Chapter 6

Bioinformatics for RNA-Seq Data Analysis

Shanrong Zhao, Baohong Zhang, Ying Zhang, William Gordon, Sarah Du, Theresa Paradis, Michael Vincent and David von Schack

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/63267

Abstract

While RNA sequencing (RNA-seq) has become increasingly popular for transcriptome profiling, the analysis of the massive amount of data generated by large-scale RNA-seq still remains a challenge. RNA-seq data analyses typically consist of (1) accurate mapping of millions of short sequencing reads to a reference genome, including the identification of splicing events; (2) quantifying expression levels of genes, transcripts, and exons; (3) differential analysis of gene expression among different biological conditions; and (4) biological interpretation of differentially expressed genes. Despite the fact that multiple algorithms pertinent to basic analyses have been developed, there are still a variety of unresolved questions. In this chapter, we review the main tools and algorithms currently available for RNA-seq data analyses, and our goal is to help RNA-seq data analysts to make an informed choice of tools in practical RNA-seq data analysis. In the meantime, RNA-seq is evolving rapidly, and newer sequencing technologies are briefly introduced, including stranded RNA-seq, targeted RNA-seq, and single-cell RNA-seq.

Keywords: data analysis, gene quantification, pipeline, RNA-seq, workflow

1. Introduction

In recent years, RNA sequencing (RNA-seq) has emerged as a powerful technology for transcriptome profiling [1–4]. Compared with microarrays, it not only avoids some of the technical limitations of this approach including varying probe performance and nonspecific hybridization, and dynamic range issues, but can also detect alternative splicing isoforms and subtle changes of splicing under different conditions. The overview of current RNA-seq

approaches using shotgun sequencing technologies such as Illumina and the corresponding data analysis workflow is summarized in **Figure 1**. Polyadenylated (Poly-A) RNA transcripts (for the so-called mRNA-seq) are enriched with oligo (dT) primers and then fragmented. After size selection, millions or even billions of short sequence reads are generated from a randomly fragmented cDNA library. For most RNA-seq studies, the data analyses consist of the following key steps [5, 6]: (1) quality check and preprocessing of raw sequence reads, (2) mapping reads to a reference genome or transcriptome, (3) counting reads mapped to individual genes or transcripts, (4) identification of differential expression (DE) genes between different biological conditions, and (5) biological interpretation of DE genes and functional enrichment analysis. Despite the fact that a large number of algorithms [7] have been developed for RNA-seq data analysis in recent years, there are still many open questions for accurate read mapping, gene quantification, and data normalization.

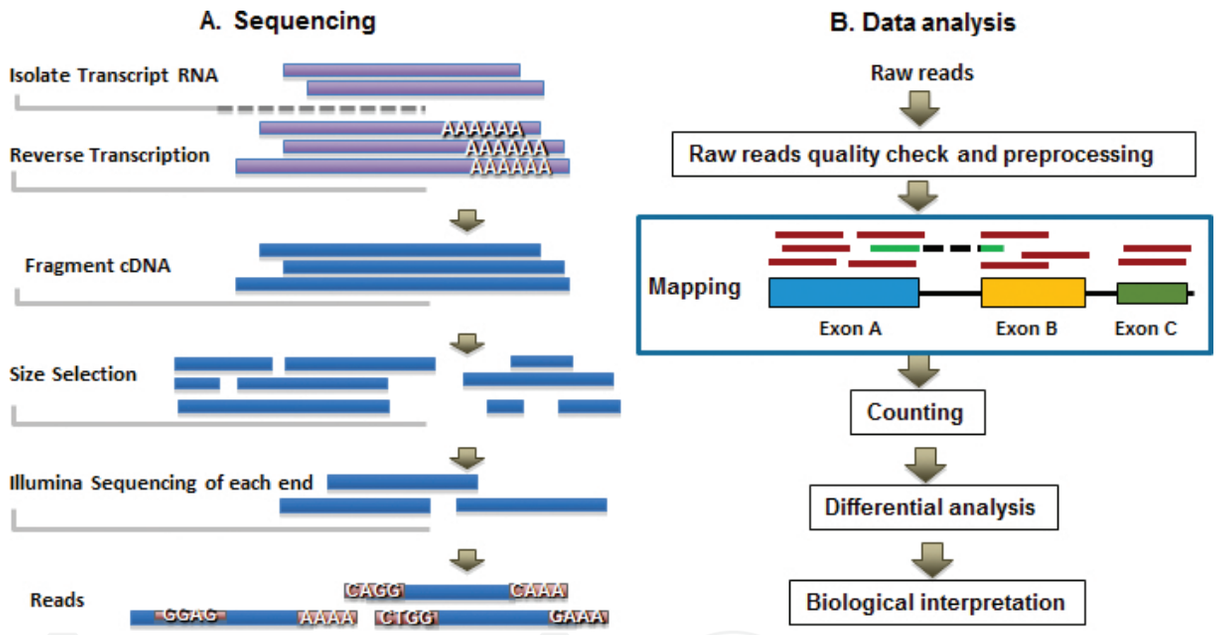

Figure 1. Overview of mRNA-seq laboratory flowchart and data analysis pipeline.

In complex mammalian genomes, both DNA strands encode genes. As a result, if two genes transcribed from opposite strands overlap, nonstranded RNA-seq cannot tell the true origin if a read falls into overlapping regions. Most recently, a stranded RNA-seq protocol has been developed for more accurate gene quantification [8, 9]. RNA-seq is a powerful tool when used to profile the entire transcriptome. However, this method can be inefficient when interested in only a small subset of genes that are involved in particular pathways, or associated with specific diseases. To meet this demand, targeted RNA-seq technology has been developed. Traditionally, gene expression measurements were performed on "bulk" samples containing populations of thousands or millions of cells. Recent advances in genomic technologies have made it possible to measure gene expression in individual cells. Accordingly, cellular properties that were previously masked in "bulk" measurements can now be observed directly [10–

12]. Single-cell RNA-seq (scRNA-seq) introduces more challenges for data analyses due to technical noise, including coverage nonuniformity, data sparsity, amplification biases, and so on. In this chapter, we cover the topics related to RNA-seq technology and data analyses.

2. RNA-seq versus microarray for gene expression profiling analysis

Microarrays and RNA-seq have been the two technologies of choice for large-scale studies of gene expression, and the side-by-side comparisons of RNA-seq and hybridization-based arrays have been performed [12–15]. Malone et al. [16] compared its ability to identify differentially expressed genes with existing array technologies and found that RNA-seq data are highly reproducible with relatively little technical variation. The DE genes identified from RNA-seq had a high overlapping with those identified by microarray. Fu et al. [17] designed a study in which they evaluated the accuracy of both microarrays and RNA-seq for mRNA quantification by using protein expression measurements as ground truths. In that study, they assessed the relative accuracy of the two transcriptome quantification approaches with respect to absolute transcript level measurements and found that RNA-seq provides better estimates of transcript expressions.

Previously, we performed a side-by-side comparison of RNA-seq and microarrays in investigating T-cell activation [18]. A comparison of data sets derived from RNA-seq and Affymetrix platforms using the same set of samples revealed a very high concordance between expression profiles generated by the two platforms. In the meantime, it was also demonstrated that RNA-seq is superior in differentiating biologically critical isoforms, detecting low abundance transcripts, and allowing the identification of genetic variants. Analysis of the two data sets also showed the benefit derived from avoidance of technical issues inherent to microarray probe performance such as cross-hybridization, nonspecific hybridization, and limited detection range of individual probes. In addition, RNA-seq has a much broader dynamic range than microarray technologies, which allows for the detection of more differentially expressed genes with higher fold-change. Thus, RNA-seq delivers both less biased and previously unknown information about the transcriptome. Because RNA-seq does not rely on a predesigned complementary sequence detection probe, it is not limited to the interrogation of selected probes on an array and can also be applied in species, for which the whole reference genome is not yet assembled.

RNA-seq allows for the detection of novel transcript species in well-studied organisms, such as unique transcripts in certain tissues or in rare cell types, and has been instrumental to catalog the diversity of novel transcript species including long noncoding RNA, miRNA, siRNA, and other small RNA classes [19]. Additionally, RNA-seq technology proves to be an invaluable tool for deciphering the extensive alternative splicing of the transcriptome [20, 21]. Alternative splicing creates two to potentially hundreds of variants in more than 90% of human genes. Furthermore, RNA-seq can identify allele-specific expression and gene fusion events [22, 23].

3. RNA-seq library preparation and sequencing platforms

3.1. General considerations for RNA-seq

The quality and quantity of the starting RNA material are likely the most important aspects to consider when deciding on the methods to generate RNA-seq libraries. For high-quality samples with mostly intact RNA, a wide variety of sequencing library preparation methods is available. For lower quality samples with partially or highly degraded RNA, there are considerably fewer methods to choose from. The amount of RNA may limit the choice of library preparation as well, since the majority of standard RNA-seq kits require a minimum of 10–100 ng of the total RNA. If the RNA amount is below this threshold it will require the use of a more specialized kit and/or the sample may require some form of amplification.

