Rapid report

AtRTD – a comprehensive reference transcript dataset resource for accurate quantification of transcript-specific expression in Arabidopsis thaliana

Author for correspondence: John W. S. Brown
Tel: +44 1382 568777
Email: j.w.s.brown@dundee.ac.uk
Received: 19 March 2015
Accepted: 5 June 2015

Runxuan Zhang1*, Cristiane P. G. Calixto2*, Nikoleta A. Tzioutziou2, Allan B. James3, Craig G. Simpson4, Wenbin Guo1,2, Yamile Marquez5, Maria Kalyna6, Rob Patro7, Eduardo Eyras8,9, Andrea Barta5, Hugh G. Nimmo3 and John W. S. Brown2,4

1Informatics and Computational Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK; 2Plant Sciences Division, College of Life Sciences, University of Dundee, Invergowrie, Dundee, DD2 5DA, UK; 3Institute of Molecular, Cell and Systems Biology, College of Medical, Veterinary and Life Sciences, University of Glasgow, Glasgow, G12 8QQ, UK; 4Cell and Molecular Sciences, The James Hutton Institute, Invergowrie, Dundee, DD2 5DA, UK; 5Max F. Perutz Laboratories, Medical University of Vienna, Dr Bohrgasse 9/3, 1030 Vienna, Austria; 6Department of Applied Genetics and Cell Biology, University of Natural Resources and Life Sciences, Muthgasse 18, 1190 Vienna, Austria; 7Computer Science Department, Stony Brook University, Stony Brook, NY 11794-4400, USA; 8Computational Genomics, Universitat Pompeu Fabra, 08002 Barcelona, Spain; 9Catalan Institution of Research and Advanced Studies (ICREA), 08010 Barcelona, Spain

Summary

RNA-sequencing (RNA-seq) allows global gene expression analysis at the individual transcript level. Accurate quantification of transcript variants generated by alternative splicing (AS) remains a challenge. We have developed a comprehensive, nonredundant Arabidopsis reference transcript dataset (AtRTD) containing over 74,000 transcripts for use with algorithms to quantify AS transcript isoforms in RNA-seq.

The AtRTD was formed by merging transcripts from TAIR10 and novel transcripts identified in an AS discovery project. We have estimated transcript abundance in RNA-seq data using the transcriptome-based alignment-free programmes SAILFISH and SALMON and have validated quantification of splicing ratios from RNA-seq by high resolution reverse transcription polymerase chain reaction (HR RT-PCR).

Good correlations between splicing ratios from RNA-seq and HR RT-PCR were obtained demonstrating the accuracy of abundances calculated for individual transcripts in RNA-seq.

The AtRTD is a resource that will have immediate utility in analysing Arabidopsis RNA-seq data to quantify differential transcript abundance and expression.

Introduction

Alternative splicing (AS) plays a key regulatory role in the growth, development and behaviour of eukaryotic organisms. AS generates multiple transcript isoforms from a gene through variable selection of different splice sites in the precursor mRNA (Black, 2003; Stamm et al., 2005; Nilsen & Graveley, 2010). The two main consequences of AS are an increase in proteome complexity by generation of different protein isoforms often with different functionality (Black, 2003; Stamm et al., 2005; Nilsen & Graveley, 2010) and the regulation of gene expression through degradation of specific transcripts by the nonsense-mediated decay (NMD) pathway (Nicholson & Mühlleman, 2010; Kalyna et al., 2012; Schweingruber et al., 2013).

In higher plants, AS has been implicated in a wide range of developmental and physiological processes (Syed et al., 2012;
Carvalho et al., 2013; Reddy et al., 2013; Staiger & Brown, 2013). The functional importance of AS in plants has been illustrated in, for example, organ development, flowering time, and the circadian clock (Zhang & Mount, 2009; Sanchez et al., 2010; Airoldi & Davies, 2012; James et al., 2012; Jones et al., 2012; Posé et al., 2013), dark-light retrograde signalling from chloroplast to nucleus (Petrillo et al., 2014), and zinc tolerance (Remy et al., 2014). AS can be tissue-specific (Kriebchaumber et al., 2012) or regulated by miRNAs and long noncoding RNAs (Jia & Rock, 2013; Bardou et al., 2014). The estimated frequency of AS in plants has increased significantly over the last 10 years (Syed et al., 2012).

