Chapter 2

Transcriptome Sequencing: RNA-Seq

Hong Zhang, Lin He, and Lei Cai

Abstract

RNA sequencing (RNA-seq) can not only be used to identify the expression of common or rare transcripts but also in the identification of other abnormal events, such as alternative splicing, novel transcripts, and fusion genes. In principle, RNA-seq can be carried out by almost all of the next-generation sequencing (NGS) platforms, but the libraries of different platforms are not exactly the same; each platform has its own kit to meet the special requirements of the instrument design.

Key words Next-generation sequencing, RNA sequencing, Messenger RNA, Library construction, Data analysis

1 Introduction

In a broad sense, transcriptome refers to the collection of all transcripts under certain physiological condition, including messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and noncoding RNA (ncRNA), while in a narrow sense, it refers to collection of all mRNA transcripts [1]. Transcriptome sequencing, also called RNA-seq or whole-transcriptome shotgun sequencing (WTSS), uses high-throughput sequencing technology to rapidly and comprehensively obtain the transcriptional status of biological samples at a specific time [2]. At present, RNA-seq is mainly used in the study of mRNA, small RNA, noncoding RNA, or microRNAs. Different types of RNA can be obtained by adding additional separation and enrichment steps before cDNA synthesis. Illumina TruSeq is a method using conjugated magnetic beads (oligo-dT) to capture ploy A+ from total RNA and then contract mRNA library. During the ploy A+ enrichment process, non-ploy A+ RNA, including miRNA, rRNA, and other noncoding RNA, were removed [3, 4]. The mRNA library preparation steps contain five steps: (1) RNA fragmentation, (2) reverse transcription, (3) adapter ligation, (4) library cleanup and amplification, and (5) library quantification, quality control [5] (Fig. 1). Here, we show the method of

Tao Huang (ed.), *Computational Systems Biology: Methods and Protocols*, Methods in Molecular Biology, vol. 1754, https://doi.org/10.1007/978-1-4939-7717-8_2,

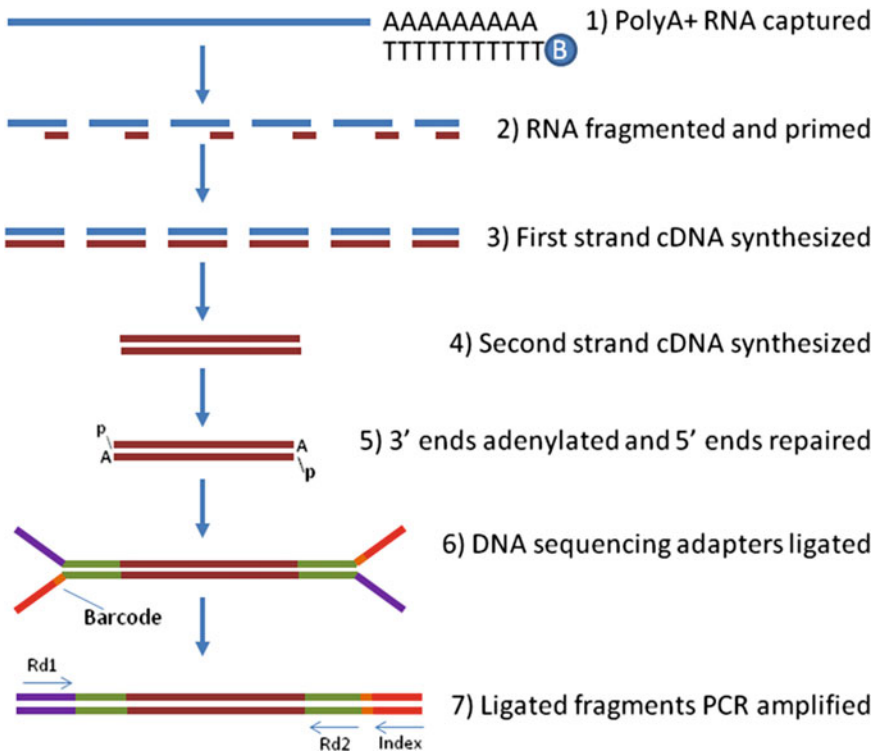

Fig. 1 mRNA library construction workflow for Illumina (from David Corney 2013)

the RNA-seq from total RNA extraction, library construction, and data analysis.

