Genome **Biology**

REVIEW

From RNA-seq reads to differential expression results

Alicia Oshlack[1]*, Mark D Robinson[1,2] and Matthew D Young[1]

Abstract

Many methods and tools are available for preprocessing high-throughput RNA sequencing data and detecting differential expression.

High-throughput sequencing technologies are now in common use in biology. These technologies produce millions of short sequence reads and are routinely being applied to genomes, epigenomes and transcriptomes. Sequencing steady-state RNA in a sample, known as RNA-seq, is free from many of the limitations of previous technologies, such as the dependence on prior knowledge of the organism, as required for microarrays and PCR (see Box 1: Comparisons of microarrays and sequencing for gene expression analysis). In addition, RNA-seq promises to unravel previously inaccessible complexities in the transcriptome, such as allele-specific expression and novel promoters and isoforms [1-4]. However, the datasets produced are large and complex and interpretation is not straightforward. As with any high-throughput technology, analysis methodology is critical to interpreting the data, and RNA-seq analysis procedures are continuing to evolve. Therefore, it is timely to review currently available data analysis methods and comment on future research directions.

Making sense of RNA-seq data depends on the scientific question of interest. For example, determining differences in allele-specific expression requires accurate determination of the prevalence of transcribed single nucleotide polymorphisms (SNPs) [5]. Alternatively, fusion genes or aberrations in cancer samples can be detected by finding novel transcripts in RNA-seq data [6,7]. In the past year, several methods have emerged that use RNA-seq data for abundance estimation [8,9],

detection of alternative splicing [10-12], RNA editing [13] and novel transcripts [11,14]. However, the primary objective of many biological studies is gene expression profiling between samples. Thus, in this review we focus on the methodologies available to detect differences in gene level expression between samples. This sort of analysis is particularly relevant for controlled experiments comparing expression in wild-type and mutant strains of the same tissue, comparing treated versus untreated cells, cancer versus normal, and so on. For example, comparison of expression changes between the cultured pathogen *Acinetobacter baumannii* and the pathogen grown in the presence of ethanol - which is known to increase virulence - revealed 49 differentially expressed genes belonging to a range of functional categories [15]. Here we outline the processing pipeline used for detecting differential expression (DE) in RNA-seq and examine the available methods and open-source software tools to perform the analysis. We also highlight several areas that require further research.
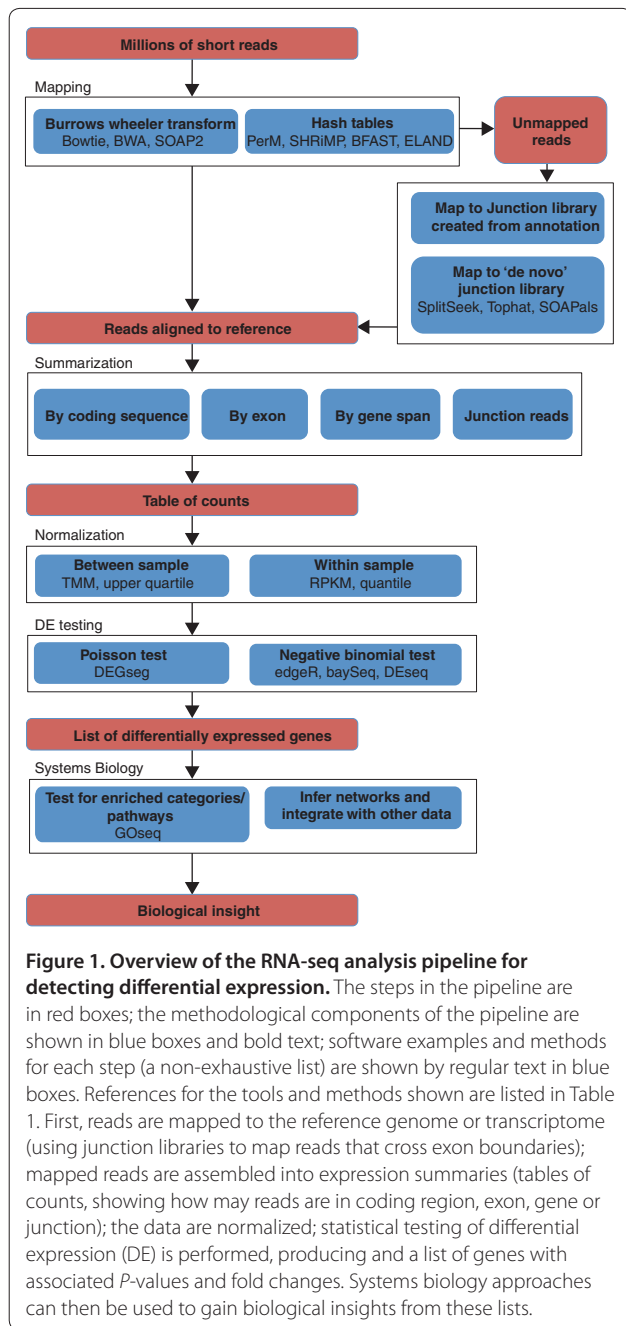
Most RNA-seq experiments take a sample of purified RNA, shear it, convert it to cDNA and sequence on a high-throughput platform, such as the Illumina GA/ HiSeq, SOLiD or Roche 454 [16]. This process generates millions of short (25 to 300 bp) reads taken from one end of the cDNA fragments. A common variant on this process is to generate short reads from both ends of each cDNA fragment, known as 'paired-end' reads. The platforms differ substantially in their chemistry and processing steps, but regardless of the precise details, the raw data consist of a long list of short sequences with associated quality scores; these form the entry point for this review.

