Techniques & Applications

RNA-Seq: revelation of the messengers

Marcel C. Van Verk[1*], Richard Hickman[1*], Corné M.J. Pieterse[1,2], and Saskia C.M. Van Wees[1]

[1] Plant–Microbe Interactions, Department of Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands
[2] Centre for BioSystems Genomics, PO Box 98, 6700 AB Wageningen, The Netherlands

Next-generation RNA-sequencing (RNA-Seq) is rapidly outcompeting microarrays as the technology of choice for whole-transcriptome studies. However, the bioinformatics skills required for RNA-Seq data analysis often pose a significant hurdle for many biologists. Here, we put forward the concepts and considerations that are critical for RNA-Seq data analysis and provide a generic tutorial with example data that outlines the whole pipeline from next-generation sequencing output to quantification of differential gene expression.

Why choose RNA-Seq?

Gene expression analysis is widely used to unravel regulatory mechanisms that control cellular processes in plants, animals, and microbes. Microarrays have been highly instrumental in profiling the global expression of genes, although this hybridization-based technology is largely restricted to known genes and has a limited range of quantification. RNA-Seq extends the possibilities of transcriptome studies to the analysis of previously unidentified genes and of splice variants [1]. Moreover, RNA-Seq offers an unlimited dynamic range of quantification at reduced technical variability. These advantages, coupled with the declining cost of sequencing, make RNA-Seq an increasingly attractive method for whole-genome expression studies in many biological systems, including species with unsequenced genomes.

A typical RNA-Seq experiment starts with mRNA that is subsequently converted into cDNA to form an RNA-Seq library. By sequencing the millions of DNA fragments in the library (known as 'reads') with next-generation sequencing, an accurate measure of the relative abundance of each transcript and splice variants can be obtained. In recent years, a wide array of bioinformatics tools have been developed to process the individual steps required to translate next-generation sequencing output into information on gene expression levels. Here we present a comprehensive description of how to sequentially implement these tools, including critical considerations at each step of a standard plant RNA-Seq workflow (Figure 1) for plants with a reference genome. In addition, a supplementary tutorial (see supplementary material online) provides directions on how to analyze data from a plant RNA-Seq experiment.

Step 1: RNA-Seq library preparation

The quality of starting RNA can heavily influence the success of an RNA-Seq experiment. RNA quality can be assessed using techniques such as the Lab-on-Chip assay (Agilent) that calculates an RNA integrity number (RIN) with a value in the range 1–10 (fully degraded RNA – intact RNA) [2]. Typically, RNA-Seq protocols require 100 ng to 4 μg of the total RNA with a RIN value of at least 8. Subsequently, mRNA is selected, fragmented, and reverse transcribed to generate the RNA-Seq cDNA library. Alternative library preparation protocols have been developed that yield strand specific reads, which facilitates the discrimination between sense and anti-sense RNA. Prior to next-generation sequencing, it is often desirable to introduce unique 6 bp indices to the cDNA fragments of each library. This allows sequencing of multiple samples in the same reaction, as the reads can be traced back to the correct sample. Some sequencing platforms tend to give relatively high error rates at the beginning of a read making this a less suitable position to append the indices. Multiplexing of libraries in multiple reactions provides a cost-effective means of reducing technical variability.