2 Materials

Prepare all solutions using ultrapure water and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise).

2.1 Total RNA Extraction

1. Liquid nitrogen.
2. 70% ethanol.
3. Tissue: keep the tissue in the liquid nitrogen until the procedure is completed.
4. TRIzol Reagent (Invitrogen).
5. DEPC-treated water (Ambion).
6. Chloroform: trichloromethane.
7. Isopropanol.
8. Thermo Scientific NanoDrop 2000 spectrophotometer: RNA quantification
9. Agilent 2100 Bioanalyzer system: RNA quality control.

2.2 mRNA Library Construction

1. RNA Purification Beads: purifying the poly-A containing mRNA molecules using oligo-dT attached magnetic bead, stored at 4 °C (Illumina, San Diego, CA).
2. Bead Washing Buffer (BWB), Elution Buffer (ELB), Bead-Binding Buffer (BBB): 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
3. Elute, Prime, Fragment Mix (EPF): 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
4. First-Strand Master Mix (FSM): 1 tube, stored at −20 °C (Illumina, San Diego, CA).
5. SuperScript II Reverse Transcriptase: 1 tube, stored at −20 °C.
6. Second-Strand Master Mix (SSM): 1 tube per 48 reactions, stored at −25 °C to −15 °C (Illumina, San Diego, CA).
7. AMPure XP beads: stored at 4 °C.
8. 80% ethanol.
9. Resuspension Buffer (RSB): 1 tube, stored at −20 °C.
10. End-Repair Mix: add 5′-phosphate groups needed for downstream ligation, 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
11. A-Tailing Mix: make fragments compatible with adapters and prevent self-ligation by adding a 3′-A overhang, 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
12. Ligation Mix: join 3′-T overhang adapters to 3′-A overhang inserts, 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
13. Stop Ligation Buffer: inactivate the ligation. 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
14. Resuspension Buffer (RSB): 1 tube, stored at −20 °C (Illumina, San Diego, CA).
15. PCR Master Mix (PMM): 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
16. PCR Primer Cocktail (PPC): 1 tube per 48 reactions, stored at −20 °C (Illumina, San Diego, CA).
17. Sequencing chip: flow cell.
18. Illumina HiSeq system.

2.3 Data Analysis

1. Raw data processing: Trimmomatic.
2. Mapping: TopHat (Bowtie).
3. Quality control: RSeQC.
4. Differentially expressed gene analysis: htseq-count, DEseq, DAVID, KEGG.

5. Differential alternative splicing analysis: MISO (a mixture of isoforms).
6. Fusion gene analysis: TopHat-Fusion.

3 Methods

3.1 Total RNA Extraction

1. Remove the tissue sample from −80 °C refrigerator, and immediately put it in the thermos cup with liquid nitrogen (*see* **Note 1**).
2. Remove the sample from the liquid nitrogen and put into a 1.5 mL EP tube; add 300 μL TRIzol reagent, fully grinding with an electric tissue grinder; then add 700 L TRIzol; and place the tube on the ice for 30 min to ensure that sufficient crushing of the cells.
3. Add 200 μL chloroform, vortex, and then centrifuge at 13,000 × *g* for 10 min.
4. Remove supernatant to a new EP tube (*see* **Note 2**).
5. Add 500 μL isopropanol, vortex, place at −20 °C for 20 min, and then centrifuge at 13,000 × *g* for 10 min.
6. Discard supernatant; add 1 mL 70% ethanol solution, mild concussion for 10s; and then centrifuge at 8000 × *g* for 2 min.
7. Discard supernatant, and repeat **step 6** one time.
8. Discard supernatant, centrifuge at 8000 × *g* for 15 s, remove excess liquid, and place the EP tube on ice for 2 min to make ethanol fully volatile.
9. According to the precipitation size, add 30–200 μL ultrapure water.
10. Determine the concentration of RNA solution by using NanoDrop 2000 spectrophotometer.
11. Use the Agilent 2100 Bioanalyzer system to detect the RNA integrity (*see* **Note 3**).
12. RNA solution should be stored in the −80 °C refrigerator.

