Chapter 2

Practical Data Processing Approach for RNA Sequencing of Microorganisms

Toshitaka Kumagai and Masayuki Machida

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.69157

Abstract

The rapid evolvement of sequencing technology has generated huge amounts of DNA/RNA sequences, even with the continuous performance acceleration. Due to the wide variety of basic studies and applications derived from the huge number of species and the microorganism diversity, the targets to be sequenced are also expanding. The huge amounts of data generated by recently developed high-throughput sequencers have required highly efficient data analysis algorithms using recently developed high-performance computers. We have developed a highly accurate and cost-effective mapping strategy that includes the exclusion of unreliable base calls and correction of the reference sequence through provisional mapping of RNA sequencing reads. The use of mapping software tools, such as HISAT and STAR, precisely aligned RNA-Seq reads to the genome of a filamentous fungus considering exon-intron boundaries. The accuracy of the expression analysis through the refinement of gene models was achieved by the results of mapped RNA-Seq reads in combination with ab initio gene finding tools using generalized hidden Markov models (GHMMs). Visualization of the mapping results greatly helps evaluate and improve the entire analysis in terms of both wet experiment and data processing. We believe that at least a portion of our approach is useful and applicable to the analysis of any microorganism.

Keywords: RNA sequencing, computational analysis, microorganisms, gene modeling, alternative splicing

1. Introduction

RNA sequencing (RNA-Seq) is currently one of the most powerful methods for the comprehensive analysis of the transcriptional expression of the entire genes of a particular organism. Due to recent extreme improvements in sequencing technology in terms of throughput and cost, large amounts of data have been accumulated, and the amount of data is increasing in

an accelerating manner. Multiplexing by so-called bar coding facilitates the flexible utilization of the high output capacity of sequencers for large numbers of samples without a significant increase in the overall sequencing cost. This technical improvement greatly contributes to the application of RNA-Seq to various microorganisms.

The purposes of using RNA-Seq are basically divided into two categories. One of these objectives is counting the number of tags to analyze the intensity of gene expression, and the other is determining the transcript sequences for various purposes, such as annotating the genome of non-model organisms and analyzing splice variants.

In a typical RNA-Seq expression analysis, once sequence reads, which are generally 10^7–10^9 reads with a length of 50–300 bases, are accumulated, they are mapped to the reference sequence, namely, a genome sequence corresponding to the organism that the RNA is prepared from Refs. [1–3]. The mapping can be achieved using a sequence similarity search between the reads and the reference sequence with a general purpose computer. Although this procedure is highly suitable for current high-throughput computing (HTC) accelerated by parallel processing, the amount of sequence reads is too large to analyze the sequence similarity in a conventional manner, even using current high-throughput computers, due to the balance of costs between sequencing and data analysis. This issue is the most important when a large number of samples are obtained in a short period of time at low cost, which is often the case in research and development using microorganisms.

The DNA sequencers developed even with the most recent technologies cannot avoid errors in sequence reads. The RNA quality might be reduced by difficult sample preparation due to a small number of samples (cells) and low RNA extraction efficiency from cells grown under particular cultivation conditions. This effect might increase the sequence errors and reduce the amount of data obtained, further complicating the mapping. Although sample preparation might often be improved by finding better conditions and/or better methods for RNA preparation, optimization generally requires time and money. Thus, a bioinformatics method with higher accuracy, higher efficiency, and lower cost is desired based on the balance of time and cost between wet experiments and computational analyses. Accuracy is the most important factor, which increases the motivation to improve the sample and computational analysis qualities, but the necessary quality of sequence reads is often unknown.

The sequencers currently available include those manufactured by Illumina [1], Life Technologies [2], Pacific Bioscience [4], and Oxford [5], and these have different specifications in terms of the number of reads, read length, accuracy, and cost. The choice of platform depends on the purpose of the experiment. A search for genes that cause phenotypic differences under different culture conditions might require a search for differentially expressed genes (DEGs) with high sensitivity among the conditions, and a sequencing platform that yields a higher number of reads rather than longer read lengths should be selected. In contrast, revealing the complete transcribed sequence of a gene of a higher eukaryote that has various isoforms would require a platform that outputs long sequences.

In addition to the various characteristics and output data formats, because sequencing technologies and their performance are continuously under development, it is also necessary to maintain current knowledge of the progress of the methods and software used for analysis. The important issue in such a fast-paced world is to not treat methods and software as complete "black boxes" but to understand the type of information included in a file of a certain format and the statistical nature of the data being processed.

Nearly 10,000 complete microorganism genomes have been published to date according to GOLD [6], and the number is increasing in an accelerating manner. Therefore, a genome sequence used as a reference for a particular species of interest might be found in the database. However, the strain to be analyzed is often not exactly the same. Sequence variations between strains cause serious problems in mapping, similar to the problem due to sequencing errors, as described above. Even if the reference and the experimental sample are from the same strain, the sequences might have variations due to multiple rounds of cultivation and/or long-term storage without appropriate freezing conditions during the distribution process.

The quality of a reference sequence in terms of nucleotide assignment accuracy, length of contigs or scaffolds, assembling reliability (artificial assembling rearrangement), and gene modeling reliability also affects the reliability of RNA-Seq results. Nucleotide assignment errors cause issues similar to sequencing errors and the variation (mutation) problems described above. Low-quality reference sequences might cause problems when calculating the expression of each gene. One of the advantages of gene expression analysis by RNA-Seq is to obtain precise information regarding the location of the transcripts, e.g., an intron-exon boundary, without preparation of probes considering various possibilities in the case of DNA microarray. This advantage is highly advantageous for the expression analysis of microorganisms for which no genomic information has been accumulated.

Although sequencing topics derived from sequencing platforms (chemistry, base calling method, hardware, etc.) and assembling are not addressed in this chapter, gene modeling, which defines CDSs (from coding DNA sequences), will be discussed because (i) RNA-Seq includes information that is important for correcting gene models and (ii) the calculation of expression levels from RNA-Seq depends on the gene model.

