000 bp or more, is considered to be chimeric, because most genes in Arabidopsis are much less than 10 000 bases long (Supplementary Fig. S1).

Among the 30 transcriptome assemblies, the number of chimeric transcripts ranged from 741 to 40 640, and the percentage of chimeric transcripts (calculated by number of chimeric transcripts divided by number of predicted transcripts in each assembly) ranged from 0.23% to 13.95% (Supplementary Table S4). Oases-Velvet produced the highest number of chimeric transcripts at different *k-mer* sizes, especially in its meta-assembly (Fig. 1e). At a *k-mer* size of 25, BinPacker and Bridger had the highest percentage of chimeras over 9% (Fig. 1f, Supplementary Table S4). All other programs tend to produce similar percentages of chimeras from 1% to 3% (Fig. 1f, Supplementary Table S4).

We expected to see that the use of larger *k-mer* sizes in the assemblies would result in fewer chimeras, but this was only observed in the Oases-Velvet and Trinity assemblies. Interestingly and surprisingly, IDBA-tran, Trans-ABySS and SOAPdenovo-Trans showed the opposite trend, the percentage of chimeras increased with *k-mer* size (Fig. 1f, Supplementary Table S4), although IDBA-tran, Trans-ABySS and SOAPdenovo produced fewer chimeric transcripts than Oases-Velvet at all *k-mer* sizes. We found that IDBA-tran, SOAPdenovo-Trans and Trans-ABySS all tend to produce a stable number of chimeras for all *k-mer* sizes. Due to stable number of chimera produced, this results in a relatively stable percentage of chimeras in these three programs.

3.4 Number of recovered full-length transcripts and detonate score

We also calculated the number of recovered full-length transcripts by matching each assembled transcript to the Arabidopsis cDNA using BlastN. An assembled transcript that covers a reference transcript length over 80% of its length, with 97% identity, and alignment length over 100 bases, is considered to be a full-length recovered transcript. In total, there are 41392 of transcripts in the Arabidopsis cDNA file. The number of recovered full-length transcripts ranged from 13600 to 28967, TransABySS and IDBA-tran performed best, SSP was the worst (Fig 2a).

The Detonate program (Li *et al.*, 2014) has recently been introduced to evaluate transcriptome quality; we applied Detonate (version 1.10) to all of the assemblies. Detonate calculates an overall score, the higher the score, the better the assembly. From the result, we can see that Trinity rank highest, and SSP worst (Fig 2b). The results are highly consistent with ours: Trinity and SOAPdenovo-trans performed best, and SSP was the worst. Unfortunately the overall DETONATE score gives little insight to why SSP was the worst. However, from our evaluation metrics, we can clearly see that SSP predicts a smaller number of long transcripts and has low coverage of the known transcripts.

3.5 Alignment in *de novo* transcriptome assembly

After completing transcriptome assembly, the next step in a typical analysis is to align reads either to the reference genome or *de novo* assembled transcriptome (Trapnell and Salzberg, 2009). We used RSEM to align the reads to both *de novo* transcriptome assemblies and reference genome. Thus, read alignment was performed using the reference genome and 19 *de novo* transcriptome assemblies (the intermediate *k-mer* size 31 and 51 were not used), including BinPacker, Bridger, IDBA-tran (*k-mer* 25, 41, 61), Oases-Velvet (*k-mer* 25, 41, 61 and Meta-assembly), SOAPdenovo-Trans (*k-mer* 25, 41 and 61), SSP, Trans-ABySS (*k-mer* 25, 42, 62 and Meta-assembly) and Trinity (*k-mer* 25, 32).

The fraction of RNA-Seq reads that align to the reference is called the alignment rate. From the RSEM alignment result, we obtained the alignment rate for each sample and transcriptome assembly. Overall, the alignment rate increased when *k-mer* size increased (Fig. 2c). The *Arabidopsis thaliana* reference genome had the best alignment rate, about 90%. At *k-mer* 25, BinPacker, Bridger and Trinity performed similarly to the reference genome. Across alignments at all *k-mer* sizes, Trans-ABySS had the highest alignment rate (93.44% for the Trans-ABySS meta-assembly). SSP performed the worst with only a 1% alignment rate (Supplementary Table S5). Due to this, we did not further analyze the SSP assembly in downstream analyses.

