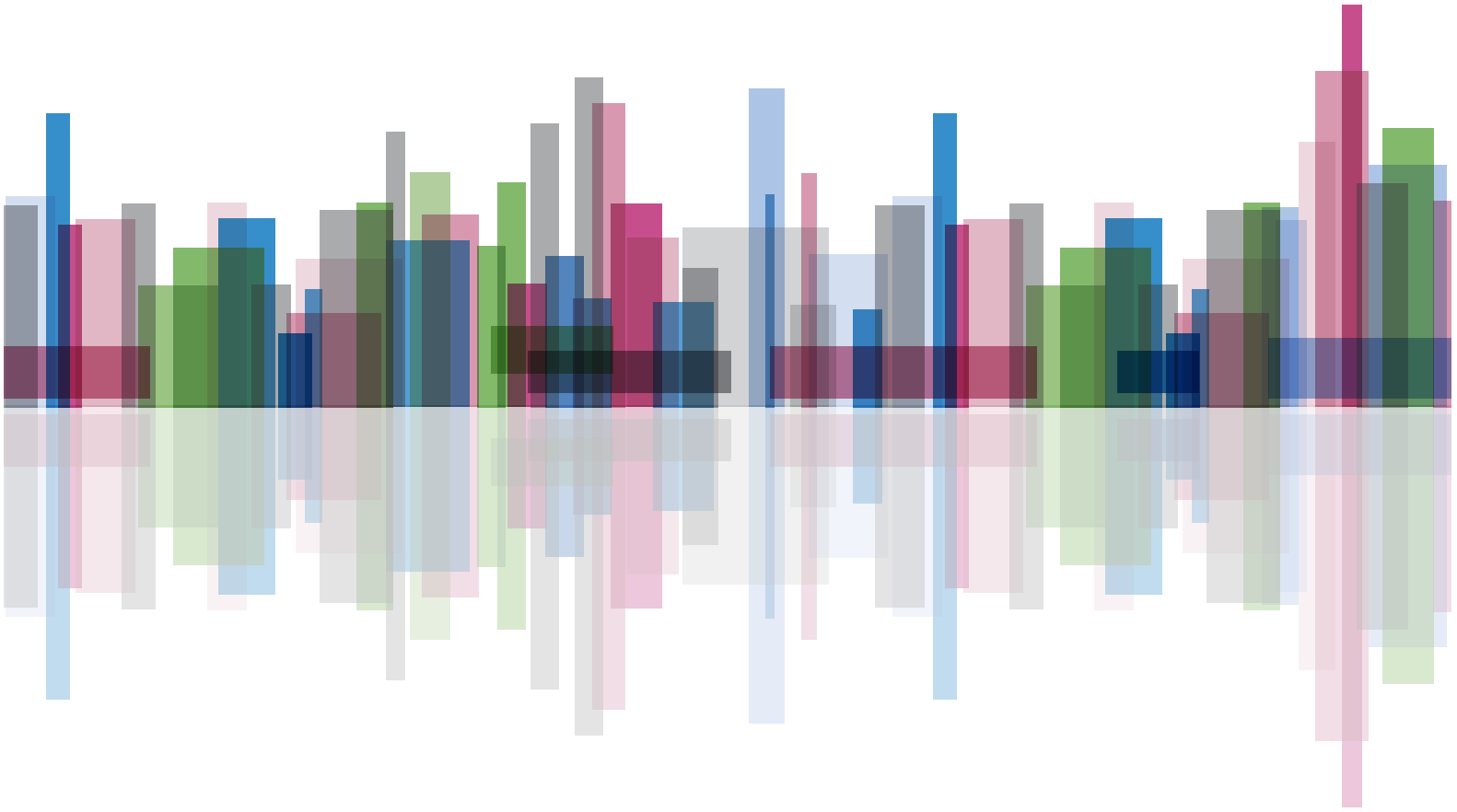
A GUIDE TO
RNA-Seq

GENEWIZ®
A Brooks Life Sciences Company

Contents

Overview

High-throughput sequencing of RNA (RNA-Seq) has revolutionized biological research by providing a new means for quantitative measurement of transcription, a highly dynamic process that controls many cellular functions. This guide highlights the *what, why*, and *how* of RNA-Seq to help you understand your options as you embark on your RNA-Seq project. It is intended for an audience with a solid background in molecular biology, and a beginner-to-intermediate understanding of next generation sequencing (NGS).



Anatomy of the Transcriptome

A transcriptome encompasses the full range of **coding RNA** and **non-coding RNA** transcripts expressed by an organism, also referred to as **total RNA** (Table 1). The term "transcriptome" can also be used to describe the array of mRNA transcripts produced in a particular cell or tissue type. In contrast with the genome, the transcriptome actively changes due to many factors, including the organism's developmental stage and environmental conditions. RNA-Seq can be used to simultaneously measure expression in thousands of genes under one condition or compare it across multiple conditions, the latter is known as **differential gene expression (DGE)**.

Table 1. Types of RNA in the Transcriptome

	CODING	NON-CODING					
			SMALL RNA				
Type	Messenger RNA (mRNA)	Long non-coding RNA (lncRNA)	MicroRNA (miRNA)	Short interfering RNA (siRNA)	Piwi-interacting RNA (piRNA)	Ribosomal RNA (rRNA)	Transfer RNA (tRNA)
Length	>200 nt	>200 nt	21-25 nt	20-25 nt	26-31 nt	120-5000 nt	75-95 nt
Modifications	Eukaryotes: 5′ cap, 3′ poly(A) tail	N^6-methyl-adenosine (m^6A)	5′ phosphate, 3′ hydroxyl	5′ phosphate, 3′ hydroxyl	5′ phosphate, 3′ end 2′-O-methylation	2′-OMe of sugar	5′ phosphate, 3′ hydroxyl
Organisms	All	Eukaryotes	Animals Plants	Eukaryotes Viruses	Animals	All	All

Coding RNA: or messenger RNA (mRNA), is an RNA molecule that is translated into proteins.

Non-coding RNA: an RNA molecule that is not translated into a protein. Non-coding RNA typically includes, but is not limited to, rRNA, tRNA, lncRNA, miRNA, piRNA, and snRNA.

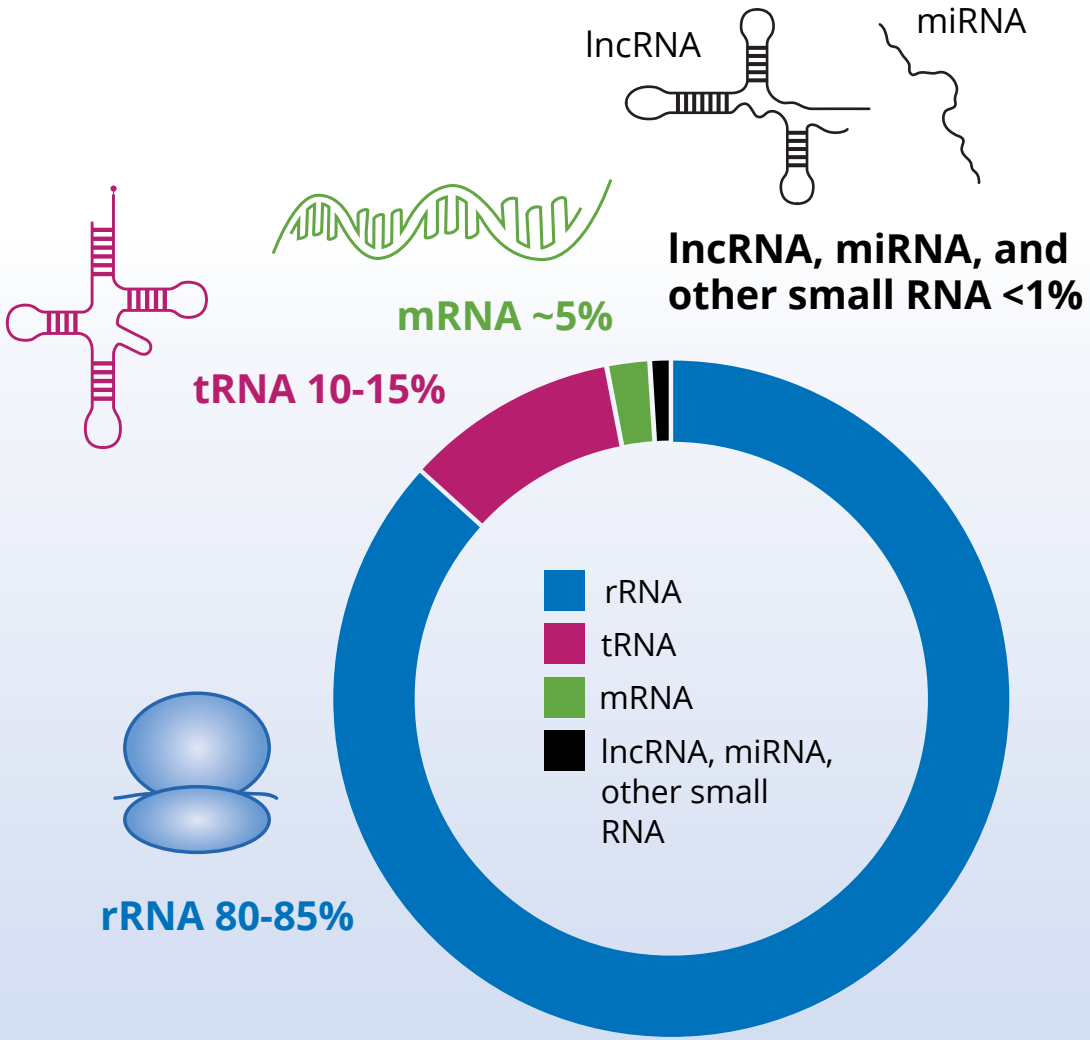
Total RNA: the complete set of RNA in a cell. Also referred to as the raw, unbiased fraction of RNA molecules obtained after extraction.