2. Factors affecting accuracy and efficiency

2.1. Quality control of sequence reads

If a reference sequence is available, a computational RNA-Seq analysis typically consists of mapping to the corresponding reference sequence and successive processes. The processes of removing unreliable reads and trimming unreliable segments of the reads are often applied without much consideration. Excluding bases with a lower quality score from the RNA-Seq reads improves the average quality score of the reads, which clearly improves the quality of the reads from the left to the right panel, as shown in **Figure 1A**. The upper panel of **Figure 1B**

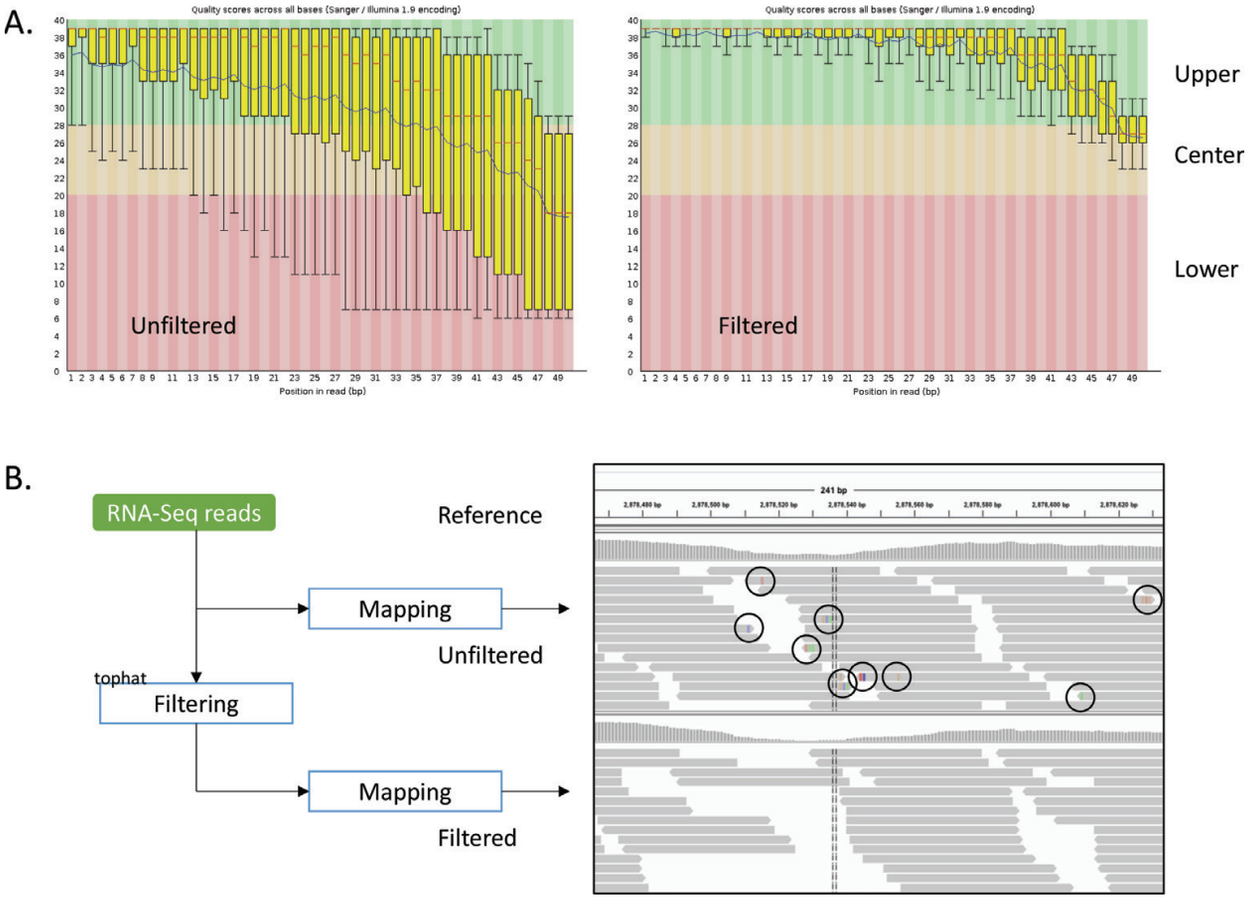

Figure 1. (A) Filtering of reads using quality values. The *Escherichia coli* genome was sequenced using a SOLiD 5500xl sequencer with a 50-bp read length and generated 5,869,272 reads. Quality score distribution of unfiltered (left) and filtered (right) reads visualized by FastQC (see **Table 1** for reference). The average quality values for each sequence position are indicated by a thin curved line. The right panel was obtained by the application of bases with a quality value ≥ 20 for more than or equal to 95%. "N" is less than or equal to 1. The number of reads after filtering was reduced to 2,697,082. For each position, a Box-Whisker-type plot, in which the central red line, yellow box, upper and lower whiskers, and blue line represent the median value, interquartile range (25–75%), 10 and 90% points, and mean quality, respectively. The *Y*-axis on the graph shows the quality scores. A higher score reflects better base call. The background of the graph divides the Y-axis into high (Upper), moderate (Center), and poor (Lower) quality calls. (B) Effect of filtering sequence reads. The sequence reads obtained before and after filtering, as indicated in A, were mapped to the reference genome and visualized. Mismatches are indicated by black circles.

shows a mapping result using unfiltered reads with the quality shown in the left panel of **Figure 1A**, indicating the presence of a significant number of bases mismatched to the reference sequence. However, using the reads in the right panel in **Figure 1A**, the mismatches are significantly decreased, as shown in the lower panel of **Figure 1B**. The filtering process requires only a relatively small calculation time but is thought to significantly improve reliability, which solves various problems derived from mismatches between reads and the reference.