To obtain reliable and reproducible RNA-seq data, RNA quality is of paramount importance. Because of the inherent instability of the RNA molecule, the quality of RNA from samples collected in clinical settings and field studies is often impacted by tissue necrosis, which quickly degrades RNA if the sample is not frozen or chemically preserved within minutes after surgery. For assessing RNA quality, most labs utilize an electrophoretic-based system, including the Agilent BioAnalyzer, the Agilent TapeStation, or the Advanced Analytics Fragment Analyzer. All of these instruments produce an RNA integrity score. On the BioAnalyzer, the score ranges from 1 to 10 (10 being perfect) and most labs would consider a RIN > 6 to be acceptable for standard RNA-seq methods. For degraded samples where the RIN is below 6, it is beneficial to calculate the DV200 score (the percent of the sample that is larger than 200 bp in size). Illumina has shown this to be an important metric for successful library preparation with kits designed for degraded RNA.

After quality and quantity of the RNA samples have been addressed, one must then choose whether to profile the total RNA space or the mRNA space only. This choice determines the main split between most RNA-seq library kit types, those which target the removal of ribosomal RNA (rRNA) versus those which utilize Poly-T beads to isolate the Poly-A tail of intact mRNAs. This choice will affect the downstream sequencing, influencing the depth to be targeted per sample, with rRNA depletion libraries requiring significantly deeper sequencing than Poly-A-based libraries in order to generate sufficient reads to capture all of the different RNA species in these samples. Additionally, if you are dealing with degraded RNA samples, you will not be able to utilize the Poly-A method, but the rRNA depletion methodologies will work for these samples. However, if you still want to focus only on the mRNAs, another option would be the more recent exon capture kits such as Illuminas RNA Access, which depend on hybridization with probes designed against known exons.

When considering the sequencing depth, a few considerations should be taken into account. mRNA libraries can be sequenced shallower than the total RNA libraries as they have less diversity and are focused only on the mature transcripts. In general a total RNA prep will only have 15–30% of the sample as mRNA, thus to get the same level of read depth for the mRNAs you will need to sequence those libraries between 3 and 10× as deep as a mRNA only library. Additionally, whether the goal of the experiment is targeted toward DE, isoform discovery,

http://dx.doi.org/10.5772/63267

or novel RNA discovery will impact the choice of read depth to target, with the latter requiring more reads. In general, targeting 30–40 million paired reads for an mRNA-seq and 70–100 million for a total RNA-seq would be the starting recommendation.

3.2. From RNA sample to raw sequence reads

The most common workflow for RNA-seq is stranded mRNA-seq. To generate a stranded mRNA-seq library, RNA from different sources (blood, tissue, and cell lines) should be purified utilizing a consistent methodology (kits with columns or manually using Trizol are both acceptable). The RNA should be treated either in solution or on-column (depending on your extraction choice) with DNASE to remove traces of genomic DNA and to prevent contamination of RNA-seq libraries by DNA. DNA-free RNA should then be assessed for both quality and quantity. RNA of sufficient quality can then be passed over Oligo-dT beads to capture mRNAs removing the non-Poly-A RNA species. The captured mRNA is then fragmented (enzymatically or mechanically) through ultrasonic shearing (with devices such as a Covaris ultrasonicator), followed by a two-step conversion to cDNA using first-strand synthesis and then second-strand synthesis (during which the cDNA is marked retaining the strand information) protocols. After cDNA conversion, the ends are repaired making them amenable to adapter ligation. Indexing of the libraries can be utilized at this point allowing for sample pooling before sequencing and a final enrichment of the indexed cDNA fragments by PCR (Illumina recommends 10–15 cycles of PCR depending on the RNA input), which is performed to generate the final RNA-seq library.

After the libraries are complete, their quality is assessed by an electrophoretic assay, if a peak at the size of dimerized adapters (80–100 bp) is observed, the libraries should be repurified before sequencing (adapter dimers will cluster very efficiently and many reads can be lost if they are not fully removed). The sequencing-ready libraries are quantified to allow for equimolar pooling typically through quantitative PCR. Method such as KAPA Library Quantification kits are recommended for best results. After quantification, libraries with different indices are pooled. The number of samples in the pool will vary depending on the read depth desired and sequencer used. Sequencing is then performed on a final diluted sample (following the manufacturer's recommendations). After the run is complete, the raw reads can be converted to fastq files (if Illumina sequencing was performed, Illumina offers an algorithm for this step called bcltofastq) and passed on for QC, alignment, and analysis.

4. Algorithms for RNA-seq data analysis

Millions, or even billions, of short reads are the starting point of RNA-seq computational data analyses [5, 6]. First, reads are QC checked and then mapped to a reference genome or transcriptome. The mapped reads for each sample are subsequently counted on gene, transcript, or exon level to assess the abundance of each category depending on the experimental purpose. The summarized data are then assessed by statistical models to identify differentially

expressed genes. Finally, pathway or network level analyses are performed to gain biological insight through systems biology approaches.

4.1. Quality check and preprocessing of raw reads

Poor-quality read data can arise from problems in the library preparation or from sequencing itself. Additionally, PCR artifacts, untrimmed adapter sequences, sequence-specific bias, and other possible contaminants can also lead to poor data quality. The presence of poor quality or technical sequences can affect the downstream analysis and data interpretation, and thus give inaccurate results. In order to assess quality of raw sequenced data, several tools such as PRINSEQ [24] and FastQC [25] have been developed. FastQC aims to provide a simple way to do some quality control checks on raw sequence data. It provides a modular set of analyses that can be used to give a quick impression of whether raw sequencing data have any problems that one should be aware of before doing any further analysis. Once the data are checked for quality, they should be processed to remove reads with low-quality bases, adapter sequences, and other contaminating sequences. Tools such as Cutadapt [26] and Trimmomatic [27], which trim adapter or other contaminating sequences based upon user-provided parameters, can be used for performing these operations. After going through the aforementioned steps, the sequencing data are ready for downstream analysis.

4.2. Read mapping

Short sequence reads generated by sequencers must first be mapped or aligned to a reference transcriptome or genome assembly to discover their true locations (origins) with respect to that reference. A large number of read mapping algorithms have been developed in recent years, including TopHat2 [28], STAR [29, 30], GSNAP [31], OSA [32], and MapSplice [33]. To assess the performance of current mapping software, Engström et al. [34] compared 26 mapping protocols based on 11 programs and pipelines and found there were major performance differences among different methods on numerous benchmarks, including basewise accuracy, alignment yield, gap placement, mismatches, and exon junction sites.

Indeed, some features of a reference genome such as repetitive regions, assembly errors, and assembly gaps render this objective impossible for a subset of reads. Furthermore, because RNA-seq libraries are constructed from transcribed RNA, intronic sequences are not present in exon-exon spanning reads. Therefore, when aligning the sequences to a reference genome, reads that span exon-exon junctions have to be split across potentially thousands of bases of intronic sequence. Many RNA-seq alignment tools use reference transcriptomes to inform the alignment of junction reads. The benefits of using a reference transcriptome to map RNA-seq reads have been demonstrated clearly in previous reports [35–37] and our own comprehensive evaluation [38] of RefGene (RefSeq Gene) [39], UCSC Known Genes [40], and Ensembl [41] in mapping of RNA-seq reads and gene quantifications.

The benefits of using a reference transcriptome in mapping of RNA-seq reads are illustrated in **Figure 2**. In **Figure 2A**, 19 junction reads can be uniquely mapped to gene HSP90AB1 when RefGene annotation [36] is provided in the alignment step. However, four reads indicated by

http://dx.doi.org/10.5772/63267

the red arrow are mapped to the same gene HSP90AB1 as nonjunction reads with mismatches at one end without the assistance of a reference transcriptome. In **Figure 2B**, those exon-exon spanning reads are mapped to gene TCEA3 with the exact same start and end positions regardless of the use of a transcriptome, but spliced differently. Both mappings are equal in terms of alignment scores and gaps between exons. It is therefore difficult, if not impossible, to tell which alignment is correct without the assistance of a reference transcriptome. Collectively, the two examples in **Figure 2** illustrate the importance of appropriate gene annotations in the correct alignment of junction reads.