3.2 Library Construction

1. Add 2 μg total RNA samples (less than 50 μL) to a 200 μL EP tube, dilute to 50 μL, then add 50 μL RNA Purification Beads (*see* **Note 4**), and gently pipette the entire volume up and down eight times to mix thoroughly.
2. Place the EP tube on PCR thermal cycler (65 °C for 5 min, 4 °C hold) to denature the RNA.
3. Place the EP tube at room temperature for 5 min to facilitate binding of the polyA RNA to the beads.

4. Place the EP tube on the magnetic stand for 5 min to separate the polyA RNA beads from the solution.
5. Discard the liquid, wash the beads by adding 200 μL Bead Washing Buffer, gently pipette the entire volume up and down eight times to mix thoroughly, and place the EP tube on the magnetic stand for 5 min.
6. Discard the liquid, add 50 μL of Elution Buffer, gently pipette, and place the EP tube on PCR thermal cycler (80 °C for 2 min, 25 °C hold).
7. Add 50 μL Bead-Binding Buffer, gently pipette, place the EP tube at room temperature for 5 min, then place the EP tube on the magnetic stand for 5 min, and discard the liquid (*see* **Note 5**).
8. Add 200 μL Bead Washing Buffer, gently pipette for eight times, and place the tube on the magnetic stand for 5 min.
9. Discard the liquid; add 19.5 μL Elute, Prime, Fragment Mix; gently pipette for eight times; and place the EP tube on PCR thermal cycler (94 °C for 8 min, 4 °C hold) (*see* **Note 6**).
10. Place the tube on the magnetic stand for 5 min, and remove 17 μL solution into a new EP tube.
11. Add 1 μL SuperScript II to 79.6 μL First-Strand Master Mix, and mix thoroughly (*see* **Note 7**).
12. Add 8 μL solution configured in **step 11** to the EP tube in **step 10**, and mix thoroughly.
13. Place the EP tube on PCR thermal cycler (25 °C for 10 min, 42 °C for 50 min, 70 °C for 15 min, 4 °C hold).
14. Add 25 μL Second-Strand Master Mix to the EP tube in **step 13**, mix thoroughly, and place the EP tube on PCR thermal cycler (16 °C for 1 h, 4 °C hold).
15. Add 90 μL AMPure XP purification beads, gently pipette for eight times, place the EP tube at room temperature for 15 min, and place the tube on the magnetic stand for 5 min.
16. Discard the liquid, add 200 μL 80% ethanol solution with the EP tube on the magnetic stand, and incubate the EP tube at room temperature for 30s.
17. Repeat **step 16** one time.
18. Discard the liquid, let the EP tube at room temperature for about 15 min till the full evaporation of the ethanol, and then remove the EP tube from the magnetic stand.
19. Add 62.5 μL Resuspension Buffer, place the EP tube at room temperature for 2 min, and then place it on the magnetic stand.
20. Remove 60 μL supernatant to a new EP tube.
21. Add 40 μL End-Repair Mix, mix thoroughly, and incubate the EP tube at 30 °C for 30 min.