3.6 Differential gene expression analysis

CDK8 and Med18 are components of the mediator complex in plants and are expected to affect response to pathogens (Lai *et al.*, 2014; Zhu *et al.*, 2014). We expect that genes regulated by the plant mediator complex or those involved in pathogen response pathways would be identified as differentially expressed in this RNA-Seq analysis since the experiment involves a pathogen challenge. We define significant DEG by two criteria: false discovery rate less than 0.1, and $\log_2$ fold-change larger than ± 2. The number of DEG identified with each program vary widely (Supplementary Table S6). However, simply comparing the numbers of significant genes among different assemblies is not sufficient to explain the differences seen in predicted DEG with the *de novo* assembly programs. We are more interested in whether the DEG identified by the *de novo* assemblies were the same as those identified in the reference genome-based analysis. Thus, we used BlastN to match each significantly differentially expressed predicted transcript from the *de novo* assemblies to the *Arabidopsis thaliana* reference genome. Then, we determined the number of matched genes identified as significant DEG in each *de novo* transcriptome assembly (Supplementary Table S6). In general, *de novo* assemblies tends to identify more significant DEG than the reference genome-based approach, but around 70% of significant DEG were found by both reference and *de novo* approaches. The overlap with the reference genome-based DEG increases with *k-mer* size for assemblies made with Trinity, Oases-Velvet and SOAPdenovo-Trans, suggesting that larger *k-mer* size may improve both the assemblies and downstream analyses. However, this trend was not observed in the IDBA-tran and Trans-ABySS transcriptome assemblies. This could be due to the relationship between genes and isoforms is not clearly indicated by these two programs, which leads to a less accurate estimation of the expression level for a ‘gene’.

3.7 Unique DEG (uDEG) identified by *de novo* transcriptome assembly but not by reference genome-based approach

More importantly, we found that some DEG were uniquely identified by *de novo* transcriptome assembly programs, but not by the reference genome-based approach. Some of these genes appear to be biologically relevant, and were identified as being significant differentially expressed in multiple *de novo* assemblies and at multiple *k-mer* sizes (Supplementary Table S7). We examined the uDEG (total number: 271) to determine why the *de novo*-based and reference genome-based approaches differ (Supplementary Fig. S2). The first reason is a threshold effect. Because we use two criteria ($\log_2(\mathrm{FC}) \geq \pm 2$ and $\mathrm{FDR} \leq 0.01$), some genes are close to one or both thresholds. The expression of these genes can easily fall on one side or the other of the threshold in analyses based on different reference transcriptomes. These genes account for 27% of the uDEG. Furthermore, 24 genes (9%) have overlap with the neighboring genes, which causes a labeling difference that makes them appear unique in the *de novo* assemblies. In addition to the technical issues described above, the other common issue in RNA-Seq is low coverage observed for certain genes; 25% of uDEG show low coverage with observed counts smaller than 4 CPM across all samples (Supplementary Fig. S2). A few transcripts corresponding to transposable elements and pseudo genes (7%) were also identified in the *de novo* assemblies.

The remaining 32% of uDEG do not differ for merely technical or methodological reasons, that is, they may be biologically relevant. We visually inspected the reads mapped to these transcripts, and the corresponding genomic region, using the IGV program (Robinson *et al.*, 2011; Thorvaldsdóttir *et al.*, 2013) to determine the cause of the differences between the reference and *de novo* based approaches. Four major reasons that explain the differences between the reference genome and *de novo* transcriptome based analyses were observed. First, incomplete annotation of the reference genome, especially of UTR regions is a significant factor. We observed that many UTR regions were clearly transcribed, resulting in significantly longer assembled transcripts. For example, AT4G14365, is annotated as beginning at base 8 271 464, and ending at base 8 273 765 (2302 bases) on chromosome 4. But the *de novo* assembly programs identify a transcript covering bases 8 271 397 to 8 273 889 (2493 bases) (Fig. 3a). Due to the larger size of the gene in the *de novo* transcriptome reference, the expression level is observed to be higher than that in the analysis based on the reference genome. The second reason for differences between reference and *de novo* assembly based analyses are exon level differences in expression level. For example, AT1G64790 (ILITHYIA, ILA), a plant immunity gene with 66 exons (18 492 base), was identified as significantly differentially expressed in 16 of the *de novo* transcriptome based analyses (Supplementary Table S7). This is due to a very large difference in the expression of the second to last exon (Fig. 3b). The reference genome-based analysis includes the entire 18 kb transcript region; expression levels for most of the exons are low in both wt and mutant samples. The large differences in expression of this exon are essentially averaged over the much longer 18 kb region, resulting in it