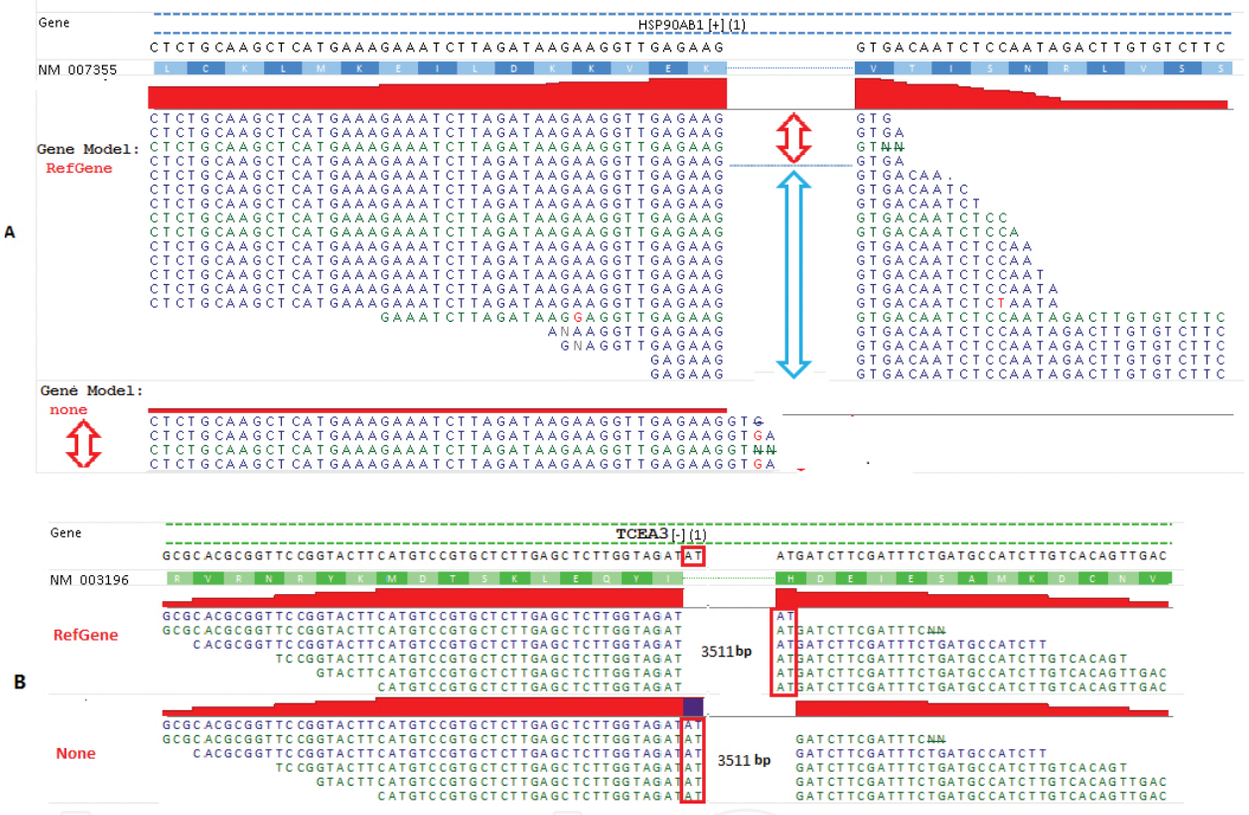

Figure 2. The impact of a reference transcriptome on the mapping of junction reads. Some exon-exon spanning reads are mapped incorrectly without the help of a reference transcriptome. Note the reads colored in blue are mapped to "+" strand, and in green when mapped to "-" strand. The mismatch nucleotide bases are colored in red.

Another problem in read mapping is that of polymorphisms, which occur when sequence reads align to multiple locations of the genome, abbreviated as multireads. Polymorphisms are especially common for large and complex transcriptomes, and the number of multireads for a mammalian genome is estimated to be between 10 and 40% [1, 3]. Generally speaking, there are three common strategies to deal with multireads in practice. The first strategy is to ignore or discard them completely, which we have demonstrated is not ideal for accurate gene quantification [36]. This practice not only discards potentially useful information, but also introduces an underestimation bias in quantifying expression of genes with highly redundant sequences (e.g., young duplicated genes). The second strategy implemented in most mapping software is to randomly assign a position from the possible matches. This practice assumes a

multiread can originate from these genomic locations equally, but this assumption is often not valid. The third strategy is to report all mapped locations for a multiread as long as the number of possible matches is below a user-defined cutoff, let's say 10. The problem with this strategy is that the cutoff is somewhat arbitrary. For an accurate detection and quantification of transcripts, it is important to resolve the mapping ambiguity for RNA-seq reads that can be mapped to multiple loci [42].

4.3. Read counting and gene quantification

Since RNA-seq has become a common technology in molecular biology laboratories, a number of methods have been developed for the inference of gene and isoform abundance, including RSEM [43], Cufflinks [44], IsoEM [45], featureCounts [46], and HTSeq [47]. The algorithms featureCounts [46] and HTSeq [47] are comparable in terms of counting results, but featureCounts is considerably faster than HTSeq by an order of magnitude for gene-level summarization and requires far less compute memory. However, neither featureCounts nor HTSeq can count reads at the transcript level due to their implementation. Andelini et al. [48] carried out a simulation study to assess the performance of five widely used counting tools and concluded that performance was heavily dependent upon the true abundance of the isoforms. Lowly expressed isoforms are poorly detected regardless of the methods.

Most recently, Kanitz et al. [49] have evaluated the accuracy of 14 methods for estimating isoform abundance and found that these tools vary widely in memory and runtime requirements. The algorithms for gene quantification can be broadly divided into two categories: transcript-based approaches (such as RSEM [43]) and "union-exon"-based approaches (such as featureCounts [46]). Because different isoforms of a gene typically have a high proportion of genomic overlap, it is intrinsically more difficult to estimate the expression of individual isoforms. Union-exon-based methods are much simpler, in which all overlapping exons of the same gene are merged into union exons. A read is counted to the gene as long as it has sufficient overlap with any of its union exons. Compared with isoforms, reads can be assigned to genes with much higher confidence. Therefore, the union-exon-based counting method is commonly used in RNA-seq, though gene-level counts cannot distinguish Isoforms [50].

A gene can be expressed in one or more transcript isoforms; accordingly, its expression level should be represented as the sum of its isoforms. We carried out a side-by-side comparison between union-exon-based approach and transcript-based method in RNA-seq gene quantification [40], and found that gene expression levels were significantly underestimated when the union-exon-based approach was used. We also discovered that the quantification of gene expression is more accurate if gene expression levels are computed by cumulating expression levels of transcript isoforms than by ignoring the transcript structures.

A gene model that hypothesizes the structure of transcripts produced by a gene also affects the analysis. Among multiple genome annotation databases, RefGene, Ensembl, and the UCSC annotation databases are the most popular. The choice of genome annotation directly affects gene expression estimation. Recently, we systematically characterized the impact of genome annotation on read mapping and transcriptome quantification [38]. Surprisingly, among the 21,958 common genes shared between RefGene and Ensembl annotations, only 16.3% of genes

http://dx.doi.org/10.5772/63267

obtained identical quantification results. Approximately 28.1% of genes' expression levels differed by >5% when using different annotation, and of those, the relative expression levels for 9.3% of genes differed by at least 50%. Our study revealed that the difference in gene definition frequently results in inconsistency in gene quantification (**Figure 3**). In Ensembl, the annotation of PIK3CA is much longer than its corresponding definition in RefGene. As a result, much more reads are counted toward PIK3CA if Ensembl annotation is used. According to the mapping profile of RNA-seq reads in **Figure 3**, the PIK3CA gene definition in Ensembl should be more accurate than the one in RefGene.

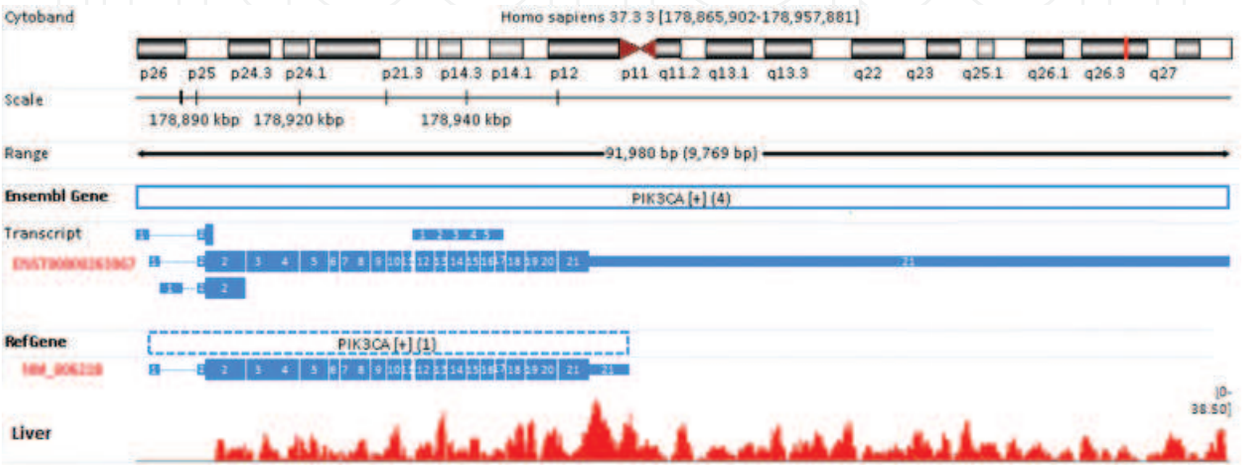

Figure 3. Different gene definitions for PIK3CA give rise to differences in gene quantification. PIK3CA in the Ensembl annotation is much longer than its definition in RefGene, which explains why 1094 reads are mapped to PIK3CA in Ensembl while only 492 reads are mapped using RefGene.

