Chapter 8

Quality Control of RNA-Seq Experiments

Xing Li, Asha Nair, Shengqin Wang, and Liguo Wang

Abstract

Direct sequencing of the complementary DNA (cDNA) using high-throughput sequencing technologies (RNA-seq) is widely used and allows for more comprehensive understanding of the transcriptome than microarray. In theory, RNA-seq should be able to precisely identify and quantify all RNA species, small or large, at low or high abundance. However, RNA-seq is a complicated, multistep process involving reverse transcription, amplification, fragmentation, purification, adaptor ligation, and sequencing. Improper operations at any of these steps could make biased or even unusable data. Additionally, RNA-seq intrinsic biases (such as GC bias and nucleotide composition bias) and transcriptome complexity can also make data imperfect. Therefore, comprehensive quality assessment is the first and most critical step for all downstream analyses and results interpretation. This chapter discusses the most widely used quality control metrics including sequence quality, sequencing depth, reads duplication rates (clonal reads), alignment quality, nucleotide composition bias, PCR bias, GC bias, rRNA and mitochondria contamination, coverage uniformity, etc.

Key words Quality control, RNA-seq, High-throughput sequencing, Next-generation sequencing

1 Introduction

RNA-seq has led to a better understanding of the RNA universe by providing unprecedented opportunities to interrogate the transcriptome from different perspectives. These include gene expression profiling [1–4], new isoforms or alternative splicing identification and quantification [5–7], novel transcripts such as lincRNAs discovery [8–10], aberrant transcripts such as gene fusion identification [11–13], and variant calling [14–17]. However, current library preparation protocols of RNA-seq are still developing and possess several intrinsic biases and limitations, such as nucleotide composition bias, GC bias, and PCR bias, which could directly detriment many RNA-seq applications [18, 19].

In general, quality of RNA-seq experiments can be assessed at two different levels. Raw sequence based metrics, which check RNA-seq experiments at a "low level" because they do not require sequence alignments. These assessments include read (i.e., a consecutive sequence of nucleotides) quality, read duplication rate

Ernesto Picardi (ed.), *RNA Bioinformatics*, Methods in Molecular Biology, vol. 1269,
DOI 10.1007/978-1-4939-2291-8_8,

(clonal reads), GC content, nucleotide composition bias, etc. However, raw sequence-based metrics largely focuses on evaluating the success of sequencing technologies and themselves alone cannot ensure the usability (biologic accuracy) of RNA-seq data. Therefore, checking RNA-seq data at a "higher level" is also imperative. These metrics include mapping statistics, coverage uniformity, saturation of sequencing depth, reads distribution over gene structure, ribosomal RNA contamination, reproducibility between biological replicates, etc.

1.1 Raw Sequence Quality

Phred quality score (Q) was originally developed by the program Phred to measure base-calling reliability from Sanger sequencing chromatograms [20, 21]. It is defined as $Q = -10 \times \log_{10}(P)$ where P is the probability of erroneous base calling. For example, a Phred quality score of 30 means the chance that this base is called incorrectly is 1 in 1,000. Although the Phred program is rarely used in next-generation sequencing field, Phred or Phred-like quality score has become widely accepted to characterize the quality of DNA sequences (*see* **Note 1**). Most often, Phred scores are reported as their corresponding ASCII characters (33–126 or "!" to "V") (*see* **Note 2** for FASTQ format), but SOLiD still uses numbers to represent quality scores.

There is no gold standard to tell if the quality of a particular sequence is good or bad, as this is really depending on the purpose of the study. For example, compared to expression profiling, variants calling tasks require much higher sequence quality. In general, scores over 30 indicate very good quality, 20–30 indicate reasonable good and <20 indicate poor quality. Parallel boxplots visualize "per nucleotide quality score" by summarizing Phred qualities for all reads at each position (Fig. 1a) [22, 23]. In addition, one can also calculate the average quality score per read ("per sequence quality score") and check the quality score distribution of all sequences (Fig. 1b).