An overview of the typical RNA-seq pipeline for DE analysis is outlined in Figure 1. First, reads are mapped to the genome or transcriptome. Second, mapped reads for each sample are assembled into gene-level, exon-level or transcript-level expression summaries, depending on the aims of the experiment. Next, the summarized data are normalized in concert with the statistical testing of DE, leading to a ranked list of genes with associated *P*-values and fold changes. Finally, biological insight from these lists can be gained by performing systems biology approaches, similar to those performed on microarray

Figure 1. Overview of the RNA-seq analysis pipeline for detecting differential expression. The steps in the pipeline are in red boxes; the methodological components of the pipeline are shown in blue boxes and bold text; software examples and methods for each step (a non-exhaustive list) are shown by regular text in blue boxes. References for the tools and methods shown are listed in Table 1. First, reads are mapped to the reference genome or transcriptome (using junction libraries to map reads that cross exon boundaries); mapped reads are assembled into expression summaries (tables of counts, showing how may reads are in coding region, exon, gene or junction); the data are normalized; statistical testing of differential expression (DE) is performed, producing and a list of genes with associated *P*-values and fold changes. Systems biology approaches can then be used to gain biological insights from these lists.

experiments. We critique below the currently available methodologies for each of these steps for RNA-seq data analysis. Rather than providing a complete list of all available tools, we focus on examples of commonly used open-source software that illustrate the methodology (Table 1). For a complete list of RNA-seq analysis software, see [17,18].

Mapping

To use RNA-seq data to compare expression between samples, it is necessary to turn millions of short reads

into a quantification of expression. The first step in this procedure is the read mapping or alignment. At its simplest, the task of mapping is to find the unique location where a short read is identical to the reference. However, in reality the reference is never a perfect representation of the actual biological source of RNA being sequenced. In addition to sample-specific attributes such as SNPs and indels (insertions or deletions), there is also the consideration that the reads arise from a spliced transcriptome rather than a genome. Furthermore, short reads can sometimes align perfectly to multiple locations and can contain sequencing errors that have to be accounted for. Therefore, the real task is to find the location where each short read best matches the reference, while allowing for errors and structural variation.

Although research into how best to align reads to a reference is ongoing, all solutions by necessity involve some compromise between the computational requirements of the algorithm and the fuzziness allowed in matching to the reference. Almost all short read aligners use a strategy of a first pass 'heuristic' match, which quickly finds a reduced list of possible locations, followed by thorough evaluation of all candidate alignments by a complex 'local alignment' algorithm. Without this initial heuristic search to reduce the number of potential alignment locations, performing local alignment of millions of short reads would be computationally impossible on current hardware.

Current aligners enable fast heuristic matching by using either hash tables [19-22] or the Burrows Wheeler transform (BWT) [23-25]. Hash-table aligners have the advantage of being easily extendable to detect complicated differences between read and reference, at the cost of ever increasing computational requirements. Alternatively, BWT-based aligners can map reads that closely match the reference very efficiently but are prohibitively slow once more complex misalignments are considered. A detailed explanation of these techniques is beyond the scope of this review, but can be found in [23,26-30].

Aligners also differ in how they handle 'multimaps' (reads that map equally well to several locations). Most aligners either discard multimaps [25], allocate them randomly [29] or allocate them on the basis of an estimate of local coverage [31,32], although a statistical method incorporating alignment scores has also been proposed [33]. Paired-end reads reduce the problem of multimapping, as both ends of the cDNA fragment from which the short reads were generated should map nearby on the transcriptome, allowing the ambiguity of multimaps to be resolved in most circumstances.

When considering reads from genomic DNA, mapping to a relevant reference genome is all that is needed. However, RNA-seq is sequencing fragments of the transcriptome. This difference is dealt with in several ways. Given

Table 1. Software methods and tools for differential expression analysis of RNA-seq

Analysis step	Method	Implementation	References
Mapping	General aligner	GMAP/GSNAP	[91]
		BFAST	[20]
		BOWTIE	[25]
		CloudBurst	[92]
		GNUmap	[93]
		MAQ/BWA	[23]
		PerM	[19]
		RazerS	[94]
		Mrfast/mrsfast	[22]
		SOAP/SOAP2	[24,95]
		SHRiMP	[21]
	De novo annotator	QPALMA/GenomeMapper/PALMapper	[37]
		SpliceMap	[96]
		SOAPals	[95]
		G-Mo.R-Se	[97]
		TopHat	[40]
		SplitSeek	[36]
	De novo transcript assembler	Oases	[98]
		MIRA	[99]
Summarization	Isoform-based	Cufflinks	[11]
		ALEXA-seq	[10]
	Gene-based	Count exons only	For example, [34,45]
		Exon junction libraries	[34,44]
Normalization	Library size		For example, [34]
	RPKM	ERANGE	[32]
	TMM	edgeR	[48]
	Upper quartile	Myrna	[45,47]
Differential expression	Poisson GLM	DEGseq	[100]
		Myrna	[47]
	Negative binomial	edgeR	[57]
		DESeq	[46]
		baySeq	[58]
Systems biology	Gene Ontology analysis	GOseq	[68]

Abbreviations: GLM, generalized linear model; RPKM, reads per kilobase of exon model per million mapped reads; TMM, trimmed mean of M-values.

that the transcriptome is 'built from' the genome, the most commonly used approach (at least initially) is to use the genome itself as the reference. This has the benefit of being easy and not biased towards any known annotation. However, reads that span exon boundaries will not map to this reference. Thus, using the genome as a reference will give greater coverage (at the same true expression level) to transcripts with fewer exons, as they will contain fewer exon junctions. Longer reads are more likely to cross exon boundaries, thus causing the fraction of junction reads to increase [2].

In order to account for junction reads, it is common practice to build exon junction libraries in which reference sequences are constructed using boundaries between annotated exons [2,32,34,35]. To map reads that cross exon boundaries without relying on existing annotations, it is possible to use the dataset itself to detect splice junctions *de novo* [36-41]. Another option is the *de novo* assembly of the transcriptome, for use as a reference, using genome assembly tools [42,43]. All *de novo* methods can identify novel transcripts and may be the only option for organisms for which no genomic reference or annotation

is available. However, *de novo* methods are computationally intensive and may require long, paired-end reads and high levels of coverage to work reliably. For example, Trapnell *et al.* [11] used over 430 million paired-end reads for *de novo* assembly of the mouse myoblast transcriptome in order to quantify expression during cell differentiation.