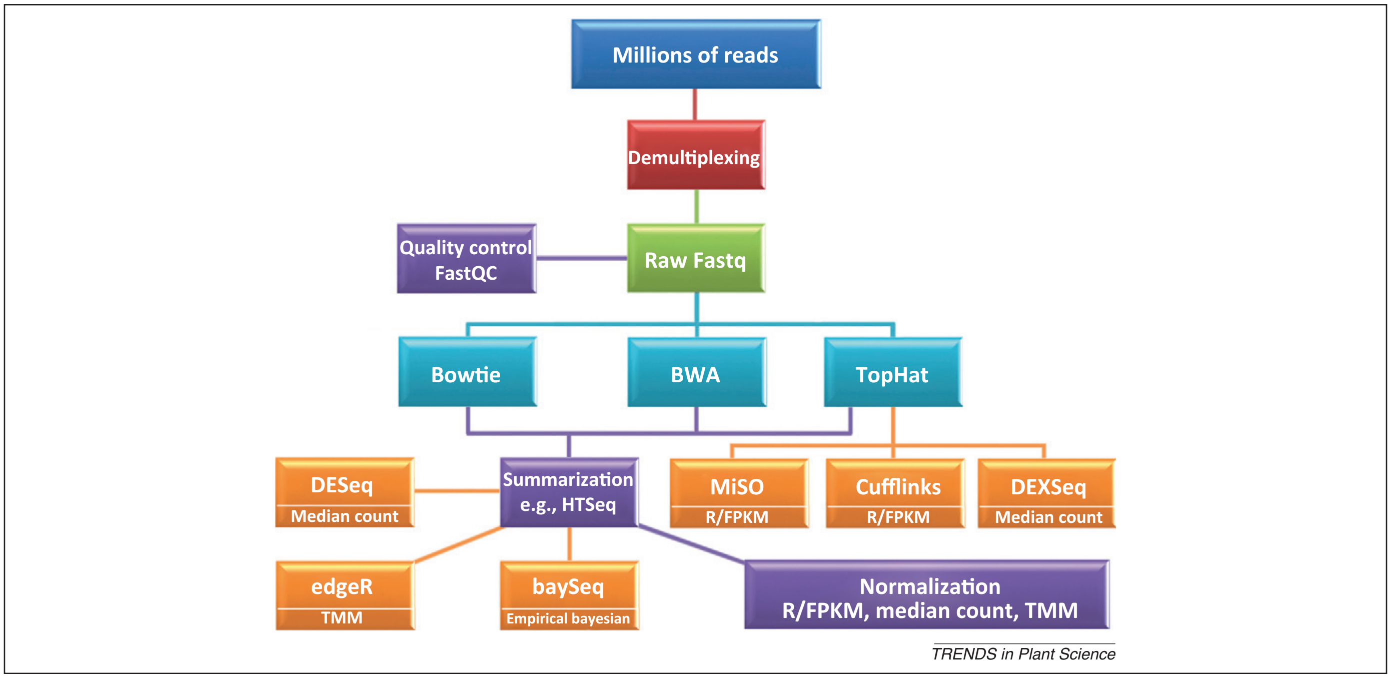
Step 2: Next-generation sequencing

At present, the Illumina HiSeq platform is the most commonly applied next-generation sequencing technology for RNA-Seq, currently yielding up to 3 billion reads per sequencing run. The Illumina HiSeq platform has two sequencing units (flow cells), each providing eight separate sequencing reactions (lanes), which depending on the total read length, takes between 1.5 and 11 days to complete. Although this platform follows a standard sequencing protocol, specific research questions and the organism under study should determine which variation of the RNA-Seq protocol is most appropriate. When profiling gene expression of relatively nonrepetitive genomes, such as that of *Arabidopsis* (*Arabidopsis thaliana*), the most suitable option is to sequence 50 bp from one end of the cDNA fragments (single end). For *Arabidopsis*, a sequencing depth of 50 million reads yields near-saturated coverage of expressed genes (Figure 2a). When studying alternative splicing of mRNA, the sequence information per gene should be increased, which is achieved by sequencing 50 or 100 bp from both ends of the cDNA fragments (paired end). Expression and/or splice variant analysis of large or highly repetitive genomes, such as that of rice (*Oryza sativa*), requires paired-end sequencing reads of 100 or 150 bp to accurately map the reads to the genome.

Step 3: Alignment of RNA-Seq reads

To perform a complete RNA-Seq analysis from reads to differential gene expression, a 64-bit computer, running

Corresponding author: Van Wees, S.C.M. (s.vanwees@uu.nl)
[*] These two authors contributed equally to this work.