4.4. Data normalization and differential analysis

After calculating the read counts, data normalization is one of the most crucial steps of data processing, and this process must be carefully considered, as it is essential to ensure accurate inference of gene expression and subsequent analyses thereof. First, the *sequencing depths* or *library sizes* (the total number of mapped reads) typically vary for different samples, which means that the observed counts are not directly comparable between samples. The most straightforward way of normalizing the difference in sequencing library sizes is to rescale the total read counts. However, such normalization is quite often not enough because RNA-seq counts inherently represent *relative* abundances of genes in a sample. The number of reads mapped to a gene is not only dependent on the expression level and length of the gene, but also the composition of the RNA population that is being sampled. A few highly expressed genes may consume a very large portion of the total reads in a sample, and accordingly, the counts for all other genes are repressed. As a result, in comparison to a sample where the reads are more evenly distributed, those repressed genes seem to have lower expression which could give rise to a lot of falsely "differentially expressed" genes.

A fundamental research aim in many RNA-seq studies is to identify differentially expressed genes between distinct sample groups. Many algorithms have recently been introduced

specifically for the identification of differentially expressed genes (DEGs) from RNA-seq data, including DESeq [51], edgeR [52, 53], GENE-Counter [54], NOISeq [55], NBPSeq [56], and Cuffdiff2 [57]. However, there is a lack of consensus on how to approach an optimal study design and choice of suitable software for the analysis of an RNA-seq data sets. Recently, numerous groups [58–66] have performed a variety of comprehensive comparisons of different statistical methods for differential RNA-seq data analysis. Still, no general consensus has been reached after so many head to head comparisons and evaluations.

Zhang et al. [61] evaluated the performance of three widely used software tools: DESeq, edgeR and Cuffdiff2. They took a number of important metrics into consideration, including sequencing depth and the number of replicates, and the set of identified DEGs was evaluated with ground truths from either quantitative RT-PCR or microarray. They concluded that no single method is always superior in all DE analyses. It was noted that edgeR performs slightly better than DESeq and Cuffdiff2 in terms of the ability to uncover true positives and that Cuffdiff2 is not recommended for gene-level DE analysis, particularly if sequencing depth is low. Seyednasrollah et al. [65] also carried out a systematic comparison of the state-of-the-art methods in RNA-seq differential analysis to guide the selection of a suitable package. In general, there can be large differences between the algorithms. Similar to the evaluation performed by Zhang et al. [63], it was observed that no single method is likely to be optimal under all circumstances. They also demonstrated how the data analysis tool utilized can markedly affect the outcome of a differential analysis and highlighted the importance of the choice of software. Soneson and Delorenzi [61] have conducted extensive comparisons of 11 methods for DE analysis of RNA-seq data and concluded that very small sample size, which is still common in RNA-seq experiments, imposes problems for all evaluated methods and any results obtained under such conditions should be interpreted with caution.

We have applied edgeR to several whole blood RNA-seq data sets (unpublished results), and the calculated normalization factors might range from 0.4 to 1.6. We were puzzled by such an unreasonably low or high scaling factor. The normalization method implemented in edgeR is based on the premise that most genes are not differentially expressed. Instead of normalizing the raw counts directly, edgeR scales the library size by a factor so that the adjusted abundance (i.e., the ratio of read counts divided by the scaled library size) for many genes is not DE. TMM (Trimmed Mean of M-values) is calculated for each sample in a data set with one sample being considered as the reference sample and the others as test samples. For each test sample, TMM is computed between this test and the reference after exclusion of the most highly expressed genes (5% by default) and genes with the largest fold-changes (30% on each side of up and down by default). Ideally, TMM should be close to 1, but in cases where it is not, its value provides an estimate of the correction factor that must be applied to the library sizes (but not the raw counts) for normalization. However, TMM becomes problematic when the library composition between test and reference samples differs significantly. On the other hand, whether the default parameters in edgeR are appropriate for a given RNA-seq data set is difficult to determine. Moreover, different sets of genes are used for calculation of scaling factors across the entire data set since the genes excluded can vary from sample to sample.

http://dx.doi.org/10.5772/63267

4.5. Pathway enrichment analysis

To obtain a list of DE genes is only the starting point of gaining biological insights into experimental systems, developmental stages, or understanding of disease or molecular mechanisms. To understand the biological context of DE genes, pathway enrichment analysis ensues. Functional enrichment analyses rely upon annotation databases such as Gene Ontology (GO) [67], Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways [68], DAVID [69], and other commercial knowledge systems such as Ingenuity Pathway Analysis (IPA). One traditional analysis starts with a gene list of interest, identified from differential RNA-seq or microarray analyses, and applies statistical methods, such as the Fisher's exact test to test for enrichment of each annotated gene set, network, and pathway.

Gene set enrichment (GSE) analysis transforms information from gene expression profiling into a pathway summary. DE genes are quite often involved in the same biological pathways, and GSE results offer greater biological interpretability over individual gene analysis. GSVA [70] extends the current GSE methods to RNA-seq data, and provides increased power to detect subtle pathway activity changes, and constitutes a starting point to build pathway-centric models of system biology. SeqGSEA [71] is a new open-source Bioconductor package for GSVA and can detect more biologically meaningful gene sets without biases toward longer and highly expressed genes.

Previous pathway analysis methods had been developed based on algorithms considering pathways as simple gene lists and ignoring pathway structures. Recently methods have been developed to incorporate various aspects of pathway topology. For example SPIA captures pathway topology through its scoring system, in which the positions and the interactions of the genes in the pathway are considered [72]. Accordingly, interacting DEG pairs are preferentially weighted over two noninteracting genes. Similarly, TAPPA [73] is a scoring method in which higher weights are automatically assigned to hub genes and interacting gene pairs. DE analysis for pathways (DEAP) [74] makes significant improvements over existing approaches by including information about pathway topological structure. It was demonstrated [74] DEAP identified 14 more important chronic obstructive pulmonary disease–related pathways that existing approaches omitted.

4.6. QuickRNASeq—an integrated pipeline for efficient RNA-seq data analysis and interactive visualization

Although the time and cost for generating RNA-seq data are decreasing, the analysis of massive amounts of RNA-seq data still remains challenging. Numerous software packages and algorithms have been developed, which has led to the need to apply these tools efficiently to obtain results within a reasonable timeframe, especially for large data sets. Based on our own experience with analyses of multiple in-house RNA-seq data sets of varying size using open source tools, the main challenges, gaps, and bottlenecks for large-scale RNA-seq data analyses can be summarized as follows:

1. It is not trivial to select appropriate software packages and set software-specific parameters since it requires an in-depth understanding of the algorithms.

2. Additional bridging scripts are often necessary to make different components work seamlessly in a pipeline.

3. In general, most algorithms are implemented to process an individual sample, and thus it is necessary to integrate and summarize results from individual samples.

4. Due to sample quality and the complicated multistep processes in RNA-seq, it is required to establish stringent RNA-seq data quality metrics to identify outliers that should be excluded from further downstream data analysis.

5. Nearly all RNA-seq data analyses are performed using Linux workstations; however, analysis results in Linux are often inaccessible to most experimental scientists. Thus, sharing results with scientists is a practical challenge.