Williams et al. showed that in RNA-Seq experiments, read trimming prior to mapping might have a substantial effect on the estimation of the gene expression level [7]. Therefore, if trimming is applied, extreme care should be taken, and other measures, such as length filtering, should be considered in the preprocessing pipeline to minimize the introduction of unwanted

Name	Category	Brief description[1]	Ref.	Link
FastQC	Quality control for raw reads	Providing a QC report to spot problems which originate either in the sequencer or in the starting library material.		https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
Sickle	Reads trimming	Detection and trimming low quality region from all reads using sliding window.		https://github.com/najoshi/sickle
Cut adapt	Reads trimming	Searching for and removing adapters in all reads.		http://cutadapt.readthedocs.io/en/stable/index.html
BWA	Mapping	Mapping reads against a large reference genome sequence.		http://bio-bwa.sourceforge.net/
TopHat2	Mapping	A splice junction mapper for RNA-Seq reads.	[12]	https://ccb.jhu.edu/software/tophat/index.shtml
HISAT2	Mapping	A spliced alignment program, a successor to TopHat2.	[13]	http://www.ccb.jhu.edu/software/hisat/index.shtml
STAR	Mapping	Spliced transcripts alignment to a reference.	[14]	https://github.com/alexdobin/STAR
Cufflinks	Transcripto me assembly, etc.[2]	Assembling of transcripts, estimation of their abundances, and testing for differential expression and regulation in RNA-Seq samples.	[30]	http://cole-trapnell-lab.github.io/cufflinks/
Kallisto	Quantificati on of gene expression	Quantification of abundances of transcripts from RNA-Seq data based on the novel idea of pseudo-alignment for rapidly determining the compatibility of reads with targets, without the need for alignment.	[16]	https://pachterlab.github.io/kallisto/
Salmon	Quantificati on of gene expression	Quantification of the expression of transcripts using RNA-seq data using new algorithms (quasi-mapping) to provide accurate expression estimates with high throughput and little memory.	[17]	https://combine-lab.github.io/salmon/
VarScan2	Variant call	A mutation caller for targeted, exome, and whole-genome resequencing data.	[11]	http://dkoboldt.github.io/varscan/
AUGUSTUS	Gene finding	Prediction of genes in eukaryotic genomic sequences using extrinsic information as hints on the gene structure.	[18]	http://bioinf.uni-greifswald.de/augustus/
BRAKER1	Gene finding	A pipeline for unsupervised RNA-Seq-based genome annotation.	[19]	http://exon. gatech.edu/braker1.html
CodingQuarry	Gene finding	A self-training gene predicting tool dedicated to fungal genome working with assembled, aligned RNA-seq transcripts.	[20]	https://sourceforge.net/projects/codingquarry/
Tablet	Genome viewer	A graphical viewer for next generation sequence assemblies and alignments.	[36]	https://ics.hutton.ac.uk/tablet/

Name	Category	Brief description[1]	Ref.	Link
Artemis	Genome viewer	A genome browser and annotation tool that allows visualization of sequence features, next generation data.	[37]	http://www.sanger.ac.uk/science/tools/artemis
IGV	Genome viewer	Interactive exploration of genomic datasets supporting various data types, including array-based and next-generation sequence data, and genomic annotations.	[38, 39]	http://software.broadinstitute.org/software/igv/
CLC Genomics Workbench	Integrated solutions	Integrated package of software tools for genomic analysis and visualization supporting various data types, including array-based and next-generation sequence data, and genomic annotations.		https://www.qiagenbioinformatics.com/products/clc-genomics-workbench
Genome Traveler	Integrated solutions	Integrated package of software tools for genomic analysis and visualization supporting various data types, including array-based and next-generation sequence data, and genomic annotations.		http://www.insilicobiology.jp/index.php?option=com_content&view=article&id= 107&Itemid=73&lang=en

[1]Functions related to the topics in this chapter are briefly summarized. Reading the references and/or accessing the web sites is required for details and other functions especially for the integrated package of software.

[2]Transcriptome assembly, quantification of gene expression and testing differential expression genes.

Table 1. Overview of software tools for transcriptome analysis.

bias. In our follow-up examination of the reads obtained using an Illumina MiSeq platform, we concluded that for relatively long sequencing reads, such as 100 or 150 bases, with low sequence errors, aggressive trimming of sequencing reads is generally no longer necessary for estimating the gene expression level. In the following section, we propose correction of the reference sequence using RNA-Seq reads in cases in which the genome sequence of the same strain used in the RNA-Seq experiment is not available to avoid mismatches between the RNA-Seq reads and the reference. The removal and trimming of unreliable sequences are necessary for this purpose.

2.2. Pipelines and peripheral tools

Figure 2A shows a typical pipeline for analyzing gene expression based on RNA-Seq reads. The pipeline effectively works for microorganisms, genome sequences, and gene models, which are reliable due to significant correction and curation by the efforts of a large number of researchers. Typical examples of such microorganism are *Escherichia coli*, *Bacillus subtilis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Neurospora crassa*, and *Aspergillus nidulans*, which are known as model organisms. Among microorganisms, filamentous fungi generally have the largest genome sizes and introns in most existing genes and are thus thought to require a pipeline with the highest performance and various functions for the analyses. Furthermore, filamentous fungi are potential producers of various