RNA sequencing (RNA-seq) analyses showed that in Arabidopsis at least 61% of intron-containing genes undergo AS (Filichkin et al., 2010; Marquez et al., 2012). RNA-seq permits the genome-wide identification of all transcript isoforms/splicing variants of a gene and the contribution that each transcript makes to expression. Currently transcript annotation in, for example, The Arabidopsis Information Resource (TAIR), is incomplete, and novel transcript variants continue to be identified in RNA-seq studies (Filichkin et al., 2010; Marquez et al., 2012). As the availability of RNA-seq data grows and AS becomes a routine part of expression analysis, the resolution of gene expression studies will increase and will require accurate methods to quantify transcript isoforms in RNA-seq. Resolution of gene expression studies will increase and will require accurate methods to quantify transcript isoforms in RNA-seq.

To construct the AtRTD, transcripts from The Arabidopsis Information Resource version 10 (TAIR10) (Lamesh et al., 2012) and from the AS discovery RNA-seq analysis, which identified c. 50k novel transcripts (Marquez et al., 2012), were merged. The AtRTD was refined and developed through iterative validation with HR RT-PCR data (see the HR RT-PCR subsection). Redundant transcripts were removed using available tools. First, to remove any confusion caused by differences in the coordinates for the same gene models in the two transcript annotation files, the start and end coordinates of the same gene models in TAIR10 and the AS discovery RNA-seq analysis (Marquez et al., 2012) were compared. Where one was contained within another, the smaller one was removed; where models overlapped partially, the coordinates of the gene model in TAIR10 were retained. Second, the GTF file that describes all the transcript isoforms from the AS discovery RNA-seq analysis (Marquez et al., 2012) was converted to GFF format using perl scripts (http://search.umd.edu/ldsa/GBrowse/bin/gff2gff3.pl version 0.1). The GFF transcript annotation files from the two sources described earlier were then sorted, checked and intron coordinates were extracted using GENOME TOOLS version 1.5.3 (Gremmle et al., 2013). Thus, when transcripts had the same intron coordinates but differed only in the length of their 5' or 3' untranslated regions (UTRs), the longest was retained. Similarly, redundant transcripts from single-exon genes with different lengths of 5' or 3' UTRs were removed. Finally, where transcript fragments occurred, probably due to low read coverage, but had the same intron coordinates of the longer or full-length transcripts, the fragments were removed. The FASTA file of the merged, nonredundant transcripts was generated using GFFREAD from the CUFFLINKS tool suite (Trapnell et al., 2010). In addition, 33 novel transcripts of core clock genes identified previously (James et al., 2012) were also added into the final transcript dataset. These transcripts were identified by HR RT-PCR; this illustrates that transcripts identified by other methods can be added to the AtRTD although we expect that the need for such manual curation will reduce as the AtRTD develops further. Thus, AtRTD represents a nonredundant, highly comprehensive reference transcript dataset that can be used with transcript quantification programmes. Here, the recently developed SAILFISH (k-mer = 30 nt) and SALMON programmes (Patro et al., 2014; R. Patro, unpublished) were used in conjunction with the AtRTD to analyse ultra-deep RNA-seq data of

Materials and Methods

Construction of AtRTD

To construct the AtRTD, transcripts from The Arabidopsis Information Resource version 10 (TAIR10) (Lamesh et al., 2012) and from the AS discovery RNA-seq analysis, which identified c. 50k novel transcripts (Marquez et al., 2012), were merged. The AtRTD was refined and developed through iterative validation with HR RT-PCR data (see the HR RT-PCR subsection). Redundant transcripts were removed using available tools. First, to remove any confusion caused by differences in the coordinates for the same gene models in the two transcript annotation files, the start and end coordinates of the same gene models in TAIR10 and the AS discovery RNA-seq analysis (Marquez et al., 2012) were compared. Where one was contained within another, the smaller one was removed; where models overlapped partially, the coordinates of the gene model in TAIR10 were retained. Second, the GTF file that describes all the transcript isoforms from the AS discovery RNA-seq analysis (Marquez et al., 2012) was converted to GFF format using perl scripts (http://search.umd.edu/ldsa/GBrowse/bin/gff2gff3.pl version 0.1). The GFF transcript annotation files from the two sources described earlier were then sorted, checked and intron coordinates were extracted using GENOME TOOLS version 1.5.3 (Gremmle et al., 2013). Thus, when transcripts had the same intron coordinates but differed only in the length of their 5' or 3' untranslated regions (UTRs), the longest was retained. Similarly, redundant transcripts from single-exon genes with different lengths of 5' or 3' UTRs were removed. Finally, where transcript fragments occurred, probably due to low read coverage, but had the same intron coordinates of the longer or full-length transcripts, the fragments were removed. The FASTA file of the merged, nonredundant transcripts was generated using GFFREAD from the CUFFLINKS tool suite (Trapnell et al., 2010). In addition, 33 novel transcripts of core clock genes identified previously (James et al., 2012) were also added into the final transcript dataset. These transcripts were identified by HR RT-PCR; this illustrates that transcripts identified by other methods can be added to the AtRTD although we expect that the need for such manual curation will reduce as the AtRTD develops further. Thus, AtRTD represents a nonredundant, highly comprehensive reference transcript dataset that can be used with transcript quantification programmes. Here, the recently developed SAILFISH (k-mer = 30 nt) and SALMON programmes (Patro et al., 2014; R. Patro, unpublished) were used in conjunction with the AtRTD to analyse ultra-deep RNA-seq data of
5-wk-old Arabidopsis plants grown at 20°C and then transferred to 4°C. Three biological samples were harvested at two time-points: dawn at 20°C (T1) and in the middle of the dark period four days after transfer to the low temperature (T2) for library preparation and sequencing. In total, c. 174 and 191 M 100 bp paired end reads were generated for T1 and T2, respectively.