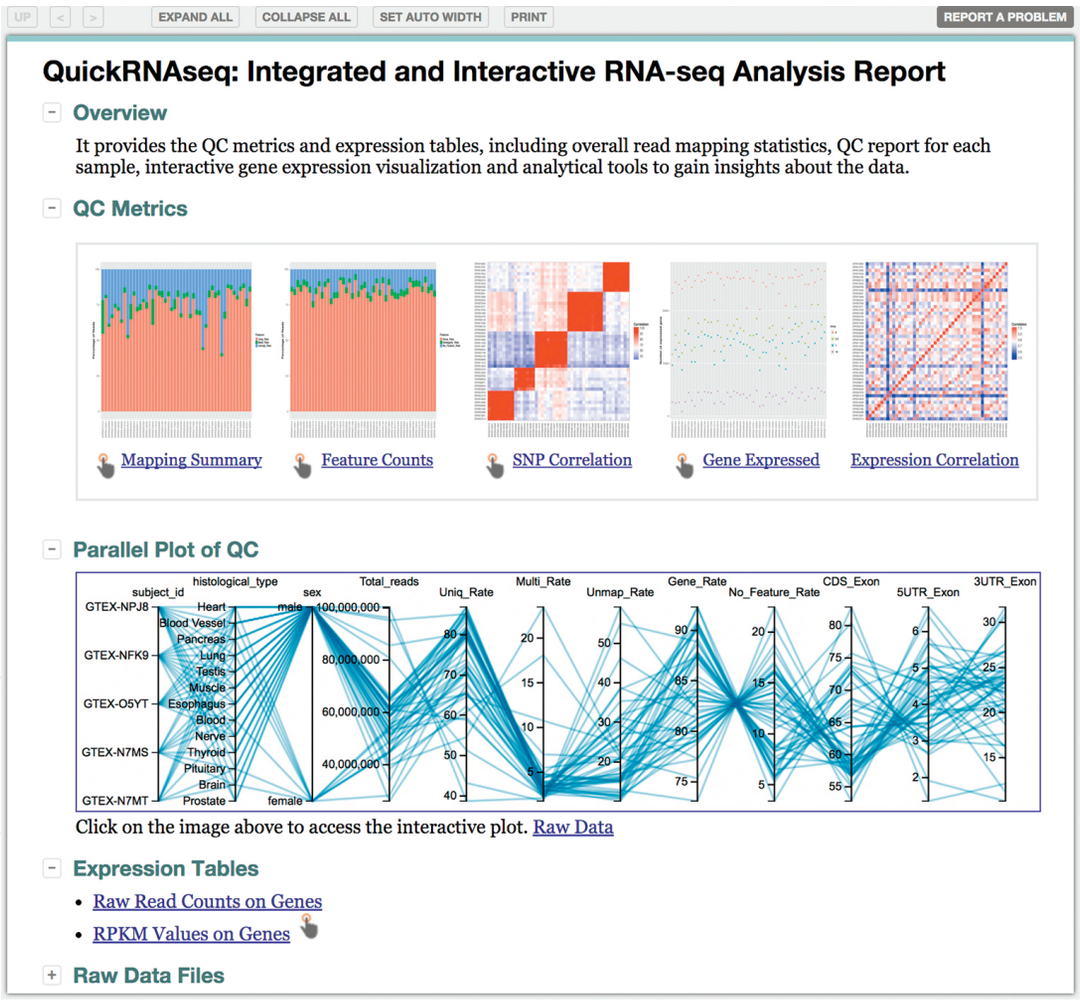

Figure 4. Representative entry webpage for a QuickRNAseq project report. The page layout and printable version of the page can be controlled by the top icons. The QC Metrics section provides QC results in plain text, static plot, and interactive plot formats accessible by clicking on the corresponding hyperlinked texts, the iconized figures, and pointing hand, respectively. The parallel plot of QC offers an integrated view of linked QC measures for a single sample or a group of samples. The expression tables section provides links to raw read counts, a normalized RPKM table, and interactive displays of gene expression levels.

http://dx.doi.org/10.5772/63267

To address these challenges, we have implemented a pipeline named QuickRNASeq to advance the automation and interactive visualizations of RNA-seq data analysis results [75]. QuickRNASeq significantly reduces data analysts' hands-on time, which results in a substantial decrease in the time and effort needed for the primary analyses of RNA-seq data before proceeding to further downstream analysis and interpretation. Additionally, QuickRNASeq provides a dynamic data sharing and interactive visualization environment for end users. All the results are accessible from a web browser without the need to set up a web server and database. The rich visualization features implemented in QuickRNASeq enable nonexpert end users to interact easily with the RNA-seq data analyses results and to drill down into specific aspects to gain insights into often complex data sets simply through a point-and-click approach. A representative entry webpage for a QuickRNASeq project report is shown in **Figure 4**. All the result files and figures are directly accessible by "point and click" from the entry webpage, which makes data navigation and visualization more convenient and intuitive, especially for experimental scientists.

5. New RNA sequencing technologies

5.1. Stranded RNA-seq

One significant shortcoming of the first-generation RNA-seq protocol was that it did not retain the strand information for each transcript. Recently, strand-specific or stranded RNA-seq protocols have been developed [76]. Previous reports [8] demonstrated that data from stranded libraries are more reliable than data from nonstranded libraries and can correctly evaluate the expression of both antisense RNA and overlapping genes. The ability to capture the relative abundance of both sense and antisense expression provides insight into regulatory interactions that might otherwise be missed [9]. With the ability to unlock new information on global gene expression, stranded RNA-seq holds the key to a deeper understanding of the transcriptome. We performed a side-by-side comparison of stranded and nonstranded RNA-seq in our whole blood RNA-seq data set, and demonstrated that stranded RNA-seq provides a more accurate estimate of gene expression compared with nonstranded RNA-seq and is therefore the recommended RNA-seq approach for all future mRNA-seq studies [77].

The advantages of stranded RNA-seq are illustrated in **Figure 5**. ICAM4 (intercellular adhesion molecule 4) shows moderate expression in whole blood. However, nonstranded RNA-seq reports no expression for this gene. As observed in **Figure 5**, ICAM4 is 100% contained within CTD-2369P2.8. In nonstranded RNA-seq, a read mapped to ICAM4 is simultaneously aligned to CTD-2369P2.8 as well. The ambiguous reads in overlapping regions are thus excluded from counting, which explains the lack of expression for ICAM4 with nonstranded RNA-seq. The ambiguous reads in overlapping genes in **Figure 5** can be perfectly resolved using stranded RNA-seq. By considering the read direction, all reads are assigned to ICAM4 (but not CTD-2369P2.8), because they are all reverse complementary to ICAM4. According to a stranded sequencing protocol, it is impossible for such reads to originate from CTD-2369P2.8.

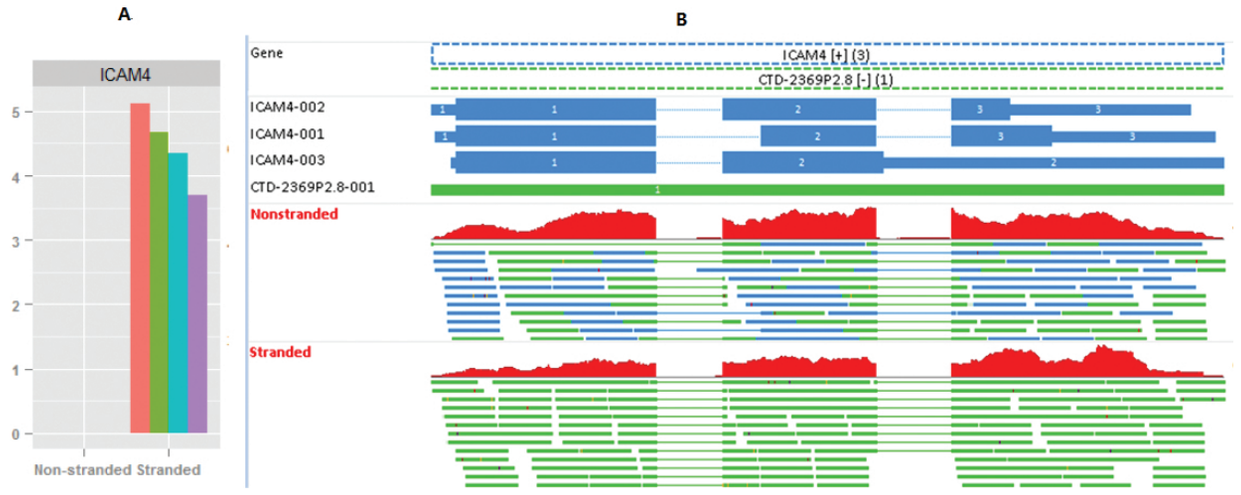

Figure 5. (A) Gene expression of ICAM4 in stranded and nonstranded RNA-seq. (B) Mapping profiles for ICAM4 (intercellular adhesion molecule 4). ICAM4 is on the "+" strand, and 100% contained within CTD-2369P2.8 in the "-" strand. In nonstranded RNA-seq, the ambiguous reads in overlapping regions are excluded from counting, which explains why there is no expression for ICAM4. However, the ambiguous reads can be perfectly resolved in stranded RNA-seq. By considering the read direction, all reads can be counted to ICAM4 but not CTD-2369P2.8. *Note:* (1) RNAs were extracted from pooled whole blood samples, and four replicates were pair-end sequenced using both stranded and nonstranded protocols. The unit of *y*-axis is RPKM in the plot to the left. (2) All genes, transcripts, and sequence reads are colored in blue if they are in the "+" strand and colored in green if in the "-" strand. According to our stranded sequencing protocol, a sequence read should be reversely complementary to its transcript origin.

5.2. Targeted RNA-seq

RNA-seq can be a powerful tool to measure gene expression, detect novel transcripts, characterize transcript isoforms, and identify sequence polymorphisms. However, this unbiased RNA-seq method can be costly and yields complex data sets that are time consuming to analyze. Often one is interested in only a small subset of genes or the goal is to study only one component of the transcriptome, such as long noncoding RNAs (lncRNAs), which constitute only a small fraction of transcripts in a total RNA sample [78–80]. A targeted quantitative RNA-seq method that is reproducible and reduces the number of sequencing reads required to measure key transcripts would be better suited to these purposes. Most recently, Tan et al. [80] describes a targeted enrichment method for the analysis of lncRNAs. Targeted RNA-seq can measure dozens to hundreds of targets simultaneously. Targeted RNA-seq gives an economical way to focus on genes of interest, and provides enhanced coverage for sensitive gene discovery, robust transcript assembly and accurate gene quantification. Common uses of this method include:

- Profiling expression of select target genes, to assess disease-associated variants and epigenetic alterations.
- Analyzing gene fusions and gene expression alterations to provide a focused view of functionally relevant changes occurring in cancer.
- Studying genes associated with a variety of key signal transduction pathways, such as NF-kB, P450, IFN response, apoptosis pathways, and many more.

http://dx.doi.org/10.5772/63267

There are several approaches for target RNA enrichment, either by hybridization only (Agilent SureSelect and Roche SeqCap system), hybridization followed by extension (Illumina Targeted RNA-seq; **Figure 6**) or PCR amplification (Thermo Fisher Ampliseq and BD Cellular Research Precise Assays). Genes of interest are enriched by custom-designed probes (**Figure 6**). However, these targeted sequencing procedures can potentially introduce biases caused by nonuniform hybridization as well as variation in amplification efficiencies across different genes and transcripts. According to our in-house pilot study (unpublished data), gene expression results are very sensitive to the choice of probes and targeted regions. Cellular Research Precise Assays [81] allow researchers to introduce so-called unique-sequence barcodes into the samples to overcome possible amplification bias and thus produces more accurate estimates of transcript abundance. By counting unique barcodes instead of reads, researchers can determine how many copies of each transcript are expressed with very high accuracy. In the meantime, targeted RNA-seq poses new challenges for data analysis. For example, most differential analysis packages assume the majority of genes are not differentially expressed, but this assumption is most likely not valid in targeted RNA-seq data sets.