Differential gene expression (DGE): quantitative differences in gene expression levels between two samples that differ by one or more factors (e.g. environmental condition, developmental stage, or cell type).

Did You Know?

Most RNA Found in a Cell is Not mRNA

Which is Typically the RNA of Interest!



In eukaryotic cells, roughly 80% of transcripts are ribosomal RNA (rRNA).

For this reason, methods to remove rRNA or enrich for mRNA are commonly used in RNA-Seq experiments.

What is RNA-Seq?

In today's research, RNA-Seq is an indispensable tool and the most frequently used technology for transcriptome analysis, enabling high-throughput profiling of coding and non-coding RNA at single-nucleotide resolution. RNA-Seq data begins with harvesting total RNA from cells and purifying the RNA molecules of interest. Single-stranded RNA is then converted into double-stranded complementary DNA (cDNA) strands in a reverse transcription reaction. Sequencing adapters and barcodes are then added to create RNA-Seq libraries that are subsequently analyzed with next generation sequencing (NGS). The data, called **'reads'**, is then mapped to the genome if the sequence is available. The number of reads aligning to a region represents the transcriptional activity of that gene (Figure 1).

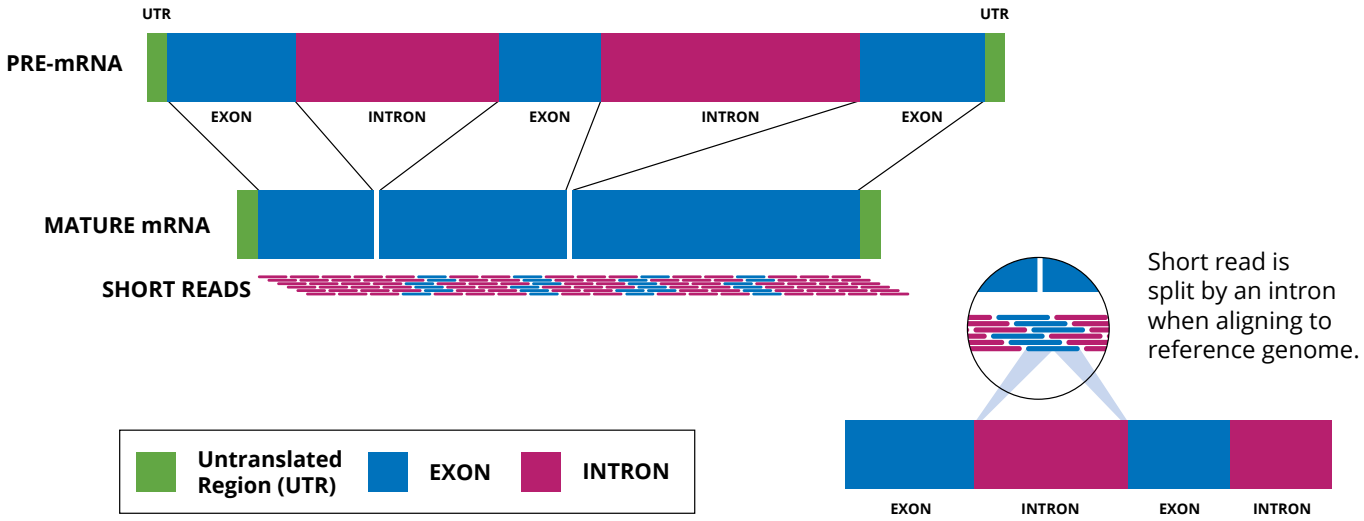


Figure 1. Different parts of a pre-mRNA showing intron (pink), exon (blue), and **splice junctions** processed into mRNA. Removal of the introns by splicing at the junctions gives rise to coding mRNA. Short reads from RNA-Seq experiments can be assembled to determine coding regions in mRNA and aligned with a **reference genome** to map read counts.

Reads: in general terms, a sequence "read" refers to the data string of A, T, C, and G bases corresponding to the sample DNA. In more specific terms, each cluster on the flow cell produces a single sequencing read. For example, 10,000 clusters on the flow cell would produce 10,000 single reads, and 20,000 paired-end reads. Existing sequencing technology on the market can generate millions of reads in a single run.

Splice junctions: the exon-intron junctions at which splicing occurs.

Reference genome: a high-quality, representative sequence of a species' genome.

Information Provided by RNA-Seq

RNA-Seq can provide a comprehensive or targeted characterization of the transcriptome. By studying transcriptomes, researchers can determine *when* and *where* genes are expressed. In general, two types of information can be obtained from RNA-Seq: qualitative (i.e. **genome annotation**) and quantitative (Table 2).

Table 2. Types of information provided by RNA-Seq

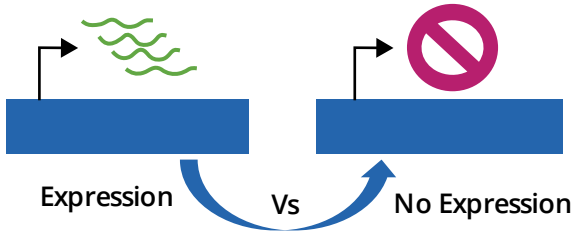
QUALITATIVE	QUANTITATIVE
• Genome annotation • Orientation of transcripts • Transcriptional start sites • Exon/intron boundaries • Polyadenylation sites • **Alternative splicing** (isoforms) • Gene fusions • Variant discovery	• Gene expression levels (relative or absolute) • Differential expression (comparing levels across two or more conditions) • Isoform expression levels

Genome annotation: the process of identifying the locations of genes in a genome.

Alternative splicing: a regulated process during gene expression that results in a single gene coding for multiple transcripts or isoforms, and thus proteins.

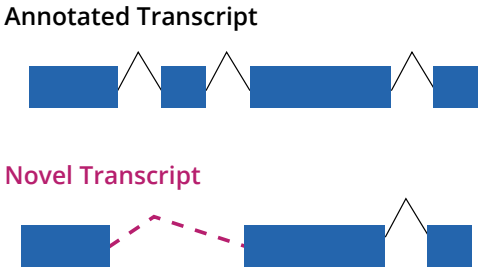
What Can You Do With This Information?

While there are a number of ways to use the data provided by RNA-Seq, here are the most common applications:



Study Changes in Gene Expression

***De Novo* Transcript Assembly**



Novel Transcript Discovery

Advantages of RNA-Seq Over Other Technologies

RNA-Seq, microarrays, and expressed sequence tag (EST) sequencing are all valuable assays for in-depth gene expression analysis. But which method should you choose? It depends on the goals of the project, your budget, and the organism of interest. Table 3 highlights some of the advantages of utilizing RNA-Seq over other transcriptomic assays.

Table 3. Comparison of RNA Analysis Technologies

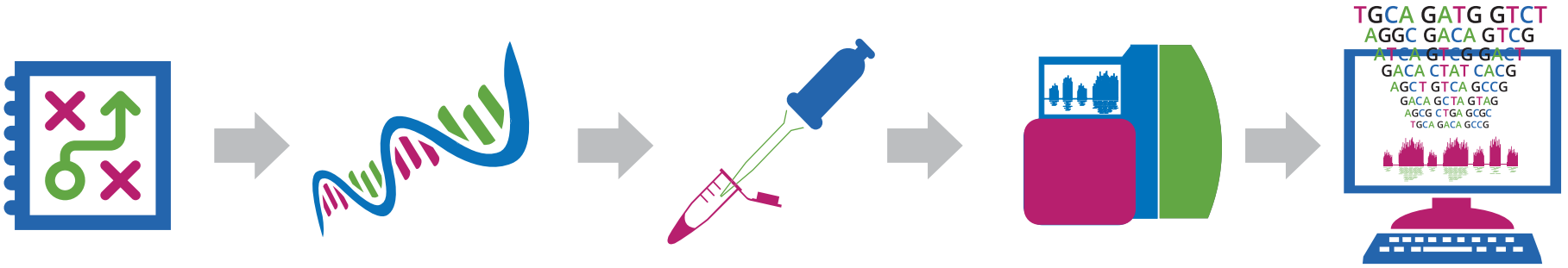
	RNA-SEQ	DNA MICROARRAYS (OR DNA CHIPS)	EXPRESSED SEQUENCE TAGS (EST)
Detection methods	High-throughput sequencing	DNA hybridization	Sanger sequencing
Throughput	High	High	Low
Resolution	Single nucleotide	Several to 100 nt (based on probe length)	Single nucleotide
Reliance on reference genome	In some cases	Yes	No
Dynamic range of expression level	>8000-fold	Few hundred-fold	Highly limited
Method of comparing gene expression	Read counts	Relative intensities	N/A
Background noise	Low	High	Low
SNP* detection in the transcribed regions	Yes	No	Yes
Identification of novel genes and isoforms	Yes	No	Yes
Ability to distinguish allelic expression	Yes	No	Yes
Cost of mapping transcriptomes of large genomes	Relatively low	High	High

*SNP = single nucleotide polymorphism

RNA-Seq Workflow

Before beginning an RNA-Seq experiment, you should understand and carefully consider each step of the RNA-Seq workflow (Figure 2), including experimental design, extraction, library preparation, sequencing, and data analysis.

Click the icons to view more details



1. Experimental Design

Define the objectives and design the assay accordingly.

2. Extraction

Isolate and purify input RNA.

3. Library Preparation

Enrich target RNA.

Convert RNA to cDNA.

Add sequencing adapters.

4. Sequencing

Sequence libraries using NGS platform.

5. Data Analysis

Trim, filter, and map reads.

Perform advanced analysis.