22. Add 160 L AMPure XP purification beads, gently pipette for eight times, place the EP tube at room temperature for 15 min, and then place the tube on the magnetic stand for 5 min.
23. Discard the liquid, add 200 μL 80% ethanol solution with the EP tube on the magnetic stand, and incubate the EP tube at room temperature for 30s.
24. Repeat **step 23** one time.
25. Discard the liquid, let the EP tube at room temperature for about 15 min till the full evaporation of the ethanol, and then remove the EP tube from the magnetic stand.
26. Add 20 μL Resuspension Buffer, place the EP tube at room temperature for 2 min, and then place it on the magnetic stand.
27. Transfer 17.5 μL of the supernatant to a new EP tube, add 12.5 μL A-Tailing Mix, gently pipette for eight times, and incubate the EP tube at 37 °C for 30 min.
28. Add 2.5 μL Ligation Mix, 2.5 μL Resuspension Buffer, and 2.5 L RNA Adapter Index, gently pipette for eight times, and incubate the EP tube at 30 °C for 10 min.
29. Add 5 μL Stop Ligation Buffer, and gently pipette for eight times.
30. Add 42.5 μL AMPure XP purification beads, gently pipette for eight times, place the EP tube at room temperature for 15 min, and then place it on the magnetic stand.
31. Discard the liquid, add 200 μL 80% ethanol solution with the EP tube on the magnetic stand, and incubate the EP tube at room temperature for 30s.
32. Repeat **step 31** one time.
33. Discard the liquid, let the EP tube at room temperature for about 15 min till the full evaporation of the ethanol, and then remove the EP tube from the magnetic stand.
34. Add 22.5 L Resuspension Buffer, placed the EP tube at room temperature for 2 min, and then place it on the magnetic stand.
35. Transfer 20 μL of the supernatant to a new EP tube.
36. Add 25 μL PCR Master Mix and 5 μL PCR Primer Cocktail.
37. Place the EP tube on PCR thermal cycler (98 °C for 30s n; 15 cycles of 98 °C for 10 s, 60 °C for 30 s, 72 °C for 30 s; 72 °C for 5 min; 10 °C hold).
38. Remove the EP tube from the PCR thermal cycler, adding 50 μL AMPure XP purification beads, gently pipette for eight times, place the EP tube at room temperature for 15 min, and then place it on the magnetic stand.
39. Discard the liquid, add 200 μL 80% ethanol solution with the EP tube on the magnetic stand, and incubate the EP tube at room temperature for 30s.

40. Repeat **step 39** one time.
41. Discard the liquid, let the EP tube at room temperature for about 15 min till the full evaporation of the ethanol, and then remove the EP tube from the magnetic stand.
42. Add 27.5 L Resuspension Buffer, place the EP tube at room temperature for 2 min, and then place it on the magnetic stand.
43. Transfer 25 μL of the supernatant to a new EP tube.
44. Dilute 2 μL library to 20 μL, and then use the 2% E-gel gel to detect the quality of the library; the final library size should be a band 200–400 bp.
45. Take appropriate library samples according to the requirements of the Illumina sequencing platform.

3.3 Data Analysis

3.3.1 Raw Reads Processing and Mapping

1. According to a base-calling pipeline, the images generated by the sequencing system (Illumina) are translated into nucleotide sequences. The raw reads are saved in fastq format, and Trimmomatic [6] can be used to filter the raw data prior to analyzing the data. There are three criteria: (1) discard reads shorter than 36 bases; (2) remove reads containing sequencing adaptors; and (3) remove bases with a quality score less than 15. The commands are as follows.