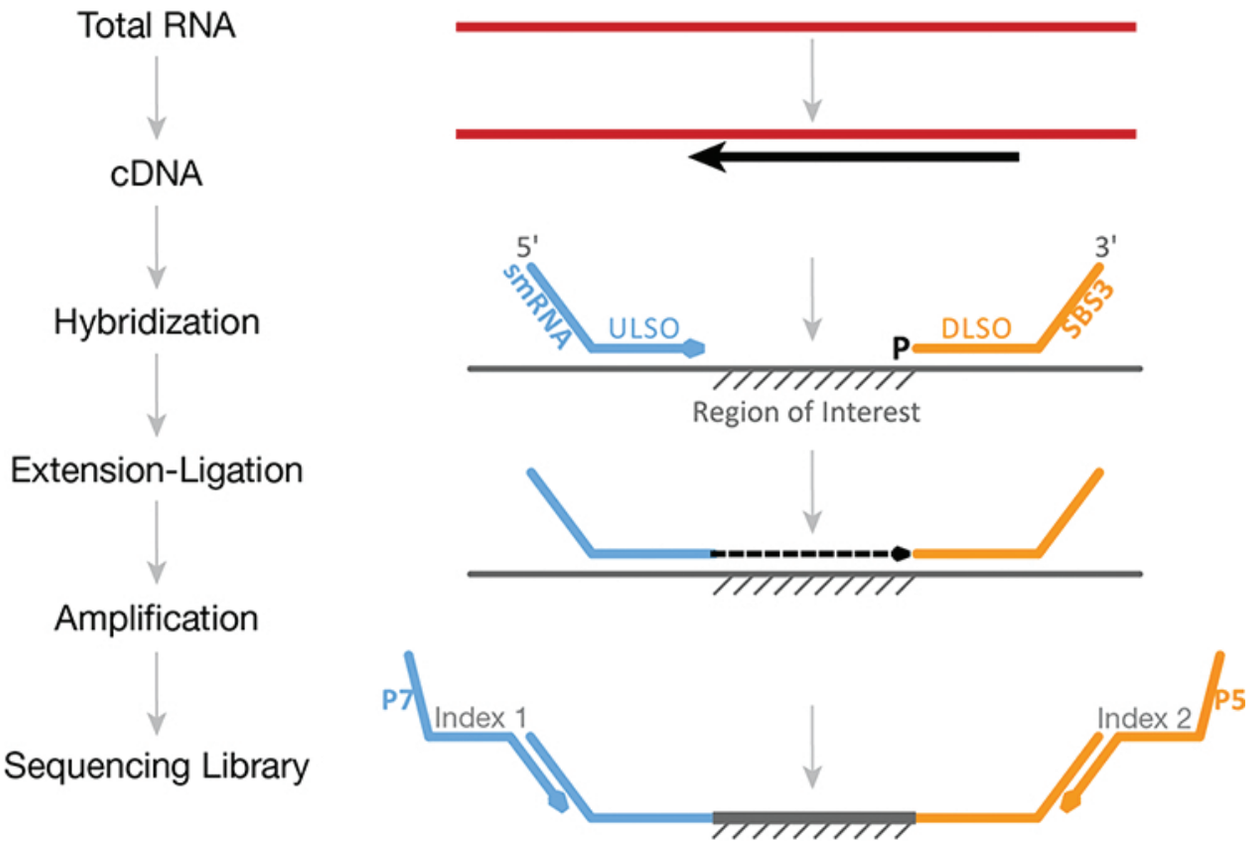

Figure 6. TruSeq targeted RNA expression workflow. Two custom-designed oligonucleotide probes with adapter sequences hybridize up and downstream of the region of interest. ULSO stands for upstream locus-specific oligo and DLSO stands for downstream locus-specific oligo.

5.3. Single-cell RNA-seq

Cell identity and function can be characterized at the molecular level by unique transcriptomic signatures. At the organismal level, different tissues have distinct gene expression profiles, and even cells in consecutive stages of embryonic development have highly divergent transcriptomic landscapes. Until recently, molecular 'fingerprints' were generated using profiling of gene expression levels from bulk populations of millions of input cells. These Ensemble-based approaches meant that the resulting expression value for each gene was an average of its expression levels across a large population of input cells. In many contexts, such bulk expression profiles are sufficient. However, there are also important questions for which bulk measures of gene expression are insufficient, for instance heterogeneity in immune cells. Besides, Ensemble measures do not provide insights into the stochastic nature of gene expression.

Recently, RNA-seq has achieved single-cell resolution, and scRNA-seq enables unbiased, high-throughput, and high-resolution transcriptomic analysis of individual cells [82, 83]. This provides an additional dimension to transcriptomic information relative to traditional methods that profile bulk populations of cells. What is currently still missing is an effective way to routinely isolate and process large numbers of individual cells for quantitative in-depth RNA-seq. Klein et al. [84] and Macosko et al. [85] have independently developed a high-throughput droplet-microfluidic approach for barcoding of RNA from thousands of individual cells for subsequent analysis by next-generation sequencing. Droplet-based scRNA-seq will be an attractive method for many laboratories because of its seemingly unlimited scalability and relatively low cost. By combining sophisticated RNA-seq technology with a new device that isolates single cells and their progeny, MIT researchers can now trace detailed family histories for several generations of cells descended from one "ancestor" [86].

Alongside the large-scale generation of single-cell transcriptomic data, it is important to consider the specific computational and analytical challenges that still have to be overcome [87]. Although some tools for bulk RNA-seq analysis can be readily applied to single-cell data, many new computational strategies are required to fully exploit this new data type. For instance, many genes at single cells are expressed in a stochastically-bursting fashion and their abundance exhibits a bimodal distribution in cell populations. Another main problem we face is that each cell can be in a different cell cycle phase, and they might vary in size and RNA content. The traditional RNA-seq data analysis does not take transcriptional bimodality into consideration, and implicitly assume the total RNA amount is the same or at least comparable. Additionally, compared with bulk RNA-seq, scRNA-seq data have much larger technical variations mainly because starting amount of RNA in a single cell is much lower and requires many more cycles of amplification. Recently, Korthauer et al. [88] have developed scDD, a differential analysis method particularly suited for scRNA-seq, to characterize differences in expression in the presence of distinct expression states within and among biological conditions. Using simulated and case study data, they demonstrate that the modeling framework is able to detect DE patterns of interest under a wide range of settings. Compared with existing approaches, scDD has higher power to detect subtle differences in gene expression distributions that are more complex than a mean shift, and is able to characterize those differences.

6. Concluding remarks

In much the same way that the advent of NGS technologies transformed our approaches in DNA-sequencing, RNA-seq approaches have dramatically changed our abilities to analyze the transcriptome of cells and tissues with a new level of detail and sensitivity. But we must also recognize that RNA-seq analysis is vulnerable to the general biases and errors inherent to next-generation sequencing (NGS) technology upon which it is based.

While RNA-seq technology is considered unbiased, it is important to note that the preparation and fragmentation of RNA and the library construction (which includes size selection) can introduce biases. Fragments are not uniformly sampled and sequenced, as there is variability in sequencing depth across the transcriptome due to preferential sites of fragmentation, variable primer, and tag nucleotide composition effects [89, 90]. RNA-seq is a complicated multistep process that involves sample collection and stabilization, RNA extraction, fragmentation, cDNA synthesis, adapter ligation, amplification, purification, and sequencing. Any step in this complex sequence of protocols can result in biased data.

Data normalization is one of the most crucial steps of RNA-seq data processing and has a profound effect on the results of the analysis. In practice, normalization of high-throughput data still remains an important topic and has received a lot of attention in the literature. The increasing number of normalization methods makes it difficult for scientists to decide which method should be used for which particular data set [58–66]. Even worse, different research groups draw contradictory conclusions. For example, Zyprych-Walczak et al. [66] concluded that the use of TMM method in most cases is displayed poorly, whereas the study by Dillies et al. [64] indicated that the use of TMM method led to good performance on simulated RNA-seq data sets. Considering the confusion and the numbers of methods available, we believe RNA-seq data normalization is a fundamental question that is far from being solved, and RNA-seq community will have to develop more sophisticated and robust algorithms to tackle this problem.