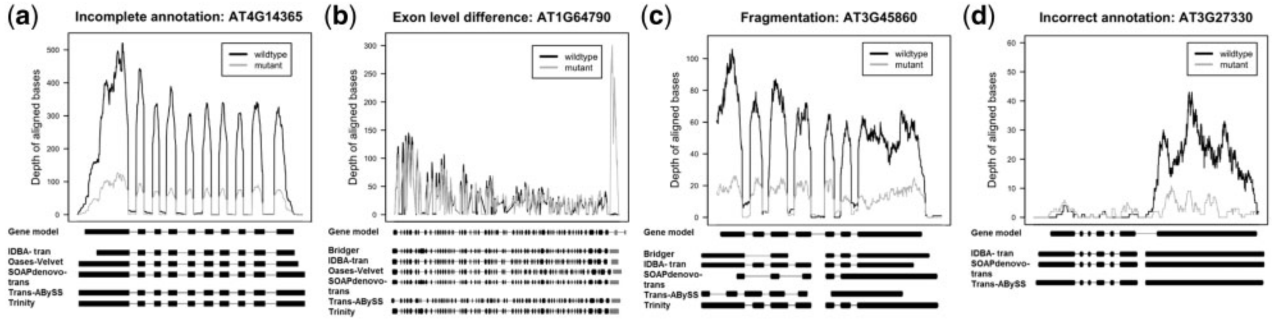

Fig. 3. Differences between reference genome and *de novo* assembly based analyses. The upper panel in each figure shows the depth of aligned bases for a region of the reference genome. The lower panel shows gene model from reference genome annotation and predicted transcripts from *de novo* assemblies. Predicted transcripts are shown only for selected, relevant, *de novo* assemblies. (**a**) Incomplete annotation: AT4G14365. UTR regions are clearly present in the *de novo* assembled transcripts, thus the predicted transcripts are longer than the gene model from reference genome. (**b**) Exon level differences: AT1G64790. The second to last exon and neighboring regions are highly expressed. *De novo* assembly programs assembled this region (highlighted in grey) as a distinct transcript. (**c**) Fragmentation: AT3G45860. The first four exons and the last three exons, consistently assembled as separate transcripts. (**d**) Incorrect gene model: AT3G27330. The left region with low expression contains a glycosyl transferase domain, and the more highly expressed region on the right matches a zinc finger motif

not being identified as a significant DEG. But in the *de novo* assemblies this exon generally assembled as a unique predicted transcript; the remainder of the annotated exons are not included because of their low read coverage. Thus in the *de novo* based analyses this gene appears differentially expressed. It is not clear what is going on in this region: there may be a second promoter that drives transcription from an internal promoter, or this region of the primary transcript may be stabilized by an unknown mechanism. In any case, this gene is known to be involved in systemic acquired resistance induced by bacterial pathogens (Monaghan and Li, 2010), exactly the challenge that was used in this experiment. The third reason for differences is fragmentation, in which a gene, such as AT3G45860 (Fig. 3c), is assembled into two or more separate transcripts. Due to apparently random differences in read coverage, the separately assembled 3′ transcript region is detected as differentially expressed, while the entire gene region examined in the reference genome based analysis is not. This difference is therefore partly due to a threshold effect; shorter assembled transcripts necessarily have fewer mapped reads and random variability is higher for low read counts. The fourth reason is poor or incorrect annotation of the reference genome. For example, AT3G27330 is identified as DEG in 9 different *de novo* transcriptome based analyses (Fig. 3d). This region is annotated as a zinc finger gene, however, in the *de novo* assemblies, it is consistently assembled as two separate genes. These predicted transcripts match to separate pfam (Finn *et al.*, 2014) motifs; the left transcript matches a glycosyltransferase domain (PF01697) and the right transcript matches a zinc finger motif (PF00097). Except for genomes whose annotation is based directly on the Arabidopsis gene annotation, these domains do not appear to occur in the same protein. Together with the transcription pattern, this strongly suggests that these are, in fact, two separate genes. In the *de novo* transcriptome analysis, one is clearly seen to be differentially expressed while in the reference genome based analysis, averaging over a longer region makes the gene appear to be not DE.