1.2 Nucleotide Composition and GC Content

GC content (or guanine-cytosine content) is the percentage of bases in a DNA sequence that are either guanine or cytosine. It is a simple way to measure nucleotide composition of DNA. The reason to use GC rather than AT (or AU in RNA) is that GC content carries more direct biologic meaning. GC pairs are more stable than AT (3 vs. 2 hydrogen bonds) which has implications in PCR experiments where the GC content of primers predicts their annealing temperature. Further, exons have much higher GC content than introns and intergenic regions and cytosine is the target of DNA methylation. People have found the dependence between read coverage and the GC content of reference genome in high-throughput sequence data. Therefore, evaluating GC content bias in RNA-seq data is of great importance to both transcript detection and abundance quantification [18].

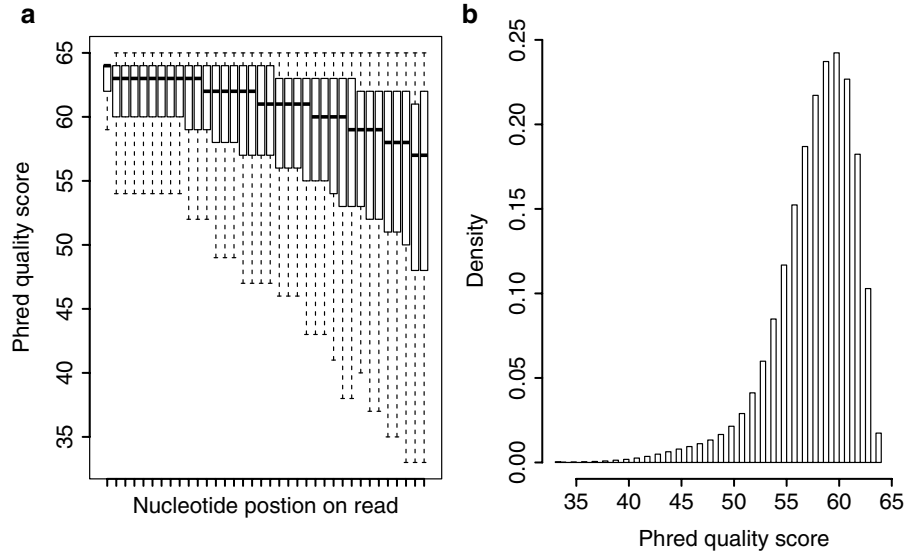

Fig. 1 (**a**) Parallel boxplot showing "per nucleotide quality score." All reads are overlaid together, and then summarize Phred quality score (*Y*-axis) for each position of read from 5′ to 3′ end (*X*-axis). (**b**) "per sequence quality score" distribution. For each read, "per sequence quality score" is calculated as the average Phred quality score (*X*-axis) across all nucleotides

Assume RNA-seq reads were randomly sampled from expressed transcripts, when we pileup reads together and calculate the nucleotide composition (percentage of A, C, G, and T) at each positions or column, we expect little differences between columns. Random fluctuations will be cancelled out because the large sample size (i.e., hundreds of millions of reads). If we visualize nucleotide composition versus nucleotide positions in a diagram [22] the lines should be roughly flat at a value of 0.25 (Fig. 2a). In practical, the first 12 bases starting from 5′ end of reads exhibit large deviation from 0.25, this is due to the random hexamer priming during PCR amplification [19]. A serious bias indicates the existence of over-represented sequences, and such bias will influence coverage uniformity as well as transcripts abundance estimation. Per sequence GC content can be roughly used to measure the randomness of sequencing library as GC content of reads from random sequence library follows normal distribution with the mean equals to the overall GC content of the transcriptome (Fig. 2b). A poorly prepared or contaminated library will exhibit a skewed distribution.

1.3 Duplicate Sequences (PCR Duplication)

Read duplication rate is affected by read length, sequencing depth, transcript abundance and PCR amplification. Supposing the sequencing library is purely random and read length is 36 bp, the chance to get a duplicated read is $1/4^{72}$ (or 4.5×10^{-44}), this chance is still slim even if the sequencing depth reaches hundreds of millions.