TRENDS in Plant Science

Figure 1. Overview of the workflow for analysis of RNA-Seq data. Steps 1 and 2: An average RNA-Seq experiment will yield millions of sequence reads in a Fastq file. Indexing of RNA-Seq libraries with 6 bp barcodes allows for sequencing of multiple samples in the same sequencing reaction. The indexed reads need to be demultiplexed in order to assign each read to the corresponding sample. Step 3: The quality of the obtained raw Fastq files can be checked with FastQC and adjusted when needed. Next, the reads are mapped against a reference transcriptome and/or genome; Bowtie or BWA can be used for direct alignments, and TopHat also for gapped and/or spliced alignments. Step 4: If differences in splice variants are not of interest, the total number of genome-mapped reads per gene can be summarized based upon existing gene models using HTSeq. Step 5: For downstream applications of the summarized genome and/or transcriptome-mapped reads, the count data need to be normalized according to the RPKM/FPKM, median count or TMM standard. Step 6: The summarized genome and/or transcriptome-mapped count data can be tested for significant differences in transcript abundance between samples, using the popular tools DESeq, edgeR, and baySeq. To test for differences in alternative splicing of mRNA, the splice-aligned data derived from TopHat are subjected to MiSO, Cufflinks, or DEXSeq analysis. Light blue boxes, read alignment tools; purple boxes, packages that organize the data; orange boxes, test tools for differential expression (including the normalization method incorporated in the differential expression tool). Abbreviations: RNA-Seq, next-generation RNA-sequencing; RPKM/FPKM, reads or paired-end fragments per kilobase of exon model per million mapped reads; TMM, trimmed mean of M-values.

a UNIX-like operating system such as Mac OS X or Linux, and a minimum of 4 Gb RAM (preferably 16 Gb), is recommended. After obtaining the RNA-Seq reads, their quality should be evaluated with a tool such as FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/), which flags any potential abnormalities that may have occurred during library preparation or the sequencing reaction. Subsequently, reads are aligned to a reference genome or transcriptome. Alignment to the genome is slower than to the transcriptome because mapping of intron-spanning reads requires greater computer processing time. However, alignment to the genome can provide important information on the presence of different mRNA splice variants. For transcriptome alignments, software tools such as Bowtie [3] and BWA [4] are commonly used, whereas the splice aligner TopHat [5] is a popular choice for mapping reads to an annotated reference genome (Figure 1; Sections 4 and 5 of the tutorial in the supplementary material online). For an extensive review of available read alignment tools, see [6]. When using splice aligners to map reads to plant genomes, it is important to note that default parameters have often been optimized for read alignments to mammalian genomes. This is particularly relevant for parameters associated with intron size, because default values are often considerably larger than should be used for plants (Section 5 of the tutorial in the supplementary material online).

The multiread problem

The majority of reads will align to a unique location within the genome or transcriptome, whereas a significant proportion of reads can map to more than one location. These so-called multireads originate from gene duplications, repetitive regions, or overlapping genes. Assignment of multireads to specific genes is problematic and can lead to mistakes in expression analysis, especially as read aligners often randomly allocate multireads to the different mapped locations or distribute equal fractions of a multiread to each location. As such, the regions within a gene that are covered by multireads have a different (higher or lower) per base coverage compared with regions that are covered by unique reads (Figure 2b,c). Hence, inclusion of multireads in RNA-Seq analysis will result in an under- or over-estimation of the expression level of this gene. Multireads are usually flagged in the alignment output files and therefore can be easily identified and omitted from further analysis. However, discarding reads does lead to loss of information and underestimation of the expression levels of the affected genes. This is not a problem when the expression of the same gene is compared across different samples, because the same inaccuracies will apply to all samples, although it will reduce sensitivity when testing for differential expression. Considering this problem, alternative strategies have been developed that probabilistically allocate multireads between affected genes, minimizing loss of quantification power [7–9].