**HR RT-PCR**

To validate the quantification of transcript isoforms, HR RT-PCR was performed on RNA from the same plant material as RNA-seq. Primer pairs covering 50 AS events in 29 genes, where the upstream primer was end-labelled with a fluorescent tag, were used in RT-PCR reactions with 24 cycles of PCR as described previously (Simpson et al., 2008; James et al., 2012; Kalyna et al., 2012; Marquez et al., 2012) and separated on an Illumina 3700 automatic DNA sequencing machine. The abundance of RT-PCR products was analysed with Genemapper and splicing ratios were calculated from peak areas of each product. We analysed the RNA-seq data from the same regions and used the TPM values of the transcripts to calculate splicing ratios for comparison with those generated by HR RT-PCR, following a similar approach to Alamancos et al. (2015).

**Results and Discussion**

AtRTD was constructed by merging transcripts from TAIR10 and the AS discovery project (Marquez et al., 2012). The latter study used a normalized library for RNA-seq to increase the depth of sequencing and significantly increased the number of transcript isoforms in Arabidopsis despite using 10-d-old seedlings and flower tissue (Marquez et al., 2012). The number of genes and transcripts contained in TAIR10, the AS discovery data and AtRTD are shown in Table 1. The gene number in TAIR10 and AtRTD differed by only 23 genes; the lower gene number in Marquez et al. (2012) reflects the number of expressed genes in two developmental stages. By contrast, the number of transcripts increased from 41 671 in TAIR10 (that also contains redundant transcripts which differ only by different lengths of 5' and 3' UTRs) to 74 216 in the merged and curated AtRTD dataset. The number of transcripts per gene, which reflects transcript isoform complexity, increased from 1.24 in TAIR10, to 2.40 and 2.21 in the AS discovery data and AtRTD, respectively (Table 1). Although the average transcripts per gene values are similar, the number of genes differs significantly (23 905 and 33 625, respectively). The increased transcript complexity in AtRTD compared with TAIR10 is also shown by the number of genes with higher numbers of transcripts (Fig. 1; Supporting Information Table S1). This is illustrated by the reduced number of genes with a single transcript in AtRTD (18 948) compared with TAIR10 (27 717). Thus, AtRTD represents a nonredundant transcript dataset that is highly enriched in AS transcripts. There are 14 499 intron-containing genes with more than one transcript suggesting that 63.22% of intron-containing genes undergo AS. It is, however, important to note that because not all developmental stages or environmental conditions are included, it is likely that AS transcripts are still not completely represented and new versions of AtRTD will be generated as other high quality RNA-seq data becomes available.

To demonstrate the utility of the AtRTD, we have used this set of transcripts to quantify transcripts in RNA-seq data with the Sailfish and Salmon quantification tools. To validate the quantification of the resulting transcript abundances from RNA-seq, the TPM of individual transcripts for the genes used in HR RT-PCR were extracted from the RNA-seq data. Transcript structures were compared to the AS events covered by the primers in HR RT-PCR and used to calculate splicing ratios for each of the AS events in that region. The splicing ratio for this comparison is the percentage of transcripts with a particular AS event expressed as a function of the level of total transcripts (Fig. 2). Figure 2 shows histograms of splicing ratios of two gene/AS examples demonstrating the high degree of similarity between the RNA-seq Sailfish and Salmon outputs and HR RT-PCR.