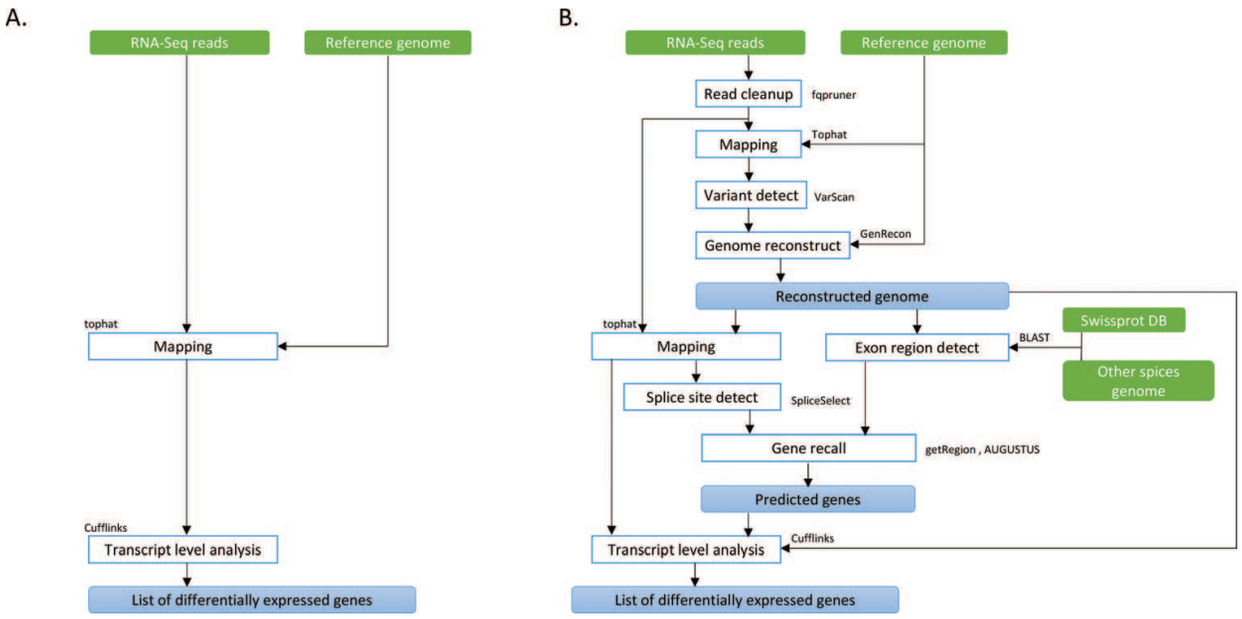

Figure 2. Example of the RNA-Seq analysis pipeline. (A) Typical simple pipeline. First, RNA-Seq reads are mapped to a reference sequence using a mapping tool such as TopHat. Next, tools such as Cufflinks count the number of reads mapped to each genomic feature and extract differentially expressed genes (DEGs). (B) Proposed pipeline for microorganisms whose reference sequence and gene models are not extensively corrected or curated. Fqpruner, GenRecon, SpliceSelect, and getRegion in the figure are in-house scripts. Fqpruner is a program written in C++ to trim the 3′-end of low-quality reads and has almost the same function as the combination of sickle (see **Table 1** for reference) and Cutadapt [10]. GenRecon is a Perl script that outputs a consensus sequence based on output by the variant detect tool, VarScan [11]. SpliceSelect is a Perl script that integrates splice site positions from multiple TopHat output files. GetRegion is a Perl script that receives the BLAST results and outputs genomic regions to execute AUGUSTUS for the prediction of genes from genomic loci involving homologous genes with known amino acid sequences.

secondary metabolites, which are economically important and have a large number of highly diverse secondary metabolism-related genes. Thus, the genomes of filamentous fungi and actinomycetes remain attractive targets in this field. To effectively and accurately analyze RNA-Seq reads from filamentous fungi without publicly available genomic information, we have developed several tools and introduced into the pipeline, as shown in **Figure 2B**.

RNA-Seq reads can be analyzed without their corresponding genome sequence as a reference through the de novo assembly of the reads. Long-read technologies, such as PacBio RS II (Pacific Bioscience) and MinION (Oxford Nanopore Technologies), should lead to better results than sequencers that generate short reads using this approach. However, we do not include the de novo assembly of RNA-Seq reads in this chapter because sequencing the genome of a microorganism using next-generation sequencers, such as Illumina technology, is relatively inexpensive in terms of cost and time. For example, we have used the improved pipeline for the analysis of the genome sequences obtained from a short-read sequencer, SOLiD 500xl, in combination with the de novo assembly pipeline that the manufacturer developed for mate-paired sequences [8] with successive automatic annotation. Illumina and Life Technologies platforms, such as HiSeq/MiSeq/NextSeq and Ion Torrent/Ion PGM, respectively, might also generate a reference genome sequence that is adequate for this purpose in an easy and cost-effective manner. Based on the assumption that the genome sequence is

available as a reference for the microorganism, the strategy of mapping the transcriptome is not included in this chapter.

The sequencing platforms described above are widely used, and bioinformatics tools have been extensively developed for each platform. The characteristics of the errors depend on the sequencing platform, such as those manufactured by Illumina, Life Technologies, and Pacific Bioscience. The number of reads, read length, and data format also varies by platform. Furthermore, more than one platform, such as a combination of Illumina and Pacific Bioscience or Life Technologies and Illumina [9], might be used, which also requires a specific methodology for obtaining reasonable results.

2.3. Basic mapping problems

The mapping of RNA-Seq reads to the reference genome has been a serious problem in RNA-Seq analysis due to the extremely large data size (e.g., more than 500 Gb are obtained from a single run of HiSeq 2500) and sequence errors in both the RNA-Seq reads themselves and the reference. Most of the mapping tools search the nucleotide sequences with a similarity greater than a certain threshold value in the reference sequence for each RNA-Seq read. Multiple mapping algorithms are widely used to accurately identify the most homologous positions on the reference sequence. However, a shorter read length than the repetitive elements in the reference sequence and sequencing errors complicates the problem.

A typical RNA-Seq experiment consists of the sequencing of both ends of a cDNA fragment to generate two reads (a read pair) separated by a sequence of variable length. The accurate alignment of these read pairs is essential to the downstream analysis of an RNA-Seq experiment, but RNA-Seq read alignment is challenging due to the noncontiguous nature of mRNA transcripts resulting from the existence of introns in eukaryotic genes. Recently developed mapping tools, such as TopHat [12], STAR [13], and HISAT [14], perform spliced alignment by considering an exon-intron boundary for the RNA-Seq reads. Software programs that support splice alignment use different strategies from several perspectives [15]. The method of determining the position on the reference sequence where a read is mapped can be roughly classified into two groups: exon first and seed and extend.