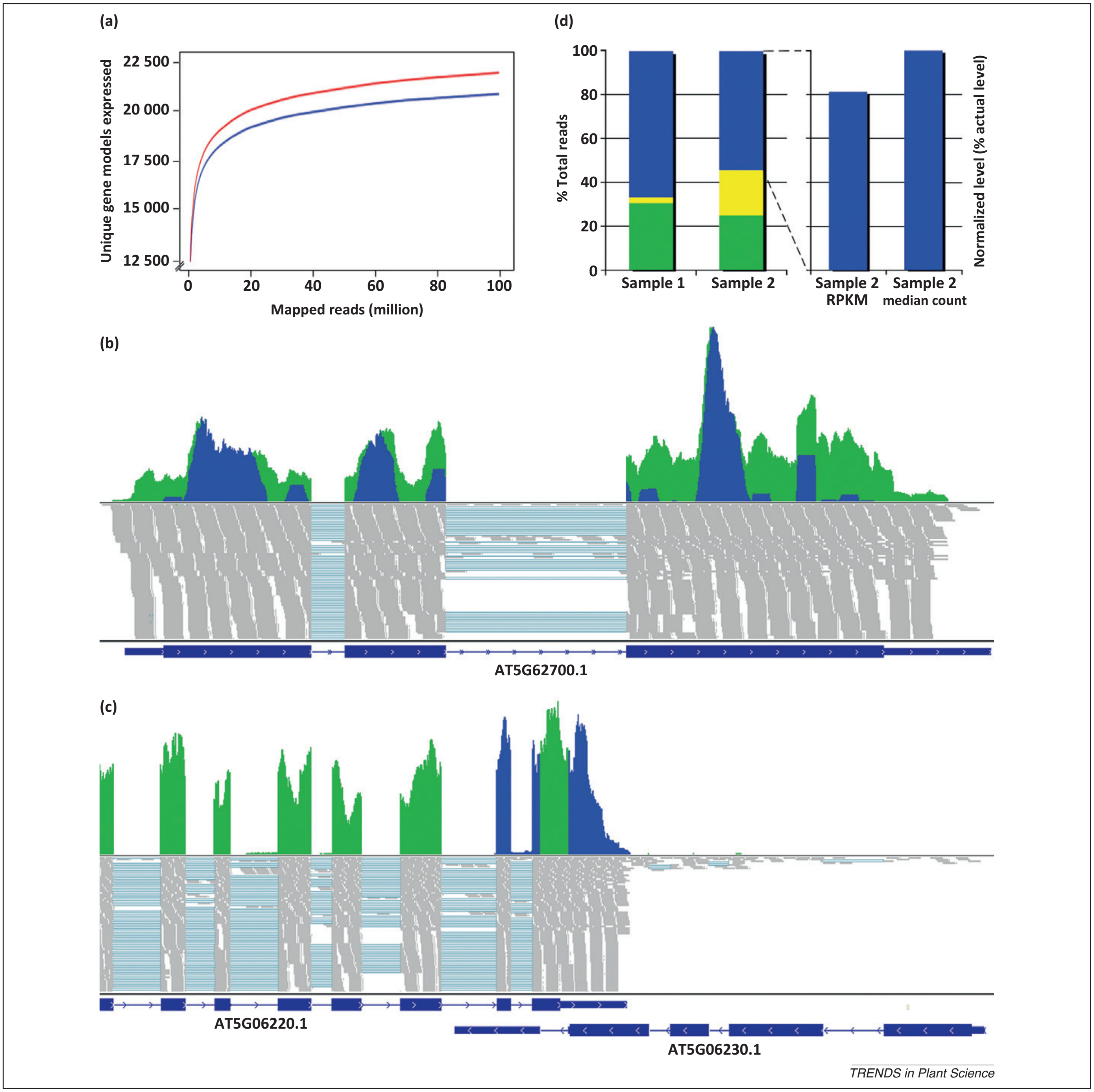


Figure 2. Considerations for next-generation RNA-sequencing (RNA-Seq) data analysis. **(a)** Expression coverage versus total number of mapped reads. Reads (50 bp single end) from *Arabidopsis* leaf tissue were mapped to the genome. Read counts were summarized over annotated *Arabidopsis* gene models, with counts of four or more per gene defining expression when considering (red line) or omitting (blue line) multireads. Approximately 5% of aligned reads are multireads. In both scenarios, the number of genes that can be detected reaches a plateau at approximately 50 million reads. When multireads are considered 21 139 genes are detected. When multireads are omitted 20 162 genes are detected. In the latter case the 10th percentile gene has 19 reads. **(b)** Multireads can cause inaccurate expression estimates. Shown is the read coverage for the *TUBULIN BETA CHAIN 3* gene (AT5G62700). Gray lines indicate direct read alignments to the reference *Arabidopsis* genome and light blue lines indicate intron-spanning alignments of reads. The histograms show the combined contribution of unique reads and multireads that cover each base of this gene. Visualization was performed by the Integrative Genome Viewer (IGV) [18]. Per base coverage that originates from reads that uniquely map to this gene region are indicated in green, whereas in blue the per base coverage that is derived from alignments of reads that map to multiple regions in the genome (multireads) are indicated. In this example, inclusion of multireads would artificially inflate read coverage. **(c)** Overlap multireads can cause inaccurate expression estimates. Shown is read coverage for two genes that share a genomic region by overlap of their 3′ ends on opposite strands. Reads that map to the region of overlap may originate from either transcript. In this example, the expression of the gene on the plus strand is markedly higher than of the gene on the minus strand, indicating that the vast majority of reads that align to the region of overlap originate from the gene on the plus strand. Random allocation of multireads would result in a severe overestimation of expression of the gene on the minus strand. **(d)** Normalization. The histogram in the left panel indicates the percentage of reads assigned to the 75 highest expressed genes (green + yellow) and to the remaining (relatively lower expressed) genes (blue) in two samples. There are no differences in read counts for every single gene in each sample except for only one of the highly expressed genes (yellow), which is 10-fold higher expressed in sample 2 compared with sample 1; this results in a higher contribution of the highly expressed genes to the total read count in sample 2. The choice of normalization method (explained in the main text) affects the analysis of differential expression levels between samples, as visualized in the histogram in the right panel. Normalized read counts for the individual genes in the relatively lower expressed gene set (blue) are correctly reflecting the actual counts when the median count normalization method is used. When using the RPKM normalization method, the normalized data are skewed due to the high expression difference of one gene (yellow), causing an underestimation of the expression level of all other genes of sample 2 to 80% of their actual expression levels, implying that all genes (in *Arabidopsis* over 30 000) would be falsely downregulated in sample 2 by 20% compared with sample 1. Abbreviation: RPKM, reads per kilobase of exon model per million mapped reads.

Step 4: Quantification of gene expression

Following alignment of RNA-Seq reads, the data need to be translated into a quantitative measure of gene expression, which can be achieved by counting the number of reads that map to each gene. Reads that are aligned to annotated transcripts comprising the transcriptome can be summarized relatively easily (an example is given in Section 6.1 of the tutorial in the supplementary material online). Reads that are aligned to annotated coding regions of the genome can be summarized using tools such as those available in the HTSeq package (Figure 1; Section 6 of the tutorial in the supplementary material online) (http://www-huber.embl.de/users/anders/HTSeq/). Estimating the expression levels of individual splice variants of a gene is more complex, as they typically share a set of exons, and thus only a minor fraction of the reads will align uniquely to the distinct regions of a particular splice variant. Methods such as Cufflinks and MiSO use statistical models to estimate the proportion of reads that can be assigned to individual splice variants [9,10]. In some cases, determination of the splice variants of a gene present within a sample could require very deep sequencing or simply be impossible with current technology.

Step 5: Normalization of RNA-Seq count data

Following quantification of expression levels, a common objective is to identify genes that are differentially expressed between different conditions. To enable an accurate comparison of expression levels between different samples, the count data must first be normalized. Two important types of sequencing bias need to be considered when normalizing count data: within-sample bias, which is primarily caused by differences in transcript length, and between-sample bias, which results mainly from differences in sequencing depth. Differential expression software packages (mentioned in Step 6) often implement their preferred method of normalization, as indicated in Figure 1.