Figure 2. Simplified RNA-Seq Workflow

1. Experimental Design

The design stage of your experiment is arguably the most critical step in ensuring the success of an RNA-Seq experiment. Researchers must make key decisions at the start of any NGS project, including the type of assay and the number of samples to analyze. The optimal approach will depend largely on the objectives of the experiment, hypotheses to be tested, and expected information to be gathered.

SET YOUR PROJECT UP FOR SUCCESS

There are many factors to take into account when planning your experiment. We've put together a guide that will help you through the process.

10 Questions to Consider When Designing Your RNA-Seq Experiment

JUMP TO SECTION

2. Extraction

The first step in characterizing the transcriptome involves isolating and purifying cellular RNA. The quality and quantity of the input material have a significant impact on data quality; therefore, care must be taken when isolating and preparing RNA for sequencing. Given the chemical instability of RNA, there are two major reasons for RNA degradation during experiments:

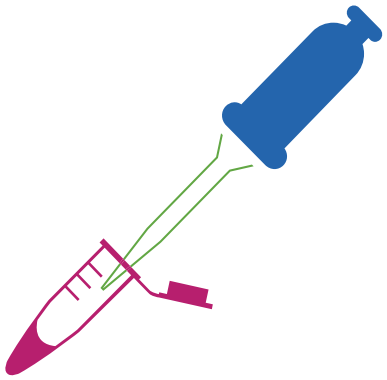
(1) RNA contains ribose sugar and is not stable in alkaline conditions because of the reactive hydroxyl bonds. RNA is also more prone to heat degradation than DNA.

(2) **Ribonucleases** (RNases) are ubiquitous and very stable, so avoiding them is nearly impossible. It is essential to maintain an RNase-free environment by wearing sterile disposable gloves when handling reagents and RNA samples, employing RNase inhibitors, and using **DEPC**-treated water instead of PCR-grade water. Additionally, proper storage of RNA is crucial to avoid RNA degradation.

In the short term, RNA may be stored in RNase-free water or TE buffer at -80°C for 1 year without degradation. For the long term, RNA samples may be stored as ethanol precipitates at -20°C. Avoid repeated freeze-thaw cycles of samples, which can lead to degradation. RNA of high integrity will maximize the likelihood of obtaining reliable and informative results.

Ribonuclease (RNase): a type of nuclease that catalyzes the degradation of RNA into smaller components.

Diethyl pyrocarbonate (DEPC): an effective RNase inhibitor, which covalently modifies certain amino acids. Treated water is autoclaved prior to use to inactivate leftover DEPC, rendering it safe for downstream applications.



3. Library Preparation

Library preparation involves generating a collection of RNA fragments that are compatible for sequencing. The process involves enrichment of target (non-ribosomal) RNA, fragmentation, reverse transcription (i.e. cDNA synthesis), and addition of sequencing adapters and amplification (Figure 3). The enrichment method determines which types of transcripts (e.g. mRNA, lncRNA, miRNA) will be included in the library. In addition, the cDNA synthesis step can be performed in a such a way as to maintain the original strand orientation of the transcript, generating what is known as 'strand-specific' or 'directional' libraries.

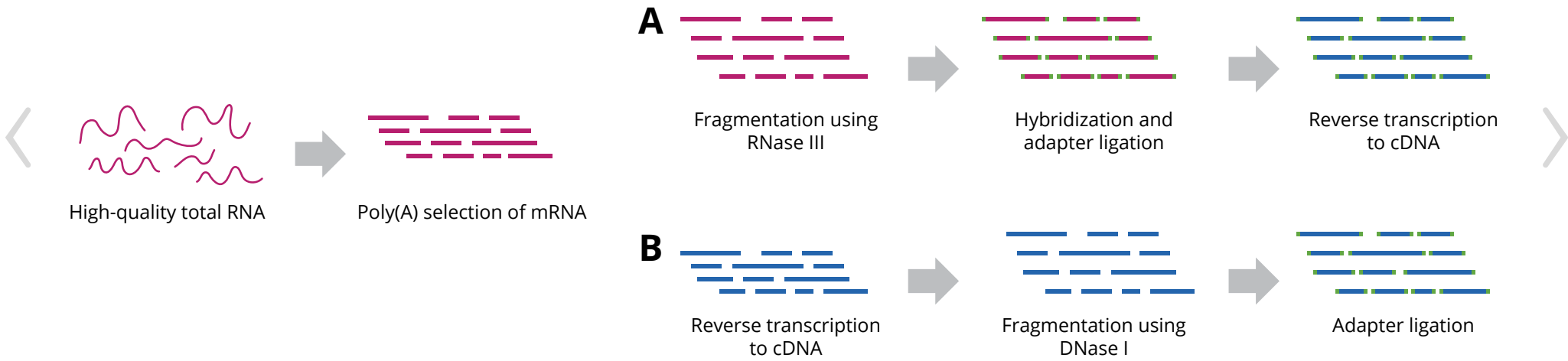


Figure 3. Typical library preparation for eukaryotic mRNA. Polyadenylated RNA is enriched from total RNA. For short-read sequencing, RNA or cDNA must be fragmented prior to addition of sequencing adapters. In panel A, fragmentation is done using RNase III, whereas panel B uses DNase I to fragment cDNA, followed by hybridization and adapter ligation.

4. Sequencing

Parameters for sequencing—such as read length, configuration, and output—depend on the goals of your project and will influence your choice of instrument and sequencing chemistry (Table 4). The main NGS technologies can be grouped into two categories: short-read (or 'second generation') sequencing, and long-read (or 'third generation') sequencing. Both have distinct benefits for RNA-Seq.

Table 4. Select Sequencing Platforms and Configurations

PLATFORM*	TYPE	READ CONFIGURATION**	NUMBER OF READS	DATA OUTPUT
Illumina NovaSeq™ 6000	Short-read	2x150 bp (PE150)	2.0-2.5 billion per S4 lane	600-750 Gb per S4 lane
Illumina HiSeq® X	Short-read	2x150 bp (PE150)	300-350 million per lane	90-110 Gb per lane
Illumina MiSeq™ 550	Short-read	2x150 bp (PE150)	350-400 million per flow cell	100-120 Gb per flow cell
PacBio® Sequel®	Long-read	Up to 500 kb (average read lengths up to 30 kb)	Up to 600,000	Up to 20 Gb

*Other technologies available are beyond the scope of this guide.

**Other configurations are available with different outputs. The 2x150 bp configuration is one of the most common for short-read sequencing. It is also known as 'paired-end 150 bp sequencing', or 'PE150'.

Short-Read Sequencing

Short-read sequencing is relatively inexpensive on a per-base basis and can generate billions of reads in a massively parallel manner (Figure 4), with single-end read lengths ranging between 50 and 300 bp. The high-throughput nature of this technology is ideal for quantifying the relative abundance of transcripts or identifying rare transcripts. Several platforms available on the market offer flexible outputs using roughly similar chemistry.

Multiple samples are typically multiplexed, or run together, on a **flow cell** to make experiments more cost-effective. The choice of platform (and flow cell) together with the number of samples multiplexed determines the **sequencing depth**. Each cDNA fragment can be sequenced from only one end, called **single-end (SE) sequencing**, or both ends, called **paired-end (PE) sequencing**. The former is generally less expensive and faster than the latter. However, paired-end sequencing helps detect genomic rearrangements and repetitive sequence alignments better than the single-end configuration, since more information is collected from each fragment.

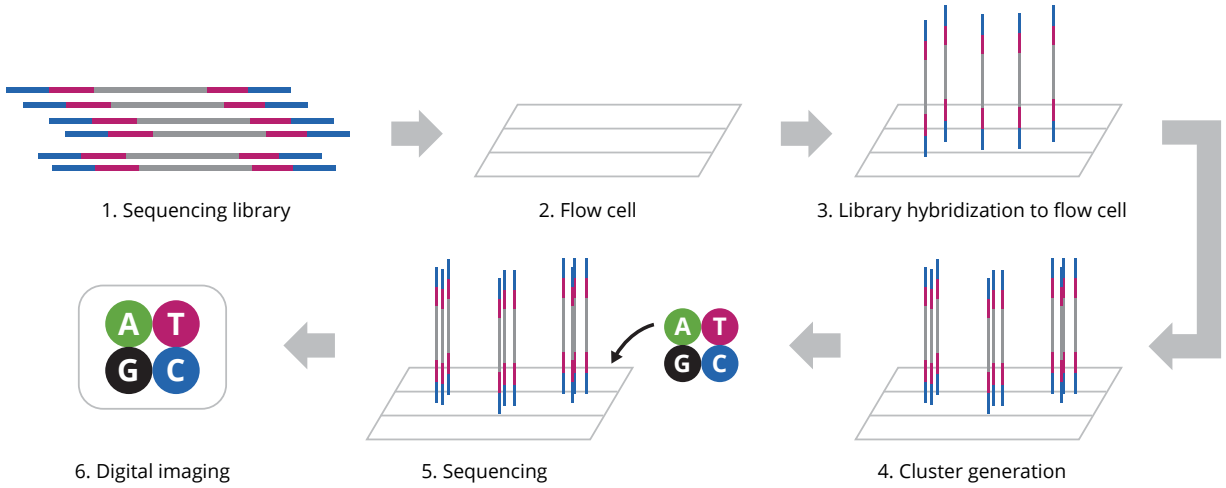


Figure 4. Illumina sequencing overview. Libraries are loaded onto flow cells. Fragments bind to immobilized oligos on the surface and are then amplified, creating clustered clones. Sequencing cycles allow sequential identification of nucleotides within each cluster.

Flow cell: a glass slide with small fluidic channels containing short oligonucleotides complementary to the adapter sequence, through which polymerases, dNTPs, and buffers can be pumped. Sequencing reactions occur on the surface of the flow cell.

Sequencing depth: typically expressed as the number of reads per sample for RNA-Seq experiments.