A commonly used approach for transcriptome mapping is to progressively increase the complexity of the mapping strategy to handle the unaligned reads [44]. For example, in a large study investigating expression variation in 69 Nigerian HapMap samples, Pickrell *et al.* [35] found that for 46 bp Illumina reads, 87% mapped to the reference genome with two mismatches using MAQ (a hash-table-based aligner) [23]. An additional 7% could be mapped to an exon-exon junction library, constructed from all possible combinations of Ensembl exons. The remaining unmapped reads were examined for evidence of the sequencer having erroneously sequenced the poly(A) tail. If a read began or ended with at least four As or Ts, these bases were trimmed and the rest of the read was mapped to the reference, resulting in a further 0.005% of reads being mapped. This large dataset enabled the annotation of over 100 new exons and identified more than a thousand genes in which genetic variation influences overall expression levels or splicing. This would not have been possible without a method for handling reads that cross exon boundaries.

Summarizing mapped reads

Having obtained genomic locations for as many reads as possible, the next task is to summarize and aggregate reads over some biologically meaningful unit, such as exons, transcripts or genes. The simplest and most common approach counts the number of reads overlapping the exons in a gene (for example, [32,34,45]). However, a significant proportion of reads map to genomic regions outside annotated exons, even in well-annotated organisms, such as mouse and human. For example, Pickrell *et al.* [35] found that about 15% of mapped reads were located outside annotated exons for their Nigerian HapMap samples and these extra-exonic reads were more likely to be cell-type-specific exons. Similarly, Figure 2a shows an example of transcription occurring outside annotated exons in the RNA-binding protein 39 (*RBM39*) gene in LNCaP prostate cancer cells. Reads from other normal tissue cell types are more limited to known exons, but also show evidence for transcription outside of known exons.

One alternative summarization is to include reads along the whole length of the gene and thereby incorporate reads from 'introns'. This will include unannotated exons in the summary and account for poorly annotated or variable exon boundaries. However, including introns

might also capture overlapping transcripts, which share a genomic location but originate from different genes. There are many other possible variations that could be used for summarization, such as including only reads that map to coding sequence or summarizing from *de novo* predicted exons [40]. Junction reads can also be added into the gene summary count or be used to model the abundance of splicing isoforms [11]. These different possibilities are illustrated schematically in Figure 2b. With these options, the choice of summarization has the potential to change the count for each gene as substantially as, or more substantially than, the choice of mapping strategy. Despite this, little research has been carried out on which summarization method is the most appropriate for DE detection.

Normalization

Normalization enables accurate comparisons of expression levels between and within samples [2,32,34]. It has been shown that normalization is an essential step in the analysis of DE from RNA-seq data [45-48]. Normalization methods differ for between- and within-library comparisons.

Within-library normalization allows quantification of expression levels of each gene relative to other genes in the sample. Because longer transcripts have higher read counts (at the same expression level), a common method for within-library normalization is to divide the summarized counts by the length of the gene [32,34]. The widely used RPKM (reads per kilobase of exon model per million mapped reads) accounts for both library size and gene length effects in within-sample comparisons. To validate this approach, Mortazavi *et al.* [32] introduced several *Arabidopsis* RNAs into their mouse tissue samples, across a range of gene lengths and expression levels. These non-native RNAs are known as 'spike-ins' and demonstrated that RPKM gives accurate comparisons of expression levels between genes. However, it has been shown that read coverage along expressed transcripts can be non-uniform because of sequence content [49] and RNA preparation methods, such as random hexamer priming [50]. Incorporating this understanding into the within-library normalization method may improve the ability to compare expression levels. Using RNA-seq data to estimate the absolute number of transcripts in a sample is possible, but it requires RNA standards and additional information, such as the total number of cells from which RNA is extracted and RNA preparation yields [32].