4 Discussion

In this study, we focus on evaluating *de novo* transcriptome assembly quality and the effect of transcriptome quality on downstream DE analysis. In the upstream analysis of assembly quality, it is obvious that some metrics used in genome assembly are not suitable for transcriptome assembly, in particular, N50, NG50 and assembly size are inappropriate because longer sequences or larger total assembly size do not indicate a better transcriptome assembly; indeed, these may indicate a high level of overjoining or chimerism. We observed that base coverage of the reference genome decreases as *k-mer* size is increased (Fig. 1b, c). When *k-mer* size is increased in *de novo* assembly, some low expression transcripts may be lost because some reads cannot be overlapped due to sequencing errors, reducing the already limited overlap and read coverage. A similar argument may be made with respect to gene coverage (Fig. 1d). However, at larger *k-mer* sizes, assembly programs should produce more accurate predicted transcripts due to the longer overlapping regions between reads. Thus, some misassembled transcripts produced at lower *k-mer* sizes should be removed. This is likely to increase the coverage of the transcriptome by the reference genome. Thus, increasing *k-mer* size is one of the most important approaches to improving assembly quality in all programs. Somewhat surprisingly, the number of chimeric transcripts identified using the latest version of Trinity was much smaller than seen previously (Yang and Smith, 2013), which suggests that Trinity has been significantly improved. While all of the transcriptome assembly programs perform well by some quality metrics, SOAPdenovo-Trans ($k\text{-}mer = 25$), which produces a less excessive number of predicted transcripts, but with high reference genome base coverage, reference gene coverage, assembly base coverage and read alignment rate appears to be the best general choice for the Arabidopsis data used here (Table 1). Trinity, while ranking below SOAPdenovo-Trans in most categories, also performs consistently, and has the highest read alignment rate of all the *de novo* assembly programs. In addition, although there are packages that can evaluate assemblies using statistical models, it is difficult to understand the biological differences between assemblies based on these summary analyses.

In the downstream DE analysis, through our comparisons between reference genome and *de novo* assembly, genes identified in the reference but NOT in *de novo* assembly represent around 10% of the differentially expressed genes. These genes generally have either poor annotation, overlap with their neighboring genes, or are not biologically interesting (Supplementary Table S8). Thus, we focused more on the genes identified in the *de novo* but NOT in the reference assemblies. Some DEG are identified only by the *de novo* based analyses, and about a third of these uDEG appear to be non-trivial and, based on the gene annotation, of possible biological relevance. Four major reasons for differences between the reference genome and *de novo* transcriptome based analyses were identified, including incomplete reference annotation, exon level expression differences, fragmentation of low coverage transcripts and incorrect reference annotation. Many

Table 1. Summary of *de novo* assembly programs

Evaluation	Bin-packer	Bridger	IDBA-tran	Oases-velvet	SOAPdenovo-Trans	SSP	Trans-ABySS	Trinity
K-mer choice	25	25	25,31,41,51,61,71	25,31,41,51,61,71,meta	25,31,41,51,61,71	25	25,32,42,52,62,meta	25,32
Number of transcripts	7	11	1,3,4,6,9,10	2,5,12,17,19,20,26	8,13,15,18,21,22	24	23,25,27,28,29,30	14,16
Gene and isoform relationship	Yes	Yes	No	Yes	Yes	Yes	No	Yes
Reference base coverage[a]	15	8	12,13,16,17,18,21	10,14,22,23,24,25,27	**1**,**2**,**3**,**4**,11,20	30	7,9,19,26,28,29	**5**,6
Assembly base coverage[a]	21	22	**2**,6,9,10,12,14,	**1**,**3**,**5**,7,11,13,16	**4**,8,15,18,20,23,	30	17,19,25,26,28.29	24,27
Reference gene coverage[a]	18	12	14,15,17,19,21,26	22,23,24,25,28,29,30	**5**,6,9,13,16,20	27	**1**,**2**,**3**,**4**,7,11	8,10
Chimerism[a]	30	28	6,10,11,13,18,19	22,23,24,25,26,27,29	8,9,14,15,16,17	**3**	**1**,**2**,**4**,**5**,7,12,	20,21
Number of recovered genes	8	6	10,11,13,14,16,19	9,20,21,22,23,25,26	**1**,**2**,**3**,7,15,18	30	12,17,24,27,28,29	**4**,**5**
Alignment[b] rate	8	9	15,17,18	11,12,14,16	**5**,7,13	19	**1**,**2**,6,10	**3**,**4**

Numbers indicate the ranking of an assembly in each category (rows); multiple numbers indicate the multiple assemblies made with differing *k-mer* sizes. The five best ranked results for each evaluation are shown in bold.

[a] The number indicates the rank among 30 assemblies. 1 is the best, 30 is the worst.

[b] The number indicates the rank among 19 assemblies 1 is the best, 19 is the worst.

genomes, particularly of less common model organisms or non-model organisms, are less complete and more poorly annotated than that of Arabidopsis. The presence of these effects/concerns in an Arabidopsis study suggest that it would be generally useful to conduct both reference genome and *de novo* transcriptome based analyses, even when a reference genome is available.