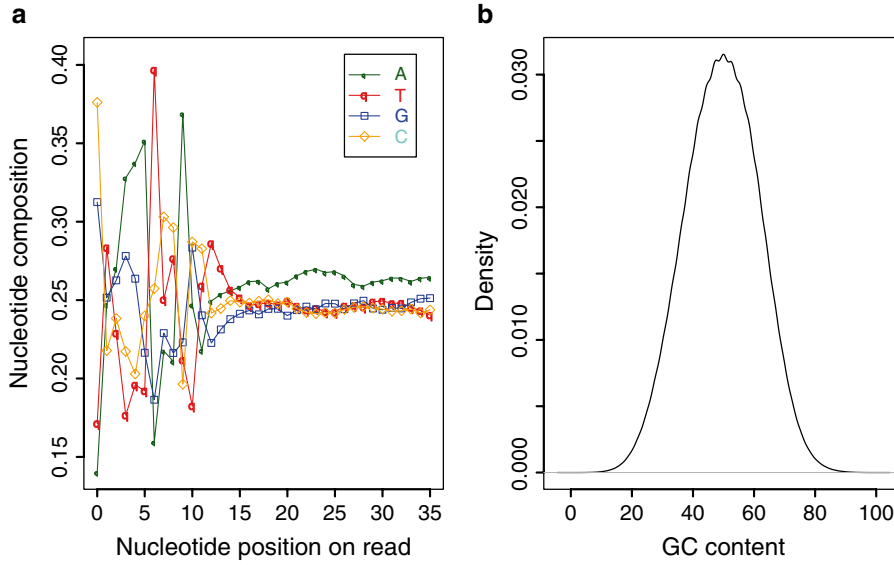

Fig. 2 (**a**) Diagram showing nucleotide composition bias at the beginning of reads. All reads are overlaid together, and then calculate nucleotide frequency (*Y*-axis) for each position of read (*X*-axis). Four nucleotides were indicated using different colors. (**b**) "per sequence GC content" distribution

Therefore, the majority of duplicated reads were artificially generated from PCR amplification [22]. And because of this, duplication rate is one way to check PCR amplification bias. To circumvent the huge memory requirement, tools such as FastQC will only track the first 200,000 short reads in each file for duplication level and creates a graph plotting the count of sequences with different degrees of duplication. By default, FastQC will raise a warning if there are more than 20 % of duplicated sequences in total and a failure if this number reaches over 50 % as the sequencing library is seriously biased and may not randomly sampling the target sequence.

1.4 Descriptive Statistics

Mapping statistics are the simplest and most intuitive way to assess if RNA sequencing was successful. These include mappability (number of reads aligned to reference genome), number of reads aligned to unique locations in the genome, and the number of splice mapped reads and number of reads mapped to mitochondria. It is difficult to derive reasonable or even empirical thresholds to determine if a particular RNA sequencing was successful or not, because these metrics really depend on read length, sequencing depth, bioinformatic analysis parameters, sample preparation protocol, and tissue type. For example, compared with shorter reads, longer reads will have better mappability, lower duplication rate, higher proportion that aligned to unique genome location, more spliced reads given the same sequencing depth. For the same sequencing depth and same read length, number of splice reads

may be dramatically different between two RNA-seq datasets simply because tissue origins are different. Muscle and heart tissues usually have much more mitochondria than other tissue types, and therefore mitochondria reads could be a problem if they account for a large proportion (i.e., >30 %), as this makes actual sequencing depth much lower than expected.

1.5 rRNA/tRNA Contamination

The goal of most RNA-seq studies is to interrogate functional message RNA (mRNA). However, structure RNAs such as Ribosomal RNA (rRNA) and transfer RNA (tRNA) are the most abundant RNA species and constitute 60–90 % of total RNA in a cell. To avoid having these RNAs dominate the sequencing data, it is necessary to remove these RNA species before preparing libraries for deep sequencing. Two approaches have been used to enrich mRNA. The first approach starts with total RNA that has been depleted of rRNA by using a set of oligos that binds to rRNA (such as RiboMinus™), and the second method selects for transcripts by isolating poly-A RNA as the staring materials for the construction of sequencing libraries.