Exon-first methods, such as TopHat, utilize a two-step process. First, they map reads to the reference sequence without allowing large gaps. Subsequently, the unmapped reads are divided into short segments, and each is independently aligned to the reference sequence. The discontinued region on the genome where contiguous segments are mapped is treated as a candidate of two connected exons obtained by splice alignment. The exon-first approach is the most effective in cases in which a majority of the reads can be mapped without gaps. If retrotransposed genes or pseudogenes originating from transcripts with multiple exons are present in the genome sequence, software that employs the exon-first approach might preferentially map the reads to the retrotransposed region. In seed-and-extend methods, such as STAR, reads are divided into short seeds (k-mers), the positions where they are present in the genome are searched, and alignments are built and extended using this information.

Seed-and-extend methods are generally considered more sensitive but slower than exon-first methods. However, with great efforts, excellent software programs using seed-and-extend or hybrid methods have been developed in recent years. Substantial effort has been spared, and software using the seed-and-extend method has become sufficiently fast. In a typical expression analysis of microorganisms using RNA-Seq, the computational processing time required for mapping reads to the reference genome sequence is no longer a major problem.

For transcript quantification, software such as Kallisto [16] and Salmon [17], which use newer algorithms that do not require the pre-mapping of reads to a reference sequence, has become increasingly faster. A very large-scale expression analysis with RNA-Seq could be performed using this type of software.

2.4. Mapping problems caused by mutation

Our analysis of RNA-Seq data from *S. cerevisiae* encountered another type of problem, which derived from the accumulation of mutations in the genome. Widely distributed strains, such as *S. cerevisiae* BY4741 and W303, can undergo a large number of mutations possibly during the distribution process due to relatively long-term storage without freezing and multiple rounds of inoculation and successive cultivation. The mutation frequency can be decreased by careful handling, such as decreasing the number of inoculation processes and avoiding stressful conditions. However, the introduction of mutations cannot be completely prevented due to spontaneous mutation, which is a natural characteristic of all organisms. The basic procedure for resolving this problem is to sequence the genome of the strain for which RNA-Seq is performed. However, because the sequencing strategy, including sample preparation, for genome sequencing is different from that used for RNA-Seq and because of the cost- and time-saving requirements, RNA-Seq data sometime have to be analyzed using the reference sequence deposited in a public database. To overcome this problem without losing reliability, we have addressed the correction of the reference sequence using RNA-Seq reads based on two methods: (1) RNA-Seq reads are mapped to the reference sequence using the spliced mapper mentioned in the previous section, and the reference sequence is corrected using the consensus of the mapped reads. (2) The de novo transcriptome assembly of RNA-Seq reads is aligned to the reference genome. The former method was almost completely automatable and worked well for small variations, such as single-base substitution. With the latter method, it was necessary to process a number of isoform candidates at the same loci of the reference genome outputted by the transcriptome assembler, which required time and effort to tune the various parameters and threshold values. Unless the genome has undergone a complicated structural change from the reference sequence, the former method is sufficient. After correcting the reference sequence, the reads were again mapped to the corrected reference sequence. This strategy worked fairly well.

2.5. Gene finding using RNA-Seq

Typical examples of the gene modeling problem are found by analyzing filamentous fungi. Industrially important fungi are often isolated due to their production of useful secondary

metabolites. Because their genomes are generally unknown, sequencing and successive gene modeling are indispensable but are performed by a limited number of researchers with a limited amount of knowledge. In such cases, RNA-Seq reads can be used to correct gene models prior to expression analysis to obtain accurate expression levels.

Several researchers have attempted to improve the accuracy of predicting protein-coding genes, and these attempts have included the use of RNA-Seq. AUGUSTUS is a gene prediction program that uses a generalized hidden Markov model (GHMM) [18], which is widely used for eukaryote genome sequencing projects. AUGUSTUS can incorporate hints of the gene structure from extrinsic sources. After RNA-Seq reads are mapped to the genome, spliced mapped reads can be used as valuable information for gene finding.

In recent years, gene prediction software using RNA-Seq for both model training and gene prediction with the trained model has been developed and has demonstrated high accuracy for gene structure prediction [19, 20]. The training of conventional gene finding depends on the gene models in the genomes of species other than the target one. However, the gene models of the species already deposited in public databases have not always been experimentally confirmed but are the results of predictions based on the results of other genomes. Thus, the use of the results of RNA-Seq read mapping, which provides direct information of the CDSs of the target species, in combination with recent gene finding algorithms, enables significant improvement in gene modeling.

We used an internally developed pipeline that performs training with RNA-Seq read mapping and ab initio gene prediction (**Figure 2B**). In this pipeline, exon-intron boundary information is predicted using mapped RNA-Seq, and coding sequence candidates is obtained by homology searches between the genome sequence and protein sequence databases, such as the Swiss-Prot database. Subsequently, AUGUSTUS was trained using these pieces of information, and all of the genes in the genome were predicted. This pipeline worked well for gene prediction of non-model organisms and has been used for the genome analysis of various filamentous fungi. The improvements in the predicted gene structures are thought to contribute to more accurate RNA-Seq expression quantification as transcript references.

In the case of bacteria, which do not have poly-A tails, the degradation of ribosomal RNA is required for the extraction of mRNA. Because the degradation will not be complete, the ribosomal RNA sequences have to be removed after sequencing by searching the consensus sequence in the reads. Another problem is that bacterial genes are sometimes overlapped on the genome and might be transcribed even in different orientations. This can be problematic for identifying CDSs based on the RNA-Seq mapping results. To solve this problem, strand-specific RNA-Seq has the advantage of obtaining useful information for gene modeling. However, because bacterial mRNA does not have poly-A tails, as described above, preparation of a strand-specific library is more difficult than the preparation of eukaryotic mRNA. A strand-specific library for bacteria can be prepared basically by two methods [21]: (i) adapter ligation to the first strand synthesized in the cDNA preparation [22] and (ii) chemical modification of RNA or the second strand of the cDNA [23–25].

2.6. Quantification of gene expression and identification of differentially expressed genes

Expression analysis with RNA-Seq typically begins by counting the number of reads mapped to reference transcript sequences. We can resolve the various mapping problems mentioned above and perform mapping to the genome with accurately predicted gene structures or assembled transcript sequences using transcriptome assembly software.