   For paired-end reads, the command is:

```
java -classpath trimmomatic.jar org.usadellab.trimmomatic.TrimmomaticPE
(see Note 8) -t threads $p -phred33 $input1 $input2 $output1_paired.fq.gz
$output1_unpaired.fq.gz $output2_paired.fq.gz $output2_unpaired.fq.gz IL-
LUMINACLIP:$WORKPATH/adapter.fa:2:40:15 SLIDINGWINDOW:4:15 MINLEN:36
```

   For single-end reads, the command is:

```
java -classpath trimmomatic.jar org.usadellab.trimmomatic.TrimmomaticSE
(see Note 8) -t threads $p -phred33 $input1 $output1_paired.fq.gz
$output1_unpaired.fq.gz ILLUMINACLIP:$WORKPATH/adapter.fa:2:40:15
SLIDINGWINDOW:4:15 MINLEN:36
```

2. Then, the clean sequencing reads should be aligned with UCSC hg19 reference genome using TopHat [7], which incorporates the Bowtie to perform the alignment. The command is:

```
tophat -p $p -G genes.gtf -o $tophat_out
$Reference/Sequence/BowtieIndex/genome
$output1_paired.fq.gz $output2_paired.fq.gz
```

3. RSeQC is a RNA-seq quality control package; it provides a number of useful modules that can comprehensively evaluate

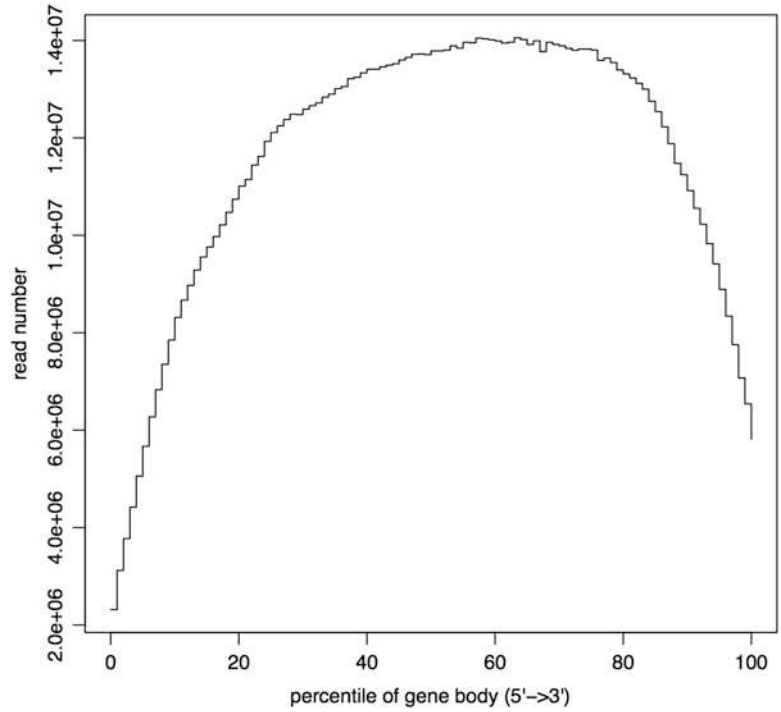

Fig. 2 Coverage uniformity over gene body without bias. Abscissa means the relative position of gene

RNA-seq data. geneBody_coverage.py can be used to calculate the RNA-seq reads coverage over gene body. The command is:

```
geneBody_coverage.py -r $Reference_BED -i $input -o $output
```

This module scales all transcripts to 100 nucleotides and calculates the number of reads covering each nucleotide position. Finally, it generates a plot illustrating the coverage profile along the gene body [8]. An example for the output is shown in Fig. 2, the reads coverage is uniform, and there is no 5′/3′ bias.

The commands for other packages of RSeQC are:

```
clipping_profile.py -i $input -o $output
infer_experiment.py -r $Reference_BED -i $input -o $output
inner_distance.py -r $Reference_BED -i $input -o $output
junction_annotation.py -r $Reference_BED -i $input -o $output
junction_saturation.py -r $Reference_BED -i $input -o $output
read_distribution.py -r $Reference_BED -i $input >$output_read_distribution.txt
read_duplication.py -i $input -o $output
read_GC.py -i $input -o $output
read_NVC.py -i $input -o $output
read_quality.py -i $input -o $output
RPKM_count.py -d '1++,1--,2+-,2-+' -i $input -o $output
RPKM_saturation.py -d '1++,1--,2+-,2-+' -r $Reference_BED -i $input -o $output
```

```
split_bam.py -r $Reference_BED -i $input -o $output
```