Single-end sequencing: a strategy in which fragments are only read from one end, generating one sequence per fragment.

Pair-end sequencing: a strategy in which fragments are read from both ends, generating two sequences per fragment. Compared to single-end sequencing, this approach provides greater physical coverage and alleviates several limitations of short-read NGS platforms.

Long-Read Sequencing

Long-read sequencing can resolve inaccessible regions of the genome and read through the entire length of RNA transcripts, allowing precise determination of specific isoforms. Two of the leading long-read sequencing platform providers include Pacific Biosciences (PacBio), and Oxford Nanopore Technologies®. The PacBio platform uses Single-Molecule Real-Time (SMRT®) sequencing to generate exceptionally long reads reaching up to 10 kb in length. This method, called 'isoform sequencing' or 'Iso-Seq', produces full-length transcripts[1]. The Oxford Nanopore platform uses a unique sequencing technology for direct, real-time analysis of exceedingly long transcripts, generating reads up to >20 kb in length[2]. In non-model species with no genomic references available, long reads can provide valuable information to detect full-length transcripts accurately (Figure 5). However, if cost reduction is paramount and/or high data output is required, short-read sequencing is a better choice.

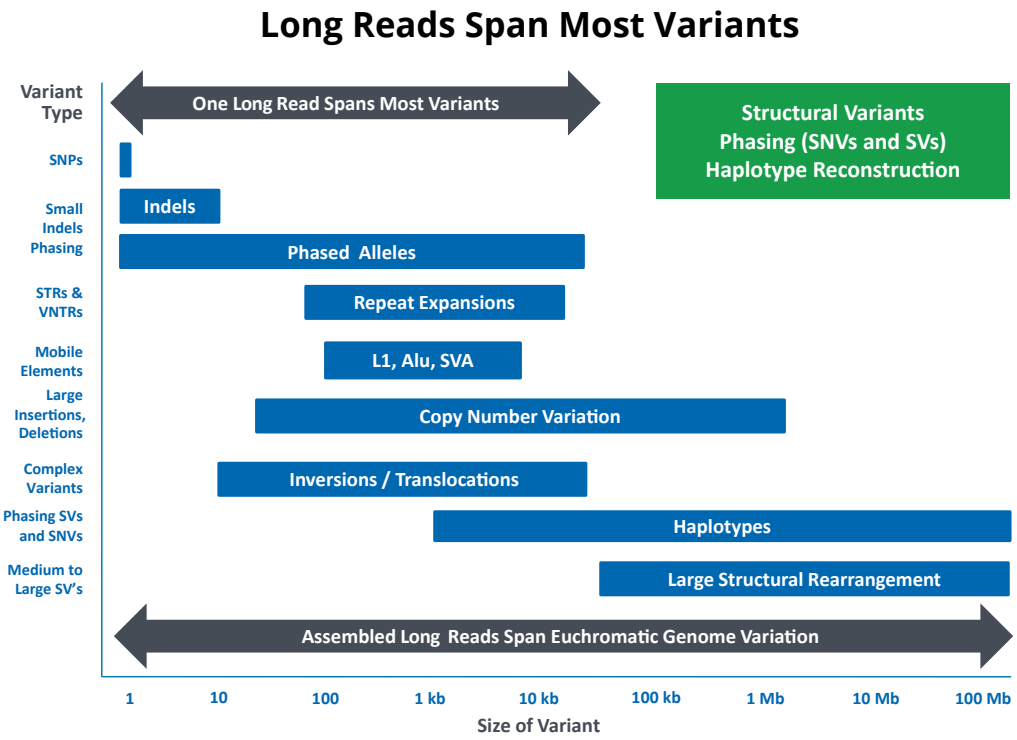


Figure 5. Capabilities of long-read sequencing. A single long read can discover a wide range of variants from SNPs and small Indels to large insertions and deletions. Assembled long reads also allow for identification of large structural rearrangements and phasing of variants, even in highly repetitive sequences.

Return to RNA-Seq Workflow

5. Data Analysis

Evaluating your data quality and extracting biologically relevant information is the final and most rewarding step in an RNA-Seq experiment. It is important to discuss your project with an experienced bioinformatician to find the best analysis pipeline for your data. One pipeline does not fit all approaches.

In general, raw sequencing data is pre-processed or 'trimmed' to remove adapter sequences and low-quality reads. If a reference genome is available, reads are typically aligned (or mapped) to the reference to discover the genomic origin of the RNA molecule. In samples lacking a reference genome, *de novo* transcriptome assembly can be performed using overlapping reads. Once reads are mapped, more sophisticated analyses such as differential gene expression (DGE) and isoform variant discovery provide researchers with detailed views of dynamic gene expression profiles (Figure 6).

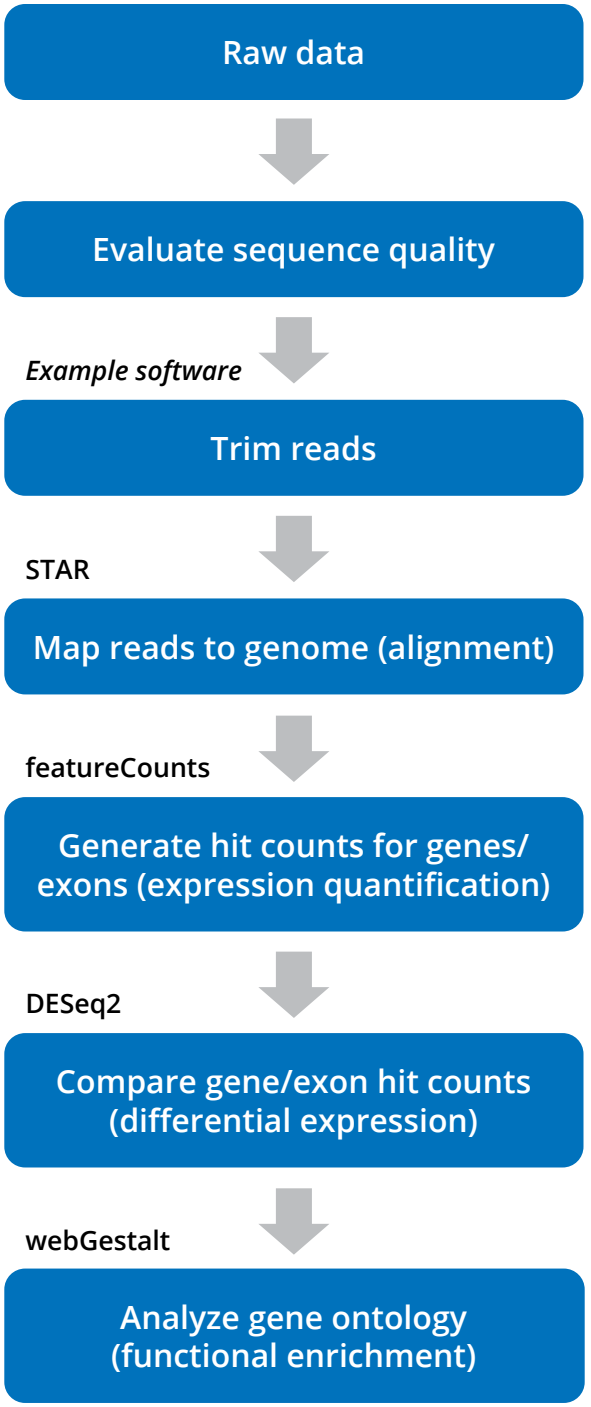


Figure 6. Example of an RNA-Seq bioinformatic pipeline and analysis software. Following read quality assessment and adapter trimming, reads are aligned to a reference genome, if available using STAR[3], or assembled via a *de novo* approach. Next, the expression level of each gene is estimated by counting the number of reads that align to each exon or full-length transcript by featureCounts[4]. Downstream analyses with RNA-Seq data include differential gene expression using DESeq2[5], and functional enrichment analysis using WebGestalt[6].

Alignment

Many reads map across splice junctions; therefore, aligning RNA-Seq reads to the genome can be more challenging than mapping genomic (DNA-Seq) reads. Quantifying transcript levels using RNA-Seq data requires aligning the reads to a reference genome or transcriptome, counting the reads per feature, and then performing DGE analysis. Many free software packages and read mapping algorithms (e.g. Bowtie2, Tophat, Cufflinks) are available.

Annotation

In addition to genes annotated in current databases, many novel genes/transcripts can still be discovered using RNA-Seq. High-throughput generation of transcriptomic data has been essential to gene annotation. With accurately assembled transcripts, RNA-Seq can precisely define exon-intron boundaries in the genome. Without the guidance of a reference genome, *de novo* assembly of transcriptomes remains a complex task requiring efficient management of large datasets.

Normalization

Normalization is essential to correct biases that may impact within-sample or between-sample comparisons. Factors such as gene length, GC-content, and sequencing depth can affect the raw read counts, potentially having a significant impact on downstream analysis. Each normalization method (e.g. **RPKM** and Sailfish) relies on certain assumptions to adjust the data, and it is imperative to select one with assumptions valid for your experiment. Although a variety of normalization schemes have been devised to combat inherent variability in RNA-Seq data, the scientific community has yet to reach a consensus on the best ones.

Differential Gene Expression (DGE) Analysis

Similar to microarray experiments, identification of differentially expressed genes between two or more groups is one of the most common applications of RNA-Seq. DGE analysis involves statistical comparison of normalized read count data to discover quantitative changes in expression levels between experimental groups. Normalized read counts are usually expressed as the number of reads or fragments per kilobase per million reads (RPKM and **FPKM**, respectively)[7]. The FPKM transformation also permits direct comparison of transcript expression between two libraries with different sequencing depth, or between two or more transcripts in the same library.