One of the most widely used normalization methods is RPKM/FPKM (reads or paired-end fragments per kilobase of exon model per million mapped reads) [7], which adjusts raw counts to the total gene length and the number of reads mapped within a sample, thereby simultaneously performing between- and within-sample normalization. Although RPKM/FPKM is a popular normalization strategy, it has been demonstrated to bias subsequent calls for differential expression in favor of longer transcripts [11]. However, RPKM/FPKM values are useful for analyzing differences in the abundance of alternative splice variants between samples, as correction for the length of each splice variant is essential for this type of analysis.

When comparing the expression of the same gene between different samples, only between-sample normalization is necessary. In this instance, the simplest approach is to adjust counts according to the total number of reads mapped in the sample (as in RPKM/FPKM normalization). Although intuitive, this approach introduces inaccuracies if there is a minor fraction of genes that comprise the majority of the total read count (Figure 2d). Consequently, normalization methods that are more robust to extreme values, such as the median count ratio [12] and trimmed mean of M-values (TMM) [13], have been developed and are recommended over methods based on the total read count [14].

Step 6: Differential gene expression

Many statistical methods have been developed to test for differential gene expression with microarray data. However, these methods are not appropriate for use with RNA-Seq data, due to the different nature of the data; microarray data are continuous, because they are based on fluorescence intensity, in contrast to RNA-Seq data that are discrete as they are derived from read counts.

These differences have led to the development of software tools to statistically test for differential gene expression between samples that are analyzed by RNA-Seq. When the total number of reads per gene has been summarized, methods such as DESeq, edgeR, and baySeq [12,13,15] that model count data using a negative binomial distribution have been shown to perform well (Figure 1) [14,16]. Comparing the relative abundance of different mRNA splice variants is recognized as a more challenging task, due to the aforementioned problems with estimating the relative expression of each splice variant (Step 4). Methods such as MiSO, DEXSeq, and CuffDiff (implemented in the Cufflinks package) can handle testing for differences in alternative splicing between two samples (Figure 1; Sections 7 and 8 of the tutorial in the supplementary material online) [9,10,17].

The future of RNA-Seq

The past few years witnessed a rise in the use of RNA-Seq for genome-wide transcriptome studies and will soon become standard practice. Accordingly, many methods have been developed to analyze the RNA-Seq data, but improvements are needed to deal with problems associated with multireads and estimating the abundance of splice variants. Upcoming third-generation sequencing, also known as single-molecule sequencing, has the potential to sequence complete transcripts at once. If indeed transcripts can be sequenced without the need of fragmentation of the cDNA and without PCR amplification, this would drastically reduce computing time and improve significantly the correct assignment of sequencing reads. This opens possibilities for whole-genome expression profiling at an unprecedented level of detail.

Acknowledgements

This work was supported by Vidi grant no. 11281 of the Technology Foundation STW, and by ERC Advanced Investigators Grant no. 269072 of the European Research Council.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tplants.2013.02.001.

References

1 Wang, Z. *et al.* (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63
2 Schroeder, A. *et al.* (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Mol. Biol.* 7, 3
3 Langmead, B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25

4 Li, H. *et al.* (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760

5 Trapnell, C. *et al.* (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105–1111

6 Froncesca, N.A. *et al.* (2012) Tools for mapping high-throughput sequencing data. *Bioinformatics* 28, 3169–3177

7 Mortazavi, A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* 5, 621–628

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

9 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

10 Katz, Y. *et al.* (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat. Methods* 7, 1009–1015

11 Oshlack, A. and Wakefield, M. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct.* 4, 14

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

13 Robinson, M. *et al.* (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26, 139–140

14 Dillies, M. *et al.* (2012) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinf.* (http://dx.doi.org/10.1093/bib/bbs046)

15 Hardcastle, T. and Kelly, K. (2010) baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinf.* 11, 422

16 Kvam, V. *et al.* (2012) A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am. J. Bot.* 99, 248–256

17 Andres, S. *et al.* (2012) Detecting differential usage of exons from RNA-Seq data. *Genome Res.* 22, 2008–2017

18 Robinson, J.T. *et al.* (2011) Integrative Genome Viewer. *Nat. Biotech.* 29, 24–26

http://dx.doi.org/10.1016/j.tplants.2013.02.001 Trends in Plant Science xx (2013) 1–5