3.3.2 Differentially Expressed Genes Detection

1. The pipeline for the analysis of differentially expressed genes (DEGs) is shown in Fig. 3.
2. Use TopHat to align the reads on the genome, and then count reads in features with htseq-count; the commands are:

```
samtools view -h -o $tophat_out/accepted_hits.sam
$tophat_thout/accepted_hits.bam
htseq-count -s no $tophat_thout/accepted_hits.sam $genes.gtf > $htseq-count.out
```

3. Use DEseq or edgeR to detect the DEGs. An example for R commands of DEseq is:

```
    #!/user/bin/R/bin/Rscript
datafile = system.file("htseq-coun.txt",package="pasilla")
pasillaCountTable = read.table("htseq-coun.txt",header=TRUE, row.names=1)
pasillaDesign = data.frame(
row.names = colnames(pasillaCountTable),
condition = c("condition1"," condition2"),
libtype = c("paired-end","paired-end"))
condition = factor ( c( "control","case"))
library( "DESeq" )
cds = newCountDataSet( pasillaCountTable,condition )
```

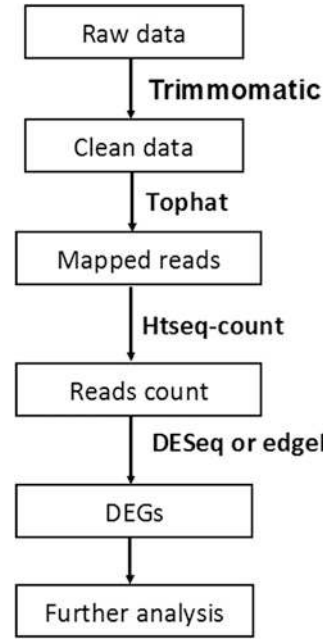

Fig. 3 The pipeline for the analysis of differentially expressed genes

```
cds = estimateSizeFactors( cds )
sizeFactors( cds )
head(counts( cds, normalized=TRUE )
cds = estimateDispersions( cds )
str( fitInfo(cds) )
plotDispEsts( cds )
res = nbinomTest( cds, "control", "case")
plotMA(res)
addmargins( table(res_sig = res$padj < .1, res_sig = res$padj < .1 ) )
write.csv (res, file=" diff_exp.csv")
```

4. Then we can use DAVID [9] to detect the functional enrichments and KEGG [10] to identify the significantly changed pathways for the differentially expressed genes.

3.3.3 Differential Exon-Skipping Event Detection

A mixture-of-isoform (MISO) [11] analysis adopted Bayesian inference algorithm to calculate the probability that a read came from a specific isoform. The software computes the percentage of transcripts that are spliced to include exons and is called the PSI (Ψ):

$$\text{Percentage splicing in (PSI)} = \frac{\#\ \text{of reads supporting inclusion}}{\#\ \text{of reads supporting inclusion} + \text{skipping}}$$

1. Compute the insert length:

```
misopy/pe_utils.py --compute-insert-len tophat_out/accepted_ hits.bam
ensGene.min_1000.const_exons.gff --output-dir insert-dist
```

2. Run MISO:

```
misopy/run_events_analysis.py --compute-genes-psi ref/hg19_miso_v2.0/
indexed_SE_events $tophat_thout/accepted_ hits.bam
--output-dir $OUTDIR/${SAMPLE}_output --read-len 101 --paired-end 157 33
```

3. Summarize MISO inferences:

```
misopy/run_miso.py --summarize-samples $OUTDIR/${SAMPLE}_output
$OUTDIR/${SAMPLE}_summaries/
```