Reads per kilobase per million (RPKM): a normalized unit of transcript expression, calculated by the following formula: (number of reads mapped to a gene sequence) / (gene length in kb) × (total number of reads) / 1,000,000).

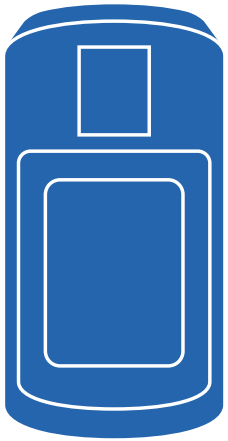
Fragments per kilobase per million (FPKM): a normalized unit of transcript expression used for paired-end sequencing. It is very similar to RPKM except it takes into account that two reads can map to one fragment; it does not count the fragment twice in this situation.

Return to RNA-Seq Workflow

Don't Forget!

Quality Control is Essential at Each Step of the RNA-Seq Workflow

The below instruments can be used at different stages of your experiment to ensure quality is maintained throughout.

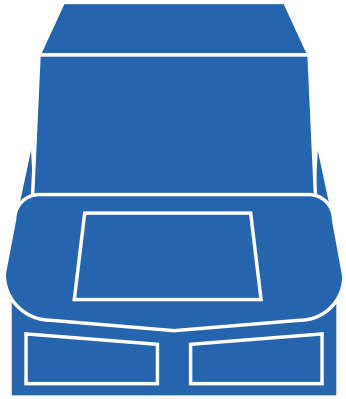


Qubit® or NanoDrop™

Indicates quantity of RNA and/or DNA present in samples

Agilent® Bioanalyzer™ or TapeStation™

Analyzes fragment length, quality, and integrity



qPCR

Reveals how much of the library can be sequenced

What Types of RNA-Seq Are Available?

Several RNA-Seq assays have been developed depending on the target RNA species, starting material, and objective of the experiment (Table 5).

Table 5. Common Types of RNA-Seq

ASSAY TYPE	TARGET RNA	RNA SELECTION METHOD	STARTING MATERIAL	RELATIVE COST
mRNA-Seq	mRNA	Poly(A) selection	Total RNA, cells, tissues	$
Total RNA-Seq	mRNA + lncRNA	rRNA depletion	Total RNA, cells, tissues	$$
Strand-Specific RNA-Seq	mRNA and/or lncRNA	Poly(A) selection or rRNA depletion	Total RNA, cells, tissues	$$
Small RNA-Seq	Small non-coding RNAs (miRNA, siRNA, piRNA)	Size fractionation (with adapter ligation to 5' phosphate)	Total RNA, cells, tissues	$$
Single-Cell RNA-Seq	mRNA	Poly(A) selection after fractionation of individual cells	Cells	$$$$
Ultra-Low Input RNA-Seq	mRNA	Poly(A) selection	Total RNA (<20 ng) or cells (<1000)	$$$
Iso-Seq	mRNA	Poly(A) selection	Total RNA	$$$

mRNA-Seq

In eukaryotes, mRNA transcripts contain **polyadenylated** tails, which are used to enrich mRNA molecules through poly(A) selection. In this process, total RNA is isolated and subjected to either hybridization with oligo(dT)-conjugated beads/columns or reverse transcription with oligo(dT) primers. Polyadenylated RNA molecules make up just 1-5% of total RNA in many species, meaning sample concentrations after poly(A) selection are typically reduced by a factor of 20-100x. Purified mRNA is converted into cDNA libraries and amplified via PCR to enrich library concentration. Poly(A) selection is the most common and the most cost-effective option for eukaryotic mRNA library preparation.

Polyadenylated: refers to mRNA molecules having a stretch of repeating adenine nucleotides, which can be used to select mRNA from total RNA.

Total RNA-Seq

Total RNA-Seq is a method used for comprehensive analyses of protein-coding and long non-coding RNAs (lncRNAs). The latter have important regulatory functions in the genome and are of interest to molecular biologists due to their capacity for epigenetic regulation of transcriptional activity. Since many lack a poly(A) tail, lncRNA molecules are often excluded from poly(A) selection. In total RNA-Seq, oligos complementary to single-stranded rRNAs are used to capture and deplete these molecules prior to sequencing.

Strand-Specific RNA-Seq

Transcript polarity is important for correct annotation of genes. Since there are many genomic regions that generate transcripts from both strands, identifying the polarity of a given transcript provides essential information about the possible function of a gene. However, the polarity of transcripts can be lost during cDNA synthesis and subsequent amplification. Strand-specific RNA-Seq, also known as 'stranded' or 'directional' RNA-Seq, preserves this information during library preparation, allowing researchers to determine the orientation of the gene on the DNA template. It can be used in conjunction with mRNA-Seq and total RNA-Seq. There are at least two methods for creating stranded RNA-Seq libraries (Figure 7).

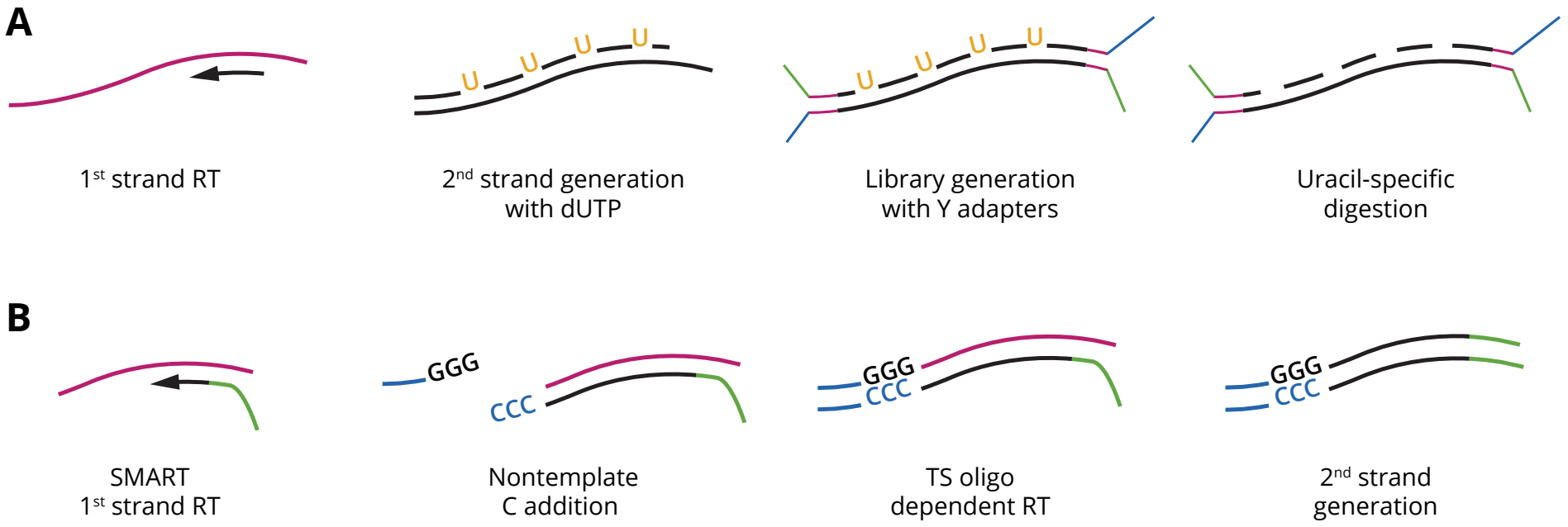


Figure 7. Strand-specific RNA library preparation. (A) Incorporation of dUTP during second-strand synthesis and subsequent uracil-specific digestion selects for the first-strand cDNA. (B) In SMART cDNA synthesis, unique adapters can be specifically attached to the 5' and 3' ends during cDNA synthesis, preserving orientation.