Even with ribosome depletion, a fair amount of ribosomal sequences may still remain in the raw data. Small amounts of rRNA contamination will not be a detriment to downstream analyses. However, a larger amount of ribosomal reads usually suggests rRNA depletion was inefficient or failed and additional sequencing may be necessary. Assessing rRNA contamination is straightforward; aligning reads to reference genome and then counts how many reads mapped to ribosome genes, or aligning reads directly to ribosomal RNA sequences.

1.6 Saturation Test of Sequencing Depth

RNA-seq experiments are diverse in their aims and design goals. The amount of sequencing needed for a given sample is determined by the goals of the experiment. For gene expression profiling, where we are interested to find quantitative differences of known genes between groups, modest sequencing depth is good enough (e.g., 30 million pair-end reads with length >30 bp for mammalian genomes). But for studies that involve investigation of alternative splicing, gene fusion detection and novel transcript identification, deeper read depths is required to be able to adequately cover not just the exons but also exon–exon junctions. It is recommended by ENCODE consortium that a minimum of 100–200 million 2×76 bp or longer reads is needed for mammalian genomes.

The saturation test is an approach to determine if current sequencing depth is deep enough to satisfy a particular purpose. It is fundamentally important because if sequencing was unsaturated, estimated gene expression metrics such as RPKM (Reads Per Kilobase exon per Million mapped reads) will be unstable and low abundant isoforms will be undiscovered. In practical, we resample 5, 10, …, 100 % of the total mapped reads and RPKMs are

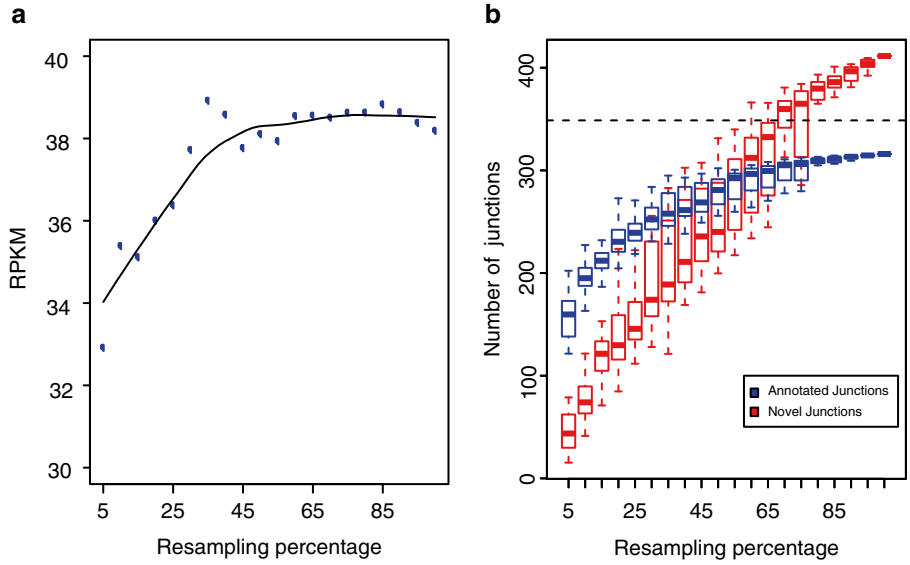

Fig. 3 Saturation test of sequencing depth. (**a**) Saturation test using RPKM (gene expression measurement). RPKMs were recalculated for each resampled subset (*blue dots*) to test if RPKM values enter a steady state (or saturated). (**b**) Saturation test using detected splice junctions (*blue*: annotated junction, *red*: novel junction). *Horizontal dashed line* indicates all annotated junctions encoded in reference gene models

subsequently recalculated using each subset. For a particular transcript, RPKM values may vary at the beginning with very small sample sizes, but finally could reach a plateau. The stable RPKM values indicate a saturated sequencing depth; otherwise, sequencing depth should be increased until RPKM values enter a stationary stage (Fig. 3a). Saturation test for splice junction is similar; splice junctions are detected for each resampled subset, and number of detected splice junctions will increase as the resample percentage increases, but finally will reach a fixed value. The junction saturation test is very important for alternative splicing analysis, as unsaturated sequencing depth would miss many low-abundance yet bona fide splicing junctions. Due to the sensitivity of RNA-seq, the number of identified novel splice junctions will increase as sequencing depth goes deeper, and therefore saturation rarely occurs for novel splice junctions even with billons of reads in mammalian genome (Fig. 3b).