Microarrays are widely used for the quantification of the abundance of mRNAs corresponding to genes. In microarray experiments, the gene expression level is measured as a continuous value, intensity. RNA-Seq differs from microarrays in that it addresses nonnegative discrete values, i.e., the number of reads mapped to the gene, in order to measure the expression of a gene. Analytical methods for microarray data that assume a Gaussian distribution, such as linear discriminant analysis, might not perform as well for RNA-Seq data with a discrete distribution.

Let us consider the problem of quantifying gene expression levels using discrete RNA-Seq data and a related problem, namely, the identification of differentially expressed genes (DEGs) between conditions. In RNA-Seq experiments, transcribed mRNA is fragmented into a certain length, cDNA is subsequently synthesized, and sequencing is performed. Thus, the total number of observed reads for a transcript is proportional to the number of expressed mRNAs for the transcript multiplied by the length of the transcript. To compensate for this bias, it is a common practice to divide the number of mapped reads by the transcript length. RPKM (Reads Per Kilobase transcript per Million mapped reads) is the most commonly used method for length and sample size normalization.

Unfortunately, this correction is not sufficient to test whether gene expression differs between conditions. Oshlack and Wakefield showed that the power of a *t*-test of the count data, regardless of whether it is divided by the length of the transcript, is proportional to the square root of the length of the transcript [26]. Therefore, for a given expression level, the test becomes more significant for longer transcripts.

Many methods have been developed for assessing differential expression from RNA-Seq data. Count data, such as the counts of mapped fragments of RNA-Seq data, are often modeled as a Poisson distribution. The Poisson distribution has equal mean and variance values, and DEGs can be identified by conducting a likelihood ratio test between conditions. Real RNA-Seq data often exhibits overdispersion. The count data measured via RNA-Seq often has a variance that is larger than the mean due to various biases and errors as well as length bias. A negative binomial distribution is widely used for modeling such cases. Several RNA-Seq data analysis software packages incorporating these models have been developed. Soneson and Delorenzi evaluated eleven software packages that implemented various methods to model count data for differential expression analyses of RNA-Seq data [27]. When designing experiments to analyze differential expressions using RNA-Seq, it is necessary to carefully consider the type of method used for DEG extraction and the amount of biological replications that are needed. Three replicates often give reproducible results in successive independent experiments in

terms of the assignment of a gene(s) with the expression of interest, although a single experiment often fails to yield reproducible results.

The comparison of the transcriptome for each condition often shows a large number of DEGs. Therefore, outlining the changes in the expression profile by extracting features common to genes whose expression intensity has changed is a common approach. Gene set enrichment analysis (GSEA) is a popular method for condensing information from gene expression profiles into a summary of pathways or functional groups. GSEA was developed for microarray data and can also be used for RNA-Seq data. However, most RNA-Seq data obtained so far have only small replicates, which enforces application of the gene-permuting GSEA method (or preranked GSEA), resulting in a great number of false positives due to the inter-gene correlation in each gene set. Yoon et al. demonstrated that the incorporation of the absolute gene statistic in one-tailed GSEA considerably improves the false-positive control and the overall discriminatory ability of the gene-permuting GSEA methods for RNA-Seq data [28].

2.7. Alternative splicing

As shown recently, RNA-Seq also enables the detection of alternative slicing from various fungi and higher organisms, such as mammals and plants. Alternative splicing from RNA-Seq can also be performed using bioinformatics software, such as GESS (graph-based exon-skipping scanner) [29] and Cufflinks [30]. Both tools can detect isoforms of transcripts based on mapping information generated by TopHat using a graph-based method. The former outputs all isoforms detected in the GTS format and requires MISO [31] to calculate the RPKM values for each isoform, whereas the latter is able to calculate the values. These tools are widely used for the analysis of higher organisms, such as mammals and plants, but not fungi.

Splicing variants have been found in various fungi, including *Aspergillus oryzae* [32], *Magnaporthe grisea* [33], *Cryptococcus neoformans* [34], and *Trichoderma longibrachiatum* [35], by deep RNA-Seq despite their significantly lower frequency compared with that found in higher organisms. Alternative splicing might affect the calculation of the FPKM (Fragments Per Kilobase of exon per Million fragments mapped)/RPKM values; however, because of the relatively low frequency (less than 10% of the entire genes on a genome) and abundance of "intron retention" [35], the results might not be significant without specific measures. Isoforms might also be detected through an inaccurate mapping of RNA-Seq reads resulting from base call errors and incorrect exon-intron boundaries. Thus, the calculation of RPKM values for the entire CDSs could be performed, particularly for the initial analysis.

2.8. Visualization and evaluation of the analysis

Visualization of RNA-Seq results is useful and strongly recommended during the analysis process for a rapid evaluation of the reliability of the analysis. Typical views of the results, including mapping, models, and nucleotide sequences, are shown in **Figure 3B** using Genome Traveler/in silico Molecular Cloning (GT/IMC) available from in silico biology, Inc. Various software tools, such as Tablet [36], Artemis [37], Integrative Genome Viewer (IGV) [38, 39], and CLC Genomics Workbench, were developed by the James Hutton Institute, Sanger Institute, Broad Institute, and CLC Bio, respectively. Some of these tools are operating system