Small RNA-Seq

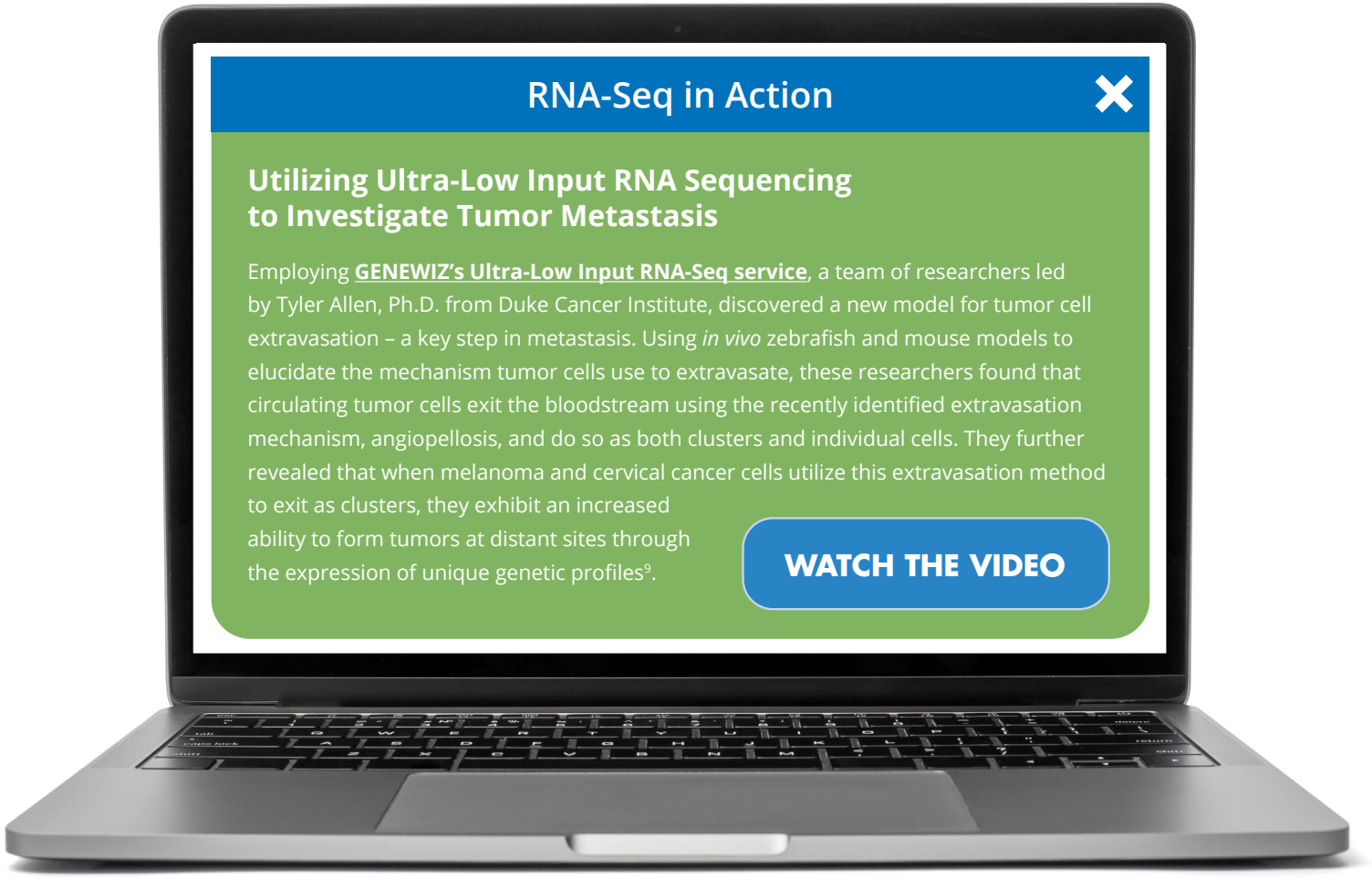
Short RNA transcripts, such as microRNAs (miRNA) and small interfering RNAs (siRNA), also play important gene regulatory functions in the cell[8]. In small RNA-Seq, RNA species are selected by size fractionation from total RNA. Library preparation typically includes ligation of sequencing adapters to 5' phosphate ends, which are found in small RNAs but absent in degraded fragments of larger RNA molecules, such as mRNA. The RNA fragments are then converted to cDNA libraries prior to sequencing. Since small RNA molecules can be as short as 21 nucleotides in length, sequencing configurations with fewer cycles (e.g. 1x50 bp) can be used.

Single-Cell RNA-Seq

Distinct from traditional "bulk" RNA-Seq methods, single-cell RNA-Seq (scRNA-Seq) allows researchers to capture the transcriptome of individual cells and uncover heterogeneous patterns of gene expression in complex cellular populations. Microfluidics or microwells are typically used to isolate single cells before library preparation. These methods preserve cellular information by adding unique barcodes to each transcript during isolation, which can be bioinformatically traced back to the cell of origin.

Ultra-Low Input RNA-Seq

Since faithful characterization of the transcriptome depends largely on the quality and quantity of the input RNA, standard RNA-Seq approaches call for an ample amount (>500 ng) of intact RNA. Samples producing lower yields and degraded RNA typically require additional amplification steps, as well as higher depths of sequencing to boost data output. These samples are prone to transcriptional bias and poor read mapping to exons. Ultra-low input methods have been developed to selectively amplify full-length transcripts with minimal bias, allowing researchers to perform RNA-Seq on samples containing as few as 10 pg of RNA or just a single cell.



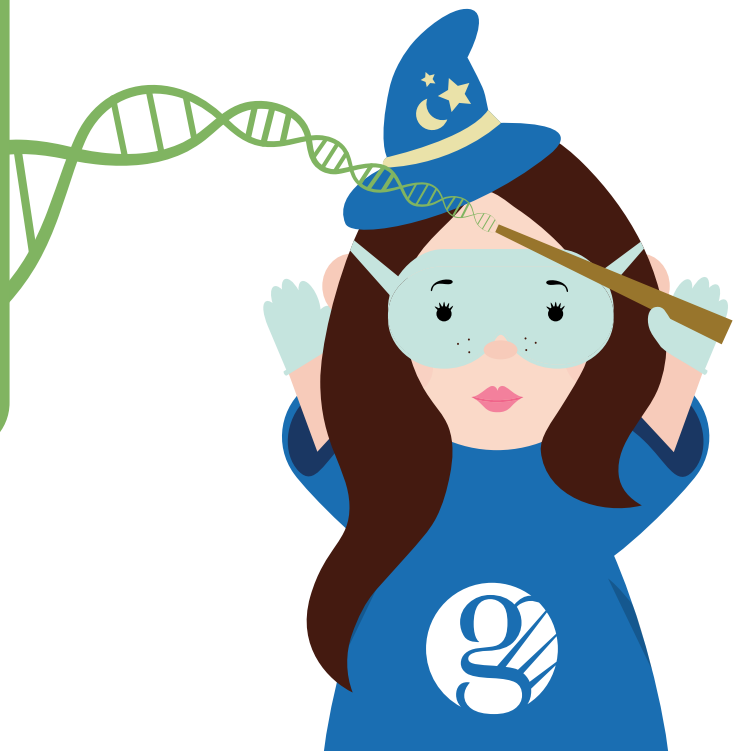
Isoform Sequencing

Alternative splicing results in multiple isoforms being encoded by a single gene, which can be effectively analyzed by isoform sequencing (Iso-Seq). Developed by PacBio, Iso-Seq uses long-read technology to sequence transcripts contiguously from end-to-end, eliminating the need for reconstruction. The result is unambiguous information about a transcript's start, polyadenylation, and splice sites from a single read. Iso-Seq characterizes the full complement of isoforms across the transcriptome, with potential applications including better annotated genomes, detection of gene fusions, and discovery of novel isoforms.

Detecting full-length mRNA sequences greatly simplifies genome annotation efforts as well as revolutionizes the discovery of novel RNA isoforms.

Learn how GENEWIZ has further optimized the capabilities of the PacBio Sequel to exceed the manufacturer's benchmarks in output and read length.

LEARN MORE



10 Questions to Consider When Designing Your RNA-Seq Experiment

Click on the question to get the answer!

Experimental Design

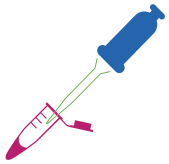
1. Which RNA-Seq assay should I use?

2. How many replicates are needed for RNA-Seq?

Extraction

3. How should I isolate RNA?

4. How should I measure RNA quality and quantity?



Library Preparation

5. Poly(A) selection or ribosomal depletion?

6. Stranded or non-stranded library preparation?

Sequencing

7. Which sequencing platform should I use?

8. Paired-end or single-end sequencing?

9. What read length and depth should I use?

Data Analysis

10. What types of data analyses are available?

Return to RNA-Seq Workflow

1. Which RNA-Seq assay should I use?

Several factors will influence which RNA-Seq assay to use, most importantly the experimental objectives. A decision tree is presented below as a starting point to help guide your RNA-Seq assay selection (Figure 8).

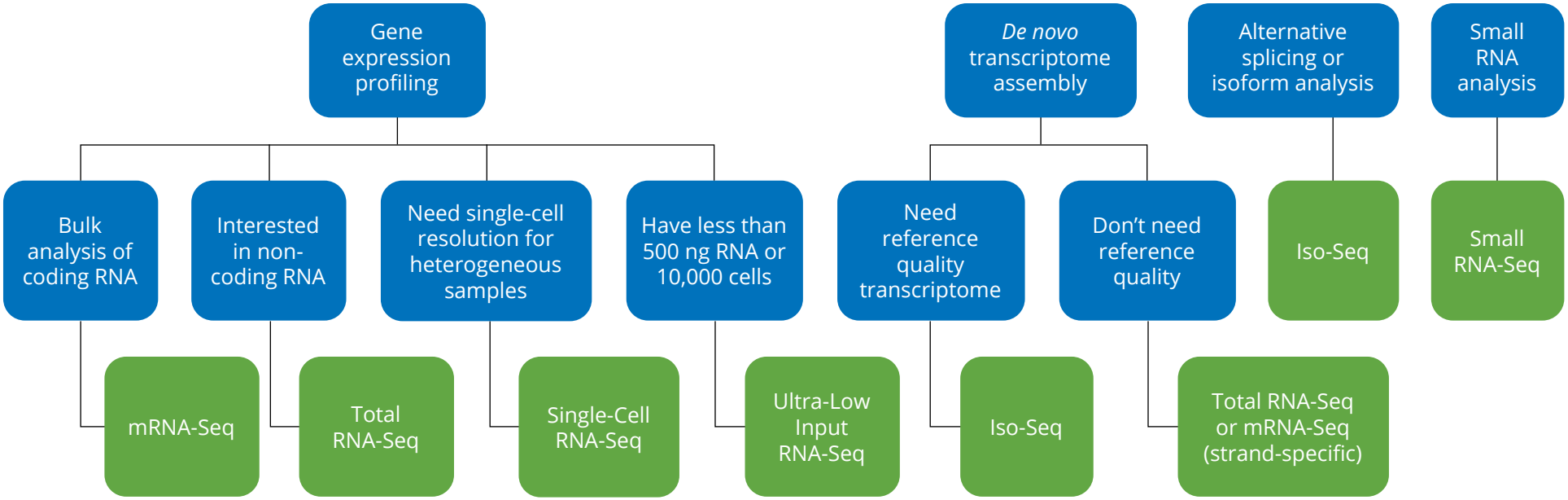


Figure 8. Simplified decision tree for selecting an RNA-Seq assay. The blue boxes indicate the objectives of the RNA-Seq study, and the green boxes represent the recommended RNA-Seq assay.

Return to 10 Questions

2. How many replicates are needed for RNA-Seq?

The number of samples depends largely on how many controls and replicates are to be included in the study. A minimum of 2 biological replicates are recommended to ensure valid biological interpretation of the results, and up to 3-5 replicates per condition for statistical analysis of the data.