Additionally, for RNA-seq data collected with multiple sequencing platforms or at multiple sites, other normalization methods are required to remove site- or technology-specific effects [91]. Several methods for such normalization have been developed, including sva [92], RUV2 [93], and PEER [94], but they vary in their ability to remove these systematic biases [91]. As the cost of NGS continues to decrease, it is likely that additional studies will be conducted to readdress the same biological question. Therefore, there is an increasing need for analyzing data from multiple studies and different labs [95].

In this chapter, we focused on data analysis workflow for mRNA-seq, but have not covered RNA-seq experimental design. However, a crucial prerequisite for any successful RNA-seq study is a good experimental design, making sure the data generated have the potential to answer the biological questions of interest. For other aspects of RNA-seq data analysis, including experimental design, alternative splicing, gene fusion, and eQTL mapping, please refer to the most recent review by Conesa et al. [96]. In practice, RNA-seq is not used as a stand-alone technology platform in research. The integration of RNA-seq data with other types of

genome-wide data allows us to connect the regulation of gene expression with specific aspects of molecular physiology and functional genomics. Integrative analyses that incorporate RNA-seq data with other genomic or proteomic experiments are becoming increasingly prevalent. For instance, the combination of RNA and DNA sequencing can be used to explore RNA-editing or expression quantitative trait loci (eQTL) mapping. Pairwise DNA-methylation and RNA-seq integration can reveal the correlation between gene expression and methylation patterns. Additionally, integration of RNA-seq and miRNA-seq data has the potential to unravel the regulatory effects of miRNAs on transcript steady-state levels. All these integrative analyses pose additional challenges that are beyond the scope of this chapter.

Author details

Shanrong Zhao*, Baohong Zhang, Ying Zhang, William Gordon, Sarah Du, Theresa Paradis, Michael Vincent and David von Schack

*Address all correspondence to: shanrong.zhao@pfizer.com

PharmaTherapeutics Clinical R&D, Pfizer Worldwide Research and Development, Cambridge, MA, USA

References

[1] Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat Methods. 2008;5:621–628. DOI: 10.1038/nmeth.1226

[2] Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63. DOI: 10.1038/nrg2484

[3] Costa V, Angelini C, De Feis I, Ciccodicola A. Uncovering the complexity of transcriptomes with RNA-seq. J Biomed Biotechnol. 2010;2010:853916. DOI: 10.1155/2010/853916

[4] Mutz KO, Heilkenbrinker A, Lönne M, Walter JG, Stahl F. Transcriptome analysis using next-generation sequencing. Curr Opin Biotechnol. 2013;24:22–30. DOI: 10.1016/j.copbio.2012.09.004

[5] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nat Methods. 2011;8:469–477. DOI: 10.1038/nmeth.1613

[6] Capobianco E. RNA-seq data: a complexity journey. Comput Struct Biotechnol J. 2014; 11:123–130. DOI: 10.1016/j.csbj.2014.09.004

[7] Han Y, Gao S, Muegge K, Zhang W, Zhou B. Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights. 2015;9(Suppl 1):29–46. DOI: 10.4137/BBI.S28991

[8] Mills JD, Kawahara Y, Janitz M. Stranded RNA-seq provides greater resolution of transcriptome profiling. Curr Genomics. 2013;14:173–181. DOI: 10.2174/1389202911314030003

[9] Sigurgeirsson B, Emanuelsson O, Lundeberg J. Analysis of stranded information using an automated procedure for strand specific RNA sequencing. BMC Genomics. 2014;15:631. DOI: 10.1186/1471-2164-15-631

[10] Junker JP, van Oudenaarden A. Single-cell transcriptomics enters the age of mass production. Mol Cell. 2015;58:563–564. DOI: 10.1016/j.molcel.2015.05.019

[11] Saliba AE, Westermann AJ, Gorski SA, Vogel J. Single-cell RNA-seq: advances and future challenges. Nucleic Acids Res. 2014;42:8845–8860. DOI: 10.1093/nar/gku555

[12] Wilson NK, Kent DG, Buettner F, Shehata M, Macaulay IC, Calero-Nieto FJ et al. Combined single-cell functional and gene expression analysis resolves heterogeneity within stem cell populations. Cell Stem Cell. 2015;16:712–724. DOI: 10.1016/j.stem.2015.04.004

[13] Mantione KJ, Kream RM, Kuzelova H, Ptacek R, Raboch J, Samuel JM, Stefano GB. Comparing bioinformatic gene expression profiling methods: microarray and RNA-seq. Med Sci Monit Basic Res. 2014;20:138–142. DOI: 10.12659/MSMBR.892101

[14] Hurd PJ, Nelson CJ. Advantages of next-generation sequencing versus the microarray in epigenetic research. Brief Funct Genomic Proteomic. 2009;8:174–183. DOI: 10.1093/bfgp/elp013

[15] McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. RNA-seq: technical variability and sampling. BMC Genomics. 2011;12:293. DOI: 10.1186/1471-2164-12-293

[16] Malone J, Oliver B. Microarrays, deep sequencing and the true measure of the transcriptome. BMC Biol. 2011;9:34. DOI: 10.1186/1741-7007-9-34

[17] Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, et al. Estimating accuracy of RNA-seq and microarrays with proteomics. BMC Genomics 2009;10:161. DOI: 10.1186/1471-2164-10-161

[18] Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9:e78644. DOI: 10.1371/journal.pone.0078644

[19] Hangauer MJ, Vaughn IW, McManus MT. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. PLoS Genet. 2013;9:e1003569. DOI: 10.1371/journal.pgen.1003569

[20] Wang ET, Sandberg R, Luo S, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456:470–476. DOI: 10.1038/nature07509

[21] Griffith M, Griffith OL, Mwenifumbo J, et al. Alternative expression analysis by RNA sequencing. Nat Methods. 2010;7:843–847. DOI: 10.1038/nmeth.1503

[22] 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. DOI: 10.1093/nar/gkq202

[23] Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X et al. Transcriptome sequencing to detect gene fusions in cancer. Nature. 2009;458:97–101. DOI: 10.1038/nature07638

[24] Schmieder R, Edwards R. Quality control and preprocessing of metagenomic datasets. Bioinformatics. 2011;27:863–864. DOI: 10.1093/bioinformatics/btr026

[25] FastQC [Internet]. 2016. Available from: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed: 2016-02-23].

[26] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. J. 2011;17:10–12. DOI: 10.14806/ej.17.1.200

[27] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics. 2014;30:2114–2120. DOI: 10.1093/bioinformatics/btu170

[28] Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14:R36. DOI: 10.1186/gb-2013-14-4-r36

[29] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21. DOI: 10.1093/bioinformatics/bts635

[30] Dobin A, Gingeras TR. Mapping RNA-seq Reads with STAR. Curr Protoc Bioinformatics. 2015;51:11.14.1–11.14.19. DOI: 10.1002/0471250953.bi1114s51

[31] Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010;26:873–881. DOI: 10.1093/bioinformatics/btq057

[32] Hu J, Ge H, Newman M, Liu K. OSA: a fast and accurate alignment tool for RNA-seq. Bioinformatics. 2012;28:1933–1934. DOI: 10.1093/bioinformatics/bts294

[33] Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucl Acids Res. 2010;38:e178. DOI: 10.1093/nar/gkq622

[34] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods. 2013;10:1185–1191. DOI: 10.1038/nmeth.2722

[35] Wu P-Y, Phan JH, Wang MD. Assessing the impact of human genome annotation choice on RNA-seq expression estimates. BMC Bioinformatics. 2013;14:S8. DOI: 10.1186/1471-2105-14-S11-S8

[36] Chen G, Wang C, Shi L, Qu X, Chen J, Yang J, et al. Incorporating the human gene annotations in different databases significantly improved transcriptomic and genetic analyses. RNA. 2013;19:479–489. DOI: 10.1261/rna.037473.112

[37] Zhao S. Assessment of the impact of using a reference transcriptome in mapping short RNA-seq reads. PLoS One. 2014;9:e101374. DOI: 10.1371/journal.pone.0101374

[38] Zhao S, Zhang B. A comprehensive evaluation of Ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16:97. DOI: 10.1186/s12864-015-1308-8

[39] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucl Acids Res. 2007;35(Database):D61–D65. DOI: 10.1093/nar/gkl842

[40] Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC known genes. Bioinformatics. 2006;22:1036–1046. DOI: 10.1093/bioinformatics/btl048

[41] Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, et al. Ensembl 2014. Nucl Acids Res. 2014;42(Database issue):D749–D55. DOI: 10.1093/nar/gkt1196

[42] Dao P, Numanagić I, Lin YY, Hach F, Karakoc E, Donmez N, et al. ORMAN: optimal resolution of ambiguous RNA-seq multimappings in the presence of novel isoforms. Bioinformatics. 2014;30:644–651. DOI: 10.1093/bioinformatics/btt591

[43] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323. DOI: 10.1186/1471-2105-12-323

[44] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–515. DOI: 10.1038/nbt.1621

[45] Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-seq data. Algorithms Mol Biol. 2011;6:9. DOI: 10.1186/1748-7188-6-9