http://dx.doi.org/10.5772/intechopen.69157

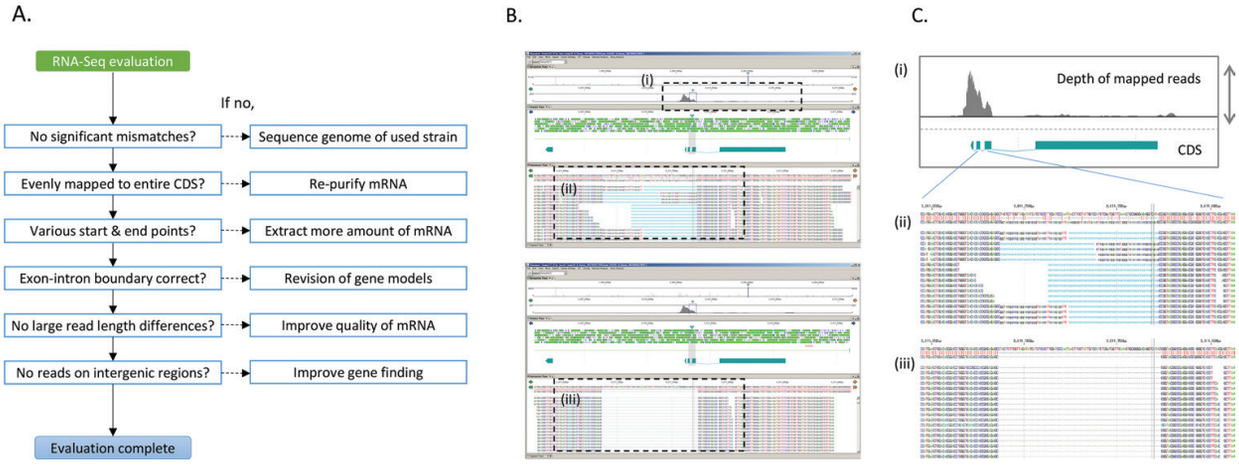

Figure 3. Visualization of the results. The RNA-Seq reads of *Aspergillus flavus* NRRL3357 (NCBI BioProject Accession: PRJNA299060) were mapped to the corresponding reference genome sequence with annotations (http://genome.jgi.doe.gov/pages/dynamicOrganismDownload.jsf?organism=Aspfl1) analyzed by the Joint Genome Institute (JGI) [41]. Genome Traveler (GT) from in silico biology is used to visualize the read mapping, gene models, and nucleotide sequences in a single window. (A) Schematic diagram of RNA-Seq analysis evaluation. (B) Mapping result using BWA (upper) and HISAT (lower). Each panel shows the depth of the RNA-Seq reads (top panel), a gene model (middle panel), and the nucleotide sequences of the mapped reads (bottom panel). The top of the middle panel shows termination codons in six frames with vertical lines and relatively long ORFs with solid rectangles. The bottom of the middle panel shows the predicted exons. The width of the bottom panel corresponds to the region indicated by brackets with triangles in the top and middle panels. (C) Magnified version of the regions indicated by the dotted rectangle in (B).

specific, but the others are executable on multiple platforms, e.g., by using Java. Because recent sequencing platforms output huge amounts of data, the operating system should be 64 bits with a memory size of at least 16 Gbytes. Differently from de novo assembly for genome sequencing, mapping requires less memory and lower CPU performance. Introduction of a small-scale server equipped with eight CPUs and 32 Gbytes of memory might help reduce the required time with a relatively low cost. The sequencing quality can also be validated by the read lengths and their variation, particularly when the reads are trimmed based on the quality.

Figure 3A presents a schematic diagram of how RNA-Seq analysis is achieved in combination with visual evaluation. The read mapping and alignment are displayed as shown in **Figure 3B and C**. When the reads have sufficient quality for the subsequent analyses, they are aligned without a significant number of mismatches. The read lengths aligned to the reference might sometimes have large differences in length, even after using a platform of fixed read length, such as Illumina and SOLiD. This effect occurs due to the low-quality values of the nucleotides at the end of a read sequence, which are removed by a trimming process, as discussed above. High-quality reads have nearly the maximum or indicated read length of the sequencing platform used. It is important that each read does not have the same starting and ending positions on the reference to confirm that excess PCR amplification, which often occurs when the RNA quality is low, was not applied.

Another important indicator of experimental quality is the depth of reads inside CDSs. High AT or GC proportions, such as 70% and greater, in a particular region might cause a lower depth of coverage depending on the sequencing platform due to insufficient amplification during emulsion PCR. The depth of the reads should be roughly the same throughout the entire CDS.

Deeper coverage at the 3′ end than at the 5′ end indicates low mRNA quality, probably due to partial degradation, when poly-A-tailed RNA capture is applied in the preparation process.

In the case of fungi, introns might not be clearly displayed by a simple mapping approach without considering the exon-intron boundary because of the short intron length (typically in the range of 5–100 nt), even when using short reads of 50 bp. The predicted CDS at the center of **Figure 3B** and **C** shows two short exons close to the 5′-end. Mapping by BWA [40], which does not consider the intron-exon boundary, aligned some reads to the intron, introducing mismatches (the upper panel of **Figure 3B** and **C—(ii)**). By referring to the mismatches between the reference and the consensus of the mapped reads, the location of the intron can be assumed to be the region where gray asterisks instead of red vertical bars are clustered at the top of the bottom panel. In contrast, read mapping using HISAT2 (the lower panel of **Figure 3B** and **C—(iii)**) and STAR (data not shown), both of which consider the intron-exon boundary, fairly accurately mapped the reads connecting two adjacent exons, introducing an intron between the exons.

The above CDS has another long intron-predicted upstream of the two short introns mentioned above, although this third intron might be too long for a gene from a filamentous fungus. Furthermore, the depth of reads for the first exon is much lower than those for the second and third exons (**Figure 3C—(i)**). Considering the precipitous change in depths between the first and second exons and the almost even distribution of the depth in the first exon despite its large size, the large difference in depth is not thought to result from partial mRNA degradation. Consequently, it is believed that the first exon should be separated from the other exons, resulting in two CDSs. In agreement with this consideration, RNA-Seq reads are also mapped to the region of the long intron with a depth similar to that of the first exon (the upstream part of the two CDSs after division) after a short intron is detected by HISAT2 (data not shown).

2.9. Perspective

Recently developed long-read sequencers, such as PacBio RS II, PacBio Sequel, and Oxford Nanopore MinION, promise to deliver more complete genome assemblies with fewer gaps. Higher error rates, low yields per cost, and stringent DNA requirements might be concerns. Short-read sequencers have an advantage for measuring transcriptional expression due to the production of a greater number of reads. In contrast, long-read sequencers have the potential to accurately analyze the structure of transcripts, including the linkage between multiple splicing variations [42]. The selection and combination of appropriate bioinformatics tools as well as sequencing platforms should be a key issue depending on the purpose of the analysis.