1.7 Reproducibility Between Replicates

For RNA sequencing technology, depending on the goal of the experiment, replicates can be of two kinds, technical and biological. Technical replicates are replicates obtained from the same sample for purposes of studying batch effects and evaluating the technology, including background noise, differences in sequencing chemistry, instrument-to-instrument differences, etc. Biological replicates are the most desired form of replicates, as these provide us with the true variation among biological samples.

The use of such replicates comes into play for experiments that involve comparison of two or more groups for differential expression analysis. It is recommended to have at least two biological replicates per group in order to statistically determine the significantly differentially expressed genes.

Evaluating the reproducibility between replicates is straightforward. Most often scatter plots are used to visualize the reproducibility between expression measurements such as RPKM or FPKM (Fragment Per Kilobase exon per Million reads). Logarithm transformation of RPKM is necessary because of the large dynamic range of RPKM values. After logarithm transformation, expression values roughly follow a normal distribution and have a high Pearson's correlation coefficient (Fig. 4).

1.8 Coverage Uniformity

Gene body coverage describes the overall reads density over the mRNA regions (both UTR exon and CDS exon). Ideally, each base has the same chance to be sequenced, and each site within gene body has similar coverage. However, read density profiles can be affected by library preparation protocol, PCR amplification, RNA degradation, genome complexity and the underlying gene

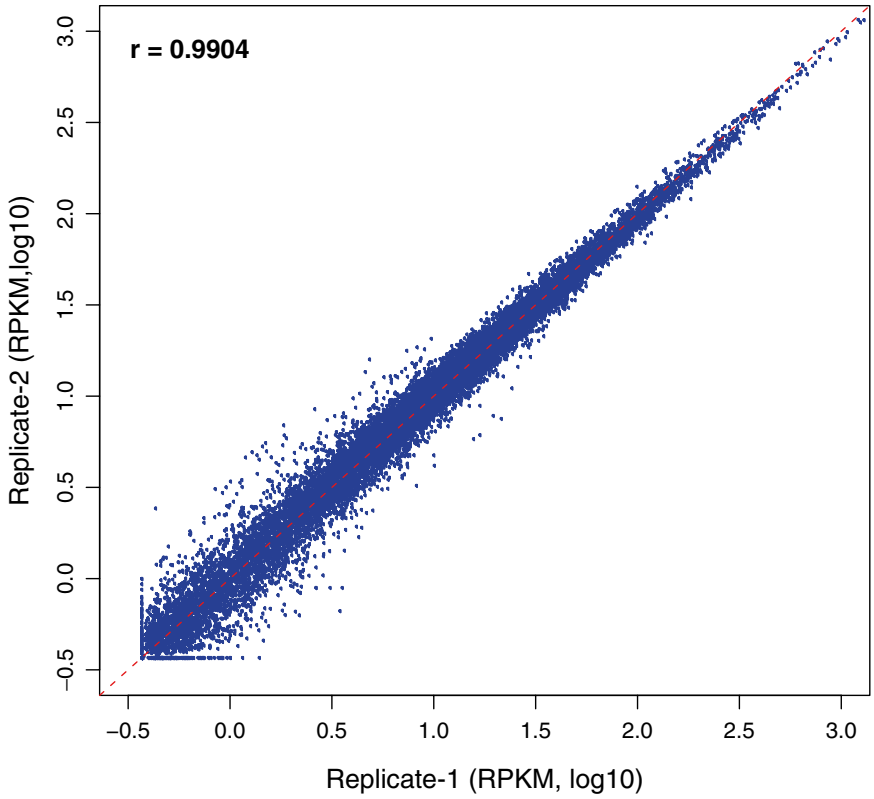

Fig. 4 Scatter plot showing reproducibility between two RNA-seq datasets (technical replicates). Each *blue dot* represents a gene, and the *red dashed line* is linear regression line

model used. For example, RNA-seq data using poly-A selection usually have higher coverage at 3′ end. PCR amplification efficiency could be different for different DNA fragments site, this also introduce uneven coverage. Low DNA complexity (or repetitive) regions usually have higher coverage but this will depend on how multi-hit reads were processed. For RNA-seq data that are not strand specific, coverage profile is also affected by underlying reference gene model; for example, an uneven coverage occurs when two genes are overlapped in the genome and express differently. Coverage profile is the most intuitive way to check uniformity, by normalizing all annotated genes into the same scale, and then calculating coverage for each position (Fig. 5).