When testing individual genes for DE between samples, technical biases, such as gene length and nucleotide composition, will mainly cancel out because the underlying sequence used for summarization is the same between samples. However, between-sample normalization is still essential for comparing counts from different libraries
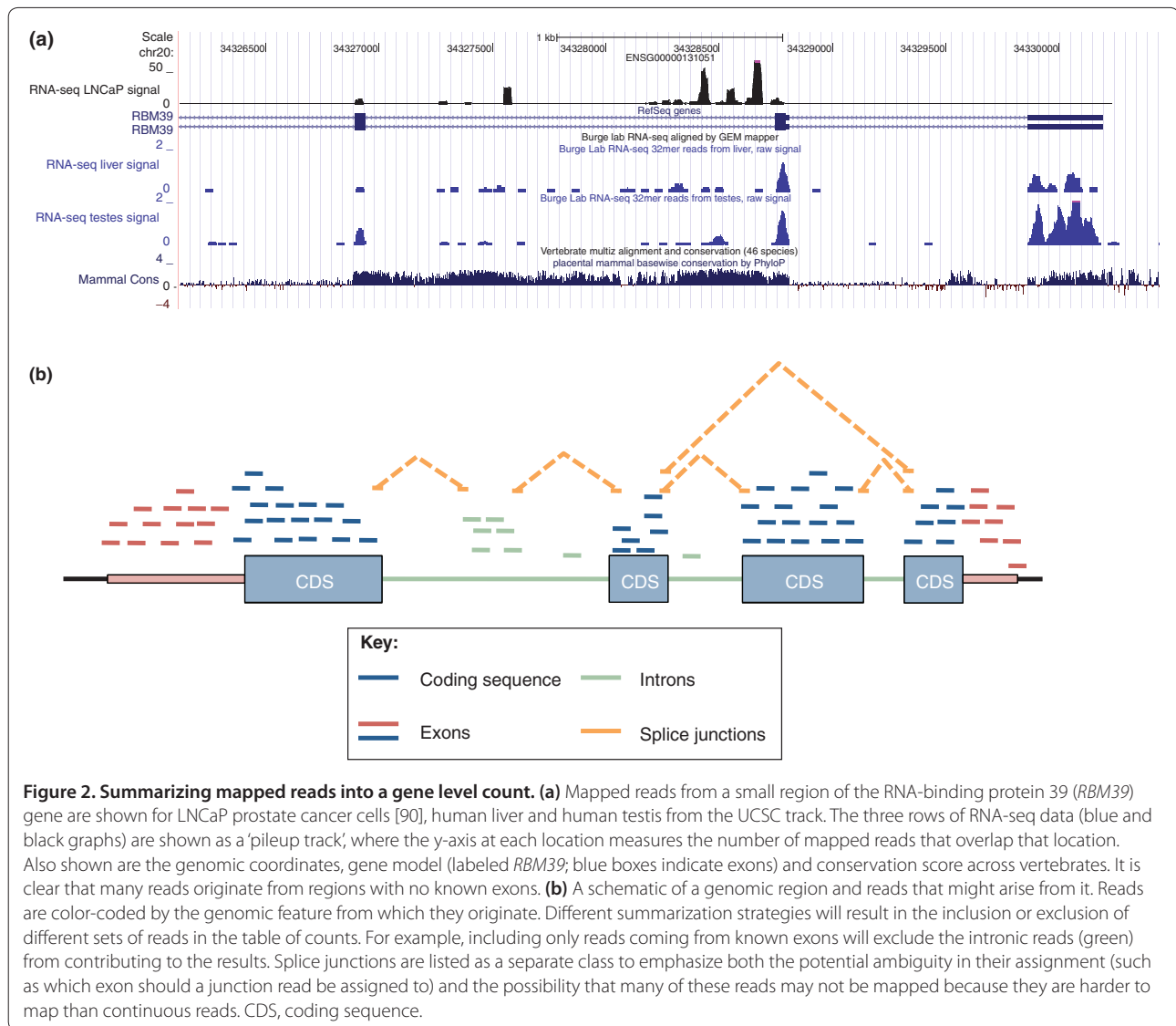
Figure 2. Summarizing mapped reads into a gene level count. (a) Mapped reads from a small region of the RNA-binding protein 39 (*RBM39*) gene are shown for LNCaP prostate cancer cells [90], human liver and human testis from the UCSC track. The three rows of RNA-seq data (blue and black graphs) are shown as a 'pileup track', where the y-axis at each location measures the number of mapped reads that overlap that location. Also shown are the genomic coordinates, gene model (labeled *RBM39*; blue boxes indicate exons) and conservation score across vertebrates. It is clear that many reads originate from regions with no known exons. **(b)** A schematic of a genomic region and reads that might arise from it. Reads are color-coded by the genomic feature from which they originate. Different summarization strategies will result in the inclusion or exclusion of different sets of reads in the table of counts. For example, including only reads coming from known exons will exclude the intronic reads (green) from contributing to the results. Splice junctions are listed as a separate class to emphasize both the potential ambiguity in their assignment (such as which exon should a junction read be assigned to) and the possibility that many of these reads may not be mapped because they are harder to map than continuous reads. CDS, coding sequence.

relative to each other. The simplest and most commonly used normalization adjusts by the total number of reads in the library [34,51], accounting for the fact that more reads will be assigned to each gene if a sample is sequenced to a greater depth. However, it has been shown that more sophisticated normalization is required to account for composition effects [48], or for the fact that a small number of highly expressed genes can consume a significant amount of the total sequence [45]. To account for these features, scaling factors can be estimated from the data and used within the statistical models that test for DE [45,46,48]. Scaling factors have the advantage that the raw count data are preserved for subsequent analysis. Alternatively, quantile normalization and a method using matching power law distributions [52,53] have also been proposed for between-sample normalization of RNA-seq. The non-linearity of both of

these transformations removes the count nature of the data, making it unclear how to appropriately test for DE. So far, quantile normalization does not seem to improve DE detection to the same extent as an appropriate scaling factor [45] and it is not clear that the power law distribution applies to all datasets [48].

Differential expression

The goal of a DE analysis is to highlight genes that have changed significantly in abundance across experimental conditions. In general, this means taking a table of summarized count data for each library and performing statistical testing between samples of interest.

Many methods have been developed for the analysis of differential expression using microarray data. However, RNA-seq gives a discrete measurement for each gene whereas microarray intensities have a continuous intensity

distribution. Although microarray intensities are typically log-transformed and analyzed as normally distributed random variables, transformation of count data is not well approximated by continuous distributions, especially in the lower count range and for small samples. Therefore, statistical models appropriate for count data are vital to extracting the most information from RNA-seq data.

In general, the Poisson distribution forms the basis for modeling RNA-seq count data. In an early RNA-seq study using a single source of RNA, sequenced on multiple lanes of an Illumina GA sequencer, goodness-of-fit statistics suggested that the distribution of counts across lanes for the majority of genes was indeed Poisson distributed [34]. This has been independently confirmed using a technical experiment [45] and software tools are readily available to perform these analyses [54]. However, biological variability is not captured well by the Poisson assumption [47,51]. Hence, Poisson-based analyses for datasets with biological replicates will be prone to high false positive rates resulting from the underestimation of sampling error [46,47,55]. Despite the low background and high sensitivity of the RNA-seq platform, designing experiments with biological replication is still critical for identifying changes in RNA abundance that generalize to the population being sampled. Design of RNA-seq experiments in general, including the fundamental considerations of blocking, randomization and replication, has recently been discussed in depth [56].

In order to account for biological variability, methods that have been developed for serial analysis of gene expression (SAGE) data have recently been applied to RNA-seq data [57]. The major difference between SAGE and RNA-seq data is the scale of the datasets. To account for biological variability, the negative binomial distribution has been used as a natural extension of the Poisson distribution, requiring an additional dispersion parameter to be estimated. A few variations of negative-binomial-based DE analysis of count data have emerged, including common dispersion models [55], sharing information over all genes using weighted likelihood [51], empirical estimation of the mean-variance relationship [46] and an empirical Bayesian implementation using equivalence classes [58]. Extensions to the Poisson model to include overdispersion have also been proposed, through the generalized Poisson distribution [59] or a two-stage Poisson model, which tests for differential expression in two modes depending on the evidence for overdispersion in the data [60]. Several tools for either simultaneous transcript discovery and quantification [11] or alternative isoform expression analysis [10] also perform DE analysis. However, it is worth noting that these methods use either the Poisson distribution or Fisher's exact test, neither of which explicitly deal with the biological variation discussed above.