Acknowledgements

This work was supported by the commission for the Development of Artificial Gene Synthesis Technology for Creating Innovative Biomaterial from the Ministry of Economy, Trade and Industry (METI), Japan. This work was also supported by the project focused on developing key technology of discovering and manufacturing drug for the next-generation treatment and

diagnosis from the Ministry of Economy, Trade and Industry (METI) and the Japan Agency for Medical Research and Development (AMED). We thank the American Journal Experts 479 for proofing the manuscript.

Author details

Toshitaka Kumagai and Masayuki Machida*

*Address all correspondence to: m.machida@aist.go.jp

Fermlab Inc., National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

References

[1] Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008;**456**(7218):53-59

[2] Perkel J. Making contact with sequencing's fourth generation. Biotechniques. 2011;**50** (2):93-95

[3] Cloonan N, Forrest AR, Kolle G, Gardiner BB, Faulkner GJ, Brown MK, et al. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 2008;**5**(7):613-619

[4] Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;**323**(5910):133-138

[5] Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: Delivery of nanopore sequencing to the genomics community. Genome Biology. 2016;**17**(1):239

[6] Mukherjee S, Stamatis D, Bertsch J, Ovchinnikova G, Verezemska O, Isbandi M, et al. Genomes OnLine Database (GOLD) v.6: Data updates and feature enhancements. Nucleic Acids Research. 2017;**45**(D1):D446-D456

[7] Williams CR, Baccarella A, Parrish JZ, Kim CC. Trimming of sequence reads alters RNA-Seq gene expression estimates. BMC Bioinformatics 2016;**17**:103

[8] Umemura M, Koyama Y, Takeda I, Hagiwara H, Ikegami T, Koike H, et al. Fine de novo sequencing of a fungal genome using only SOLiD short read data: Verification on *Aspergillus oryzae* RIB40. PLoS One. 2013;**8**(5):e63673

[9] Ikegami T, Inatsugi T, Kojima I, Umemura M, Hagiwara H, Machida M, et al. Hybrid de novo genome assembly using MiSeq and SOLiD short read data. PLoS One. 2015;**10**(4):e0126289

[10] Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. 2011;**17**(1):10-12

[11] Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Research. 2012;**22**(3):568-576

[12] 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 Biology. 2013;**14**(4):R36

[13] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: Ultrafast universal RNA-seq aligner. Bioinformatics. 2013;**29**(1):15-21

[14] Kim D, Langmead B, Salzberg SL. HISAT: A fast spliced aligner with low memory requirements. Nature Methods. 2015;**12**(4):357-360

[15] Garber M, Grabherr MG, Guttman M, Trapnell C. Computational methods for transcriptome annotation and quantification using RNA-seq. Nature Methods. 2011;**8**(6):469-477

[16] Bray NL, Pimentel H, Melsted P, Pachter L. Near-optimal probabilistic RNA-seq quantification. Nature Biotechnology. 2016;**34**(5):525-527

[17] Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods. 2017;**14**:417-419

[18] Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics. 2006;**7**:62

[19] Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2016;**32**(5):767-769

[20] Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 2015;**16**:170

[21] Mills JD, Kawahara Y, Janitz M. Strand-specific RNA-Seq provides greater resolution of transcriptome profiling. Current Genomics. 2013;**14**(3):173-181

[22] Croucher NJ, Fookes MC, Perkins TT, Turner DJ, Marguerat SB, Keane T, et al. A simple method for directional transcriptome sequencing using Illumina technology. Nucleic Acids Research. 2009;**37**(22):e148

[23] Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, et al. Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nature Methods. 2010;**7**(9):709-715

[24] He Y, Vogelstein B, Velculescu VE, Papadopoulos N, Kinzler KW. The antisense transcriptomes of human cells. Science. 2008;**322**(5909):1855-1857

[25] Borodina T, Adjaye J, Sultan M. A strand-specific library preparation protocol for RNA sequencing. Methods in Enzymology. 2011;**500**:79-98

[26] Oshlack A, Wakefield MJ. Transcript length bias in RNA-seq data confounds systems biology. Biology Direct. 2009;**4**:14

[27] Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013;**14**:91

[28] Yoon S, Kim SY, Nam D. Improving gene-set enrichment analysis of RNA-Seq data with small replicates. PLoS One. 2016;**11**(11):e0165919

[29] Ye Z, Chen Z, Lan X, Hara S, Sunkel B, Huang TH, et al. Computational analysis reveals a correlation of exon-skipping events with splicing, transcription and epigenetic factors. Nucleic Acids Research. 2014;**42**(5):2856-2869

[30] Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011;**27**(17):2325-2329

[31] Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 2010;**7**(12):1009-1015

[32] Wang B, Guo G, Wang C, Lin Y, Wang X, Zhao M, et al. Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. Nucleic Acids Research. 2010;**38**(15):5075-5087

[33] Ebbole DJ, Jin Y, Thon M, Pan H, Bhattarai E, Thomas T, et al. Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: Analysis of expressed sequence tags. Molecular Plant-Microbe Interactions. 2004;**17**(12):1337-1347

[34] Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, et al. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science. 2005;**307**(5713):1321-1324

[35] Xie BB, Li D, Shi WL, Qin QL, Wang XW, Rong JC, et al. Deep RNA sequencing reveals a high frequency of alternative splicing events in the fungus *Trichoderma longibrachiatum*. BMC Genomics 2015;**16**:54

[36] Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics. 2013;**14**(2):193-202

[37] Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, et al. Artemis: Sequence visualization and annotation. Bioinformatics. 2000;**16**(10):944-945

[38] Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Briefings in Bioinformatics. 2013;**14**(2):178-192

[39] Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature Biotechnology. 2011;**29**(1):24-26

[40] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;**26**(5):589-595

[41] Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, et al. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research. 2014;**42**(Database issue):D26-D31

[42] Tilgner H, Grubert F, Sharon D, Snyder MP. Defining a personal, allele-specific, and single-molecule long-read transcriptome. Proceedings of the National Academy of Sciences of the United States of America. 2014;**111**(27):9869-9874