1.9 Reads Distribution (Intron, Exon, UTR, etc.)

After mapping reads to a reference genome, we can calculate the fraction of reads assigned to exons (including both UTR and CDS exons), introns and intergenic regions based on the provided gene model. In ideal conditions and for well-annotated organisms, most of reads in RNA-seq data should be mapped to exonic regions. However, in practice, a considerable amount of reads are mapped to intron or intergenic regions. Except for mapping artifacts, intergenic/intronic reads are mainly from DNA contamination, pre-mRNAs, new isoforms, or novel transcripts. Some UTR regions are overrepresented (i.e., higher reads density) because of DNA

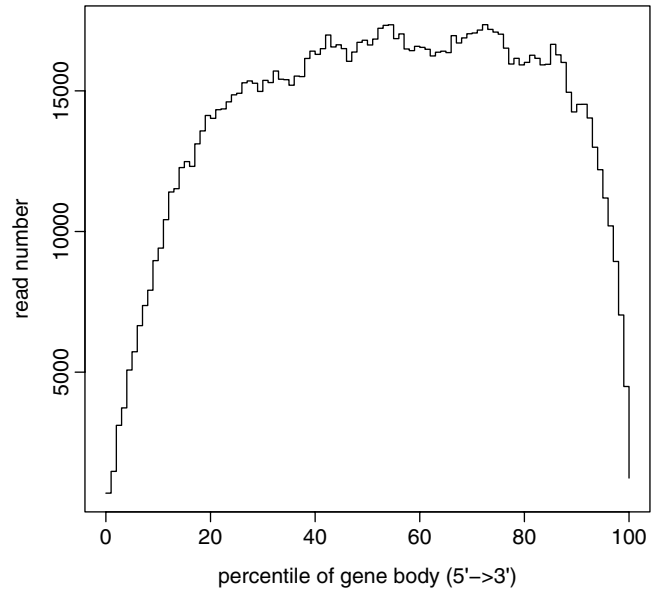

Fig. 5 Coverage uniformity over gene body. All transcripts were scaled into the same length (100 nucleotides) and then reads coverage (*Y*-axis) was calculated for each position (*X*-axis) from 5′ to 3′ end

repeats or PCR bias, but most often they have lower reads density due to RNA degradation. Specially, when poly-A RNA-seq protocol was used, reads are biased (i.e., overrepresented) in 3′UTR.

1.10 Strand Specificity

In mammalian genomes, many genes encoded on different strands are partially or completely overlapped. Therefore, conventional RNA-seq cannot decode the complex transcriptome due the lack of RNA polarity information. Strand specific RNA-seq experiments can overcome these limitations and is better suited for genome annotation (such as identify anti-sense noncoding RNA, demarcate the exact boundaries of adjacent genes transcribed on different strand), de novo transcriptome assembly, accurate digital gene expression and variant calling. Strandness is achieved by degrade one cDNA strand or by ligating distinct RNA adapters to the 5′ and 3′ ends of each RNA molecule prior to cDNA synthesis [24]. The specificity of a strand-specific protocol can be calculated by comparing the strand of reads to the strand of a reference gene. For example, if there is only one gene located in the forward strand in a particular region, one would expect the majority (>90 %) of reads mapped to this region will be forward reads (determined from protocol). Further, one can use spliced reads to measure strand specificity, because the strandness of spliced reads can be easily inferred from splicing motif (such as GT/AG).