Acknowledgements

We thank Dr. Tesfaye Mengiste, Department of Botany and Plant Pathology, Purdue University for providing the RNA-Seq reads used in this study.

Funding

This work was supported by grants from NSF (IOS- 1456594), and with the support of 'Next-Generation BioGreen 21 Program (SSAC, Project No. PJ01137902)' Rural Development Administration, Republic of Korea.

Conflict of Interest: none declared.

References

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bolger,A.M. *et al.* (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

Bradnam,K.R. *et al.* (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience*, **2**, 1–31.

Chang,Z. *et al.* (2015) Bridger: a new framework for *de novo* transcriptome assembly using RNA-seq data. *Genome Biol.*, **16**, 1–10.

Chevreux,B. *et al.* (1999) Genome sequence assembly using trace signals and additional sequence information. *Proc. Ger. Conf. Bioinf.*, **99**, 45–56.

Earl,D. *et al.* (2011) Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.*, **21**, 2224–2241.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, 222–230.

Grabherr,M.G. *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.*, **29**, 644–652.

Lai,Z. *et al.* (2014) MED18 interaction with distinct transcription factors regulates multiple plant functions. *Nat. Commun.*, **5**, 1–14.

Li,B. *et al.* (2014) Evaluation of *de novo* transcriptome assemblies from RNA-Seq data. *Genome Biol.*, **15**, 553.

Li,B. *et al.* (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, **26**, 493–500.

Li,B. and Dewey,C.N. (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, **12**, 323.

Liu,J. *et al.* (2016) BinPacker: packing-based *de novo* transcriptome assembly from RNA-seq data. *PLOS Comput. Biol.*, **12**, e1004772.

Luo,R. *et al.* (2012) SOAPdenovo2: an empirically improved memory-efficient short-read *de novo* assembler. *Gigascience*, **1**, 18.

Monaghan,J. and Li,X. (2010) The HEAT repeat protein ILITYHIA is required for plant immunity. *Plant Cell Physiol.*, **51**, 742–753.

Neil,S.T. and Emrich,S.J. (2013) Assessing *de novo* transcriptome assembly metrics for consistency and utility. *BMC Genomics*, **14**, 465.

Peng,Y. *et al.* (2013) IDBA-tran: a more robust *de novo* de Bruijn graph assembler for transcriptomes with uneven expression levels. *Bioinformatics*, **29**, 326–334.

Robertson,G. *et al.* (2010) *De novo* assembly and analysis of RNA-seq data. *Nat. Methods*, **7**, 909–912.

Robinson,J.T. *et al.* (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Robinson,M.D. and Smyth,G.K. (2007) Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, **23**, 2881–2887.

Safikhani,Z. *et al.* (2013) SSP: An interval integer linear programming for *de novo* transcriptome assembly and isoform discovery of RNA-seq reads. *Genomics*, **102**, 507–514.

Salzberg,S.L. *et al.* (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.*, **22**, 557–567.

Schulz,M.H. *et al.* (2012) Oases: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics*, **28**, 1086–1092.

Simpson,J.T. *et al.* (2009) ABySS: a parallel assembler for short read sequence data. *Genome Res.*, **19**, 1117–1123.

Smith-Unna,R. *et al.* (2015) TransRate: reference free quality assessment of *de novo* transcriptome assemblies. *BioRxiv*, 021626.

Steijger,T. *et al.* (2013) Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods*, **10**, 1177–1184.

Thorvaldsdóttir,H. *et al.* (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief. Bioinf.*, **14**, 178–192.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Trapnell,C. and Salzberg,S.L. (2009) How to map billions of short reads onto genomes. *Nat. Biotechnol.*, **27**, 455–457.

Vijay,N. *et al.* (2013) Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive *in silico* assessment of RNA-seq experiments. *Mol. Ecol.*, **22**, 620–634.

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

Xie,Y. *et al.* (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics*, **30**, 1660–1666.

Yang,Y. and Smith,S.A. (2013) Optimizing *de novo* assembly of short-read RNA-seq data for phylogenomics. *BMC Genomics*, **14**, 328.

Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.

Zhu,Y. *et al.* (2014) CYCLIN-DEPENDENT KINASE8 differentially regulates plant immunity to fungal pathogens through kinase-dependent and -independent functions in Arabidopsis. *Plant Cell*, **26**, 4149–4170.