4. Make pairwise comparisons between samples (*see* **Note 9**):

```
misopy/run_miso.py --compare-samples $OUTDIR/$control_output $OUTDIR/
$case_output $OUTDIR/
```

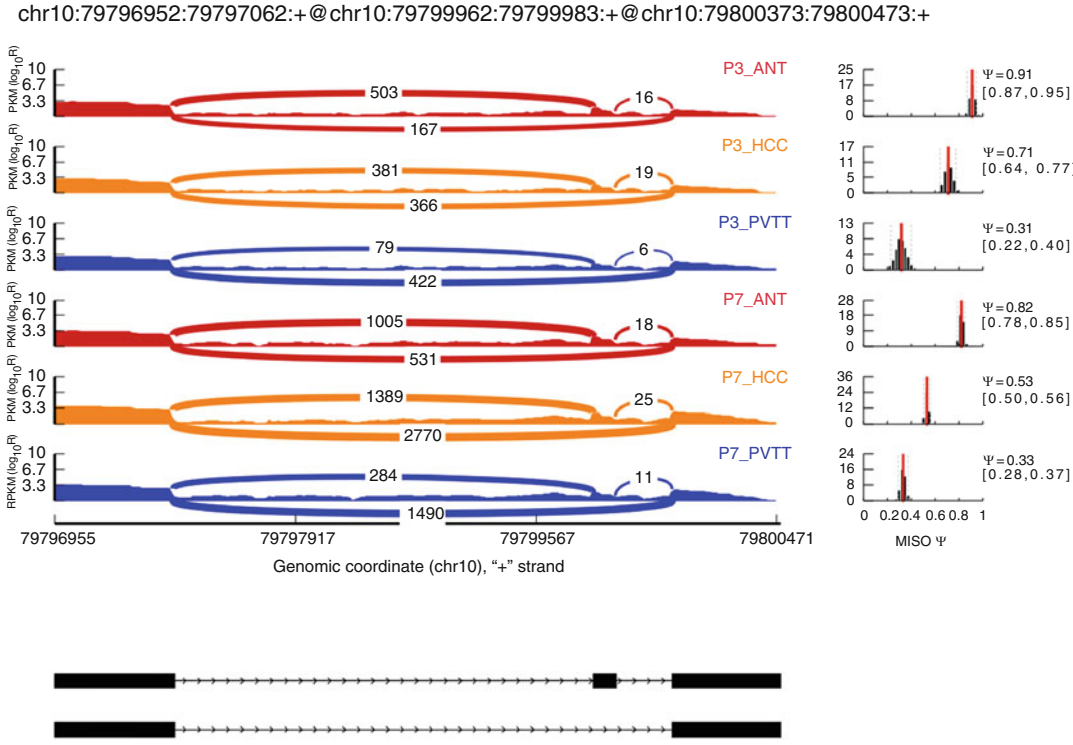

Fig. 4 RNA-seq reads coverage of the gene *RPS24* in P3 and P7. The RNA-seq reads were mapping to the UCSC reference genome (hg19) of *RPS24*. The PVTT tissue tracks were shown in blue, the HCC tissue in orange, and the ANT tissue in red. The counts of reads spanning the junction of exons were shown. Ψ = PSI (percentage splicing in)

5. Filter significant events (*see* **Note 10**):

```
misopy/filter_events.py --filter
$OUTDIR/$control_output_vs_$case_output/bayes-factors/$con-
trol_output_vs_$cas
e_output.miso_bf  --num-inc 1 --num-exc 1 --num-sum-inc-exc
10 --delta-psi 0.2
--bayes-factor 1000 --output-dir $OUTDIR/${SAMPLE}_filtered/
```

6. Visualizing and plotting MISO output by sashimi_plot, an example for *RSP24* gene [4] is shown in Fig. 4; the command is:

```
plott.py --plot-event
"chr10:79796952:79797062:+@chr10: 79799962:79799983:+@chr10:7
9800373:79800473:+"  ref/hg19_miso_v2.0/indexed_SE_events/
sample_sashimi_plot_settings.txt  --output-dir  test-plot-
RSP24/
```

3.3.4 Detecting Gene Fusions

Use TopHat to map the clean reads to the downloaded reference transcript sequences. Then, the mapped BAM files are passed to TopHat-Fusion [12] to detect the candidate fusions.