Controls

Experimental controls are recommended for almost any sequencing experiment and can help improve reproducibility or identify experimental biases. RNA-Seq is a high-throughput quantitative method, and experimental controls are beneficial for identifying and reducing error rates. This can be achieved by including a known quantity of synthetic RNA to act as an internal reference for control variation between samples and minimization of batch effects[10]. For example, an equal quantity of spike-in RNA controls can be used to bioinformatically normalize the total amount of sequenced reads between samples, and provide a measure of sensitivity and specificity in RNA-Seq experiments[11].

Replicates

Choosing the appropriate number of biological replicates for an experiment is a trade-off between cost and precision. The number of replicates needed may change based on the amount of biological variability associated with the samples of interest, and should be empirically determined. Prior transcriptomic analyses can be used as a starting point for the use of biological replicates. If the goal is to investigate differences between treatments, biological replication is necessary to generalize the results to a larger population. The ENCODE consortium recommends a minimum of two replicates for each biological condition, and replicates generally range from two to five replicates per condition. As replicates increase, so does the reproducibility, precision, and cost of the assay. Technical replicates (libraries prepared from the same RNA sample) were commonly used in early RNA-Seq experiments, but today, biological variation is understood to far outweigh technical variation when coverage of at least 5 reads per nucleotide (5x coverage) is obtained[12].

3.

How should I isolate RNA?

RNA isolation has traditionally required phenol-chloroform extraction, which necessitates the use of harsh chemicals. Modern protocols are safer and more reliable, usually relying on columns or beads to bind and isolate RNA. However, RNA isolation does not always remove all genomic DNA traces from the samples, and thus researchers should include DNase digestion during the process.

Two common sources of RNA degradation are heat and ribonucleases (RNases). Therefore, researchers should keep RNA samples at 4°C during sample processing and always work in an RNase-free environment. Preserving RNA integrity is critical to ensuring samples are not degraded before arrival and processing at the NGS facility. The following are common approaches for preserving RNA integrity:

- *Cryogenics:* Flash freezing using liquid nitrogen and subsequent storage at -80°C or transporting on dry ice slows RNase activity and auto-hydrolysis.
- *Chemical reagents:* Commercially available stabilizing reagents (e.g. RNAlater®, RNAprotect®, and **RNA stabilization tubes**) inhibit RNases and prevent damage from RNA hydrolysis and oxidation.
- *Lyophilization:* Certain RNA stabilization strategies dehydrate RNA in the presence of a stabilization matrix. This prevents heat denaturation and digestion by RNases, as well as enables samples to be stored and shipped at room temperature.

4.

How should I measure RNA quality and quantity?

Quality (Integrity)

High-quality, undegraded RNA has the best chance of producing accurate and information-rich RNA-Seq results[13]. Degradation by RNases or heat denaturation breaks up long RNA molecules into shorter fragments.

These short fragments are more difficult to assemble bioinformatically, and therefore RNA sequence information can be lost. RNA quality and integrity can be measured using commercially available automated systems, such as the Agilent 2100 Bioanalyzer or TapeStation. Researchers have traditionally used rRNA to assess overall RNA quality, as it is abundant in total RNA. For example, mammalian genomes carry the 18S and 28S ribosomal sequences, which can be analyzed on a gel or microfluidic platform to determine RNA quality. Intact RNA samples will exhibit strong rRNA bands without excessive background signal. Degradation reduces the intensity of the 18S and 28S bands while increasing the proportion of shorter fragments. Many researchers aim for an **RNA integrity number (RIN)** value of 8 or above (Figure 9). Alternatively, an estimate of RNA integrity can also be assessed by running 2-4 µg of a total RNA sample on a 1% agarose denaturing gel and staining with ethidium bromide.

RNA integrity number (RIN): an algorithm for assigning integrity values to RNA measurements, based on the distribution of RNA lengths in a sample.

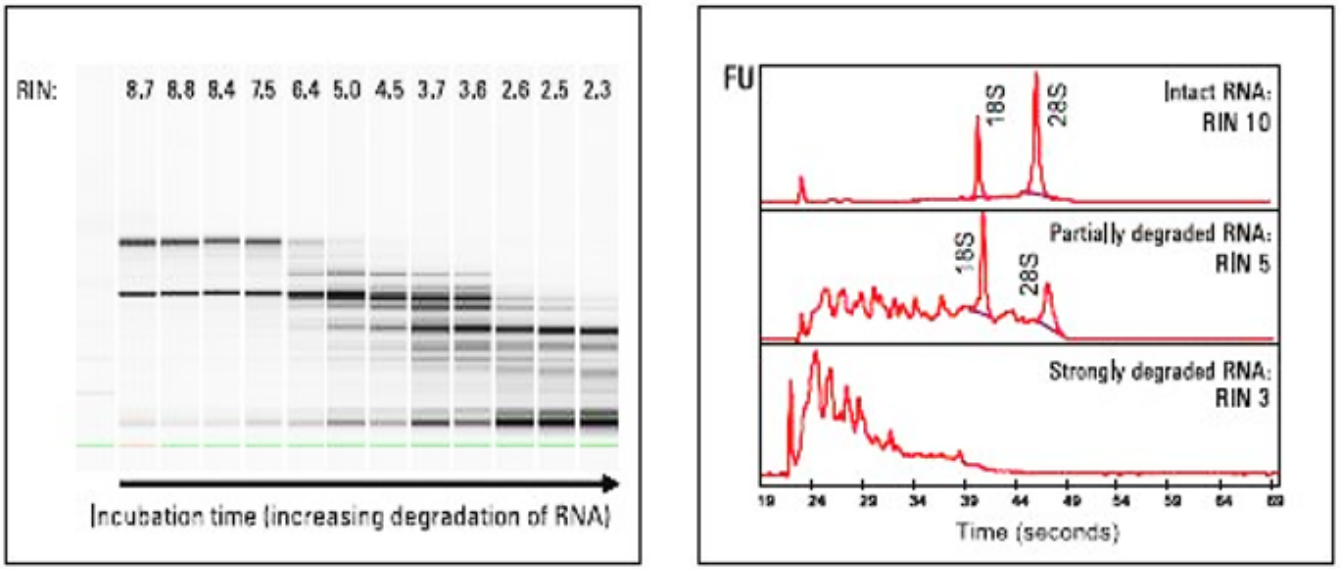


Figure 9. Results from RNA samples run on a TapeStation (left) or Bioanalyzer (right). RIN values range from 1 to 10, with 10 indicating fully intact RNA.

Quality (Purity)

Certain forms of contamination can be detected using UV spectroscopy and the calculation of absorbance ratios. Good quality RNA will have an $A_{260/280}$ ratio of 1.7 to 2.1 and $A_{260/230}$ of 2.0 to 2.2. $A_{260/280}$ lower than 1.7 may indicate protein contamination, whereas $A_{260/230}$ less than 1.8 may indicate the presence of organics (e.g. phenol or guanidine) leftover from the extraction protocol. Such contaminants may interfere with library preparation.

Quantity (Concentration)

Fluorescence-based methods, such as Qubit® or PicoGreen™, provide the most accurate measurement of RNA concentration. UV spectroscopy (e.g. NanoDrop™) is also commonly used, but because DNA and RNA both absorb light at 260 nm, DNA contamination can cause spectroscopic methods to overestimate the amount of RNA.

5.

Poly(A) selection or ribosomal depletion?

For eukaryotic systems, researchers have a choice between poly(A) selection or rRNA depletion for enriching non-ribosomal RNA during library preparation. Prokaryotes require rRNA depletion methods as their mRNA is not polyadenylated. The decision tree in Figure 10 can help guide selection of an enrichment method.

Poly(A) Selection (Eukaryotes Only)

Poly(A) selection is the most common method for library preparation, as it often provides efficient enrichment of mRNA transcripts in a cost-effective manner. One drawback is the potential for 3′ end bias. Thus, RNA of high integrity is highly recommended when using this method to ensure sequences toward the 5′ end are sufficiently represented in the data.

Ribosomal Depletion

Ribosomal depletion kits (e.g. RiboMinus™ or Ribo-Zero®) use oligos complementary to rRNA and are conjugated to a substrate to remove rRNA molecules prior to library preparation. In eukaryotes, this technique allows researchers to study both coding and non-coding transcripts.

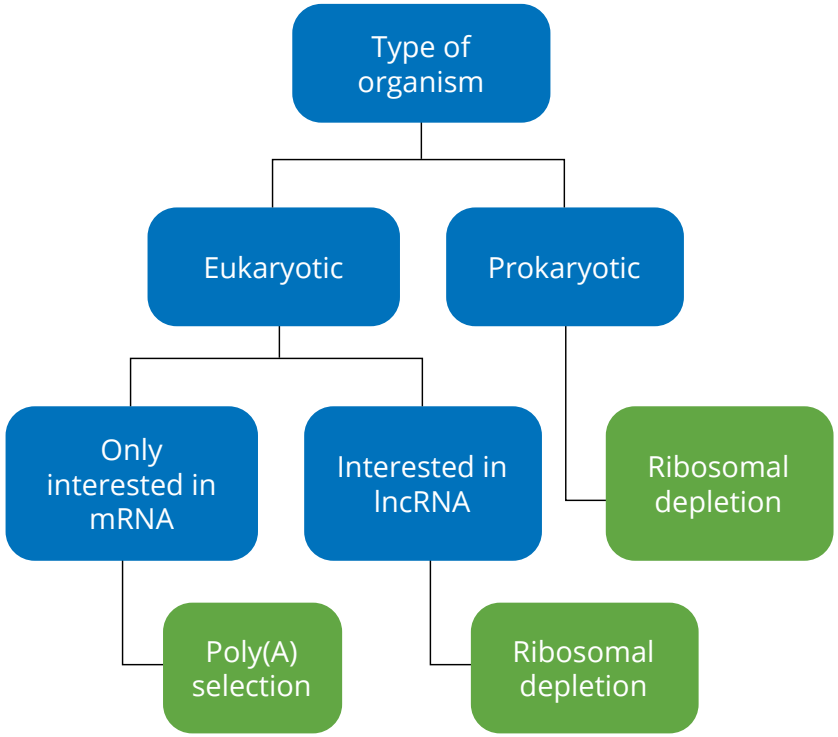


Figure 10. Simplified decision tree for RNA target enrichment

Return to 10 Questions

6.