[46] Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–930. DOI: 10.1093/bioinformatics/btt656

[47] Anders S, Theodor P, Huber W. HTSeq—A Python framework to work with high-throughput sequencing data. Bioinformatics. 2014;31:166–169.DOI: 10.1093/bioinformatics/btu638

[48] Angelini C, De Canditiis D, De Feis I. Computational approaches for isoform detection and estimation: good and bad news. BMC Bioinformatics. 2014;15:135. DOI: 10.1186/1471-2105-15-135

[49] Kanitz A, Gypas F, Gruber AJ, Gruber AR, Martin G, Zavolan M. Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data. Genome Biol. 2015;16:150. DOI: 10.1186/s13059-015-0702-5

[50] Zhao S, Xi L, Zhang B. Union exon based approach for RNA-seq gene quantification: to be or not to be? PLoS One. 2015;10(11):e0141910. DOI: 10.1371/journal.pone.0141910

[51] Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11:R106. DOI: 10.1186/gb-2010-11-10-r106

[52] Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25. DOI: 10.1186/gb-2010-11-3-r25

[53] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–140. DOI: 10.1093/bioinformatics/btp616

[54] Cumbie JS, Kimbrel JA, Di Y, Schafer DW, Wilhelm LJ, Fox SE, et al. GENE-Counter: a computational pipeline for the analysis of RNA-seq data for gene expression differences. PLoS One. 2011;6:e25279. DOI: 10.1371/journal.pone.0025279

[55] Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A. Differential expression in RNA-seq: a matter of depth. Genome Res. 2011; 21:2213–2223. DOI: 10.1101/gr.124321.111

[56] Di Y, Schafer D, Cumbie J, Chang J. The NBP negative binomial model for assessing differential gene expression from RNA-seq. Stat Appl Genet Mol Biol. 2011;10: Article 24. DOI: 10.2202/1544-6115.1637

[57] Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL, Pachter L. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013;31:46–53. DOI: 10.1038/nbt.2450.

[58] 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. DOI: 10.1186/1471-2105-11-94

[59] Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. Am J Bot 2012;99:248–256. DOI: 10.3732/ajb.1100340

[60] Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-sequencing. BMC Genomics. 2012;13:484. DOI: 10.1186/1471-2164-13-484

[61] Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, Narayanan RK, Robinson GJ, Lundberg AE, Bartlett PF, Wray NR, Zhao QY. A comparative study of techniques for

differential expression analysis on RNA-seq data. PLoS One. 2014;9:e103207. DOI: 10.1371/journal.pone.0103207

[62] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;14:91. DOI: 10.1186/1471-2105-14-91

[63] Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, Mason CE, Socci ND, Betel D. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14:R95. DOI: 10.1186/gb-2013-14-9-r95

[64] Dillies MA, Rau A, Aubert J, Hennequet-Antier C, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform. 2013;14:671–683. DOI: 10.1093/bib/bbs046

[65] Seyednasrollah F, Laiho A, Elo LL. Comparison of software packages for detecting differential expression in RNA-seq studies. Brief Bioinformatics. 2015;16:59–70. DOI: 10.1093/bib/bbt086

[66] Zyprych-Walczak J, Szabelska A, Handschuh L, Górczak K, Klamecka K, Figlerowicz M, Siatkowski I. The Impact of normalization methods on RNA-seq data analysis. Biomed Res Int. 2015;2015:621690. DOI: 10.1155/2015/621690

[67] The Gene Ontology Consortium. Gene Ontology Consortium: going forward. Nucl Acids Res. 2015;43(Database issue):D1049–D1056. DOI: 10.1093/nar/gku1179

[68] Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. Nucl Acids Res. 2004;32(Database issue):D277–D2780. DOI: 10.1093/nar/gkh063

[69] Huang D, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4:47. DOI: 10.1038/nprot.2008.211

[70] Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. DOI: 10.1186/1471-2105-14-7

[71] Wang X, Cairns MJ. SeqGSEA: a Bioconductor package for gene set enrichment analysis of RNA-seq data integrating differential expression and splicing. Bioinformatics. 2014;30:1777–1779. DOI: 10.1093/bioinformatics/btu090

[72] Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim JS et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25:75–82. DOI: 10.1093/bioinformatics/btn577

[73] Gao S, Wang X. TAPPA: topological analysis of pathway phenotype association. Bioinformatics. 2007;23:3100–3102. DOI: 10.1093/bioinformatics/btm460

[74] Haynes WA, Higdon R, Stanberry L, Collins D, Kolker E. Differential expression analysis for pathways. PLoS Comput Biol. 2013;9:e1002967. DOI: 10.1371/journal.pcbi.1002967

[75] Zhao S, Xi L, Quan J, Xi H, Zhang Y, von Schack D, Vincent M, Zhang B. QuickRNASeq lifts large-scale RNA-seq data analyses to the next level of automation and interactive visualization. BMC Genomics. 2016;17:39. DOI: 10.1186/s12864-015-2356-9

[76] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A. Comprehensive comparative analysis of stranded RNA sequencing methods. Nat Methods. 2010;7:709–715. DOI: 10.1038/nmeth.1491

[77] Zhao S, Zhang Y, Gordon W, Quan J, Xi H, Du, S, von Schack D, Zhang B. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. BMC Genomics. 2015;16:487. DOI: 10.1186/s12864-015-1876-7

[78] Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL. Targeted RNA sequencing reveals the deep complexity of the human transcriptome. Nat Biotechnol. 2011;30:99–104. DOI: 10.1038/nbt.2024

[79] Mercer TR, Clark MB, Crawford J, Brunck ME, Gerhardt DJ, Taft RJ, Nielsen LK, Dinger ME, Mattick JS. Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat Protoc. 2014;9:989–1009. DOI: 10.1038/nprot.2014.058

[80] Tan JC, Bouriakov VD, Feng L, Richmond TA, Burgess D. Targeted lncRNA sequencing with the SeqCap RNA enrichment system. Methods Mol Biol.2016;1402:73–100. DOI: 10.1007/978-1-4939-3378-5_8

[81] Cellular Research Precise Assays [Internet]. 2016. Available from: http://www.cellularresearch.com/products/precise-assays.html [Accessed: 2016-02-23].

[82] Chattopadhyay PK, Gierahn TM, Roederer M, Love JC. Single-cell technologies for monitoring immune systems. Nat Immunol. 2014;15:128–135. DOI: 10.1038/ni.2796

[83] Avital G, Hashimshony T, Yanai I. Seeing is believing: new methods for in situ single-cell transcriptomics. Genome Biol. 2014;15:110. DOI: 10.1186/gb4169

[84] Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, Peshkin L, Weitz DA, Kirschner MW. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161:1187–1201. DOI: 10.1016/j.cell.2015.04.044

[85] Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161:1202–1214. DOI: 10.1016/j.cell.2015.05.002

[86] Kimmerling RJ, Lee Szeto G, Li JW, Genshaft AS, Kazer SW, Payer KR, de Riba Borrajo J, Blainey PC, Irvine DJ, Shalek AK, Manalis SR. A microfluidic platform enabling single-cell RNA-seq of multigenerational lineages. Nat Commun. 2016;7:10220. DOI: 10.1038/ncomms10220

[87] Stegle O, Teichmann SA, Marioni JC. Computational and analytical challenges in single-cell transcriptomics. Nat Rev Genet. 2015;16:133–145. DOI: 10.1038/nrg3833

[88] Korthauer KD, Chu LF, Newton MA, Li Y, Thomson J, Stewart R, Kendziorski C. scDD: a statistical approach for identifying differential distributions in single-cell RNA-seq experiments. bioRXiv. 2016 [Epub ahead of print]. DOI: 10.1101/035501

[89] Hansen KD, Brenner SE, Dudoit S. Biases in illumina transcriptome sequencing caused by random hexamer priming. Nucl Acids Res. 2010;38:e131. DOI: 10.1093/nar/gkq224

[90] Risso D, Schwartz K, Sherlock G, Dudoit S. GC-content normalization for RNA-seq data. BMC Bioinformatics. 2011;12:480. DOI: 10.1186/1471-2105-12-480

[91] Li S, łabaj PP, Zumbo P, Sykacek P, Shi W, Shi L, et al. Detecting and correcting systematic variation in large-scale RNA sequencing data. Nat Biotechnol. 2014;32:888–895. DOI: 10.1038/nbt.3000

[92] Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28:882–883. DOI: 10.1093/bioinformatics/bts034

[93] Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2013;13:539–552. DOI: 10.1093/biostatistics/kxr034

[94] Stegle O, Parts L, Durbin R, Winn J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. PLoS Comput Biol. 2010;6:e1000770. DOI: 10.1371/journal.pcbi.1000770

[95] Rau A, Marot G, Jaffrézic F. Differential meta-analysis of RNA-seq data from multiple studies. BMC Bioinformatics. 2014;15:91. DOI: 10.1186/1471-2105-15-91

[96] Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szcześniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A. A survey of best practices for RNA-seq data analysis. Genome Biol. 2016;17:13. DOI: 10.1186/s13059-016-0881-8

IntechOpen