1. Running TopHat

```
tophat -o ${OUTDIR}/tophat_${SAMPLE} -p $p --fusion-search --keep-fasta-order --bowtie1 --no-coverage-search -r 0 --mate-std-dev 80 --max-intron-length 100000 --fusion-min-dist 100000 --fusion-anchor-length 13 --fusion-ignore-chromosomes chrM $BowtieIndex/genome $R1.fastq $R2.fastq
```

2. Run TopHat-fusion-post to filter out fusion candidates

```
tophat-fusion-post -p $p -o $OUTDIR/fusion_post_out --num-fusion-reads 1 --num-fusion-pairs 2 --num-fusion-both 5 $BowtieIndex/genome
```

4 Notes

1. The tissue should be stored immediately following sacrifice and extraction.
2. Be careful not to absorb liquids in the middle or lower part.
3. Illumina recommends that the total RNA integrity following isolation using Agilent 2100 with an RNA integrity number (RIN) value $\geq$8.
4. Vortex RNA Purification Beads tube vigorously to resuspend the oligo-dT beads.
5. This step can make mRNA combine with the beads and remove most of the rRNA.
6. The first strand of cDNA can be synthesized in this step by using reverse transcriptase and random primers.
7. The mixed solution should be prepared when used and can not be kept for too long after prepared.
8. PE, paired-end mode; SE, single-end mode.
9. To detect differentially expressed events or isoforms.
10. The default Bayes factor value is 10, and $|\Delta PSI| > 20\%$.

References

1. Villegas VE, Zaphiropoulos PG (2015) Neighboring gene regulation by antisense long non-coding RNAs. Int J Mol Sci 16 (2):3251–3266. https://doi.org/10.3390/ijms16023251

2. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, Szczesniak MW, Gaffney DJ, Elo LL, Zhang X, Mortazavi A (2016) A survey of best practices for RNA-seq data analysis. Genome Biol 17:13. https://doi.org/10.1186/s13059-016-0881-8

3. Zhang H, Weng X, Ye J, He L, Zhou D, Liu Y (2015) Promoter hypermethylation of TERT is associated with hepatocellular carcinoma in the Han Chinese population. Clin Res Hepatol Gastroenterol 39(5):600–609. https://doi.org/10.1016/j.clinre.2015.01.002
4. Zhang H, Ye J, Weng X, Liu F, He L, Zhou D, Liu Y (2015) Comparative transcriptome analysis reveals that the extracellular matrix receptor interaction contributes to the venous metastases of hepatocellular carcinoma. Cancer Genet 208(10):482–491. https://doi.org/10.1016/j.cancergen.2015.06.002
5. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, Ordoukhanian P (2014) Library construction for next-generation sequencing: overviews and challenges. Biotechniques 56(2):61.-64, 66, 68, passim. https://doi.org/10.2144/000114133
6. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170
7. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc 7(3):562–578. https://doi.org/10.1038/nprot.2012.016
8. Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28(16):2184–2185. https://doi.org/10.1093/bioinformatics/bts356
9. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. https://doi.org/10.1038/nprot.2008.211
10. Ogata H, Goto S, Fujibuchi W, Kanehisa M (1998) Computation with the KEGG pathway database. Biosystems 47(1–2):119–128
11. Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7(12):1009–1015. https://doi.org/10.1038/nmeth.1528
12. Kim D, Salzberg SL (2011) TopHat-fusion: an algorithm for discovery of novel fusion transcripts. Genome Biol 12(8):R72. https://doi.org/10.1186/gb-2011-12-8-r72