Stranded or non-stranded library preparation?

Stranded Libraries

Stranded libraries provide a complete view of the transcriptome, allowing you to identify the orientation of the sequenced transcripts and accurately quantify expression levels for genes with overlapping genomic loci that are transcribed from opposite strands. With stranded libraries, antisense transcripts can be identified as such, whereas non-stranded libraries lack information about the polarity of the originating RNA molecule. This information is important for genome annotation and novel transcript discovery (Figure 11).

Non-Stranded Libraries

Non-stranded library preparation is typically less expensive and used for measuring gene expression in organisms with well-annotated genomes. In a non-stranded protocol, sequencing reads may originate from transcripts in either the forward or reverse direction. Thus, your experimental objective and budget should be considered when deciding between stranded and non-stranded methods. Most commercial library preparation kits are strand-specific (Figure 11).

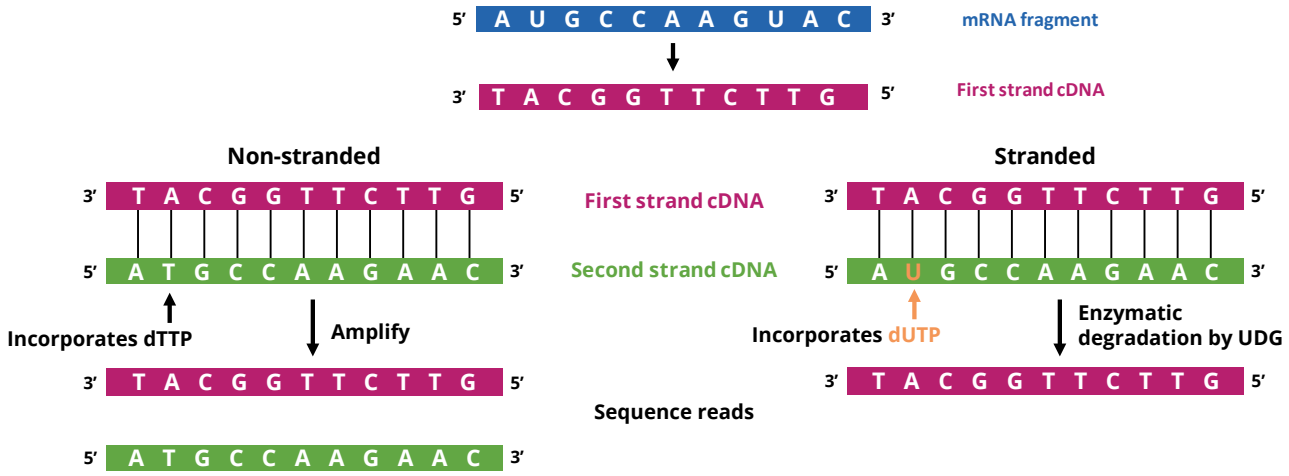


Figure 11. Difference between stranded and non-stranded protocols. During second-strand cDNA synthesis, dTTPs are incorporated in non-stranded libraries, whereas dUTPs are incorporated in stranded libraries. After library preparation, the dUTP-marked strand is selectively degraded by Uracil-DNA-Glycosylase (UDG), ensuring only the first strand survives the subsequent PCR amplification step and hence the strand information of the libraries is retained. In non-stranded libraries, the incorporation of dTTP ensures both strands of RNA are available prior to the final amplification step.

Return to 10 Questions

7. Which sequencing platform should I use?

While several NGS platforms are available for RNA-Seq, each have their own advantages and disadvantages depending upon your situation. We've created a platform comparison (Table 6) to help you determine which would be best suited for your RNA-Seq project.

Table 6. NGS Platform Comparison for RNA-Seq

TYPE	PLATFORM PROVIDER	ADVANTAGES	DISADVANTAGES
Short-read sequencing	Illumina	• High throughput • High accuracy • Low cost per base • Widely available	• Maximum single-end read length of 300 bp (600 bp paired-end)
	BGI/MGI DNBSEQ™	• Low cost per base	• Not available in US or Europe • Maximum single-end read length of 150 bp (300 bp paired-end)
Long-read sequencing	PacBio	• Contiguous, full-length transcript sequencing up to 10 kb • Lower error rate compared to other long-read sequencing technologies	• Higher cost • Low-throughput
	Oxford Nanopore	• Real-time, high-throughput transcript sequencing up to >20 kb • Cost-effective and portable	• Higher error rate compared to other long-read sequencing technologies

Return to 10 Questions

8. Paired-end or single-end sequencing?

Sequencing can be done either unidirectionally (single-end sequencing) or bidirectionally (paired-end sequencing) and then aligned to a reference genome database or assembled to obtain *de novo* transcripts (Figure 12).

Paired-end sequencing is usually recommended for most applications as it provides richer data and permits longer library insert sizes. Single-end sequencing, which is less expensive, is typically reserved for quantifying gene expression in well-annotated genomes and analyzing small RNA molecules.

Figure 12. Single-end vs. paired-end sequencing. Single-end sequencing enables sequencing from one end only, whereas paired-end sequencing enables both ends of DNA fragments to be sequenced.

9.

What read length and depth should I use?

Read Length

Most commercial providers offer paired-end 150 bp (also known as 2x150 bp or PE150) sequencing as it is appropriate for most RNA-Seq experiments. General recommendations based on application are below.

- Gene expression profiling: 1x50 bp to 2x150 bp
- Transcriptome assembly: 2x75 bp to 2x150 bp
- Small RNA analysis: 1x50 bp

Read Depth

For RNA-Seq experiments, read depth is typically expressed in number of reads per sample. With paired-end sequencing, you can express this value in terms of single reads or read pairs. For example, if you sequenced 20 million fragments (or clusters on a flow cell) with a paired-end strategy, you would generate 40 million single reads or 20 million read pairs. Most experiments will require 5 to 200 million reads per sample. In general, smaller genomes (with fewer genes) need fewer reads for quantification of gene expression.

Choosing the required read length and depth depends on the application. The guide in Table 7 refers only to short-read sequencing technology.

Table 7. Recommended Depth Based on Application

APPLICATION	GENOME SIZE	RECOMMENDED DEPTH PER SAMPLE
Gene expression profiling	≥100 Mb	20 to 30 million reads or read pairs
	<100 Mb	5 to 10 million reads or read pairs
Transcriptome assembly	N/A	100 million read pairs

10.

What types of data analyses are available?

Common types of analysis for RNA-Seq include:

- Quality optimization (trimming, filtering, error correction if possible)
- Mapping to reference genomes
- Mapping to transcriptomes
- *De novo* assembly and annotation to protein database
- Transcript and isoform quantification
- Translocation events
- SNP detection
- Indel detection
- Splicing variant analysis
- Gene set enrichment analysis (GSEA)
- GO-enrichment analysis
- Principal component analysis
- Plots and heatmaps

HERE ARE SEVERAL POPULAR OPEN SOURCE SOFTWARE PROGRAMS TO HELP YOU WITH DATA ANALYSIS:

- ***BioJupies***
- ***OMICtools***
- ***ExpVIP*** *is an interactive interface that allows easy analysis and visualization of data*

Return to 10 Questions

References

1. Wang, B., et al., *Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing.* Nat Commun, 2016. **7**: p. 11708.
2. Kono, N. and K. Arakawa, *Nanopore sequencing: Review of potential applications in functional genomics.* Dev Growth Differ, 2019. **61**(5): p. 316-326.
3. Dobin, A., et al., STAR: *ultrafast universal RNA-seq aligner.* Bioinformatics, 2013. **29**(1): p. 15-21.
4. Liao, Y., G.K. Smyth, and W. Shi, *featureCounts: an efficient general purpose program for assigning sequence reads to genomic features.* Bioinformatics, 2014. **30**(7): p. 923-30.
5. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biol, 2014. **15**(12): p. 550.
6. Wang, J., et al., *WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit.* Nucleic Acids Res, 2017. **45**(W1): p. W130-W137.
7. Mortazavi, A., et al., *Mapping and quantifying mammalian transcriptomes by RNA-Seq.* Nat Methods, 2008. **5**(7): p. 621-8.
8. Sotiropoulou, G., et al., E*merging roles of microRNAs as molecular switches in the integrated circuit of the cancer cell.* RNA, 2009. **15**(8): p. 1443-61.
9. Allen, T.A., et al., *Circulating tumor cells exit circulation while maintaining multicellularity, augmenting metastatic potential.* J Cell Sci, 2019. **132**(17).
10. Jiang, L., et al., *Synthetic spike-in standards for RNA-seq experiments.* Genome Res, 2011. **21**(9): p. 1543-51.
11. Fardin, P., et al., *Normalization of low-density microarray using external spike-in controls: analysis of macrophage cell lines expression profile.* BMC Genomics, 2007. **8**: p. 17.
12. McIntyre, L.M., et al., *RNA-seq: technical variability and sampling.* BMC Genomics, 2011. **12**: p. 293.
13. Adiconis, X., et al., *Comparative analysis of RNA sequencing methods for degraded or low-input samples.* Nat Methods, 2013. **10**(7): p. 623-9.