HEAT SHOCK FACTOR 3 (HSF3) has three different transcripts resulting from AS in a 3'UTR intron: fully spliced, use of alternative 3' splice site which removes 8 nt from exon 3 and intron 2 retention (Fig. 2a). All three are relatively abundant transcripts and show small differences in splicing ratios at the different time-points. RUBISCO ACTIVASE (RCA) undergoes AS in intron 6 (Fig. 2b). An alternative 5' splice site within the intron adds 11 nt and changes the frame of the resulting protein to give different C-terminal sequences. Proteins from these two variants differ in length by 28 amino acids and contain eight or 36

---

**Table 1 Number of Arabidopsis thaliana genes and transcripts in different datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Number of genes</th>
<th>Number of transcripts</th>
<th>Average transcripts per gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAIR10</td>
<td>33 602</td>
<td>41 671&lt;sup&gt;a&lt;/sup&gt;</td>
<td>1.24</td>
</tr>
<tr>
<td>Marquez et al. (2012)</td>
<td>23 905</td>
<td>57 408&lt;sup&gt;b&lt;/sup&gt;</td>
<td>2.40</td>
</tr>
<tr>
<td>AtRTD v3</td>
<td>33 625</td>
<td>74 216&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2.21</td>
</tr>
</tbody>
</table>

Note: TAIR10, The Arabidopsis Information Resource version 10; AtRTD, Arabidopsis reference transcript dataset. <sup>a</sup>Contains redundant transcripts which differ only by different lengths of 5' and 3' UTRs. <sup>b</sup>De novo assembled transcripts defined by splice junctions. <sup>c</sup>Merged, nonredundant transcripts.

---

![Fig. 1 Distribution of the number of transcripts per gene in The Arabidopsis Information Resource version 10 (TAIR10) and the Arabidopsis reference transcript dataset (AtRTD).](image-url)
different amino acids at their C-terminal ends. These examples demonstrate that RNA-seq detects multiple transcripts including transcripts which are present in low abundance (Fig. 2b, I6R), and distinguishes between transcripts differing by a small number of nucleotides (Fig. 2). To directly compare the output from SAILFISH and SALMON with HR RT-PCR, we analysed 50 AS events from 29 genes and three biological replicates of the two time-points (a total of 300 data points). As in some cases HR RT-PCR detected relatively low abundance AS transcripts which were not identified in RNA-seq, for this comparison we calculated a splicing ratio by comparing the abundance of the AS transcript to that of the fully spliced transcript (Fig. 3). For SAILFISH and SALMON, the Pearson’s correlation coefficient was 0.7044 and 0.9051, respectively, and the Spearman’s rank correlation coefficient was 0.712 and 0.907, respectively (Fig. 3).

In this paper we demonstrate that the combination of the AtRTD (a comprehensive nonredundant reference transcript dataset for Arabidopsis) with SAILFISH or SALMON allows accurate estimation of individual transcript abundances. The novel AtRTD resource contains a significantly higher number of transcript isoforms than TAIR10 that is still widely used as a reference to analyse RNA-seq in Arabidopsis. A high degree of correlation between splicing ratios calculated from TPM from RNA-seq data and the HR RT-PCR was observed with SALMON.
outperforming SAILFISH. The HR RT-PCR system was used previously to validate RNA-seq data qualitatively (Marquez et al., 2012) and to monitor changes in AS under different growth conditions and in a range of different mutants (Simpson et al., 2008; Raczynska et al., 2010; James et al., 2012; Jones et al., 2012; Kalyna et al., 2012; Streitner et al., 2012; Petillo et al., 2014) and here we demonstrate its utility in validating quantitative RNA-seq transcript data. RNA-seq is now widely used in transcriptome/expression analyses in plants to examine, for example, abiotic and biotic stress responses, development, light regulation and AS itself (e.g. Zhang et al., 2010; Rühl et al., 2012; Li et al., 2013; Ding et al., 2014; Thatcher et al., 2014; Wu et al., 2014; Mandadi & Scholtzhofer, 2015). The AtRTD is freely available for download at http://ics.hutton.ac.uk/atRTD/ and can now be used to analyse such Arabidopsis RNA-seq data.

Acknowledgements

This research was supported by funding from the Biotechnology and Biological Sciences Research Council (BBSRC) (BB/K006568/1 to J.W.S.B.; BB/K006835/1 to H.G.N.), the Scottish Government Rural and Environment Science and Analytical Services division (RESAS) and by the Austrian Science Fund (FWF) (P26333) to M.K. and (DK W1207) to A.B. The authors acknowledge the European Alternative Splicing Network of Excellence (EURASNET), LSHG-CT-2005-518238 for catalysing important collaborations. The authors thank Janet Laird (University of Glasgow) and Linda Milne (James Hutton Institute) for technical assistance.

References

Rapid report


Supporting Information

Additional supporting information may be found in the online version of this article.

Table S1 Transcript complexity of AtRTD: distribution of the number of transcripts per gene

Please note: Wiley Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the New Phytologist Central Office.