Many of the current strategies for DE analysis of count data are limited to simple experimental designs, such as pairwise or multiple group comparisons. To the best of our knowledge, no general methods have been proposed for the analysis of more complex designs, such as paired samples or time course experiments, in the context of RNA-seq data. In the absence of such methods, researchers have transformed their count data and used tools appropriate for continuous data [31,47,61]. Generalized linear models provide the logical extension to the count models presented above, and clever strategies to share information over all genes will need to be developed; software tools now provide these methods (such as edgeR [57]). Furthermore, the methods discussed above are predominantly aimed at summarizing expression levels at which annotation exists. Methods, such as the maximum mean discrepancy test [62], have recently been proposed to detect DE in an untargeted manner.

Systems biology: going beyond gene lists

In many cases, creating lists of DE genes is not the final step of the analysis; further biological insight into an experimental system can be gained by looking at the expression changes of sets of genes. Many tools focusing on gene set testing, network inference and knowledge databases have been designed for analyzing lists of DE genes from microarray datasets [63-65]. However, RNA-seq is affected by biases not present in microarray data. For example, gene length bias is an issue in RNA-seq data, in which longer genes have higher counts (at the same expression level) [66]. This results in greater statistical power to detect DE for long and highly expressed genes. These biases can dramatically affect the results of downstream analyses, such as testing Gene Ontology (GO) terms for enrichment among DE genes [66,67]. In order to enable gene set analyses, Bullard *et al.* [45] suggested modifying a DE *t*-statistic by dividing by the square root of gene length to minimize the effect of length bias on DE. Alternatively, GO-seq is an approach developed specifically for RNA-seq data that can incorporate length or total count bias into gene set tests [68]. As the understanding of biases in RNA-seq data grows, systems biology tools that incorporate this understanding will be critical to extracting biological insight.

There is wide scope for integrating the results of RNA-seq data with other sources of biological data to establish a more complete picture of gene regulation [69]. For example, RNA-seq has been used in conjunction with genotyping data to identify genetic loci responsible for variation in gene expression between individuals (expression quantitative trait loci or eQTLs) [35,70]. Furthermore, integration of expression data with transcription factor binding, RNA interference, histone modification and DNA methylation information has the potential for

greater understanding of a variety of regulatory mechanisms. A few reports of these 'integrative' analyses have emerged recently [71-73]. For example, Lister and co-authors [71] highlighted a striking difference in the correlations of RNA-seq expression with CG and non-CG methylation levels in gene bodies. Similarly, combinations of sequencing-based datasets are beginning to provide insights into the mono-allelic associations between expression, histone modifications and DNA methylation [74].

Outlook

In this review, we have outlined the major steps in processing the millions of short reads produced by RNA-seq into an analysis of DE between samples. In brief, the process is to map and summarize short read sequences, then normalize between samples and perform a statistical test of DE. Further biological insight can be gained by looking for patterns of expression changes within sets of genes and integrating the RNA-seq data with data from other sources.

Although many parts of this pipeline have been the focus of extensive research, there are still areas that offer the possibility of further refinements. So far, there has been little work researching which summarization metric is best suited to finding DE between samples. There is also scope for expanding existing statistical methods for DE detection to enable the analysis of more complex experimental designs. Moreover, the relative merits of the many approaches now available deserve further study, in terms of their flexibility to analyze various study designs, their performance in small and large studies, dependence on sequencing depth and the accuracy of the assumptions (such as mean-variance relationships) that are imposed. Furthermore, although there are many examples of using RNA-seq for the detection of alternative splicing, there is scope to extend current methods to detect differences in gene isoform preference [10,11] when biological variability is prominent, perhaps using the count-based statistical methods mentioned above.

Given that there are substantial differences in the protocols that generate short reads, it will be important to formally compare RNA-seq platforms and the relative merits of the many data analysis methodologies. Such investigations may reveal benefits of platform-specific DE analysis methods and will also facilitate greater data integration. As the field is still relatively young, we expect many new methods and tools for the analysis of RNA-seq data to emerge in the near future.

Box 1: Comparisons of microarrays and sequencing for gene expression analysis

Several comparisons of RNA-seq and microarray data have now been made. These include proof-of-principle demonstrations of the sequencing platform [2,31,32], dedicated comparison studies [34,75-77] and analysis methodology development [10]. The results are unanimous: sequencing has higher sensitivity and dynamic range, coupled with lower technical variation. Furthermore, comparisons have highlighted strong concordance between microarrays and sequencing in measures of both absolute and differential expression. Nevertheless, microarrays have been, and continue to be, highly successful in interrogating the transcriptome in many biological settings. Examples include defining the cell of origin for breast cancer subtypes [78] and investigating the effect of evolution on gene expression in *Drosophila* [79].

Microarrays and sequencing each have their own specific biases that can affect the ability of a platform to measure DE. It is well known that cross-hybridization of microarray probes affects expression measures in a non-uniform way [80,81] and sequence content influences measured probe intensities [82]. Meanwhile, several studies have observed a GC bias in RNA-seq data [45] and RNA-seq can suffer from mapping ambiguity for paralogous sequences. Furthermore, there is a higher statistical power to detect changes at higher counts (for example, a twofold difference of 200 reads to 100 reads is more statistically significant than 20 reads to 10, under the null hypothesis of no difference); this bias typically manifests in RNA-seq as an association between DE and gene length, an effect not present in microarray data [66,68]. Other studies indicate that specific sequencing protocols produce biases in the generated reads, which can be related to the sequence composition and distance along the transcript [49,50,83,84]. For example, library preparation for small RNAs has been found to strongly affect the set of observed sequences [85]. Furthermore, transcriptome assembly approaches are necessarily biased by expression level because less information is available for genes expressed at a low level [11,14]. Many of these biases are still being explored and clever statistical methods that harness this knowledge may be able to provide improvements on existing methods.