2 Notes

1. Two different equations, standard Sanger format and Solexa/Illumina format (prior to v1.3) have been used to calculate Q [25]. After Illumina Pipeline v1.4, the quality scheme has been changed into standard Sanger scale.

$$Q_{\text{Sanger}} = -10 \times \log_{10} P; \quad Q_{\text{Solexa(prior\_to\_v1.3)}} = -10 \times \log_{10} \frac{P}{1-P}$$

2. FASTQ file stores nucleotide sequences and Phred qualities in the same file. For example:

@read_1 ← (read identifier)

```
CCGGCCCCAGGCTCCTGTCTCCCCCCAGGTGTGTGGTGATGCCAGGCATG←sequence
```

+read_1 ← (read identifier, can be omitted)

```
@@;D?DADF=FHFBEBH+A2ADFGFH:;09?*/?BB=F)==FH3=@CH@?←quality
```

@read_2 ← (another read identifier)

```
CTGCTCTCTCTTGCTGATGGACAAGGGGGCATCAAACAGCTTCTCCTCTG
```

+read_2 ← (another read identifier, can be omitted)

```
CCCFFFFFHHHGHJJIJJJJJJJIJJJJJJIIJIJJFGIIIJJIIIIIGJ
```

References

1. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. doi:10.1038/nmeth.1226
2. Marioni JCJ, Mason CEC, Mane SMS et al (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Gene Dev 18:1509–1517. doi:10.1101/gr.079558.108
3. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. doi:10.1038/nrg2484
4. Wilhelm BT, Landry J-R (2009) RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. Methods 48:249–257. doi:10.1016/j.ymeth.2009.03.016
5. Wang ET, Sandberg R, Luo S et al (2008) Alternative isoform regulation in human tissue transcriptomes. Nature 456:470–476. doi:10.1038/nature07509
6. Katz Y, Wang ET, Airoldi EM, Burge CB (2010) Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nat Methods 7:1009–1015. doi:10.1038/nmeth.1528
7. Trapnell C, Williams BA, Pertea G et al (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515. doi:10.1038/nbt.1621
8. Cabili MN, Trapnell C, Goff L et al (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Gene Dev 25:1915–1927. doi:10.1101/gad.17446611
9. Guttman M, Garber M, Levin JZ et al (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510. doi:10.1038/nbt.1633
10. Prensner JRJ, Iyer MKM, Balbin OAO et al (2011) Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. Nat Biotechnol 29:742–749. doi:10.1038/nbt.1914
11. Kannan K, Wang L, Wang J et al (2011) Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci U S A 108:9172–9177. doi:10.1073/pnas.1100489108
12. Pflueger D, Terry S, Sboner A et al (2011) Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. Gene Dev 21:56–67. doi:10.1101/gr.110684.110
13. Edgren H, Murumagi A, Kangaspeska S et al (2011) Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 12:R6. doi:10.1186/gb-2011-12-1-r6
14. Peng ZZ, Cheng YY, Tan BC-MB et al (2012) Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat Biotechnol 30:253–260. doi:10.1038/nbt.2122
15. Bahn JHJ, Lee J-HJ, Li GG et al (2012) Accurate identification of A-to-I RNA editing in human by transcriptome sequencing. Gene Dev 22:142–150. doi:10.1101/gr.124107.111
16. Park EE, Williams BB, Wold BJB, Mortazavi AA (2012) RNA editing in the human ENCODE RNA-seq data. Gene Dev 22:1626–1633. doi:10.1101/gr.134957.111
17. Ramaswami G, Zhang R, Piskol R et al (2013) Identifying RNA editing sites using RNA sequencing data alone. Nat Methods. doi:10.1038/nmeth.2330
18. Benjamini Y, Speed TP (2012) Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res 40:e72. doi:10.1093/nar/gks001
19. Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131. doi:10.1093/nar/gkq224
20. Ewing B, Hillier L, Wendl MC, Green P (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res 8(3):175–85
21. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res 8(3):186–94
22. Babraham Bioinformatics – FastQC a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
23. Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics. Oxford, England. doi:10.1093/bioinformatics/bts356
24. Levin JZ, Yassour M, Adiconis X et al (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods. doi:10.1038/nmeth.1491
25. Cock PJ, Fields CJ, Goto N, Heuer ML, Rice PM (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 38(6):1767-71. doi:10.1093/nar/gkp1137