In addition to the larger dynamic range and sensitivity of RNA-seq, several additional factors have contributed to the rapid uptake of sequencing for differential expression analysis. First, microarrays are simply not available for many non-model organisms (for example, Affymetrix offers microarrays for approximately 30 species [86]). By contrast, genomes and sequence information are readily available for thousands of species [87]. Moreover, even when genomes are not available, RNA-seq can still be performed and the transcriptome can still be interrogated (for instance, a recent study used RNA-seq to investigate the cell origin of the Tasmanian Devil facial tumor [88]). Second, sequencing gives unprecedented detail about transcriptional features that arrays

cannot, such as novel transcribed regions, allele-specific expression, RNA editing and a comprehensive capability to capture alternative splicing. For example, a recent RNA-seq study [11] was able to show several examples of isoform switching during cell differentiation, and RNA-seq was used to show parent-of-origin expression in mouse brain [5].

Sequencing is not without its challenges, of course. The cost of the platform may be limiting for some studies. However, with the expansion in total sequencing capacity and the ability to multiplex, the cost per sample to generate sufficient sequence depth will soon be comparable to that of microarrays. However, the cost of informatics to house, process and analyze the data is substantial [89]. Researchers with limited access to computing staff and resources may elect to use microarrays because data analysis procedures are relatively mature. Finally, it is clear that data analysis methodologies for sequencing data will continue to evolve for some time yet.

Author information
All authors contributed equally to this review.

Author details
[1]Bioinformatics Division, Walter and Eliza Hall Institute, 1G Royal Parade, Parkville 3052, Australia. [2]Epigenetics Laboratory, Cancer Program, Garvan Institute of Medical Research, 384 Victoria Street, Darlinghurst, NSW 2010, Australia.

Published: 22 December 2010

References
1. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.** *Nat Genet* 2008, **40:**1413-1415.
2. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D, Schmidt D, O'Keeffe S, Haas S, Vingron M, Lehrach H, Yaspo ML: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome.** *Science* 2008, **321:**956-960.
3. Wagner JR, Ge B, Pokholok D, Gunderson KL, Pastinen T, Blanchette M: **Computational analysis of whole-genome differential allelic expression data in human.** *PLoS Comput Biol* 2010, **6:**e1000849.
4. Wang X, Sun Q, McGrath SD, Mardis ER, Soloway PD, Clark AG: **Transcriptome-wide identification of novel imprinted genes in neonatal mouse brain.** *PLoS ONE* 2008, **3:**e3839.
5. Gregg C, Zhang J, Weissbourd B, Luo S, Schroth GP, Haig D, Dulac C: **High-resolution analysis of parent-of-origin allelic expression in the mouse brain.** *Science* 2010, **329:**643-648.
6. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, Sam L, Barrette T, Palanisamy N, Chinnaiyan AM: **Transcriptome sequencing to detect gene fusions in cancer.** *Nature* 2009, **458:**97-101.
7. Berger MF, Levin JZ, Vijayendran K, Sivachenko A, Adiconis X, Maguire J, Johnson LA, Robinson J, Verhaak RG, Sougnez C, Onofrio RC, Ziaugra L, Cibulskis K, Laine E, Barretina J, Winckler W, Fisher DE, Getz G, Meyerson M, Jaffe DB, Gabriel SB, Lander ES, Dummer R, Gnirke A, Nusbaum C, Garraway LA: **Integrative analysis of the melanoma transcriptome.** *Genome Res* 2010, **20:**413-427.
8. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty.** *Bioinformatics* 2010, **26:**493-500.
9. Jiang H, Wong WH: **Statistical inferences for isoform expression in RNA-Seq.** *Bioinformatics* 2009, **25:**1026-1032.
10. Griffith M, Griffith OL, Mwenifumbo J, Goya R, Morrissy AS, Morin RD, Corbett R, Tang MJ, Hou YC, Pugh TJ, Robertson G, Chittaranjan S, Ally A, Asano JK, Chan SY, Li HI, McDonald H, Teague K, Zhao Y, Zeng T, Delaney A, Hirst M, Morin GB, Jones SJ, Tai IT, Marra MA: **Alternative expression analysis by RNA sequencing.** *Nat Methods* 2010, **7:**843-847.
11. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotechnol* 2010, **28:**511-515.
12. Wang L, Xi Y, Yu J, Dong L, Yen L, Li W: **A statistical method for the detection of alternative splicing using RNA-seq.** *PLoS One* 2010, **5:**e8529.
13. Picardi E, Horner DS, Chiara M, Schiavon R, Valle G, Pesole G: **Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing.** *Nucleic Acids Res* 2010, **38:**4755-4767.
14. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ, Griffith M, Raymond A, Thiessen N, Cezard T, Butterfield YS, Newsome R, Chan SK, She R, Varhol R, Kamoh B, Prabhu AL, Tam A, Zhao Y, Moore RA, Hirst M, Marra MA, Jones SJ, Hoodless PA, Birol I: *De novo* **assembly and analysis of RNA-seq data.** *Nat Methods* 2010, **7:**909-912.
15. Camarena L, Bruno V, Euskirchen G, Poggio S, Snyder M: **Molecular mechanisms of ethanol-induced pathogenesis revealed by RNA-sequencing.** *PLoS Pathog* 2010, **6:**e1000834.
16. Shendure J, Ji H: **Next-generation DNA sequencing.** *Nat Biotechnol* 2008, **26:**1135-1145.
17. **Software: Seqwiki: Seqanswers** [http://seqanswers.com/wiki/Software]
18. **Wikipedia: Short-Read Sequence Alignment** [http://en.wikipedia.org/wiki/List_of_sequence_alignment_software#Short-Read_Sequence_Alignment]
19. Chen Y, Souaiaia T, Chen T: **PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds.** *Bioinformatics* 2009, **25:**2514-2521.
20. Homer N, Merriman B, Nelson SF: **BFAST: an alignment tool for large scale genome resequencing.** *PLoS ONE* 2009, **4:**e7767.
21. Rumble SM, Lacroute P, Dalca AV, Fiume M, Sidow A, Brudno M: **SHRiMP: accurate mapping of short color-space reads.** *PLoS Comput Biol* 2009, **5:**e1000386.
22. Hach F, Hormozdiari F, Alkan C, Birol I, Eichler EE, Sahinalp SC: **mrsFAST: a cache-oblivious algorithm for short-read mapping.** *Nat Methods* 2010, **7:**576-577.
23. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25:**1754-1760.
24. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment.** *Bioinformatics* 2009, **25:**1966-1967.
25. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10:**R25.
26. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies.** *Nat Methods* 2009, **6:**S22-S32.
27. Li H, Homer N: **A survey of sequence alignment algorithms for next-generation sequencing.** *Brief Bioinform* 2010, **11:**473-583.
28. Ferragina P, Manzini G: **Opportunistic data structures with applications.** In *Proceedings of the 41st Annual Symposium on Foundations of Computer Science: 12-14 Nov 2000*. Redondo Beach, USA, 2000; 390-398.
29. Li H, Ruan J, Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18:**1851-1858.
30. Flicek P, Birney E: **Sense from sequence reads: methods for alignment and assembly.** *Nat Methods* 2009, **6:**S6-S12.
31. Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM: **Stem cell transcriptome profiling via massive-scale mRNA sequencing.** *Nat Methods* 2008, **5:**613-619.
32. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq.** *Nat Methods* 2008, **5:**621-628.
33. Taub M, Speed TP: **Methods for allocating ambiguous short-reads.** *Commun Inf Syst* 2010, **10:**69-82.
34. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18:**1509-1517.

35. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK: **Understanding mechanisms underlying human gene expression variation with RNA sequencing.** *Nature* 2010, **464:**768-772.

36. Ameur A, Wetterbom A, Feuk L, Gyllensten U: **Global and unbiased detection of splice junctions from RNA-seq data.** *Genome Biol* 2010, **11:**R34.

37. De Bona F, Ossowski S, Schneeberger K, Rätsch G: **Optimal spliced alignments of short sequence reads.** *Bioinformatics* 2008, **24:**i174-i180.

38. Denoeud F, Aury JM, Da Silva C, Noel B, Rogier O, Delledonne M, Morgante M, Valle G, Wincker P, Scarpelli C, Jaillon O, Artiguenave F: **Annotating genomes with massive-scale RNA sequencing.** *Genome Biol* 2008, **9:**R175.

39. Hammer P, Banck MS, Amberg R, Wang C, Petznick G, Lou S, Khrebtukova I, Schroth GP, Beyerlein P, Beutler AS: **mRNA-seq with agnostic splice site discovery for nervous system transcriptomics tested in chronic pain.** *Genome Res* 2010, **20:**847-860.

40. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq.** *Bioinformatics* 2009, **25:**1105-1111.

41. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, MacLeod JN, Chiang DY, Prins JF, Liu J: **MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery.** *Nucleic Acids Res* 2010, **38:**e178.

42. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly using de Bruijn graphs.** *Genome Res* 2008, **18:**821-829.

43. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19:**1117-1123.

44. Cloonan N, Xu Q, Faulkner GJ, Taylor DF, Tang DT, Kolle G, Grimmond SM: **RNA-MATE: a recursive mapping strategy for high-throughput RNA-sequencing data.** *Bioinformatics* 2009, **25:**2615-2616.

45. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments.** *BMC Bioinformatics* 2010, **11:**94.

46. Anders S, Huber W: **Differential expression analysis for sequence count data.** *Genome Biol* 2010, **11:**R106.

47. Langmead B, Hansen KD, Leek JT: **Cloud-scale RNA-sequencing differential expression analysis with Myrna.** *Genome Biol* 2010, **11:**R83.

48. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11:**R25.

49. Li J, Jiang H, Wong WH: **Modeling non-uniformity in short-read rates in RNA-Seq data.** *Genome Biol* 2010, **11:**R50.

50. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming.** *Nucleic Acids Res* 2010, **38:**e131.

51. Robinson MD, Smyth GK: **Moderated statistical tests for assessing differences in tag abundance.** *Bioinformatics* 2007, **23:**2881-2887.

52. Balwierz PJ, Carninci P, Daub CO, Kawai J, Hayashizaki Y, Van Belle W, Beisel C, van Nimwegen E: **Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data.** *Genome Biol* 2009, **10:**R79.

53. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA: **mRNA-Seq whole-transcriptome analysis of a single cell.** *Nat Methods* 2009, **6:**377-382.

54. Wang L, Feng Z, Wang X, Zhang X: **DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.** *Bioinformatics* 2010, **26:**136-138.

55. Robinson MD, Smyth GK: **Small-sample estimation of negative binomial dispersion, with applications to SAGE data.** *Biostatistics* 2008, **9:**321-332.

56. Auer PL, Doerge RW: **Statistical design and analysis of RNA sequencing data.** *Genetics* 2010, **185:**405-416.

57. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics* 2010, **26:**139-140.

58. Hardcastle TJ, Kelly KA: **baySeq: Empirical Bayesian analysis of patterns of differential expression in count data.** *BMC Bioinformatics* 2010, **11:**442.

59. Srivastava S, Chen L: **A two-parameter generalized Poisson model to improve the analysis of RNA-seq data.** *Nucleic Acids Res* 2010, **38:**e170.

60. Auer PL: **Statistical design and analysis of next-generation sequencing data.** *PhD Thesis*. Purdue University West Lafayette, Indiana; 2010

61. Parikh A, Miranda ER, Katoh-Kurasawa M, Fuller D, Rot G, Zagar L, Curk T, Sucgang R, Chen R, Zupan B, Loomis WF, Kuspa A, Shaulsky G: **Conserved developmental transcriptomes in evolutionarily divergent species.** *Genome Biol* 2010, **11:**R35.

62. Stegle O, Drewe P, Bohnert R, Borgwardt K, Rätsch G: **Statistical tests for detecting differential RNA-transcript expression from read counts.** *Nat Precedings* 2010, in press.

63. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A* 2005, **102:**15545-15550.

64. Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA: **DAVID: Database for Annotation, Visualization, and Integrated Discovery.** *Genome Biol* 2003, **4:**P3.

65. Kanehisa M, Goto S: **KEGG: Kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res* 2000, **28:**27-30.

66. Oshlack A, Wakefield MJ: **Transcript length bias in RNA-seq data confounds systems biology.** *Biol Direct* 2009, **4:**14.

67. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G: **Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.** *Nat Genet* 2000, **25:**25-29.

68. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-seq: accounting for selection bias.** *Genome Biol* 2010, **11:**R14.

69. Hawkins RD, Hon GC, Ren B: **Next-generation genomics: an integrative approach.** *Nat Rev Genet* 2010, **11:**476-486.

70. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, Ingle C, Nisbett J, Guigo R, Dermitzakis ET: **Transcriptome genetics using second generation sequencing in a Caucasian population.** *Nature* 2010, **464:**773-777.

71. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, Nery JR, Lee L, Ye Z, Ngo QM, Edsall L, Antosiewicz-Bourget J, Stewart R, Ruotti V, Millar AH, Thomson JA, Ren B, Ecker JR: **Human DNA methylomes at base resolution show widespread epigenomic differences.** *Nature* 2009, **462:**315-322.

72. Ouyang Z, Zhou Q, Wong WH: **ChIP-Seq of transcription factors predicts absolute and differential gene expression in embryonic stem cells.** *Proc Natl Acad Sci U S A* 2009, **106:**21521-21526.

73. Raha D, Wang Z, Moqtaderi Z, Wu L, Zhong G, Gerstein M, Struhl K, Snyder M: **Close association of RNA polymerase II and many transcription factors with Pol III genes.** *Proc Natl Acad Sci U S A* 2010, **107:**3639-3644.

74. Harris RA, Wang T, Coarfa C, Nagarajan RP, Hong C, Downey SL, Johnson BE, Fouse SD, Delaney A, Zhao Y, Olshen A, Ballinger T, Zhou X, Forsberg KJ, Gu J, Echipare L, O'Geen H, Lister R, Pelizzola M, Xi Y, Epstein CB, Bernstein BE, Hawkins RD, Ren B, Chung WY, Gu H, Bock C, Gnirke A, Zhang MQ, Haussler D, *et al.*: **Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications.** *Nat Biotechnol* 2010, **28:**1097-1105.

75. Bradford JR, Hey Y, Yates T, Li Y, Pepper SD, Miller CJ: **A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling.** *BMC Genomics* 2010, **11:**282.

76. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P: **Estimating accuracy of RNA-Seq and microarrays with proteomics.** *BMC Genomics* 2009, **10:**161.

77. 't Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT: **Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms.** *Nucleic Acids Res* 2008, **36:**e141.

78. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ; kConFab, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ: **Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers.** *Nat Med* 2009, **15:**907-913.

79. Rifkin SA, Houle D, Kim J, White KP: **A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression.** *Nature* 2005, **438:**220-223.

80. Naef F, Magnasco MO: **Solving the riddle of the bright mismatches: labeling and effective binding in oligonucleotide arrays.** *Phys Rev E Stat Nonlin Soft Matter Phys* 2003, **68:**011906.

81. Wu Z, Irizarry RA: **Stochastic models inspired by hybridization theory for short oligonucleotide arrays.** *J Comput Biol* 2005, **12:**882-893.

82. Binder H, Kirsten T, Loeffler M, Stadler PF: **Sensitivity of microarray oligonucleotide probes: variability and effect of base composition.** *J Phys*

Chem B 2004, **108**:18003-18014.

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

84. Quail MA, Kozarewa I, Smith F, Scally A, Stephens PJ, Durbin R, Swerdlow H, Turner DJ: A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 2008, **5**:1005-1010.

85. Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E: Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* 2009, **6**:474-476.

86. Affymetrix [http://www.affymetrix.com]

87. NCBI: Entrez Genome [http://www.ncbi.nlm.nih.gov/sites/genome]

88. Murchison EP, Tovar C, Hsu A, Bender HS, Kheradpour P, Rebbeck CA, Obendorf D, Conlan C, Bahlo M, Blizzard CA, Pyecroft S, Kreiss A, Kellis M, Stark A, Harkins TT, Marshall Graves JA, Woods GM, Hannon GJ, Papenfuss AT: The Tasmanian devil transcriptome reveals Schwann cell origins of a clonally transmissible cancer. *Science* 2010, **327**:84-87.

89. Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP: Computational solutions to large-scale data management and analysis. *Nat Rev Genet* 2010, **11**:647-657.

90. Li H, Lovci MT, Kwon YS, Rosenfeld MG, Fu XD, Yeo GW: Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci U S A* 2008, **105**:20179-20184.

91. Wu TD, Nacu S: Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* 2010, **26**:873-881.

92. Schatz MC: CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics* 2009, **25**:1363-1369.

93. Clement NL, Snell Q, Clement MJ, Hollenhorst PC, Purwar J, Graves BJ, Cairns BR, Johnson WE: The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics* 2010, **26**:38-45.

94. Weese D, Emde AK, Rausch T, Doring A, Reinert K: RazerS - fast read mapping with sensitivity control. *Genome Res* 2009, **19**:1646-1654.

95. Li R, Li Y, Kristiansen K, Wang J: SOAP: short oligonucleotide alignment program. *Bioinformatics* 2008, **24**:713-714.

96. Au KF, Jiang H, Lin L, Xing Y, Wong WH: Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic Acids Res* 2010, **38**:4570-4578.

97. G-Mo.R-Se: Gene MOdeling using RNA-Seq [http://www.genoscope.cns.fr/externe/gmorse/]

98. Oases: De Novo Transcriptome Assembler For Very Short Reads [http://www.ebi.ac.uk/~zerbino/oases/]

99. Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WE, Wetter T, Suhai S: Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 2004, **14**:1147-1159.

100. DEGseq: Identify Differentially Expressed Genes from RNA-seq data [http://www.bioconductor.org/packages/release/bioc/html/DEGseq.html]